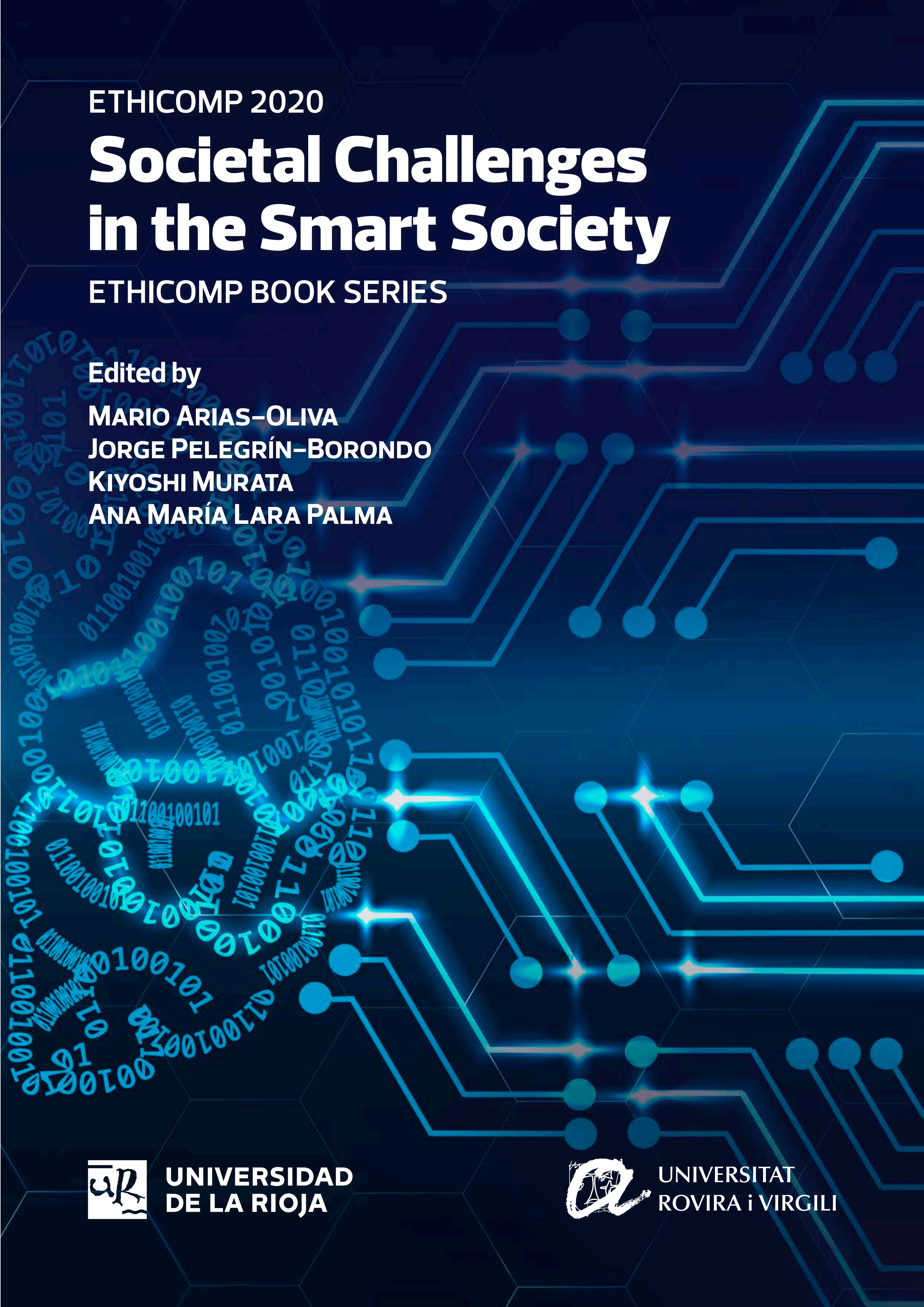
ETHICOMP 2020

Societal Challenges in the Smart Society

ETHICOMP BOOK SERIES

Edited by

Mario Arias-Oliva
Jorge Pelegrín-Borondo
Kiyoshi Murata
Ana María Lara Palma



UNIVERSIDAD
DE LA RIOJA

UNIVERSITAT
ROVIRA i VIRGILI

Edited by

Mario Arias-Oliva

Jorge Pelegrín-Borondo

Kiyoshi Murata

Ana María Lara Palma

ETHICOMP 2020

Societal Challenges in the Smart Society

ETHICOMP Book Series

UNIVERSIDAD DE LA RIOJA

UNIVERSITAT ROVIRA i VIRGILI

ETHICOMP BOOK SERIES

Title	Societal Challenges in the Smart Society
Edited by	Mario Arias-Oliva (Universitat Rovira i Virgili), Jorge Pelegrín-Borondo (University of La Rioja), Kiyoshi Murata (Meiji University), Ana María Lara Palma (University of Burgos)
ISBN	978-84-09-20273-7
Local	Logroño, Spain
Date	2020
Publisher	Universidad de La Rioja

Individual papers – authors of the papers. No responsibility is accepted for the accuracy of the information contained in the text or illustrations. The opinions expressed in the papers are not necessarily those of the editors or the publisher.

Publisher: Universidad de La Rioja, www.unirioja.es

Cover designed by Universidad de La Rioja, Servicio de Comunicación, and Antonio Pérez-Portabella.

ISBN 978-84-09-20273-7

To those who passed away due to the COVID-19 pandemic

The ETHICOMP Book series fosters an international community of scholars and technologists, including computer professionals and business professionals from industry who share their research, ideas and trends in the emerging technological society with regard to ethics. Information technologies are transforming our lives, becoming a key resource that makes our day to day activities inconceivable without their use. The degree of dependence on ICT is growing every day, making it necessary to reshape the ethical role of technology in order to balance society's 'techno-welfare' with the ethical use of technologies. Ethical paradigms should be adapted to societal needs, shifting from traditional non-technological ethical principles to ethical paradigms aligned with current challenges in the smart society.

Table of contents

1. Creating Shared Understanding of 'Trustworthy ICT'

ARTIFICIAL INTELLIGENCE: HOW TO DISCUSS ABOUT IT IN ETHICS

Olli I. Heimo, Kai K. Kimppa

University of Turku (Finland), University of Turku (Finland)

olli.heimo@utu.fi; kai.kimppa@utu.fi

ABSTRACT

In this paper we look into how several different AI technologies are addressed in the ethics literature. We claim that in many cases the technologies are not defined well enough for the moral concerns to be as relevant as they could. We propose that for AI and ethics research to be taken seriously by those designing, using and creating policy, the ethical research to AI needs to be more specific on the technologies evaluated from an ethical perspective, and descriptive understanding of the technologies in question must be presented more clearly for the normative suggestions to be considered valid.

KEYWORDS: Artificial Intelligence, Ethics, Weak AI, Strong AI, Discourse.

1. INTRODUCTION

Artificial intelligence (AI) is the buzzword for the era and is penetrating our society in levels unimagined before – or so it seems to be (see e.g. Newman, 2018; Branche, 2019; Horaczek, 2019). In IT-ethics discourse there is plenty of discussion about the dangers of AI (see e.g. Gerdes & Øhstrøm 2015) and the discourse seems to vary from loss of privacy (see e.g. Belloni et al. 2014) to outright nuclear war (See e.g. Arnold & Scheutz 2018) in the spirit of the movie *Terminator 2*.

AI is presented sometimes as a bogeyman-technology, sometimes as a saviour of our age destined to save us from climate change, overpopulation, food shortage etc. Yet it seems that with AI discussion there is a lot of space for misunderstandings and misrepresentations starting from but not limited to what is AI. In this paper therefore the AI from the ethical perspective of what we should discuss about AI is presented.

This question will become more prevalent the more AI is being used in different circumstances. Actual applications behave very differently, even with same 'base' AI technology, depending on the application area, and even individual application. Thus, understanding and describing the application and the area for which the AI is being used as a solution becomes paramount to understand the specific ethical issues raised by the application; when there are ethical issues – not all applications of AI produce ethical concerns (e.g. using AI to separate different kinds of metal, wood and plastic from waste products), but rather only practical questions. Very high level attempts at ethical analyses will necessarily prove problematic; even military applications of AI can be ethically done, even if we would agree that AI automated weapons ought not to be created. Thus, the first step offered in this paper is to divide the area to different topic areas. In future papers, this division needs to be handled in more detail in each specific area, and those more specific areas need to be analysed in turn to find the areas with more and less ethical issues; although in the end, the question is always on an individual application level.

2. WHAT IS AI

2.1. General definition

There is of course various different ways to conceptualise the difference between different kinds of things labelled as AI. Whereas the technical ones have the tendency to focus on the technical structure of the tool at hand, from the ethical point of view the focus should be more on 1) what the system can do and 2) how it does it. Moreover, we should also focus on the issue on how the bad consequences could be avoided (Mill 1863) and how the people with malicious intentions could be controlled (Rawls 1971). There of course are different motivations and (hopeful) consequences when using AI, which are duly worthy of a different discourse and study in themselves), but in this paper the issue of *definition* for the use itself is discussed. Hence, in the full paper we will discuss the following four different groups of AI:

1. Scripts (gaming and otherwise)
2. Data mining and analysis
3. Weak AI & Strong AI (In its current form: neural networks, machine learning, mutating algorithms etc.)
4. General AI (Skynet, HAL, Ex Machina, etc.)

2.2. Scripts

First of all the *scripts*, mostly advertised as “AI” in computer games are just “simple” algorithms. As these are mostly the first version of AI we meet when talking about it, we must remember that they are merely scripts and cheating (i.e. not AI at all) to make the opponents in computer games more lifelike, to make the sensation that you are playing against actual intelligent opponents. This of course is not true because the easiest, cheapest, and thus most profitable way to give the illusion of a smart enemy is to give the script the power of knowing something they should not.

Hence the idea is to give the player the illusion, but the actual implementation is much simpler (and for smarter or more experienced players also quite transparent…). That is the art of making a good computer game opponent. Hence computer game AIs are just glorified mathematical models to entertain the customers.

2.3. Traditional data mining

The second one discussed as an AI quite often is *data mining* and the related data analysis, “just” gathering specific information from a huge pile of data. Data mining is “the science of extracting useful knowledge from such huge data repositories”(Chakrabarti et al., 2006). Yet this is usually and mostly done by scripting; Patterns and mathematical models are found and tiny bits of data from the patterns are combined to find similarities, extraordinarities and peculiarities then to be analysed by humans aided by a traditional algorithm. Data mining is a multidisciplinary field of study combining broadly statistics, linear algebra, database systems, and algorithms and data structures where the information stored can be made knowledge. (Chakrabarti et al., 2006, Hand, 2017.)

Traditionally there is nothing intelligent about these algorithms except the people making them. Therefore, compared to real artificial intelligence, they too are just glorified mathematical models and smart people working with them – a massive difference to the former though. It is of course possible, and in many cases advisable, to use machine learning and mutating algorithms in data mining (Wu, 2004), but as it is not required, in this categorisation, those deserve a place of their own.

2.4. Weak AI & Strong AI

Thirdly, we discuss machine learning, mutating algorithms, neural networks and other state of the art AI research, i. e. *weak AI*. This is *the point* we should currently focus on when discussing themes related to AI. These methods make the computer better by every step the computer makes; every decision the computer makes improves the computer, not the user.

To clarify, Artificial Intelligence refers to a system, in which is a mutating algorithm, a neural network, or similar structure (also known as weak AI) where the computer program "learns" from the data and feedback it is given. Weak AI is only capable on solving certain problems in chosen platforms and cannot achieve consciousness. It can although be rather excellent in identifying text, in speech-to-text applications, translation, identifying humans, human emotions, and actions from pictures and videos, and playing chess, go, checkers and other games. (Pietikäinen & Silvén, 2019pp. 23, 104-113)

Strong AI is an AI which is close on human intelligence and has at least some idea of self. The machine can use different background information while planning and making decisions. Fully autonomous actions in chaning environment, e.g. in traffic, already require partially a strong AI. Especially in conflict situations even though the lower level decisions, noticing other road users or chaning lines, are clearly in the territory of weak AI. To duplicate a natural and believable discussion between a human and a machine a strong AI is required due to the necessity to understand the context of the discussion. Strong AI is clearly the next big step in AI development. (Pietikäinen & Silvén, 2019, pp. 23, 113)

These technologies are usually opaque (i.e. black box –design), so even their owners or creators cannot *know* how or why the AI ended up with the particular end-result. (See e.g. Covington, Adams, and Sargin, 2016). As AI has been penetrating the society in many different levels for years, e.g. banking, insurance, and financial sectors (see e.g. Coeckelbergh, 2015).

2.5. General AI

The fourth issue, *General AI*, (sometimes *Artificial General Intelligence, AGI* (see e.g. Goertzel, 2007, p. V)), often discussed in the field of AI and described in multitude of Sci-Fi is the "living" AI, the thinking AI – possibly the feeling and fearing AI. The issue with a general AI is that we seem to be nowhere near in science. There are many "general AI" studies done in specific settings, e.g. gaming, where the development is focused in the AI learning to play different video games. These however are not general AIs as such, but moreover machine learning algorithms.

There is also a general AI category Super AI (also known as superintelligence), e.g. the "Skynet", the singularity "the moment at which intelligence embedded in silicon surpasses human intelligence" (Burkhardt, 2011, Pietikäinen & Silvén, 2019, pp. 23-24, Coeckelbergh , 2020, pp. 10-13) and starts to consider itself equal or better than humans. These AIs are luckily or sadly, depending on the narrative the utopia or the dystopia, are still mere fiction and in the technological scale in a future we cannot yet even comprehend.

3. PROBLEM

When discussing technology, the possibilities of technology and possible technologies we must be aware that the first of these does already exist. The second one of these is due to exist, and the third one may exist. While it is possible that technology will exist in say 5-10 years, we also must remember that the society will not be what it is now and other technologies will exist and the society has moved on. There are numerous issues within the field of AI currently at hand, e.g. biased AI (Heimo & Kimppa 2019), liability of autonomous vehicles (see e.g. Heimo, Kimppa & Hakkala 2019), weaponizing AI systems (see e.g. Gotterbarn, 2010), facial recognition (see e.g. , Heimo & Kimppa 2019; Doffman, 2019) just to mention few. Moreover there are plenty of near-future applications of these that must be handled before they become a critical issue. Yet it is important to discuss about all the levels of AI technologies – and to tie them to their timeline!

As we know we must interpret the writings of the past for they were written in their time (see e.g. MacIntyre, 2014), we must also interpret the future which will be different in ways we cannot fully understand. Therefore to predict the AI can do in 10-20 years' time is quite different when we cannot fathom what kind of society we will have in 10 years' time. We must yet keep in mind that what we give up now in the sense of privacy, personal information, liberties etc. can and will be taken away from us more efficiently with the future AI, especially if we follow the Chinese route, which is possible. But to talk of the society now with a futuristic AI seems intellectually dishonest. We do not have flying cars, hoverboards nor the cure for cancer, things predicted and assumed by everyone in any popular culture from the 80s or 90s (see e.g. Back to the Future) yet we have Twitter, Wikipedia and cat picture memes, not something we would actually have been predicting at the time. It is not that we would say that predicting future is irrelevant, moreover we wish to encourage people, scientists and philosophers to focus be explicit when predicting the future; to emphasize their predictions of the timeline they assume technology be in use. Hence when we are talking about AI there are many possibilities for the future but a General AI is a as much of a thing of a future we cannot yet predict, as datamining is a thing of the past. Predictions as predictions, and facts as facts, that is all we can do for honest science.

4. DISCUSSION

Therefore, when analysing digitalisation via AI and it's possibilities, it is clear that we should focus on weak AI and strong AI. These are the things of now and near future whereas scripts and data mining are not AI at all and general AI is still being sci-fi which we are not yet sure shall it happen, and if, when. Yet to create valid scientific discourse we should be focused on what we know instead of what we do not. To make predictions, alert other scientists (as it is a proper task for an IT-ethicist), and to guide the scientific discourse and development, we need that knowledge.

AIs already control a lot of our daily lives, e.g. in entertainment where Netflix, YouTube, and Social media sites which content is shown to us due our preferences, how we are classified by the system, and what the media corporation wishes to promote. This however seems to be still quite dumb and does not fulfil the promises marketed to the public. Since the algorithms generate frustration in the users due poor suggestions as majority of the content one wishes to see must be acquired by searching. Yet the AIs are learning and might turn out to be the privacy endangerment predicted. (Heimo & Kimppa, 2019)

One of the key issues when talking AI is the black box –mentality of the given systems. Whereas we can understand where our solutions come from and tweak them to be ethical (e.g. not discriminatory against women, as was the case in Amazon's HR (Hamilton, 2018)). The black box feature is one of the key issues when discussing about the AI in ethics and a key reason why the definitions around AI should be clearly expressed.

Also the question of when is important. As focusing on the discussing about AI, the distinction between now, near-future and far-future should be made clear. If the discussion around time-frame obscures, the discussion itself can become obscured due to the predicted development of technology and the various other possibilities in the future. Therefore, if the time-frame of the discussion is not clear, the discussion is no longer valid as we are not discussing about the same thing anymore.

Hence the authors propose a two-stage model on evaluating AI:

- Are we talking about AI or something else? Describe the AI clerly.
- Are we talking now/near-future or far-future/sci-fi? Tell the audience roughly the time-frame, e.g. 5-10 years or 30-50 years.

A fine example on the discourse without timelines is in Coeckelbergh's (2010) esteemed article "Robot rights? Towards a social-relational justification of moral consideration" where Coeckelbergh, rises interesting arguments about robot rights and finds equally interesting questions and justifications. Yet the article lacks the depth in the description of AI development timelines on predicting the need for the change mentioning only "near-future" and "long stage", which after 10 years seem to be still that. The main goal of this article of course is not to alarm us to the imminent requirement for robot rights nor demand any action for or against the current development but moreover to participate to an academic discourse presented in the paper.

Yet the argument of this paper is that we should improve the precision when discussing future technologies – especially *with near-future applications* and at least *when rising alarm or demanding action*. The prediction of this paper is that the future of predicting future is danger if the current predictions of future are done without clearly describing the foreseeable future.

5. CONCLUSIONS

What we want to emphasize with this paper is that many authors on the ethics of AI leave the kind of AI they are discussing so unclear as to not make it clear whether they even understand the topic area at all. They have vague notions of AI, which they do not specify to the extent that the ethical questions are first of all not relevant to any specific technology currently used, nor clearly future studies on the problematic paths that we may take. This causes AI ethics not to be taken seriously by those who ought to take it seriously, namely designers of AI, companies using AI, and governments and intergovernmental organizations attempting to regulate AI development.

If we in the field of ICT and ethics are not believable, our suggestions will be ignored, and AI development may be either misdirected or left all together undirected, and thus create applications which are problematic for users, companies and the society alike. Especially considering the surprising amount of AI ethicists that have recently emerged from anonymity on the field, traditional ICT and ethics researchers who have done years, even decades of study in the field of AI and ethics need to be extremely careful to see to the validity of their claims, whilst at the same time they need to be very visible in the current discussions relating to AI and ethics in all relevant levels from concept creation to actual applications to government and intergovernmental policy creation.

REFERENCES

Arnold T. & Scheutz M. (2018) The ”big red button” is too late: an alternative model for the ethical evaluation of AI systems, Ethics and Information Technology, 20:59-69.

Belloni, A. et al. (2014) Towards A Framework To Deal With Ethical Conflicts In Autonomous Agents And Multi-Agent Systems, CEPE 2014.

Branche, P. (2019), Artificial Intelligence Beyond The Buzzword From Two Fintech CEOs, Forbes, Aug 21 2019, https://www.forbes.com/sites/philippebranch/2019/08/21/artificial-intelligence-beyond-the-buzzword-from-two-fintech-ceos/#43f741c7113d

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., ... & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. Ethics and information technology, 12(3), 209-221.

Coeckelbergh, M. (2015) The tragedy of the master: automation, vulnerability, and distance, Ethics and Information Technology, 17:219-229.

Coeckelbergh, M. (2020) AI Ethics, MIT Press, 2020.

Covington, P., Adams, J., and Sargin, E. (2016) Deep neural networks for youtube recommendations. Proceedings of the 10th ACM conference on recommender systems. ACM, 2016.

Doffman, Z. (2019) China's 'Abusive' Facial Recognition Machine Targeted By New U.S. Sanctions, Forbes, Oct 8, 2019. https://www.forbes.com/sites/zakdoffman/2019/10/08/trump-lands-crushing-new-blow-on-chinas-facial-recognition-unicorns/#52641d79283a

Gerdes, A. & Øhstrøm, P. (2015) Issues in robot ethics seen through the lens of a moral Turing test, JICES 13/2:98-109.

Goertzel, B. (2007). Artificial general intelligence (Vol. 2). C. Pennachin (Ed.). New York: Springer.

Gotterbarn, D. (2010) Autonomous weapon's ethical decsions:" I am sorry Dave; I am afraid I can't do that.". In proceedings of ETHICOMP 2010 The "backwards, forwards and sideways" changes of ICT Universitat Rovira i Virgili, Tarragona, Spain 14 to 16 April 2010.

Hamilton, I.A. (2018) Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women, Business insider, October 10th 2018, available at https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10?r=US&IR=T

Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622.

Heimo, O. I. & Kimppa, K. K. (2019) No Worries–the AI Is Dumb (for Now), Proceedings of the Third Seminar on Technology Ethics 2019 Turku, Finland, October 23-24, 2019, pp. 1-8.

Heimo, O. I., Kimppa, K. K. & Hakkala, A (2019) Automated automobiles in Society, IEEE Smart World Congress, Leicester, UK, 2019.

Horaczek, S. (2019), A handy guide to the tech buzzwords from CES 2019, Popular Science Jan 9 2019, https://www.popsci.com/ces-buzzwords/

Mill, John S. (1863) Utilitarianism, https://www.utilitarianism.com/mill1.htm, accessed 21.10.2019.

Newman, D. (2018) Top 10 Digital Transformation Trends For 2019, Forbes, Sep 11, 2018, https://www.forbes.com/sites/danielnewman/2018/09/11/top-10-digital-transformation-trends-for-2019/#279e1bca3c30

Pietikäinen, M. & Silvén, O. (2019) Tekoälyn haasteet – koneoppimisesta ja konenäöstä tunnetekoälyyn, Center for Machine Vision and Signal Analysis (CMVS), November 2019, ISBN 978-952-62-2482-4

Rawls, J. (1971) A Theory of Justice, Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Robbins S. (2018) The Dark Ages of AI, Ethicomp 2018.

Wu, X. (2004) Data mining: artificial intelligence in data analysis. In Proceedings. IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004.(IAT 2004). (p. 7). IEEE.

DEVELOPING A MEASURE OF ONLINE WELLBEING AND USER TRUST

Liz Dowthwaite, Elvira Perez Vallejos, Helen Creswick, Virginia Portillo, Menisha Patel, Jun Zhao

University of Nottingham (UK), University of Nottingham (UK), University of Nottingham (UK),
University of Nottingham (UK), University of Oxford (UK), University of Oxford (UK)

liz.dowthwaite@nottingham.ac.uk; elvira.perez@nottingham.ac.uk;
helen.creswick@nottingham.ac.uk; virginia.portillo@nottingham.ac.uk;
menisha.patel@cs.ox.ac.uk; jun.zhao@cs.ox.ac.uk

ABSTRACT

This paper describes the first stage of the ongoing development of two scales to measure online wellbeing and trust, based on the results of a series of workshops with younger and older adults. The first, the Online Wellbeing Scale includes subscales covering both psychological, or eudaimonic, wellbeing and subjective, or hedonic, wellbeing, as well as digital literacy and online activity; the overall aim is to understand how a user's online experiences affect their wellbeing. The second scale, the Trust Index includes three subscales covering the importance of trust to the user, trusting beliefs, and contextual factors; the aim for this scale is to examine trust in online algorithm-driven systems. The scales will be used together to aid researchers in understanding how trust (or lack of trust) relates to overall wellbeing online. They will also contribute to the development of a suite of tools for empowering users to negotiate issues of trust online, as well as in designing guidelines for the inclusion of trust considerations in the development of online algorithm-driven systems. The next step is to release the prototype scales developed as a result of this pilot in a large online study in to validate the measures.

KEYWORDS: wellbeing, trust, online experience, scale.

1. INTRODUCTION

As interaction with online platforms is becoming an essential part of people's everyday lives, the use of automated decision-making algorithms in filtering and distributing the vast quantities of information and content to users is having an increasing effect on society, with many people raising questions about the fairness, accuracy and reliability of such outcomes. Online users often do not know when to trust algorithmic processes and the platforms that use them, reporting anxiety and uncertainty, feelings of disempowerment, defeatism, and loss of faith in regulation (Creswick et al., 2019; Knowles & Hanson, 2018). Various other negative effects of using such online technologies have been identified, for example, concerns about hostile actors spreading online disinformation, vulnerable groups becoming victims of scams, harmful user-generated content and bullying, and addiction and excessive screen time (Chadborn et al., 2019; DCMS, 2019; Kidron et al., 2018; Livingstone et al., 2010). These issues lead to concerns about wellbeing which can affect both the user and broader society. It is therefore important that mechanisms and tools are developed to assess online wellbeing and trust with the view to support users interacting with the online world.

This paper describes the first stage of the ongoing development of an ‘Online Wellbeing Scale’ (OWS) and a ‘Trust Index’ (TI) to aid in understanding how trust (or lack of trust) relates to overall wellbeing online. There are two broad aims of the scales. For researchers, the scales will allow exploration of the relationship between wellbeing and trust, to understand how trust of algorithmic systems affects user’s online experiences across different online activities and at different levels of digital literacy. For the users, the scales will contribute to the development of a tool for self-measuring and reflecting on trust, as part of engaging in dialogue with platforms in order to jointly recover from trust breakdowns. The scales will be part of a suite of tools for (1) empowering users to negotiate issues of trust online and (2) designing guidelines for the inclusion of trust relationships in the development of algorithm-driven systems.

2. BACKGROUND

Interaction with online platforms is becoming an essential part of people’s everyday lives. As digital technology develops, people are using the Internet to carry out more and more of their everyday activities, from socialising and entertainment to financial transactions and working. News feeds, online media, search engine results and product recommendations increasingly use personalisation algorithms (i.e., sequences of instructions or commands for computers to solve a task) to help users cut through the vast amounts of available information. They use vast quantities of data, especially personal data, to provide a more personalised, appealing and engaging online experience. The use of such algorithms is therefore having an increasing effect on society, with many people raising questions about the fairness, accuracy and reliability of outcomes. Algorithmic decision-making often lacks transparency and when using online services users are generally given next-to-no information about the algorithms that are being used, or the data that is used to feed those algorithms. Online users therefore often do not know when to trust algorithmic processes and the platforms that use them, reporting anxiety and uncertainty, feelings of disempowerment, defeatism, and loss of faith in regulation (Creswick et al., 2019; Knowles & Hanson, 2018). These issues lead to concerns about wellbeing, which can affect both the user and broader society.

Other aspects of the online world have also raised concerns about wellbeing. Online media have been associated with both physical and mental harms (DCMS, 2019). This is particularly the case among children and young people; approximately 1 in 5 11-16 year-olds have been exposed to potentially harmful content online (Livingstone et al., 2010). There are also emerging challenges about designed addiction and excessive screen time (Kidron et al., 2018). An increase in time spent on social media has also been associated with decreased life satisfaction and quality of life in children (McDool et al., 2016), and may be damaging to mental health (Royal Society for Public Health, 2017). Vulnerable groups, including older adults, are often victims of financial scams or intimidation, and feel they have no choice but to access online services that they do not fully understand or trust (Chadborn et al., 2019). There are also serious concerns about hostile actors using online disinformation to undermine democratic values and principles. It is therefore important to be able to assess the effects of the online world on wellbeing to make meaningful recommendations for the responsible design of technologies. It has also been suggested that problematic internet and social media use may be linked to personality and psychological needs (Kozan et al., 2019). Whilst there are many existing measures of wellbeing, there are no measures of online wellbeing specifically, where ‘online wellbeing’ is defined as the effects of carrying out activities and tasks online on a person’s wellbeing. This paper describes the first stages of creating such a scale, the Online Wellbeing Scale (OWS).

2.1. Measuring Online Wellbeing

There are many different definitions of wellbeing, often focusing on the different dimensions of wellbeing rather than a single concept (Dodge et al., 2012). Studies of wellbeing often fall into two main traditions: eudaimonic and hedonistic (Deci & Ryan, 2008). Eudaimonic wellbeing refers to living well and flourishing as a person, and is often conceptualised as psychological wellbeing (PWB), whilst hedonistic wellbeing refers to the balance of positive and negative emotions that are experienced by an individual, conceptualised as subjective wellbeing (SWB). The OWS will measure both types in order to get a broad idea of how the online world affects overall wellbeing. The scale will rely on self-report, as such measures allow individuals to define for themselves how they experience and understand their own wellbeing, rather than forcing an objective definition from outside (Alexandrova, 2005).

The attainment of PWB is part of the grounding of Self-Determination Theory (SDT) (Ryan et al., 2006). SDT is a collection of six sub-theories which present a framework for studying motivation and personality, focussing on how social and cultural factors affect people's agency, wellbeing, and performance. One of these sub-theories, Basic Psychological Needs Theory (BPNT) states that psychological health and wellbeing are achieved by satisfying certain basic needs: *autonomy* (i.e. agency, the freedom or independence to act as desired), *relatedness* (i.e. a social connection with others), and *competence* (i.e. self-efficacy, the ability to carry out an action effectively) (Deci & Ryan, 2000; Ryan & Deci, 2000, 2017). The three basic psychological needs are universal, having been found across many cultures (Chen et al., 2015) and domains including family, friends, relationships, school, work, and hobbies (Milyavskaya & Koestner, 2011). BPNT has been explored in a few online contexts, for example social media, where different activities were linked to a lack of particular needs (Masur et al., 2014), and online crowdsourcing where different gamification mechanisms satisfied needs to greater or lesser degrees (Goh et al., 2017). It has also been suggested that consideration of basic psychological needs is vital to improving the design of user experience (Peters et al., 2018; Wang et al., 2019). The Basic Psychological Need Satisfaction (BPNS) scale (Gagné, 2003; Ryan et al., 2006; Ryan & Deci, 2000) is a widely used, strongly validated measure of need satisfaction. One study looking at online and face-to-face learning contexts found that some items on the BPNS (in this instance relatedness measures) may need modifying in order to be appropriate for use online (Wang et al., 2019). It is therefore important that research into PWB online should consider the contextual appropriateness of BPNS and a more specific version of this scale which reflects the online domain is needed. The OWS will aim to fulfil this requirement.

SWB focuses on the area of life satisfaction in terms of positive and negative emotions or affect. SWB and PWB are separate constructs, but are very closely related and experienced together; people with high levels of both SWB and PWB can be categorised as 'flourishing' (Heintzelman, 2018) Benefits to SWB tend to be greater than PWB immediately after an experience but PWB benefits are higher in long term; activities that increase PWB often also increase SWB but vice versa is not necessarily the case (Heintzelman, 2018). Therefore measuring both is of value to understanding overall wellbeing. Satisfaction of BPN have also been repeatedly found to be positively associated with life satisfaction (Diener et al., 2017). Measures such as the Positive and Negative Affect Schedule (PANAS, Watson et al., 1988) and the Scale of Positive and Negative Experience (SPANE, Diener et al., 2010) are widely used to measure SWB. There is little work on the effect of either being online or using algorithmically mediated systems, although the conscious choice to use an algorithm has been found to reduce positive mood (Alexander et al 2018).

The basic psychological needs of SDT/PWB can be both experiential outcomes of *activities* and *motives* that directly influence behaviour. Different activities have been shown to satisfy the basic needs to greater or lesser extent; Martela & Sheldon (2019) suggest motives and activities should be measured alongside both SWB and PWB to get a rounded measure of wellbeing: “SWB only answers the question of *how* the subject is feeling, but not the question of *why* the subject is feeling so, or *what* he or she is doing” (p.7). It is likely that the many different activities people carry out online, from shopping for a new television online to interacting with friends on social media will have different effects. The current overabundant flow of online information and social relationships can easily contribute to users not being able to avoid excessive multi-tasking or overconsumption of media. Considering the context of a person’s online life can speak to their motivations for being online, and relating these to wellbeing can help to identify which activities may be eudaimonic in nature or contribute to SWB.

Additionally, user understanding and digital literacy may also affect levels of online wellbeing (Gui et al., 2017). Users’ digital skills (i.e., operational, technical and formal, information/cognition, digital communication, digital content creation and strategic skills) can influence users’ online experiences and ability to cope with the side effects of over engagement, lack of transparency and agency experienced by many when online (Iordache et al., 2017). In keeping with the use of self-report measures, it may be appropriate to think of digital literacy as a measure of the users’ perceived ability to navigate and stay safe in the online world. For example, a user with a high level of confidence in their ability may feel an increased sense of autonomy and competence, and lead to more positive experiences; however equally such a user may feel less autonomy due to the nature of online decision-making algorithms, and experience greater stress or frustration during particular activities.

2.2. The relationships between online wellbeing and trust

As noted previously, the increased use of automated decision-making algorithms online may lead to users being unsure whether they can trust the platforms that use them, which can lead to concerns about wellbeing, especially amongst potentially vulnerable users (Creswick et al., 2019; Knowles & Hanson, 2018). Therefore, alongside the OWS, a second scale measuring user trust online (‘Trust Index’ or TI) is also being developed, which can be used either concurrently or as a separate measure. The first stage of TI development involves identifying particular online factors that affect user trust, as well as identifying common opinions and experiences of trust that can then be transformed into a series of statements which will be validated in future studies. Outside of specific online contexts such as e-commerce and health information, work on trust in online platforms is relatively scarce. Bhattacherjee (2002) developed a scale for individual trust in online firms, for example Amazon, but there are no existing validated measures of online trust that can be adapted to take into account the general contexts in which trust is enacted. Trust is often measured by single statements or binary, yes/no, type questions, but it has been suggested it should be conceived as a complex and multidimensional psychological state, with both affective and motivational aspects (Kramer, 1999).

Trust is often considered in two forms: interpersonal and institutional. Considering algorithmically-mediated platforms, it appears that theories of institutional trust may be most appropriate, however some aspects of interpersonal trust may come into play if for example a user treats the website as a ‘person’, anthropomorphising the system. It has been found that people use similar neurophysiological mechanisms to trust algorithms as they would people, with companies taking advantage of this by ‘personalising’ decision aids (for example Apple's Siri or Amazon's Alexa) (Alexander et al., 2018). It has also been found that people who believed that others are generally trustworthy were more than twice as likely to adopt an algorithm (Alexander et al., 2018).

The most relevant existing work on trust in this context come from the fields of organisational psychology, management, and marketing. Within such research there is a broad range of conceptualisations, antecedents and types of trust (Kramer, 1999). Often research highlights antecedents of trust such as familiarity and reputation, security and other situational assurances, and encouraging 'normal' user experiences (Gefen et al., 2003, 2008; Yoon, 2002). Differences in levels of trust and related factors have been examined cross-culturally, for example between Japan and the US (Yamagishi & Yamagishi, 1994), indicating that it is also important to consider the different perceptions about trust that users may have, and the differing levels of importance they may place on trust, rather than sticking to a rigid definition and assumptions. As such, self-report is once again of value.

It is likely that someone's trust in a situation will affect their wellbeing, but if for example, trust is not a consideration for them when they are online, their wellbeing and behaviour may be less affected than someone for whom trust is highly important. Trust has been found to be a consistent predictor of SWB (Helliwell & Wang, 2011) although in Serbia this was only the case for institutional trust (Jovanović, 2016). Both interpersonal and institutional trust have been positively associated with life satisfaction in many different countries (Elgar et al., 2011). Recently, trust has also been linked to SDT, suggesting that the satisfaction of basic psychological needs can lead to a motivation to trust, and this motivation will affect people's willingness to continue to trust someone, or to restore trust (van der Werff et al., 2019). Measuring aspects of behaviour and wellbeing alongside trust may also lead to understanding of why and how trust in online systems manifests; it may also help to explain why and when people deviate from the expected trusting behavior, i.e., trust when they objectively should not, and vice versa, which they often do (McKnight et al., 1998).

3. METHOD

The first stage of development of the OWS and TI took place as part of a larger study into online trust, comparing attitudes of younger (16-25 years old) and older (over 65) adults. The study was approved by the Ethics Review Board at all co-authors' institutions as a joint research effort.

3.1. Participants

In total, 74 participants took part in nine 3-hour workshops between April and July of 2019; 5 workshops with 40 older adults (mean age 71 years, 62.5% female) and 4 workshops with 34 younger adults (mean age 20, 58.8% female). Recruitment took place through social media, fliers, and emails to groups such as University of the Third Age. Sessions were conducted in easily accessible venues in Nottingham and Oxford. Prior to the study participants were asked to confirm that they "regularly use the internet for searching for information, making bookings, or buying products" (89.2% indicated they used the internet several times a day) and were thanked afterwards with a £20 high street voucher.

3.2. Design

The project focused on user-driven, human-centred, and Responsible Research and Innovation (Jirotka et al., 2017) approaches to investigating trust. Thus the workshop structure, including timings and ordering of activities, the tasks themselves, and practical considerations were co-created through a series of activities with members of the public in the relevant age groups, ensuring that the content was relevant, understandable, and engaging. The final workshop structure, content and duration was decided through this iterative, user-centred piloting process. The resulting workshops took a mixed-methods approach to encourage participants to think about issues in different contexts. As such the workshops consisted of four distinct activities, with all participants taking part in each activity: (1) a quasi-naturalistic experiment observing user behaviour online, involving a screen-based task in which participants were asked to carry out a common online task (booking a hotel) on two different sites; (2) scenario-based discussions of trust; (3) a paper-based group task looking at different ways of presenting information to identify user requirements for an online tool to negotiate trust; and (4) pre- and post- session questionnaires measuring wellbeing related to trust online. As such, each activity can stand alone or be combined to compare responses in different contexts.

3.3. Materials and Procedure

This paper focuses on the results of the questionnaire activity. The questionnaires were designed to explore whether there is a link between trust, motivation, digital literacy, and wellbeing factors, and how this might be measured. They consisted of a mixture of free text, multiple choice, and Likert-like items. A pre-session questionnaire asked about: *Online Activity:* how often people go online, plus the primary reason and any secondary reasons that they go online out of 5 activities: socialising, buying or booking things, finding information, watching videos or playing games, and sharing content; *Trust:* rating of the importance of trust, whether they have ever stopped using a site because of a lack of trust, and 4 trust-related statements on 7-point Likert-like scales; and *Digital Confidence:* 7 statements on 7-point Likert-like scales, related to perceived digital literacy and how confident users are in carrying out tasks online.

A post-session questionnaire began with questions about the session, then repeated statements from the pre-session questionnaire to see if there were changes in opinion, followed by: *Trust:* ratings on a 7-point Likert-like scales of how much 12 different features of websites affect their trust, and an open-ended question about which factor is most important; and *Wellbeing:* an open-text question about whether the online world affects wellbeing, plus 2 instruments for measuring wellbeing, modified to reflect online experiences. PWB was measured using the Basic Psychological Need Satisfaction (BPNS) scale (Gagné, 2003; Ryan et al., 2006; Ryan & Deci, 2000), including 18 statements on a 7-point Likert-like scale. The scale was minimally modified to reflect the online world, for example “I feel like I am free to decide for myself how to live my life” was altered to “I feel like I am free to decide for myself how to act online”, whilst other statements simply had the word ‘online’ added for context. SWB measurement used a format similar to the Scale of Positive and Negative experience (SPANE) (Diener et al., 2010), including 8 positive and 8 negative feelings on a 7-point Likert-like scale. The measure included words that are often used by Internet users to describe their experiences (Creswick et al., 2019). As suggested by Diener et al. (2010) the amount of time the state was felt rather than intensity of feeling was captured as it is stronger with regards to life satisfaction. A 7-point scale was chosen throughout in order to be consistent with the BPNS, as the most widely validated set of statements.

3.4. Analysis

The analysis of the quantitative data reported in this paper was carried out in SPSS. Cronbach’s alpha reliability analysis was carried out on the scale items to check the internal consistency of items. In general, an alpha of 0.7 to 0.8 is deemed acceptable, with 0.8 more appropriate for cognitive measures, although at feasibility phase, scores as low as 0.5 may be accepted (Field, 2013). The results make no assumptions about the unidimensionality of the scales at this stage. Qualitative analysis of questionnaire responses was carried out using an inductive, data driven approach, to identify themes that were strongly interrelated with the raw data (Braun & Clarke, 2006). A single researcher analysed and coded all the responses fully using NVivo. Another researcher checked this coding for consistency and agreement, and any discrepancies were discussed and resolved. The codes were then grouped into key themes.

4. RESULTS AND DISCUSSION

The results here look at the relevant sections of the questionnaires for development of the prototype OWS and TI. Table 1 summarises each subscale that will be included in the OWS and its properties, based on the following results. The decision to change from a 7 to a 5-point scale for all subscales was taken for consistency with the new wellbeing measures, which are discussed below.

Table 1. Online Wellbeing Scale subscales, Trust Index subscales, and properties.

OWS Subscale	**Items**	**Score range**
Online Activity	6	6 (very little activity) to 30 (a lot of activity)
Digital Confidence	6	6 (very low confidence) to 30 (very high confidence)
Psychological Wellbeing - Autonomy Dis/Satisfaction - Competence Dis/Satisfaction - Relatedness Dis/Satisfaction - Overall Dis/Satisfaction	18 - 3/3 - 3/3 - 3/3 - 9/9	3 (very low need dis/satisfaction) to 15 (very high need dis/satisfaction) -12 (max. dissatisfaction) to 12 (max. satisfaction)
Subjective Wellbeing - Positive/Negative Affect - Affect Balance	12 - 6/6 - 12	6 (very low +/- affect) to 30 (very high +/- affect) -24 (unhappiest possible) to 24 (happiest possible)
TI Subscale		
Importance	6	6 (not at all important) to 30 (highly important)
Belief	6	6 (not at all trusting) to 30 (highly trusting)
Context	16	For each item, 1 (not important to trust) to 5 (very important to trust) or 1 (not trustworthy) to 5 (very trustworthy)

4.1. The Online Wellbeing Scale

Overall, 66.2% of participants felt that the online world and their use of the internet affected their wellbeing, and of these 63.8% felt that its effect was negative, with a further 8.5% suggesting that it could be both. This shows the importance of developing ways to measure the effects of the online world on wellbeing. Participants did a wide range of online activities, with an average of 4 of the 5 options being selected; all activities were chosen by at least half of the participants. The most common activities were finding information (95.9%), socialising (85.1%) and buying or booking things (82.4%). The most common primary activity was socialising (35.1%) followed by finding information (31.1%). Other suggestions for online activities over the last four weeks were made by 18.9% showing that the original list covered most common online activities. The other suggestions were email (8.1%), banking (8.1%) and work (2.7%). This has contributed to the *Online Activity* block of the OWS, containing the items from the initial questionnaire, plus an additional item, 'financial or organisation' which covers all activities suggested by participants. This 6-item block measures the frequency of each activity in an additive way, from 1 (very rarely or never) to 5 (very often or always) for each activity. This allows for a measure of both the range and extent of online activity.

The 7 statements which related to digital confidence from the pre-session questionnaire have cronbach's alpha (α) reliability of 0.797; removing one item ("If I do not trust a website I can do something about it") improves reliability to α=0.827. Modifying one statement slightly from "I am able to tell whether a website is trustworthy" to "Your ability to tell whether or not a website is trustworthy", the *Digital Confidence* block contains 6 items to be rated from 1 (very low) to 5 (very high).

The BPNS from the post-session questionnaire had low reliability for autonomy (α=0.581) and competence (α=0.454). Only relatedness reached an acceptable level (α=0.792). As such, the scale as a whole is not considered reliable. Removing items increased reliability for each scale (α=0.636, α=0.515, α=0.812 respectively) but not enough for the whole scale to be considered acceptable. Many participants also skipped items, for example only 53 participants completed the relatedness items. It was noted that, particularly for the older participants, statements relating to interacting with people online were often either ignored or misunderstood. For the prototype OWS therefore, the *Psychological wellbeing* scale will be replaced with a modified version of the Balanced Measure of Psychological Needs (Sheldon & Hilpert, 2012). This scale uses simpler language and reduces each construct to 6 items, with the ability to calculate the overall level of satisfaction and dissatisfaction of needs. It is also easily modified for different domains. Modifications to this scale will include focusing wording on the online world and replacing specific references to 'people' with a more general interactional focus.

The SWB measure scored acceptable reliability for the positive experience scale of α=0.775, improved by removing one item ("High Mood") to α=0.785, and the negative experience scale scored a good reliability of α=0.805, improved by removing one item ("Tracked") to α=0.808. As each state was presented with an opposite state, removing the equivalent positive and negative words ("Safe" and "Low Mood") resulted in a reliability of α=0.797 for positive experience and α=0.781 for negative experience. This results in a modified *Subjective Wellbeing* scale of 6 items each for positive and negative feelings: Calm/Anxious; Creative/Apathetic; Empowered/Disempowered; Pleased/Annoyed; Powerful/Powerless. This scale is used to derive an overall affect balance score and can also be divided into positive and negative feelings scales.

4.2. The Trust Index

The Trust statements from the pre-session questionnaire do not form a coherent scale, with the highest cronbach's alpha score reaching just 0.249. It was also found that there are several different aspects to trust that apply to the online world. As such, in order to create the TI prototype, three different aspects of trust will be measures in separate subscales. Table 2 summarises the subscales and properties. The *Importance* of trust will form an individual score for how important trust is to the individual when they are online. A total of 78.4% of participants had left a website or stopped using an online service because they did not trust it, suggesting that to some trust is a highly important factor when they are online. 50% of participants agreed that it was indeed highly important, but 9.5% felt that it was not very important at all, so to those participants, trust may not affect their online wellbeing or behavior. Additionally, 83.8% agreed that "websites have a responsibility to act in a trustworthy manner towards their users". Therefore, *Importance* is worth measuring in this scale.

The trusting *Beliefs* of an individual will form an overall score for how much that individual trusts or believes in the online services they use in general. Responses to "I tend to trust things I find on the internet" were varied, with 40.6% disagreeing, 28.4% agreeing, and the remainder responding neutrally. Similarly, 32.5% agreed that "websites do enough to make sure their users trust them" with 25.7% disagreeing. If a person has a general lack of trust in the online world, this may affect their behaviour and wellbeing in different ways to if a person is generally trusting of what they do and see online (Alexander et al., 2018; Elgar et al., 2011; Helliwell & Wang, 2011). Combined with their feelings about the importance of trust this may give a good indication of the levels of trust a user is likely to have in an online service or platform. A 'general trust scale' (Kramer, 1999; Yamagishi & Yamagishi, 1994), examining beliefs about trust and honesty with regards to other people will be modified to relate to online platforms.

The remaining subscale will relate to different *Contexts* and factors related to going online and whether these affect levels of trust, plus a final item which measures whether an individual thinks that their online activities influence their sense of wellbeing. This is to allow a quick comparison of trusting beliefs with their feelings about wellbeing, to compliment the OWS. The contextual items include online reviews, financial transactions, social media, recommendations and entertainment, as the main activities that participants flagged as important, and relating to those they indicated taking part in most often. All of the factors presented to participants in the questionnaire were deemed important, with the most important being reputation (x̄=6.11) and use of personal data (x̄=5.99). Items related to ease-of use also scored highly (x̄=5.50). These also matched with the qualitative responses asking for the most important factors: "*well known firm, been there before"; "the company and its reputation and previous knowledge of it"; "the information which it is demanding for me to share with it, and the language which is used surrounding this"; "difficult to navigate"; "user friendly structure"*. The factor that emerged the most in qualitative responses was security measures: *"the padlock item in the top to let me know it is secure"; "checking whether it's a secure site".* Familiarity with the site was also important: *"if I have used it before"; "whether I know about it/I have had previous experience using it"*. The other factors that are included therefore are: brand and social reputation, personal data, reliability and ease of use, familiarity with the website, privacy policies and security measures, and algorithms. Several of these also relate strongly to factors or antecedents found in the literature on trust, for example familiarity, brand reputation, security, and reliability (Gefen et al., 2003, 2008; Kramer, 1999; Yoon, 2002); the reliance on familiarity and reputation also corresponds to the finding that social proof (i.e., knowing that other people has used it) was found to be the most effective way to persuade people to adopt algorithms (Alexander et al., 2018). This subscale will not be scored like the others, but will be used to identify areas where trust is most relevant or has most effect on users' experiences.

5. CONCLUSIONS AND NEXT STEPS

The questionnaires completed during a series of workshops with younger and older adults allowed for preliminary examination of how online trust and wellbeing might usefully be measured, in order to develop an Online Wellbeing Scale and Trust Index. Examination of the internal reliability of several groups of statements aimed at measuring different factors has led to the development of prototype scales. The 42-item OWS has 4 subscales, with each construct measured by a series of statements on 5-point Likert or Likert-like scales. The first two blocks provide a baseline for the users' online experience: *Online activity* examines the range and amount of activity the participant carries out online whilst *Digital confidence* forms a self-report measure of digital literacy. Both of these factors are likely to have an effect on both online wellbeing and trust. The second 2 blocks examine two major conceptualisations of wellbeing: *Psychological* (eudaimonic) *wellbeing* and *Subjective (*hedonic) *wellbeing.* The TI, which can be used standalone or in cohort with the OWS, consists of 29 statements covering 3 subscales measured by a 5-point Likert Scale. These cover three areas of trust which emerged as important to trust in the workshops and the questionnaire results, with reference to related literature on trust: the *Importance* of trust to the individual user*,* the *Belief* of the user whether or not the online world is trustworthy*,* and the *Context* in which trust or distrust is enacted.

The modified prototypes of the OWS and TI are currently being tested in a large online study. This will allow both validation of the scales and large-scale examination of the role of trust in online wellbeing. As well as testing the internal reliability of the subscales, a principal components analysis will be carried out with other calculations of dimensionality to ensure that the subscales measure individual factors within trust and wellbeing. This will be especially important in considering the *Context* subscale of the TI. The results of the online study will also help lead to recommendations for ways in which online platforms can build user trust into their systems. At the same time, using the TI for reflection and empowerment will be explored with users and other stakeholders, including investigating ways to present results that are meaningful and engaging, and how this and other tools can encourage meaningful dialogue between the two groups.

ACKNOWLEDGEMENTS

This work was supported by EPSRC [grant number EP/R033633/1]. Elvira Perez Vallejos also acknowledges the financial support of the NIHR Nottingham Biomedical Research Centre.

REFERENCES

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, *89*, 279-288. https://doi.org/10.1016/j.chb.2018.07.026

Alexandrova, A. (2005). Subjective Well-Being and Kahneman's 'Objective Happiness'. *Journal of Happiness Studies*, *6*(3), 301-324.

Bhattacherjee, A. (2002). Individual Trust in Online Firms: Scale Development and Initial Test. *Journal of Management Information Systems*, *19*(1), 211-241. https://doi.org/10.1080/07421222.2002.11045715

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Chadborn, N. H., Blair, K., Creswick, H., Hughes, N., Dowthwaite, L., Adenekan, O., & Pérez Vallejos, E. (2019). Citizens' Juries: When Older Adults Deliberate on the Benefits and Risks of Smart Health and Smart Homes. *Healthcare*, *7*(2), 54. https://doi.org/10.3390/healthcare7020054

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., Duriez, B., Lens, W., Matos, L., Mouratidis, A., Ryan, R. M., Sheldon, K. M., Soenens, B., Van Petegem, S., & Verstuyf, J. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, *39*(2), 216-236. https://doi.org/10.1007/s11031-014-9450-1

Creswick, H., Dowthwaite, L., Koene, A., Perez Vallejos, E., Portillo, V., Cano, M., & Woodard, C. (2019). "… They don't really listen to people": Young people's concerns and recommendations for improving online experiences. *Journal of Information, Communication and Ethics in Society*, *17*(2). https://doi.org/10.1108/JICES-11-2018-0090

DCMS. (2019). *Online Harms White Paper* [White Paper]. DCMS. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

Deci, E. L., & Ryan, R. M. (2000). The 'what' and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227-268.

Deci, E. L., & Ryan, R. M. (2008). Hedonia, eudaimonia, and well-being: An introduction. *Journal of Happiness Studies*, *9*, 1-11.

Diener, Ed, Heintzelman, S. J., Kushlev, K., Tay, L., Wirtz, D., Lutes, L. D., & Oishi, S. (2017). Findings all psychologists should know from the new science on subjective well-being. *Canadian Psychology/Psychologie Canadienne*, *58*(2), 87-104. https://doi.org/10.1037/cap0000063

Diener, Ed, Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research*, *97*(2), 143-156. https://doi.org/10.1007/s11205-009-9493-y

Dodge, R., Daly, A., Huyton, J., & Sanders, L. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing*, *2*(3), 222-235. https://doi.org/10.5502/ijw.v2i3.4

Elgar, F. J., Davis, C. G., Wohl, M. J., Trites, S. J., Zelenski, J. M., & Martin, M. S. (2011). Social capital, health and life satisfaction in 50 countries. *Health & Place*, *17*(5), 1044-1053. https://doi.org/10.1016/j.healthplace.2011.06.010

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics* (3rd ed.). Sage.

Gagné, M. (2003). The role of autonomy support and autonomy orientation in prosocial behavior engagement. *Motivation and Emotion*, *27*(3), 199-223.

Gefen, D., Benbasat, I., & Pavlou, P. (2008). A Research Agenda for Trust in Online Environments. *Journal of Management Information Systems*, *24*(4), 275-286. https://doi.org/10.2753/MIS0742-1222240411

Gefen, Karahanna, & Straub. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51. https://doi.org/10.2307/30036519

Goh, D. H.-L., Pe-Than, E. P. P., & Lee, C. S. (2017). Perceptions of virtual reward systems in crowdsourcing games. *Computers in Human Behavior*, *70*, 365-374. https://doi.org/10.1016/ j.chb.2017.01.006

Gui, M., Fasoli, M., Carradore, R., & Carradore, R. (2017). “Digital Well-Being”. Developing a New Theoretical Tool For Media Literacy Research. *Italian Journal of Sociology of Education*, *9*(1), 155-173. https://doi.org/10.14658/pupj-ijse-2017-1-8

Heintzelman, S. J. (2018). Eudaimonia in the Contemporary Science of Subjective Well-Being: Psychological Well-Being, Self- Determination, and Meaning in Life. In E Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being* (p. 14). DEF Publishers.

Helliwell, J. F., & Wang, S. (2011). Trust and Wellbeing. *International Journal of Wellbeing*, *1*(1), Article 1. https://www.internationaljournalofwellbeing.org/index.php/ijow/article/view/9

Iordache, C., Mariën, I., Baelden, D., & Baelden, D. (2017). Developing Digital Skills and Competences: A Quick-Scan Analysis of 13 Digital Literacy Models. *Italian Journal of Sociology of Education*, *9*(1), 6-30. https://doi.org/10.14658/pupj-ijse-2017-1-2

Jirotka, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M. (2017). Responsible research and innovation in the digital age. *Communications of the ACM*, *60*(5), 62-68. https://doi.org/10.1145/3064940

Jovanović, V. (2016). Trust and subjective well-being: The case of Serbia. *Personality and Individual Differences*, *98*, 284-288. https://doi.org/10.1016/j.paid.2016.04.061

Kidron, B., Evans, A., & Afia, J. (2018). *Disrupted Childhood: The Cost of Persuasive Design* (p. 47). 5Rights Foundation.

Knowles, B., & Hanson, V. L. (2018). Older Adults’ Deployment of ‘Distrust’. *ACM Transactions on Computer-Human Interaction*, *25*(4), 1-25. https://doi.org/10.1145/3196490

Kozan, H. İ. Ö., Baloğlu, M., & Kesici, Ş. (2019). The Role of Personality and Psychological Needs on the Problematic Internet Use and Problematic Social Media Use. *Addicta: The Turkish Journal on Addictions*, *6*(2), 20.3-219.

Kramer, R. M. (1999). Trust and Distrust in Organisations: Emerging Perspectives, Enduring Questions. *Annual Review of Psychology*, *50*(1), 569-598. https://doi.org/10.1146/annurev.psych.50.1.569

Livingstone, S., Haddon, L., Görzig, A., & Ólafsson, K. (2010). *Risks and Safety for children on the internet: The UK Report*. EU Kids Online.

Martela, F., & Sheldon, K. M. (2019). Clarifying the Concept of Well-Being: Psychological Need Satisfaction as the Common Core Connecting Eudaimonic and Subjective Well-Being. *Review of General Psychology*, *23*(4), 458-474. https://doi.org/10.1177/1089268019880886

Masur, P. K., Reinecke, L., Ziegele, M., & Quiring, O. (2014). The interplay of intrinsic need satisfaction and Facebook specific motives in explaining addictive behavior on Facebook. *Computers in Human Behavior*, *39*, 376-386. https://doi.org/10.1016/j.chb.2014.05.047

McDool, E., Powell, P., Roberts, J., & Taylor, K. (2016). *Social Media Use and Children's Wellbeing* (Working Paper No. 201601; Sheffield Economic Research Paper Series). Department of Economics, University of Sheffield.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. *The Academy of Management Review*, *23*(3), 473-490.

Milyavskaya, M., & Koestner, R. (2011). Psychological needs, motivation, and well-being: A test of self-determination theory across multiple domains. *Personality and Individual Differences*, *50*(3), 387-391. https://doi.org/10.1016/j.paid.2010.10.029

Peters, D., Calvo, R. A., & Ryan, R. M. (2018). Designing for Motivation, Engagement and Wellbeing in Digital Experience. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00797

Royal Society for Public Health. (2017). *#StatusOfMind: Social media and young people's mental health and wellbeing*. Royal Social for Public Health and Youth Health Movement. https://www.rsph.org.uk/uploads/assets/uploaded/62be270a-a55f-4719-ad668c2ec7a74c2a.pdf

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68.

Ryan, R. M., & Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness* (1 edition). Guilford Press.

Ryan, R. M., Huta, V., & Deci, E. L. (2006). Living Well: A self-determination theory perspective on eudaimonia. *Journal of Happiness Studies*, *9*, 139-170.

Sheldon, K. M., & Hilpert, J. C. (2012). The balanced measure of psychological needs (BMPN) scale: An alternative domain general measure of need satisfaction. *Motivation and Emotion*, *36*(4), 439-451. https://doi.org/10.1007/s11031-012-9279-4

van der Werff, L., Legood, A., Buckley, F., Weibel, A., & de Cremer, D. (2019). Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review*, *9*(2-3), 99-123. https://doi.org/10.1177/2041386619873616

Wang, C., Hsu, H.-C. K., Bonem, E. M., Moss, J. D., Yu, S., Nelson, D. B., & Levesque-Bristol, C. (2019). Need satisfaction and need dissatisfaction: A comparative study of online and face-to-face learning contexts. *Computers in Human Behavior*, *95*, 114-125. https://doi.org/10.1016/j.chb.2019.01.034

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070. https://doi.org/10.1037/0022-3514.54.6.1063

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129-166. https://doi.org/10.1007/BF02249397

Yoon, S.-J. (2002). The Antecedents and Consequences of Trust in Online-Purchase Decisions. *Journal of Interactive Marketing*, *16*(2), 47-63.

VIRTUOUS JUST CONSEQUENTIALISM:
EXPANDING THE IDEA MOOR GAVE US

Olli I. Heimo, Kai K. Kimppa

University of Turku (Finland)

olli.heimo@utu.fi; kai.kimppa@utu.fi

ABSTRACT

We find Moor's paper on just consequentialism from 1999 to be an interesting take on how developers can process ethical issues in developing software. However, we think it stops short from its goal by only mentioning virtue ethics. Thus, we integrate virtue ethics into the model in this paper. We find that habits based on compassion through virtual virtue friendship can fill the gap and create better developers; developers who habitually take into account the users, targets and organisations using the software they develop.

KEYWORDS: Virtue ethics, Just Consequentialism, Ethics, Aristotle, Moor.

1. INTRODUCTION

As an IT professional, one has power over others through the decisions one makes. These decisions do not only create possibilities to create value through work or entertainment, but also value through moral decisions by allowing or limiting the growth of the users' characters. The decisions made in the system design (e.g. UI, functionalities, and communication methods), when designed correctly, can affect the character building process of the user by allowing, denying, and most importantly of all supporting certain actions. Therefore, the developer can promote virtues or vices and affect in the development of the users' character with the choises in design. (Heimo et al., 2018)

The IT professional can hence be in the position of a virtual virtue friend (see Heimo et al., 2018) – for a person whom they will very probably never meet nor whom will never hear about the professional– and support the virtuousness of that user through the decisions they make in the design, e.g. encouraging honest and truthful actions in multiplayer computer games. Yet, as is evident, being virtuous is a sort of vague moral guideline – especially when these persons do not know much about each other, and therefore more specific instructions and short time aims should be clarified.

Hence we turn to Moor's just consequentialism which in turn can evaluate specific situations in everyday life. It is rather easy compared to the virtuous to see and examine ones motivations and consequences. Thus we want to mate Aristotle (even though through more modern interpretations) to Moor. One can make a virtue out of being Moorean by making a habit of having ones motivations just, and evaluating ones actions, whether they have just consequences. To extend this to the role of an IS/IT designer to be a "virtue friend" in a "virtual world", to support the users' possibilities in acting for just motivations and just end results (on virtual friendship elsewhere see e.g. Briggle, 2008 or Elder, 2014, although their handling is more direct than ours). However as we clarify, this does not mean forcing the motivations nor forcing the desired consequences but, rather, as a friend supporting others develop their character to a more Moorean view.

2. JUST CONSEQUENTIALISM

James Moor (1999) in his paper on just consequentialism and computing says that he is approaching the topic from both deontological and consequentialist perspectives – and thus he indeed does. He uses a Rawlsian approach of justice and consequentialist considerations to build a framework through which a developer can evaluate the function of their application. If an action is both justified and its consequences are good, the function of the application is also good. In this paper we claim this is not quite enough. We intend to show that combined with an interpretation on friendship from Aristotle the argument can be strengthened.

Moor (1999) is aiming to create a practical guideline to follow for systems developers – consider whether the system you are building is just and whether the consequences of the system are good; and this is what he does in his paper. Unfortunately, that can lead to a mechanical, rather than an internalised method of evaluating whether a solution is good. We want to go a step further, to build an Aristotelian character for the designer, which automatically does the right thing, rather than needs to stop and consider every choice based on a two-by-two table: “is this just?”, “are the consequences good?” As Moor (1999, p.65) says “Policies are rules of conduct ranging from formal laws to informal, implicit guidelines for action.” We fear the latter, implicit action, cannot be reached by just consequentialism, and thus emphasize the building of character through automated habits, which the designer will internalise as well as they have internalised the use of charts, design methods and coding. Namely, the kind of thinking Moor (1999, p. 65) calls for, “Every action can be considered as an instance of a policy – in this kind of situation such and such action is allowed or required or forbidden” we want to “automate” through a habit, and later through building a character via a virtue ethics approach.

We applaud Moor’s (1999) goal of finding a unified theory, but think virtue ethics provides a more solid basis for this than the just consequentialism he proposes, although it obviously has its merits, which we propose to be integrated into our model. We think a character based approach will solve the issues Moor raises; being a holistic perspective not tied to either the justness or the consequences alone. Of course, even the person with the best of characters is likely to make *mistakes*, but not out of malice, but out of situational limits on understanding. Thus, we do not see our proposed theory as contradicting Moor’s ideas, but rather taking the unification of ethical theory – at least when it comes to computer ethics – a step closer to the goal.

Even though Moor (1999 p. 66) takes some preliminary steps towards this thinking in his paper, merely mentioning Aristotle’s concept of “human flourishing” and claiming it necessary is not enough, we need actual tools to reach this end. In the next chapter we argue that the developer becoming a virtual virtue friend of the users, targets – even the organisations – which the systems are designed for is the necessary, and yet unstudied next step towards this goal.

3. VIRTUE ETHICS

Aristotle states that to reach Eudaimonia, one must be virtuous in their everyday life by fulfilling his or her telos by achieving virtues, avoiding vices and developing their character. This is how a person flourishes in their life. Developing character by implementing virtues as a part of everyday life, following virtues, acting virtuously and thus making the virtuous acts as a natural aspect of ones actions is a trait shared by good people. (Heimo, 2018; Heimo et al., 2018.) Vallor (2013) states that moral skills are necessary for moral virtue:

> Someone could have moral skills in the sense of practical moral knowledge but fail to be virtuous because they are unreliable in acting upon this knowledge, or because they act well only for nonmoral reasons. Still, moral skills are a necessary if not a sufficient condition for moral virtue. Without the requisite cultivation of moral knowledge and skill, even a person who sincerely wishes to do well consistently and for its own sake will be unsuccessful.

Differing from other versions of normative ethics, virtue ethics does not generate a set of norms but rather guidelines both encouraging people to avoid vice and promote the virtues. The normativity is in the level of ideas rather than a set of strict rules governing our daily actions. True value, according to Aristotle, is only generated through virtues – and vice versa with vices. In Aristotelian sense a character is not virtuous by following virtue alone, since one might follow virtue reluctantly and in the face of temptation but rather, when a person automatically aims toward all virtues, the character can become virtuous. (EN I, 9 – 10; 1098a, 15 – 21; 1098b 5 – 30; 1100a31 – 1101a21, II, 1; 1103a31 – 1103b25, 1104a10 – 1105a16; McPherson, 2013). Or, as Vallor (2009) explains:

> […] the moral development of individuals cannot be assessed or predicted simply by looking at what they think, feel or believe—we also have to know what kinds of actions they will get in the habit of doing, and whether those actions will eventually promote in such persons the development of virtues or vices.

Therefore being virtuous is a life choice rather than a situational choice and only through life choices can a happy and good life be enjoyed. Yet socially valued virtues might not equal ethical virtues (Beauchomp & Childress 2001, p. 27). Humans are often expected in e.g. at work to follow socially valued virtues even when they conflict with their personal moral virtues. (e.g., Murphy, 1999). Excelling at one's work does not equal being virtuous. The human life as a whole, not divisible into parts that can ignore other parts, must be taken in accordance and constructed virtuously (MacIntyre, 2004, pp. 240 – 241; 2007, p. xv).

Yet only higher-level abstraction can be derived from Aristotelian virtue ethics thinking – be brave and do not act cowardly, foolhardily, or rashly! Churchland (2011, p. 115) however turns the higher abstraction-level as a favourable position to virtue ethics with the higher-level concepts – ideas rather than rules – as they tend to work better with the human mind. With this Churchland implies that *there cannot be universal categorical laws* which offer us the right from wrong all the time. We are more able to understand the right from the wrong – truthful from the false – from ideas than via a complex set of norms or rules. (Churchland, pp. 114-116.)

According to Aristotle the actions, not mere words define the honesty and truthfulness but moreover the life itself. Repeated practice over time in truthfulness leads the person to the concept of honesty and to see the value in it. Or as Vallor (2013) explains, those who have cultivated themselves in the virtue of honesty have "[…] learned how to excel at truth-telling in any situation that might arise: who to tell the truth to, when and where, in what way, and to what extent."

MacIntyre (2004) states that to interpret Aristotle, humans must not just study Aristotle, but moreover comprehend that he wrote for his time. In that time the world, linguistics, and culture varied from what we know and to understand Aristotle is in connection in understanding ancient times as social norms and social actors are different. Even the meanings of the words have changed. To understand

and interpret the virtues required in relation to modern world is a key element in understanding Aristotle (MacIntyre, 2004).

To understand how the virtues in that society were seen is a key element in understanding Aristotle since he aimed to be understood by other educated Greeks, not by barbarians who live in the Northern part of the continent and over two millennia later. According to MacIntyre (2004), language and society are bound to each other and we live in rather different society than the ancient Greeks did. Hence to comprehend the virtues and vices we should focus on the translation and modernisation of the term and understand the virtue itself. We should aim to understand how to be a good person.

Interpreting Aristotle is clearly not the easiest task and the referred authors (Vallor, MacIntyre and Churchland) with many other prominent philosophers give seemingly contradictory results. Yet, the key idea behind Aristotelian virtue ethics in general – one promoting virtues and avoiding vices – works rather well, at least in this case. The limitations such as the meaning of virtue and what can be interpreted as a virtue can be kept in margins for they clearly are problems belonging to meta-ethics. The focus of this paper is in the aim of developing character, which none of these authors seem to contradict with.

4. VIRTUAL VIRTUE FRIENDSHIP

Aristotelian concept of friendship requires as bearing mutually recognized good will and wishing well for each other. [EN, VIII, 2-3, 1155b30-1156a10] This demand is hard if not impossible to pursue in the situation where the friends are not aware of the other as a person but only as an entity. However, it should not be impossible for one to act with good will and wishing well in modern times where digitalization has made everyday life increasingly virtual. Let us call that *virtual friendship;* friendship from one person to a group or an entity of people which they may or may not meet but whom their professional decisions affect.

It is sort of friendship between host and guests [see EN, VIII, 3, 1156a30-32] but not limited to it. It could be motivated by pleasure [see e.g. EN, VIII, 2-3, 1156a1-1156b32], but that is not necessary. In addition, joint benefit must not be symmetrical in status and those in better positions should act accordingly [EN VIII, 13, 1162a33-1162b3]. The benefit from it is therefore sometimes asymmetrical [EN VIII, 13, 1162b3-1162b4] – e.g. money, status, career etc. for the developer and possibilities to work, pleasant pastime and social interaction for the user – yet the friendship is beneficial to both parties. For example, the most pleasurable game developing experience can and should come when the audience is enthusiastic about the game – the developer has done good work and the pleasure affected to the gamers gives him or her joy, for then the developer knows they have succeeded in making a good product and thus being virtuous in work.

But to achieve the highest level of friendship requires the element of virtue where the friends trust the other not to intentionally harm them, act against them or aim only to benefit them even though the element of benefit is duly present. Most of all virtue friends aid each other by helping each other, offering kindness and companionship, and aiding the development of their character. [EN, IX, 11-12, 1171a22-1172a15]

Thus as an IT professional being a virtue friend to the users – those dependent on their decisions – seems just and virtuous. To support them, not to force them or be indifferent about them, to make a habit of having their motivations just, evaluating their actions through not just motivations but also through the consequences and steer their habits towards what they themselves have learned and discovered to be just. As friends treat others whose virtuous actions they try to support, so that their motivations become just and habitual, so that the consequences of their acts be just and the

Promethean values meet the Epimethian thinking to promote the habit of creating iterations of more just consequences.

Thus, it seems to us that plain just consequentialism will not, unfortunately, give us the results computer ethics aims for in practice. Rather, instead of following an algorithm in which the developers consider whether an act is both just and that its consequences are good, they must become a kind of virtual virtue friends of the users, targets and the organisations which will be using or targeted by the systems they design, lest they just "tick the boxes", of which Professor Emeritus Gotterbarn always warns us of. Building the right kind of character is of paramount importance; making it a habit to always consider the benefits – and possible draw backs – of the systems developed must become second nature to the developers. And this, we claim, can only be achieved through a virtue ethics approach where the developers truly internalise their connection with the users and targets of the systems they develop.

However, for the virtual friendship to be virtue friendship, a virtual virtue friendship, the good will and wishing well must actualise. The developer should not only do a good product, but they should treat the users as a good host does for their friends; as their esteemed guests. They should not lure them to the vices for a short time benefit but to promote the virtuous development of those they are being the host to, but to aid the friend to achieve, thus improving the character of both and making both more vigilant in their respective virtues.

5. DISCUSSION & CONCLUSIONS

Therefore it seems that virtual virtue friendship is a set of actions where a developer acts as a friend should act towards the users. The point that Aristotelian thinking in how to build character through virtues and virtues through friendship is still sound today – when it is interpreted through modern thinkers, rather than just the original, somewhat era dependent values (see e.g. MacIntyre 2004).

It seems to be obvious that we do not want to treat everyone as we treat our closest friends. We do not want to share our lives, health (see e.g. Wahlstrom, Fairweather & Ashman, 2011), possessions or time with everyone, but with a selected few we consider close and trustworthy. Virtual virtue friends are just that – a group of people we do not so much share our life with, but we treat as virtue friends. Yet, to promote the virtues of others, helping them develop their character and aiding others while doing the work we do seems to be good from the viewpoint of what is just, and from the viewpoint of where the consequences, at least in the large scale, could be beneficial.

Whereas consequentialist and deontological models treat the user more as a target of the moral action and asserts the developer with a strict set of dos and don'ts, virtue ethics brings out the deeper connection between the two. As the relationship is viewed more as of that of friends, the users become more as those we wish well and aim to help in the world – and through that the world itself becomes a better place.

Adding virtue ethics to the Moorean thinking therefore strengthens the multi-ethical approach but also makes it more complicated. It requires not only the understanding of what is just, or the capability to see the consequences behind the actions, but those two and the capability to understand the users, their needs, wants, and situation in life. Most of all it requires the understanding of the product, the market, and the user groups, which is a rather gruesome task to accomplish on its own. Yet again, a virtuous developer must aim to understand those tasks even if difficult and therefore should be able to develop their character towards the virtue of virtual virtue friend. It is not an easy task, but to reach higher understanding of virtues, by not only fulfilling the just and proper consequences, but also by treating people in such manner that they are more able to avoid vices and develop their virtues, a developer can aid other human beings to be the best versions of themselves possible. This is optimal both a priori and a posteriori.

REFERENCES

Aristotle (NE). (Circa 350 BCA). Nicomachean ethics. Several translations used.

Beauchamp, T. L., & Childress, J. F. (2001). Principles of biomedical ethics. Oxford University Press, USA.

Briggle, Adam (2008) Real friends: how the Internet can foster friendship, Ethics and Information Technology, 10:71-79.

Churchland, P. S (2011) Braintrust: What Neuroscience Tells Us about Morality, Princeton University Press, 2011.

Elder, Alexis (2014) Excellent online friendships: an Aristotelian defense of social media, Ethics and Information Technology, 16:287-297.

Heimo, Olli I. (2018). *Icarus, or the idea toward efficient, economical, and ethical acquirement of critical governmental information systems.* Ph.D. Thesis, University of Turku. https://www.utupub.fi/handle/10024/146362

Heimo, Olli I., Harviainen, J. Tuomas, Kimppa, Kai K. & Mäkilä, Tuomas (2018) *Virtual to Virtuous Money: A Virtue Ethics Perspective on Video Game Business Logic*, Journal of Business Ethics, Springer.

MacIntyre A. (2004). Hyveiden jäljillä: Moraaliteoreettinen tutkimus. (2nd ed., After Virtue: A Study in Moral Theory. Translated by N. Noponen) Gaudeamus, Helsinki.

MacIntyre A. (2007) After Virtue: A Study in Moral Theory (3rd ed., 2011 imprint), Bloomsbury Publishing Inc, London.

McPherson D. (2013). Vocational Virtue Ethics: Prospects for a Virtue Ethic Approach to Business. Journal of Business Ethics, 116(2), 283–296.

Moor, J. H. (1999). Just consequentialism and computing, Ethics and Information Technology. 1: pp. 65–69.

Murphy, P. E. (1999). Character and virtue ethics in international marketing: An agenda for managers, researchers and educators. Journal of Business Ethics, 18(1), 107-124.

Vallor, S. (2009), Social networking technology and the virtues, *Ethics and Information Technology* (2010) 12:157-170, Published online: 11 August 2009 Springer Science+Business Media B.V. 2009, DOI 10.1007/s10676-009-9202-1

Vallor, S. (2013) The future of military virtue: Autonomous systems and the moral deskilling of the military, 2013 5th International Conference on Cyber Conflict (CYCON 2013), Tallinn, 2013, pp. 1-15.

Wahlstrom, Kirsten, Fairweather, N. Ben & Ashman, Helen (2011) Brain-Computer Interfaces: a technical approach to supporting privacy, Ethicomp 2011

2. Cyborg: A Cross Cultural Observatory

ADOPTING WEARABLES AND INSIDEABLES TECHNOLOGIES: WHAT IS THE MAIN FACTOR INFLUENCING IN MEXICAN YOUNGSTERS?

Juan Carlos Yáñez Luna, Pedro Isidoro González Ramírez, Mario Arias-Oliva, Jorge Pelegrín-Borondo

Universidad Autónoma de San Luis Potosí (México), Universidad Autónoma de San Luis Potosí (México), Universitat Rovira i Virgili (Spain), University of La Rioja (Spain)

jcyl@uaslp.mx; pedro.gonzalez@uaslp.mx; mario.arias@urv.cat; jorge.pelegrin@unirioja.es

ABSTRACT

This paper analyses the ethical perception of adopting and using wearables or insideables. We collected 152 samples from youngsters in San Luis Potosí (Mexico), most of them undergraduate students. An online survey was adapted based on (Pelegrín-Borondo & Arias-Oliva, 2017) to collect students' data and were analysed in the computational software STATA 25. We analyse the effects of two factors: social influence and performance expectancy on intention to use. Also, we analyse the effect of the "consumer innovative" on gender factor. Our results show that there is positive relation between variables, but in special those that relate to wearables. That means Mexican youngsters prefer to use wearables to insideables. Also, the study shows that there is a slight difference between the degree of innovative consumer between men and women.

KEYWORDS: Social Influence, Consumer Innovative, Ethics, Wearables, Insideables.

1. INTRODUCTION

People that were born in the las century, have never thought that technology could be an important part of their lives. Nowadays, it is a fact that youngsters adopt technology earlier (digital natives) than elderly people. People uses technologies in their day to day (Smartphone, Tablet, Computers, Internet of things, etc.) as a part of their work or as a part of their leisure activities, but, why does they do?, which are the factors that have influence on them?. An important challenge of markets in the world is to identify some of the factors that have influence in the consumer behaviour. This activity sets a framework of conditions which allow businesses to compete, innovate and create jobs;(Porter, 2008) in other words, create competitive advantages.

In this respect, globalization has enabled technology industries increase production and trade throughout the world. Most of the technological gadgets are produced in developing countries such as China, Taiwan, Chile, Brazil, etc., this allows to reduce manufacturing costs and increase its production to seek greater competitiveness. In the case of technologies focused to mobility such as Smartphones, gadgets, etc., every year companies develop new and sophisticated technologies with Internet as a common media. Companies are also working in the topics of Artificial intelligence (McLean & Osei-Frimpong, 2019), also in wearables (Fröbel, Avramidis, & Joost, 2019) and insideables or implants (Haeberle et al., 2019). Most of them requires Internet or the Smartphone to work adequately, but recent technologies works with, a set of data who interpret the environment and take appropriate decision; well known as Artificial Intelligence (AI). We can imagine a near future in which the use of devices with AI will improve the human disabilities such as physical or mental defects through a set of microcircuits implanted and managed (or not) by external devices like wearables. The acceptance of technology for improve the human abilities or disabilities is a complicated topic specially in social context. In the one hand, many individuals believe in the use of technology for transform their lives and to increase their welfare, on the other, many people make their lives in a strict order based in culture, religion or others social structures. In this regard, marketers, economists or decision takers in the business and government should study that more this topic in order to take steps that affect their economy.

1.1. Technology consumerism

During the first decades of the 21st century, technologies managed to integrate into existing information processes. The paradigm of the knowledge society was consolidated in some countries (Castro-Jaramillo, Guevara-Valencia, & Jaramillo-Rojas, 2016). The prominence of technologies in everyday life generate a new reality in societies. Mobility is now a life partner of the human being practically from birth to death.

Today's society is considered as productive and recurrent to information. All this information is transformed into knowledge. For companies, the generation of knowledge is convenient since it can establish their competitive advantages in their environment or beyond their borders. However, the consumer does not always consider that it is part of the strategies of business, of marketing to turn him into an innovative consumer.

According with Fresneda Lorente, (2019), the consumerism of electronic technologies will be increased in the 2020, a total of 20% of technology revenue is expected. Most of the consumerism in technologies focuses on wearables for health (52% of sales of wearables in the world) and the 39% will represent to health gadgets, such as Fitness bracelets or Smartwatches.

According to Garibay (2018), in Mexico 51.9% of individuals adopted and uses at least 3 gadgets; Likewise, the report of Interactive Advertising Bureau México (2019) shows that the acceptance of wearables and virtual reality grew up at least 15% in relation with previous years. We can assume that the penetration of technologies especially mobile or internet based, will continue growing exponentially, and it is possible that the consumers behaviour changes in the future. In relation with previous cited the penetration of wearables in the world has increased to 38% for people between 25-34 years old (Escamilla, 2019).

The implants business in Mexico focus mainly in Cosmetic and Health context. According with Rios Montanez (2020) in Mexico, the number of "cosmetic" procedures (surgical and non-surgical) increased by approximately 32% between 2014 and 2018. The technological implants in Mexico is not common as other kind of surgical intervention, this may due to certain factors such as, expensive technology, expensive surgeries, lack of knowledge about the topic or culture and religion impediments.

The consumption of products is a fundamental part for jump-starting the markets in order to rise an economic development sustainable, however most of Mexican do not have the financial solvency for acquiring forefront technology (gadgets – wearables, implants or mobile) due principally to additional duties of importation and foreign i+D added costs. Most of the technology acquired in Mexico is imported from different countries in which has trade agreements.

In their last meeting (in México), the OECD countries established certain objectives in order to increase the digital transformation of services in each country (OECD, 2017). In the case of México, the amount for invest in Technology and Innovation is less than 1% of the Gross Domestic Product (GDP) in comparison with others OECD countries that invest more than 20% (Camhaji, 2017). This situation leads to economic stagnation and under development. Consequently, the growth of the country will diminished for a lack of knowledge development; and as is augmented in Cabrero Mendoza (2017) "The knowledge-based economy refers to the ability to generate scientific and technological knowledge, which allows to be more competitive, grow more, and transform the economy to achieve higher levels of social welfare".

Mexican government approved fiscal incentives to facilitate the consumerism of technology in all economic sectors. Those incentives could provide facilities to companies for save almost 94% of the investment in technology (Neuman, 2017). But for individuals to acquire forefront technology is still expensive, Mexican (specifically youngsters from mid-sized class) decide for purchasing cheaper technology such as low range wearables. Despite cuts in the budget, Universities and Research Institutions in Mexico have been working in i+D. The principal aims in the research is the generation of biomaterials that could impact in the individuals' needs (Manjarrez Nevárez et al., 2017).

The proposal of this research is to analyse the perceptions about the acceptance of wearables or technological implants (insideables) by Mexican citizens. We also consider how the transhumanism concept influences in the consumer consumption of technologies and how influences in their ethical behaviour.

2. ACCEPTANCE OF TECHNOLOGY

The acceptance of products developed by an industry is an important element for the improvement of market strategies. The technology industry is characterized by having high demand; however, it is not an industry characterized by having affordable prices, at least during its first years on the market. This is because in many cases technological products can be considered as luxury products, that is, they are not necessary goods. The economy explains through the law of supply and demand the behaviour of the markets (the interaction between buyers and sellers). This model describes the effects that exist between the price of a good and the relationship between the availability and the degree of demand for that good in the market. However, for there to be a demand, that product may need to be well diffused.

According to the theory of diffusion of innovations (Rogers, 1983), an innovation will depend on various factors, such as ease of use, price, return on investment, expected benefits, among others. The importance of investigating these factors is to determine the degree of adoption that a product will have in the market and in each time. But it is particularly complicate to measure the degree of adoption between Period 1 (actual) and Period 2 (near future). Most organizations aim to market their products or services. However, not all companies are conceptualized in this regard. The most widely accepted classification in the literature is in accordance with the degree of innovation or adoption of technologies, such as pioneers, followers and reluctant (Aguilar Jimenez, Gamboa Pico, & Rueda Díaz, 2011). Currently, the pioneering companies are threatened by the dynamism of the market. This dynamism considerably reduces the market leadership of pioneers (Audretsch & Callejón Fornielles, 2007), therefore, achieving rapid profits is fundamentally aware of the challenges in the markets.

On the other hand, the production of technologies and the interaction with telecommunications have opened new business environments. The health and cosmetic market is one of the majority covered in terms of implants (Foerster, Cantu, Wildman, & Tuck, 2019; Wei, 2014). From a business perspective, companies must take advantage of the aspect that covers social thought, in which human beings always seek to satisfy their needs for well-being and comfort (Velázquez Fernández, 2009).

In this regard, there are certain factors that influence consumer behaviour and that have direct implications for their decision to adopt a technology. Various studies have been carried out on the acceptance of technology under different contexts, for example, the social influence, utility / application and ease of use are the most common in this type of study (Fred D. Davis, 1989; Venkatesh & Bala, 2008), other researchers seek to observe hedonic aspects (McLean & Osei-Frimpong, 2019), ethical and moral aspects (Gauttier, 2019; Pelegrín-Borondo, Arias-Oliva, Murata, & Souto-Romero, 2018), consumer perceptions of risk (M.-C. Lee, 2009; Shin, 2010), and innovation (Murata, Arias-Oliva, & Pelegrín-Borondo, 2019).

The academic literature has covered different fields of study of consumer behaviour. Y.-H. Lee, Hsieh, & Hsu (2011) considered an early introduction of new technologies may generate several degrees of uncertainty in the individual. The uncertainty is an element for measure the perceived risk and trust in the individual in the purchase decision (Pelaez, Chen, & Chen, 2019). In this regard, it is usually observed that the perceived risk can be a factor that have implications with those considered important, such as the Perceived trust, perceived utility, or the intention of use (M.-C. Lee, 2009; Shin, 2010). Although the perceived risk is directly related with the previous factors, it is important for the theoretical literature to develop studies of how morality and ethics influences in the perceived risk.

As discussed above, understanding consumer influencing factors is important in behavioural studies. Other areas such as psychology and sociology have also contributed to the theoretical context arguing intrinsic and extrinsic motivators. According to F. D. Davis, Bagozzi, & Warshaw (1989) extrinsic motivators influence behaviour with a proportional amount of effort, intrinsic motivators have influences on the performance of an activity and there is no effort present. In this type of model, the variables are regularly classified according to their order of motivation, pouring their influence on the perceived utility or on the intention of use.

Sociology, in general, indicates that the behaviour of an individual lies in the degree of satisfaction of their needs, such as social needs and the influence with the regulations of the individual. Social norms, Image, Social Influence, are factors that are regularly applied in the same context, Venkatesh, Morris, Davis, & Davis (2003) indicate that Social Influence is defined as: "the degree to which an individual perceives that important others believe he or she should use the new system ". Acceptance studies have been carried out of technologies that evaluate the relationship between social influence and other factors such as facilitating conditions (Koo & Chung, 2014; Sugarhood, Wherton, Procter, Hinder, & Greenhalgh, 2014; Zhou, Lu, & Wang, 2010),their results indicate a positive influence on relationship with the intention of use.

The theory of technology acceptance indicates that the individual expects that the adopted technology will fully satisfy his need. This satisfaction will generate a useful perspective or a perception of improvement in the performance of their activities. One of the factors that measure this satisfaction is the expectation of performance, which is defined as the rate at which an individual considers that using a technology will help them to obtain profits in their job performance (Venkatesh et al., 2003). The implications of adopting a technology will have to be denoted. This factor can be studied from a social perspective and a business perspective. In both perspectives, it can be considered a common element that is the development of a competitive advantage of those who are adopters against those who are not. However, from a social perspective, it is possible that other factors have positive or negative implications in the relationship, for example Social Influence or, ethical, moral, religious values, traditional beliefs, etc. Reinares-Lara, Olarte-Pascual, & Pelegrín-Borondo (2018) did not find a moderating relationship on ethical and moral factors on performance expectation. However, for this investigation a direct relationship will be determined. Zhou et al., (2010) found a direct relationship between technological adjustment activities and the expectation of performance towards the intention to use a technology, Oliveira, Faria, Thomas, & Popovič (2014) also point out that the expectation of performance is a factor that explains that the consumer is to seek extra value on the use of a technology.

The diffusion theory of innovations indicates that some individuals tend to adopt an innovation before others. However, the concept of innovation has not been fully conceptualized, and its measurement is complicated due to its hypothetical nature. Still, marketing scholars often segment individuals into "innovators" and "non-innovators" (Agarwal & Prasad, 1998). The information generated by consumers is of great importance to companies for the development of new products. Despite this, the dynamics of the markets has given way to a new type of consumer, who is referred to as "active" in the academic literature. This new paradigm describes the consumer as innovative (Hippel, Ogawa, & Jong, 2011). Based on the theory, we assume the existence of a direct relationship between the degree of innovation and the intention to use a technology, so the relationship between the factors of innovation and gender of the consumer and their intention to use will be analysed. According to Rogers (1983, p. 242), innovatively refers to the degree to which an individual is relatively earlier in adopting new ideas than other members of a system.

3. METHODOLOGY AND HYPOTHESIS

For this study we analyse specific information of 152 youngster of the city of San Luis Potosí (Mexico). Most of them are undergraduate students. The data analysed were collected by a survey instrument (Pelegrín-Borondo & Arias-Oliva, 2017). The survey was designed to obtain information about several topics such as, intention to use, performance expectancy, effort expectancy, social influence, hedonic motivation, facilitating conditions, perceived risk, ethical awareness and innovativeness.

The items in the survey were measured in most of the cases using a Likert Scale from 0 strongly agree to 10 strongly disagree. The survey was developed online by the Google docs engine and was applied online (e.g. emailed and sent through social networks). The descriptive analysis of the data was carried out using the statistical software SPSS v.25. For the comparison of groups, the R statistical software, the RStudio interface and the PLSPM packages were used (Sanchez, 2013).

With the collected data and following the state of the art we will analyse the following hypothesis:

> **H1**. The subjective norms have a significant influence in youngsters to use wearables or insideables.
>
> **H2**. The performance expectancy has a significant influence in youngster to adopt wearables or insideables.
>
> **H3**. The innovativeness of the consumer has a significant influence in youngster to use wearables or insideables.
>
> **H4**. There are differences in the innovativeness degree between male and female adopting wearables or insideables.

4. RESULTS

For the first section of the survey we introduce some demographic data to know how our sample is distributed. We identified that most representative gender were females with the 55.9% of surveyed and the 44.1% were males. The gender balance of respondents is show in Table 1.

Table 1. Contingency table of age and gender.

Gender	Age			
	18-24	25-30	30+	Total
Female	69 (57.0%)	5 (41.7%)	11 (57.9%)	85 (55.9%)
Male	52 (43.0%)	7 (58.3%)	8 (42.1%)	67 (44.1%)
Total	121 (100.0%)	12 (100.0%)	19 (100.0%)	152 (100.0%)

Source: self-elaboration-based survey data

According to the data analysis, the perception of surveyed youngsters about the adoption and use of technologies shown a high preference for wearables devices instead insideables. For example, the average of the intention to use (IU) of wearables is 7.73 and 7.46 and for insideables the average for both variables are 5.63 and 5.29. In Table 2, we can observe that the means for most of the variables applied (intention of use, performance expectancy (PE), effort expectancy (EE), social influence (SI), hedonic motivation (HM) and facilitating conditions (FC)) are greater in the wearables section than insideables section. Additionally, we can note that the variable perceived risk is the only one in where statistical mean are switched, it is probably a social factor influencing (negatively) in the perception of acceptance of insideable. Furthermore, in Table 2 we made a t-test for paired samples means between wearables and insideables, and we found statistical significance for all variables with a p-value<0.001, except in variables PE2 and PE4 with a p-value<0.1. It shows that the perception of wearables is strongly than insideables, in all variables.

Table 2. t-test paired samples IU-PE-EE-SI-HM-FC-PR.

Variables	Wearables		Insideables		t-test (paired samples)
	mean	S.D.	mean	S.D.	Pr(|T| > |t|)
IU1. I intend to use wearables/insideables	7.7303	2.26687	5.6382	3.07640	0.000
IU2. I predict that I would use wearables/insideables	7.4671	2.57064	5.2961	3.00956	0.000
PE1. I believe wearables/insideables will be useful in my daily life	7.5592	2.38298	6.2961	2.93605	0.000
PE2. Using wearables/insideables will increase my chances of achieving things that are important to me	6.3684	2.47821	6.2237	2.93460	0.510
PE3. Using wearables/insideables will help me accomplish things more quickly	7.5724	2.06068	6.6842	2.77506	0.000
PE4. Using wearables/insideables will increase my productivity	6.9605	2.33568	6.5921	2.88259	0.085
EE1. Learning how to use wearables/insideables will be easy for me	8.1645	2.05050	6.3289	2.81852	0.000
EE2. My interaction with wearables/insideables will be clear and understandable	7.7961	2.06304	6.1974	2.88646	0.000
EE3. I will find wearables/insideables easy to use	7.9671	2.20601	6.1842	2.84342	0.000
EE4. It will be easy for me to become skillful at using wearables/insideables	7.2039	2.29407	5.6250	2.84433	0.000

SI1. People who are important to me will think that I should use wearables/insideables	5.1447	2.87120	4.3750	3.06402	0.002
SI2. People who influence my behavior will think that I should use wearables/insideables	5.4079	2.68761	4.4079	2.99968	0.000
SI3. People whose opinions that I value will prefer that I use wearables/insideables	5.1974	2.80973	4.3092	3.00382	0.000
HM1. Using wearables/insideables will be fun	7.9868	2.05548	6.3355	2.69398	0.000
HM2. Using wearables/insideables will be enjoyable	7.9145	2.03917	6.0132	2.87254	0.000
HM3. Using wearables/insideables will be very entertaining	7.9868	2.11268	6.3750	2.73997	0.000
FC1. I will have the resources necessary to use wearables/insideables	6.3224	2.38817	5.2500	2.82198	0.000
FC2. I will have the knowledge necessary to use wearables/insideables	7.5921	2.12633	5.7829	2.76189	0.000
FC3. Wearables/insideables will be compatible with other technologies I use	8.1974	1.89121	6.5855	2.62551	0.000
FC4. I will be able to get help from others when I have difficulties using wearables/insideables	8.0263	2.02941	6.4474	2.55221	0.000
PR1. Using wearables/insideables is risky	4.8355	3.03935	6.7632	2.49698	0.000
PR2. There is too much uncertainty associated with using wearables/insideables	5.9276	2.74309	7.4276	2.54913	0.000
PR3. Compared to other technologies, wearables/insideables are riskier	5.2632	2.71063	7.2039	2.50123	0.000
Ho: mean(difference)= 0 Ha: mean(difference) ≠0 degrees of freedom =151					

Source: self-elaboration-based survey data

We analyse the means difference of variables for the ethical, morality, traditional and cultural opinions of surveyed. Table 3 summarize the results for a t-test for paired samples. As the Table 2, we can observe that respondents prefer wearables (higher mean) than insideables (lower mean). In this case, excluding EA8, we can reject the null hypothesis of equals means.

Table 3. t-test paired samples Ethical Awareness.

Variables	**Wearables**		**Insideables**		**t-test (paired samples)**
	mean	**S.D.**	**mean**	**S.D.**	**Pr(|T| > |t|)**
EA1. Unethical / Ethical	7.3882	2.3587	5.9803	2.6830	0.0000
EA2. Unjust / Just	7.4737	2.2492	6.2566	2.5540	0.0000
EA3. Unfair / Fair	7.3553	2.3000	5.6447	2.8177	0.0000
EA4. Not morally right / Morally right	7.3618	2.4752	5.9803	2.5436	0.0000
EA5. Not acceptable to my family /Acceptable to my family	7.5329	2.2900	5.7763	2.8054	0.0000
EA6. Culturally unacceptable / Culturally acceptable	7.5132	2.2345	5.5066	2.3303	0.0000
EA7. Traditionally unacceptable / Traditionally acceptable	6.5921	2.4423	5.0526	2.5366	0.0000
EA8. Not self-promoting for me / Self-promoting for me	6.6842	2.2856	6.1908	2.6388	0.0195
EA9. Not personally satisfying for me / Personally satisfying for me	7.5000	2.4522	6.2697	2.6091	0.0000

EA10. Produces the least utility / Produces the greatest utility	8.0132	2.2077	7.3026	2.5320	0.0003
EA11. Minimizes benefits while maximizes harm / Maximizes benefits while minimizes harm	7.4145	2.2091	6.4671	2.4681	0.0000
EA12. Violates an unwritten contract / Does not violate an unwritten contract	7.0526	2.6163	5.8947	2.5218	0.0000
EA13. Violates an unspoken promise / Does not violate an unspoken promise	7.2368	2.6488	6.0066	2.6432	0.0000
Ho: mean(difference)= 0 Ha: mean(difference) ≠0 degrees of freedom =151					

Source: self-elaboration-based survey data

Table 4 shows the correlation matrix between the variables Intention of use (IU) and Table 5 shows the Performance Expectancy (PE) and Intention of use and Social Influence (SI) correlations. From the results, we can conclude that there is a positive and significant relationship between the IU and PE, and IU and SI, both for wearables and insideables. However, we observed a stronger correlation in these variables for insideables than wearables, which suggests that although the mean of wearables is greater than insideables for each variable (as seen in Table 2 and Table 3) SI and PE determinate in a greater way the use of the insideables.

Table 4. Correlation test IU-PE.

	Wearables		**Insideables**	
	IU1	**IU2**	**IU1**	**IU2**
IU1	1	.785**	1	.866**
IU2	.785**	1	.866**	1
PE1	.587**	.601**	.771**	.728**
PE2	.455**	.534**	.791**	.736**
PE3	.552**	.611**	.716**	.665**
PE4	.486**	.551**	.669**	.645**
**. La correlación es significativa en el nivel 0,01 (bilateral).				

Source: self-elaboration-based survey data

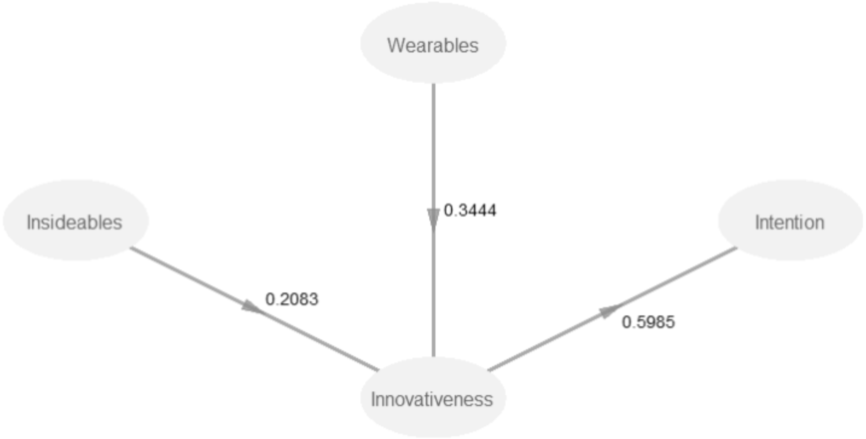
Table 5. Correlation test IU-SI.

	Wearables		**Insideables**	
	IU1	**IU2**	**IU1**	**IU2**
IU1	1	.785**	1	.866**
IU2	.785**	1	.866**	1
SI1	.345**	.353**	.588**	.611**
SI2	.404**	.391**	.563**	.609**
SI3	.373**	.380**	.588**	.609**
**. La correlación es significativa en el nivel 0,01 (bilateral).				

Source: self-elaboration-based survey data

At the top section of this work, we proposed the hypothesis H3 and H4. Based on the state of the art, we assume that some ethical, cultural, and moral factors may have implications in the degree of innovativeness of individuals. In this case, we developed a simple model based on Path Analysis theory, in which four factors are evaluated. Wearables, that contains the individuals' ethical and morality perceptions to use wearables. Similarity, the factor Insideable comprises the individuals' ethical and morality perceptions to use insideables. The third factor includes the degree of innovativeness of the individual, that means, she or he is more or less an innovative consumer of technologies, the last one covers some variables that determine the degree of intention to use a technology. Figure 1 shows the path results of the model. Even if the path analysis showed a small weight, but we can observe that there is a positive relation between wearables and insideables factors and innovativeness.

Figure 1. Global Inner model with path coefficients.



Source: self-elaboration-based survey data

To assess how relevant these results are we should examine the bootstrapped path coefficient.

Table 6. Bootstrapped path coefficients.

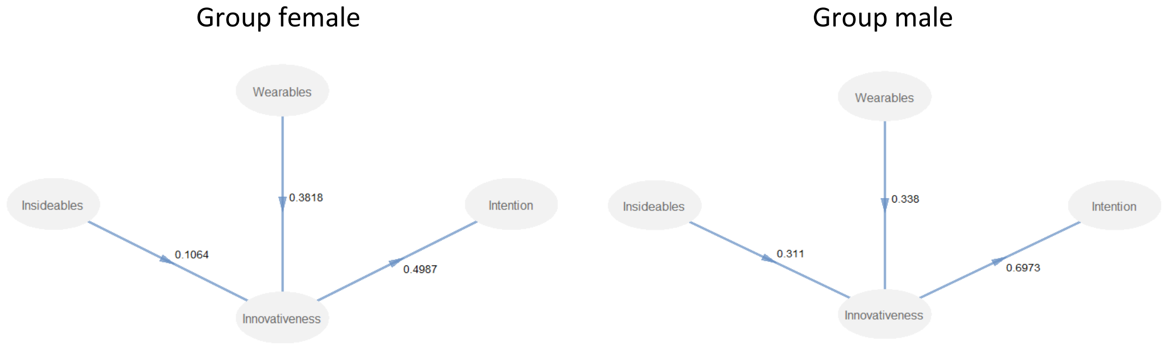
Path	Original	Mean.Boot	Std.Error	perc.025	perc.975
Wearables → Innovativeness	0.3444	0.3465	0.0897	0.1581	0.4780
Insideables → Innovativeness	0.2083	0.2187	0.0817	0.0626	0.3612
Innovativeness → Intention	0.5985	0.6018	0.0566	0.4933	0.7057

Source: self-elaboration based survey data

Table 6 summarize the Boostrap analysis for the proposed model. It shows that all path coefficients are significantly different of zero. We can observe the path Insideables → Innovativeness shows a coefficient of 0.0626 in the lowest percentile, it could indicate a less importance in the relation of ethical and moral with the innovativeness degree and it may be due the less adoption of insideables by youngsters in comparison with wearables. In general, we can say from obtained results that youngsters' innovativeness has not that much to do with the ethical/moral of using wearables or insideables. The intention to use has an important influence in the path with the innovativeness degree.

In order to evaluate the proposed H5, we calculate a new PLS regression by grouping Female and Male. We aim to know is there are significant differences between gender and the degree of innovativeness in youngsters. Sanchez (2013) points out that Path Models should be calculated separately as is shown in Figure 2.

Figure 2. Path Coefficients of Female and Male Students.



Source: self-elaboration-based survey data

Both of models shows different path coefficients as we expected due to the data collected. As in any model comparison, researchers must measure those differences detailly; that means, which is the real difference between models. There are two methods to comparing groups: the Bootstarp t-test and the Bootstrap permutation test. Matthews (2017) points out that the permutation tests is the most recommended due it has a better control of error type 1. The author suggests the use of 5000 iterations for this method. The Table 7 shows the results of the permutation analysis.

Table 7. Group comparison in PLS-PM for path coefficients.

Path	**Global**	**Group female**	**Group male**	**Diff.abs**	**p.value**	**Sig.05**
Wearables → Innovativeness	0.3444	0.3818	0.338	0.0438	0.8122	No
Insideables → Innovativeness	0.2083	0.1064	0.311	0.2046	0.2705	No
Innovativeness → Intention	0.5985	0.4987	0.6973	0.1985	0.0952	No
Based in centroid weighting scheme and 5000 iterations						

Source: self-elaboration-based survey data

As we can see from the obtained results, none of the path coefficients between females and males are significantly different.

5. CONCLUSION

In this study we evaluate some cases of perceptions in youngsters regarding the adoption and use of technologies. The aim study was to know if the youngster social values, such as ethics, culture, moral, etc have significant influence on the perception to use wearables or insideables. Our results showed that youngsters are more likely to accept and use wearables (like electronic gadgets) than insideables (like electronic implants). The study also points out that there are some factors associated to the intention to use; such as, Social Influence, Performance Expectancy among others as we mentioned before, the group study shows that there are no relation in the youngsters' innovativeness and the believes or social values, that means, as males as females are technology pioneers similarly.

We also pointed out that human being will seek to meet their needs and their wants. Commonly, technology addresses the need and the wants, but individuals normally wants all that is considered good, beneficial for them (Thomson, 1998) or whit a perceived value (Bustamante, 2015). In this study, was found that youngsters are more likely to accept and use Wearables, a possible circumstance of this fact is that technology addresses their needs towards wants more and not for a real perceived value. However, the market of gadgets is very wide and a study on specific gadget must be carried out.

One more finding in this study is the perception of using insideables is not very good, this can be due to ethical, moral, or religious reasons that today are concepts very ingrained in the country. An implication in business is that the companies developing these devices should define their market segment based on these findings and make a good marketing strategy using social influence to be able to cover a wider market.

Finally, an important finding concerns the perception of utility of the devices (implants or not). In the study it can be observed that the assertions of the respondents suggest that they would be willing to use technological implants only if they find a perceived utility. This finding has an important implication, if on the one hand the social influence determines the use of the wearables for the use of technological implants is the performance. Therefore, the market must be guided more by the sense of usefulness than fashionable.

REFERENCES

Agarwal, R., & Prasad, J. (1998). A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology. *Information Systems Research*, *9*(2), 204–215.

Aguilar Jimenez, A. S., Gamboa Pico, L. P., & Rueda Díaz, V. C. (2011). Adopción de tecnologías de información y comunicaciones en pequeñas y medianas empresas manufactureras en Bucaramanga y su área metropolitana. Una aproximación al sector de la confección. *Iteckne*, *9*(1), 42–50. https://doi.org/10.15332/iteckne.v9i1.59

Audretsch, D., & Callejón Fornielles, M. (2007). La política industrial actual: conocimiento e innovación empresarial. *Economía industrial*, (363), 33–46.

Bustamante, J. C. (2015). Use of mediating and moderating variables in explaining consumer loyalty in service environments. *Estudios Gerenciales*, *31*, 299–309. https://doi.org/10.1016/j.estger.2015.05.002

Cabrero Mendoza, E. (2017). ¿Dónde está México en ciencia y tecnología? Recuperado de http://www.jornada.unam.mx/2017/10/02/opinion/030a1pol

Camhaji, E. (2017). La ciencia, la oportunidad que México ha dejado pasar. Recuperado el 6 de enero de 2018, de https://elpais.com/elpais/2017/12/01/ciencia/1512157927_534452.html

Castro-Jaramillo, Á. M., Guevara-Valencia, S., & Jaramillo-Rojas, C. A. (2016). Análisis sociojurídico del surgimiento y expansión de las redes sociales en internet y la intimidad en Colombia. *Revista Criterio Libre Jurídico*, *13*(2), 67–78. https://doi.org/10.18041/crilibjur.2016.v13n2.26201. 67

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, *35*(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

Davis, Fred D. (1989). Perceived Usefulness , Perceived Ease of Use , and User Acceptance of lnformation Technology. *MIS Quaterly*, *13*(3), 319–340.

Escamilla, O. (2019). ¿Cómo se encuentra el mercado de los Wearables? Recuperado de https://www.merca20.com/mercado-de-los-wearables/

Foerster, A., Cantu, L. R., Wildman, R., & Tuck, C. (2019). Current Market for Biomedical Implants. En *Polymer-Based Additive Manufacturing* (pp. 97–119). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-24532-0_5

Fresneda Lorente, C. (2019). El gasto en tecnología de la información crecerá más de un 3 % hasta 2020. Recuperado de https://es.weforum.org/agenda/2017/03/el-gasto-en-tecnologia-de-la-informacion-crecera-mas-de-un-3-hasta-2020

Fröbel, F., Avramidis, E., & Joost, G. (2019). Workshop on Wearables and Machine Learning: Applications of Artificial Intelligence , Approaches on Textile Technology. En *Cooperation International Conference in HCI and UX* (pp. 177–181).

Garibay, J. (2018). ¿Será 2018 un buen año para los Wearables? Recuperado el 5 de enero de 2018, de https://www.merca20.com/sera-el-2018-un-buen-ano-para-los-wearables/

Gauttier, S. (2019). 'I've got you under my skin' – The role of ethical consideration in the (non-) acceptance of insideables in the workplace. *Technology in Society*, *56*(August 2018), 93–108. https://doi.org/10.1016/j.techsoc.2018.09.008

Haeberle, H. S., Helm, J. M., Navarro, S. M., Karnuta, J. M., Schaffer, J. L., Callaghan, J. J., … Ramkumar, P. N. (2019). Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. *Journal of Arthroplasty*, 3–5. https://doi.org/10.1016/j.arth.2019.05.055

Hippel, E. Von, Ogawa, S., & Jong, J. P. J. De. (2011). The Age of the The Age of the Consumer. *MIT Sloan Management Review*, *53*(1), 0–16.

Interactive Advertising Bureau México. (2019). *Estudio de Consumos de Medios y Dispositivos entre internautas Mexicanos*.

Koo, C., & Chung, N. (2014). Examining the eco-technological knowledge of Smart Green IT adoption behavior: A self-determination perspective. *Technological Forecasting and Social Change*, *88*, 140–155. https://doi.org/10.1016/j.techfore.2014.06.025

Lee, M.-C. (2009). Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit. *Electronic Commerce Research and Applications*, *8*(3), 130–141. https://doi.org/10.1016/j.elerap.2008.11.006

Lee, Y.-H., Hsieh, Y.-C., & Hsu, C.-N. (2011). Adding Innovation Diffusion Theory to the Technology Acceptance Model: Supporting Employees' Intentions to use E-Learning Systems. *Educational Technology & Society*, *14*(4), 124–137.

Manjarrez Nevárez, L. A., Terrazas Bandala, L. P., Zermeño Ortega, M. R., De la Vega Cobos, C., Zapata Chávez, E., Torres Rojo, F. I., … Lerma Gutiérrez, R. (2017). Biomateriales como Implantes en el Cuerpo Humano. Recuperado el 8 de enero de 2018, de http://beta.uach.mx/articulo/2017/10/20/biomateriales-como-implantes-en-el-cuerpo-humano/

Matthews, L. (2017). Applying Multigroup Analysis in PLS-SEM: A Step-by-Step Process. En *Partial Least Squares Path Modeling* (pp. 219–243). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-64069-3_10

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, *99*(May), 28–37. https://doi.org/10.1016/j.chb.2019.05.009

Murata, K., Arias-Oliva, M., & Pelegrín-Borondo, J. (2019). Cross-cultural study about cyborg market acceptance: Japan versus Spain. *European Research on Management and Business Economics*, *25*(3), 129–137. https://doi.org/10.1016/j.iedeen.2019.07.003

Neuman, G. (2017). Invierte en tecnología y deduce hasta un 94 %. ¡ Deducir o no deducir , esa es la...! Recuperado el 6 de enero de 2018, de https://www.pulsopyme.com/inviertir-tecnologia-deduce/

OECD. (2017). *OECD Digital Economy Outlook 2017*. Paris: OECD Publishing. https://doi.org/10.1787/9789264276284-en

Oliveira, T., Faria, M., Thomas, M. A., & Popovič, A. (2014). Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM. *International Journal of Information Management*, *34*(5), 689–703. https://doi.org/10.1016/j.ijinfomgt.2014.06.004

Pelaez, A., Chen, C. W., & Chen, Y. X. (2019). Effects of Perceived Risk on Intention to Purchase: A Meta-Analysis. *Journal of Computer Information Systems*, *59*(1), 73–84. https://doi.org/10.1080/08874417.2017.1300514

Pelegrín-Borondo, J., & Arias-Oliva, M. (2017). Cyborg Ethics : wearables to insideables Project working fundamentals : Project basic description (track description) Team members, 1–4.

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2018). Does Ethical Judgment Determine the Decision to Become a Cyborg?: Influence of Ethical Judgment on the Cyborg Market. *Journal of Business Ethics*, *0*(0), 0. https://doi.org/10.1007/s10551-018-3970-7

Porter, M. (2008). Las cinco fuerzas competitivas que le dan forma a la estrategia. *Harvard Business Review*, *86*(1), 58–77.

Reinares-Lara, E., Olarte-Pascual, C., & Pelegrín-Borondo, J. (2018). Do you want to be a cyborg? The moderating effect of ethics on neural implant acceptance. *Computers in Human Behavior*, *85*, 43–53. https://doi.org/10.1016/j.chb.2018.03.032

Rios Montanez, A. M. (2020). Mexico: aesthetic procedures 2014-2018. Recuperado de https://www.statista.com/statistics/1088930/mexico-cosmetic-aesthetic-procedures/

Rogers, E. M. (1983). *Duffusion of Innovations* (3rd.). New York: The Free Press. A division of Collier Macmillan Publishing Co., Inc.

Sanchez, G. (2013). *PLS Path Modeling with R*. *R Package Notes*. Berkeley: Trowchez Editions. https://doi.org/citeulike-article-id:13341888

Shin, D. H. (2010). Modeling the interaction of users and mobile payment system: Conceptual framework. *International Journal of Human-Computer Interaction*, *26*(10), 917–940. https://doi.org/10.1080/10447318.2010.502098

Sugarhood, P., Wherton, J., Procter, R., Hinder, S., & Greenhalgh, T. (2014). Technology as system innovation: a key informant interview study of the application of the diffusion of innovation model to telecare. *Disability and Rehabilitation: Assistive Technology*, *9*(1), 79–87. https://doi.org/10.3109/17483107.2013.823573

Thomson, G. (1998). Deseos y necesidades. *Ideas y Valores*, *Agosto*(107), 43–55.

Velázquez Fernández, H. (2009). Transhumanismo, libertad e identidad humana. *Thémata. Revista de filosofía*, *41*, 577–590.

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478.

Wei, J. (2014). How wearables intersect with the cloud and the internet of things: Considerations for the developers of wearables. *IEEE Consumer Electronics Magazine*, *3*(3), 53–56. https://doi.org/10.1109/MCE.2014.2317895

Zhou, T., Lu, Y., & Wang, B. (2010). Integrating TTF and UTAUT to explain mobile banking user adoption. *Computers in Human Behavior*, *26*(4), 760–767. https://doi.org/10.1016/j.chb.2010.01.013

CYBORG AS A SURGEON: A THEORETICAL FRAMEWORK FOR CYBORG ACCEPTANCE IN HEALTHCARE SERVICE

Ala' Al-Mahameed, Mario Arias-Oliva, Jorge Pelegrín-Borondo

Universitat Rovira i Virgili (Spain), Universitat Rovira i Virgili (Spain), University of La Rioja (Spain)

ala.almahameed@estudiants.urv.cat; mario.arias@urv.cat; jorge.pelegrin@unirioja.es

ABSTRACT

The interest in cyborg technology keeps growing as the human desire to improve their mental and physical capabilities becomes a common dream. Even though this technology is still under development, it has received the attention of many researchers to study the extent of human acceptance to become a cyborg. On the other hand, this research finds that it is important to investigate the possibility of human acceptance to deal with the cyborg, especially in healthcare services encounter. Accordingly, the research has developed a theoretical model for accepting healthcare services provided by the cyborg. The model has been developed based on previous studies related to models and theories of social robot acceptance, being a cyborg acceptance, and new technology acceptance in general. The proposed model assumes that Perceived Usefulness, Perceived Ease of Use, Social influence, Perceived Risk, Empathy, Trust, and Emotions (Positive and Negative emotions and Anxiety) could be the key drivers of the intention to use the proposed services.

KEYWORDS: Cyborg, Nanotechnology, Robot, Technology Acceptance, Healthcare Services.

1. INTRODUCTION

Besides cultural assumptions of what may be considered “incomplete,” “normal,” or “improved”, a belief in the ability of technology to enhance body’s natural capabilities has led to a variety of body-altering techniques that doesn’t only restore functions but may exceed what is typically considered therapeutic medical intervention. In fact, it raised a strong argument about the ability to enhance the human body according to particular needs or desires (Hogle, 2005). These body-altering techniques are used to produce the “Cyborg”, which could be defined as a cognitively or bodily enhancement of humans to form a concept called "Transhumanism". The transhumanism and technology convergence is directing the scientists’ efforts toward enhancing human social skills, health, happiness, and intelligence with higher performance than before. Technological implants, brain-computer interfaces, extension and externalization of cognitive functions, and neuronal prosthesis are some examples of Cyborg technologies (Romportl, 2015).

Greguric (2014) pointed out four research areas affecting the development of human enhancement technologies: cognitive sciences, nanotechnologies, information technologies, and biotechnologies. As well, the enhancements could be categorized into the following types:

A. Cognitive abilities enhancement: such as infrared vision, memory enhancement, decision making, and sensory perception, by using technological implants or wearable technologies.

B. Physical capabilities enhancement: such as strength, stamina, and accuracy, by using bionic technology, genetic engineering, and pharmacology.

The technological implants are defined as electronic devices that can be implanted into the human body to improve one's capabilities or for the restoration of lost functions (Pelegrín-Borondo et al., 2017). Moreover, there are different implantable devices in use for different purposes. Researchers and scientists claim that brain-machine interactions will enable humans to log onto the Internet, access different databases, talk new languages fluently and help people with failing memories. It promises to make humans fundamentally different by radically changing their capabilities. Furthermore, humans could be able to control devices remotely using their thoughts (McGee & Maguire, 2007).

The boundaries between what considered human and what considered machine is getting unclear, as technology becomes close to being embedded within the human body (Britton & Semaan, 2017). Meanwhile, human interests in reinforcing one's emotional, cognitive and physical abilities are seen as a common dream among human-being individuals, which is associated with improving the quality of human life. Some types of enhancements are already available through surgeries, wearables, pharmaceutical compounds and technological implants (Gauttier, 2018). For instance, the implanted Radio Frequency Identification Device (RFID) can transmit data numbers as pulses to be used for credit cards and door access control (Warwick, 2016). Based on their use, some of these enhancements are already accepted by society, such as cosmetic surgeries, wearables, and pharmaceuticals. However, the acceptance of technological implants is still under investigation. The innovation in biomedicine, genetics, robotics, and nanotechnology is making it possible to produce hybrid bodies that combine biological and technological parts (Kostrica, 2018; Triviño, 2015). Additionally, reducing the size of the electronic components has introduced the Nanotechnology (Nanoimplants), which are small devices that can be implanted inside the human body, to improve human physical and cognitive capabilities (Pelegrín-Borondo et al., 2016; Reinares-Lara et al., 2016). It had been seen as science fiction for using insideable technologies for healthy people to increase their innate capabilities. Nowadays, the market has accepted different types of such technologies which proves that technology keeps progressing. The expectations regarding the cyborg market are promising for a reputable business with a potentially significant impact on future technologies and human societies (Pelegrín-Borondo et al., 2018). As per Pelegrín-Borondo et al. (2017), the use of physical and technological implants to compensate physical disabilities and increase attractive power is already accepted by society. Moreover, the technological implants to increase innate human capacity are partially accepted and further investigations have been established to formulate a complete picture of users' acceptance of these technologies. In other words, the acceptance of creating cyborgs is still under investigation, as the technology itself is under development (Reinares-Lara et al., 2018). Nevertheless, the aim of this research is to investigate the acceptance of cyborg as an entity in society, its development as a technology, and how humans will perceive cyborg individuals once they become a reality. In addtion, the research is investigating the acceptance of cyborg services compared to human services, especially in the healthcare sector, and proposing a theoretical model that can identify the choice criteria of cyborg services.

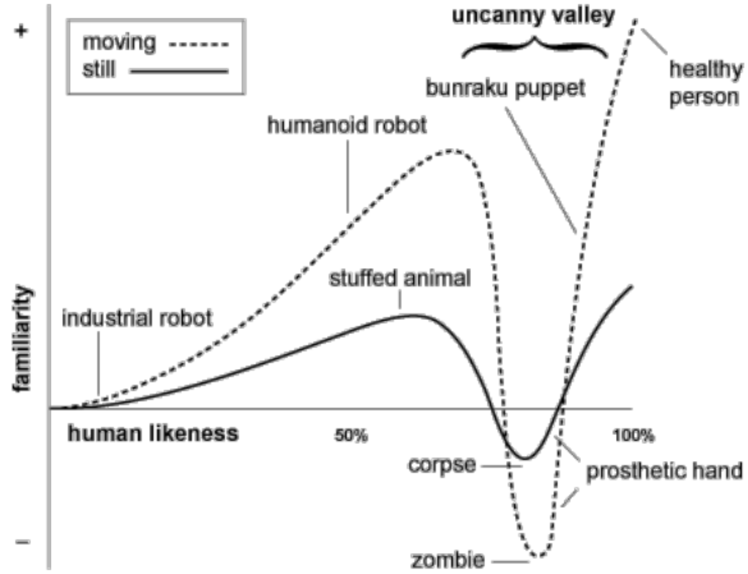
2. LITERATURE

When cyborg technology will be established, people's understanding of technology design and use must be shifted to perceive the differences between traditional technology and new technology applications (Britton & Semaan, 2017). Also, the development of these technologies is important for the future of neural prosthetic inventions. No much is known about the moral attitude of people toward the ratio between risk and benefits of using such technology and about their preferences, expectations, and needs. Furthermore, the ethical issues related to the associated risk with these technologies is an important topic that should be discussed. As well as the acceptance, which could be shifted from positive to negative state, as the use will be shifted from therapy to enhancement. For instance, the cochlear Implant could be considered a therapy device if the user has deafness. If the user has no hearing issues, then it could be considered as an enhancement. The successes of these technologies could depend on the offered benefits and people's perception of these benefits (Schicktanz et al., 2015).

The technological implants to restore physical functions and physical implants to increase seductive strength are already accepted by society (Pelegrín-Borondo et al., 2017). However, few studies have been conducted to investigate the acceptance of implants for enhancement applications and to create cyborgs (e.g. Olarte- Pascual et al., 2015; Pelegrín-Borondo et al., 2018; Pelegrin-Borondo et al., 2017; Pelegrín-Borondo, Reinares-Lara, et al., 2017; Pelegrín-Borondo et al., 2016; Reinares-Lara el al., 2018; Reinares-Lara et al, 2016). Whereas, this research is about investigating the acceptance of cyborg as an entity.

In general, humans will start to use the perceptual cues and former experiences to classify an object (e.g. Human and Cyborg) and to effectively expect their behavior. In this stage, human already recognizes the abnormality of the other human, from the physical structure (e.g. wearables) or through the behavior (e.g. implants). This stage is very important to avoid falling in “Uncanny Valley”, in which the human will feel with unfamiliarity while interacting with human-like objects (Stein & Ohler, 2017). Originally, uncanny valley theory (Fig.1) was introduced by Mori (1970) to propose the relation between human-likeness and familiarity while dealing with industrial robots. The theory proposed that, at some point (First Peak), maximum familiarity would be achieved once the robots become human-like in terms of behavior and appearance. Furthermore, the motion will enhance familiarity perception. However, the author pointed out the feel of strangeness that could drop familiarity to the negative portion, which is representing the “Uncanny Valley”. The author has mentioned an important point regarding the prosthetic hand, which is representing one of the cyborg shapes. He believed that, as the enhanced prosthetic hand looks like normal ones, humans would perceive familiarity. However, once humans figure out the abnormality of this hand, the familiarity curve will drop to the uncanny valley and humans could feel with eeriness.

Figure 1. Uncanny valley theory (Mori, 1970, p.33).



Some authors refer to the problem of trust when humans meet strangers. They mentioned the role of facial expressions in affecting trust behavior. Scharlemann et al. (2001) investigated the relationship between facial characteristics and trust while interacting with others. Their study claimed that facial expressions (e.g. smile) can stimulate trust behavior. Additionally, for life-like agents, trustworthiness could be achieved by enhancing their competence (Mulken et al., 1999). In the same context, empathy and emotions can overcome the uncanny valley's negative outcomes. Emotions have been considered as a way to distinguish humans from objects and machines. Also, the ability to express basic emotions is proof of humanity (Heisele et al., 2002). In fact, the idea is about the mismatch between human expectations and perception to avoid uncanniness. For instance, the ability of robots to express emotions will fall into "Uncanny Valley" if humans expect that a robot will not be able to do that, which in turn could produce eeriness feeling. However, their ability to experience and detect emotions without expressing them could keep them in the same area at the First Peak (Koschate et al., 2016). Moreover, as the proposed relation between humans and cyborgs will include direct interaction, it is essential to investigate the impact of anxiety on the interaction. Indeed, the expected anxiety is a reflection of the abnormality and superpower associated with cyborg technology. Factually, anxiety problem could not be related to the technology itself, rather than it could be an emergence of this negative feeling while interacting with it (Oh et al., 2017). Nevertheless, changing the attention to be toward the technology benefits could help in reducing anxiety associated with using it (Reinares-Lara et al., 2016). Meanwhile, some studies claimed that anxiety is not a significant determinant of the intention toward new technologies (Pelegrín-Borondo et al., 2017; Venkatesh et al., 2003).

As cyborg is still an outcome of technological innovations, it could be worthy to stimulate the acceptance of a cyborg throughout the acceptance of new technologies, such as robots and the acceptance to become a cyborg. In this context, different models and theories have been utilized in studying the acceptance of such technologies. The research will consider the following theories and models in studying the acceptance of Cyborg:

1. Technology Acceptance Model (TAM1) for Davis (1985) and its extensions TAM2 (Venkatesh & Davis, 2000) and TAM3 (Venkatesh & Bala, 2008).
2. The Unified Theory of Acceptance and Use of Technology (UTAUT1) for Venkatesh et al. (2003) and its extension UTAUT2 for Venkatesh et al. (2012).
3. The Cognitive-Affective-Normative Model (CAN) for Pelegrín-Borondo et al. (2016), which has been developed to study the acceptance of being a cyborg.

The Perceived Ease of Use (PEU) is one of TAM constructs that represents the effort needed to use a specific system. The second construct of the TAM model is the Perceived Usefulness (PU), which is related to the benefits associated with the use of any technology (Davis, 1985; Heijden, 2004). Humans need to perceive the usefulness of cyborg in terms of its superiority in performance if compared to human performance. Furthermore, Performance Expectancy, which corresponds to Perceived Usefulness, is related to the individual's beliefs about the system's ability to improve their job performance. And Effort expectancy, which corresponds to Perceived Ease of Use, is related to the simplicity of using the system (Venkatesh et al., 2003). It could stimulate the acceptance of dealing with cyborgs if humans find them better than the other options or stimulate the rejections if there are no differences in terms of performance and outcomes. But it is important also to consider the possibility of the low effects of these two constructs in the initial investigation of cyborg acceptance since the technology is still in its novelty stage (Pelegrín-Borondo et al., 2017).

Individuals are members of their social entity. Therefore, other members' opinions and advice toward any behavior or decision could make a difference and could direct that behavior or decision. Social influence was introduced by the Theory of Reasoned Action (TRA) for Fishbein and Ajzen (1975) and the Theory of Planned Behavior (TPB) for Ajzen (1991). And it has been used in the technology acceptance models (Davis, 1989; Venkatesh, 2000). As well, it showed a significant impact on the acceptance of Nanoimplants (Pelegrín-Borondo et al.,2015; Pelegrín-Borondo et al., 2017, 2016; Reinares-Lara et al., 2018, 2016), breast augmentation for young women (Moser & Aiken, 2011), and the acceptance of virtual customer integration (Füller et al., 2010).

In Cognitive-Affective-Normative (CAN) model, which was developed by Pelegrín-Borondo et al. (2016) to study the intention behavior toward being a cyborg, the authors used the emotional dimensions: Positive and Negative emotions. While using a specific service, customers may develop positive or negative emotions. The positive ones could be considered important for the future behavior of the customers (Pappas et al., 2013). Likewise, they could be considered important in directing the attitude of customers toward new technologies, and they could enhance the predictive power of technology acceptance models (Kulviwat et al., 2007).

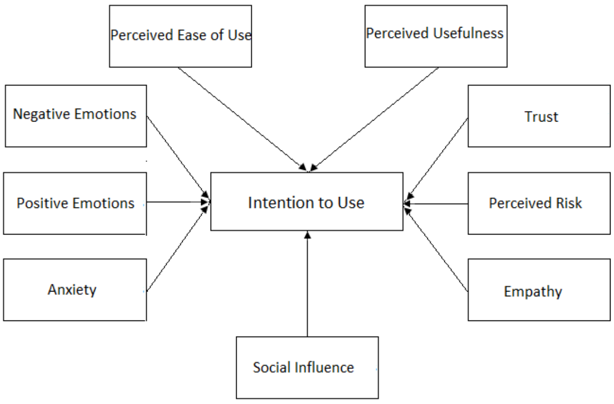
In the healthcare services sector, different studies have been investigating the acceptance of new technologies among the customers and by applying the abovementioned models and theories, such as the acceptance of electronic health systems (e-health), mobile health services (m-health) and health information systems. Some studies found the PEU as the dominant influencer on the intention behavior toward these technologies, as a direct impact (Aggelidis & Chatzoglou, 2009; Keikhosrokiani et al., 2018; Pai & Huang, 2011) or through PU and Attitude dimensions (Chow et al., 2013). However, literature is supporting the PU as the most significant determinant of the intention toward these technologies if compared to PEU (Alsharo et al., 2018; Chang et al., 2015; Chen et al., 2013; Dünnebeil et al., 2012; Hendrikx et al., 2013; Kijsanayotin et al., 2009; Lai, 2014; Dhanar et al., 2017; Phichitchaisopa & Naenna, 2013; Sezgin et al., 2017; Sun et al., 2013) and the Social Influence as well (Bawack & Kamdjoug, 2018; Chu et al., 2018; Guo et al., 2012; Hossainet al., 2019; Jaebeom Lee & Rho, 2013). Also, they have been used in studying the acceptance of wearable technologies for healthcare applications (Li et al.,2016; Nasir & Yurder, 2015; Yang et al., 2016) and in the electronic exchange of information across the healthcare sector too (Ahadzadeh et al.,2015; Chu et al., 2018; Hsieh, 2014).

Some authors have pointed out the importance of Perceived Risk in human-robot interactions. They claimed that, once users perceive risk more than benefits, they could avoid the use of robots at all (Hancock et al., 2011). However, the risk impact has been assessed through other dimensions (e.g. trust). But the need is to investigate the impact of this construct by itself and through extending the conceptual models to include Perceived Risk in assessing the intention toward such technologies (Blutet al., 2018). Because the previous studies had pointed out to risk as an outcome (or side effect) of using the new technologies, not as users' perception, and without integrating it into their research models (e.g. Destephe et al., 2015; Lilley, 2012; Matsui et al., 2018; Wirtz et al., 2018; Young et al., 2009). The same issue is also found in studying new technology acceptance in healthcare applications (e.g. Kates et al., 2015; McColl et al., 2013; Moro, 2018; Young et al., 2009).

As for cyborg technology, the term itself is immature and technology is still in the development stage, especially for human enhancement purposes. Besides, nothing is much known about its acceptance in society. However, society already accepted the use of this technology (e.g. technological and physical implants) for restoring physical functions, such as in Cochlear Implants (CI), and for increasing seductive capacities, such as in breast implants (Moser & Aiken, 2011; Pelegrín-Borondo et al., 2017). For instance, CI is seen as a hearing aid for helping deaf people to restore their hearing ability. Whereas, it could be used as an enhancement tool to increase human hearing capacity to beyond the normality. Therefore, this shift in the use of therapy to enhancement could change people's perception of these technologies (Joseph Lee, 2016). As well, some CI users for therapy purposes are introducing themselves as cyborg entities (Christie & Bloustien, 2010). In the same context, Gao et al. (2015) studied the acceptance of healthcare wearable technologies. The authors pointed out three significant factors related to the intention to use wearable technologies in terms of privacy, healthcare and technology perspectives. Their results suggested that the importance of these factors is depending on their applications (Medical or Fitness). For example, social influence is significantly important for fitness wearables. However, PU is one of the important determinants of medical wearable technology acceptance. On the other side, the ethical issues related to the associated risk with these technologies and their limits are an important topic that should be discussed in future researches. The successes of cyborg technologies could depend on the offered benefits and on people's perception of these benefits (Schicktanz et al., 2015). Ethically speaking, technological implants for therapy use are acceptable. While, it is still unclear for enhancement applications, despite what some authors mentioned the critical need for reformulating the meaning of ethics, in terms of moral judgments, to be applicable to this type of technology (Schermer, 2009). Reinares-Lara et al. (2018) studied the effect of ethics on the acceptance of technological implants. The authors mentioned the ethical problem, which is covering different areas, such as personal security and privacy, and its effect on personal identity. In fact, the study implemented the ethical construct into the CAN model, to investigate its moderating influence on the acceptance of brain implants for increasing capacities. The model is consisting of Performance Expectancy, Effort Expectancy, Emotions (Negative Emotions, Positive Emotions, and Anxiety) and Subjective Norms as the determinants of the intention to use technological implants. Even though results did not prove the moderating effect of the ethical side of the implant's acceptance, it explained the intention differences in using them. Meanwhile, the same results confirmed the impact of performance expectancy, effort expectancy, negative emotions, positive emotions and social influence on the intended behavior, where the last two constructs had the strongest impact. This is consistent with the results of Pelegrín-Borondo et al. (2017) and Pelegrín-Borondo et al. (2016) studies. Consequently, studying the acceptance of cyborg technology should consider the cultural differences during the investigations. For instance, the CAN model could be employed in different countries and integrated with the ethical dimension, to be able to generalize the results, as ethical aspects are inherently cultural (Reinares-Lara et al., 2018).

After all, these studies have been investigating the acceptance of being a cyborg. Whilst, the research purpose is to study the acceptance of cyborg as an entity in the healthcare service encounter, and if compared to the services offered by the human-being. Moreover, since cyborg represents a combination of technology and humanity, both aspects will be considered, to investigate cyborg acceptance. Accordingly, the research proposed the theoretical model, which is shown in figure 2.

Figure 2. Theoretical Model.



3. CONCLUSION

Human body enhancements to create cyborgs are increasing widely and they could reduce the fears of human being extinction because of robots. Hence, human enhancement could be more efficient than producing robots to obtain multilateral embodied intelligence. Because the natural motion skills of human-being require complex structural elasticity and massive computational resources. Nevertheless, the technology is still under development and it could be seen as a futuristic technology that will be able to produce an enhanced human body as imagined in science fiction novels and movies. In fact, it has received the attention of many researchers to study human acceptance to become a cyborg. However, this research was interested in developing a theoretical framework that can be used to investigate the acceptance of the services that could be offered by cyborg, especially in the healthcare service encounter. The research integrated different constructs from previous studies that have been interested in studying new technology acceptance, such as robot and being cyborg acceptance.

The early investigation of the potential acceptance of such technologies could direct the future efforts of technology developers and service providers toward meeting customers' expectations. Meanwhile, the investigation of the service acceptance itself could require future researches to draw more attention to customer expectations of such services. Also, the ethical impact of the proposed service could be required in future investigations, since the cyborg is representing an advance technology that could have the ability to imitate and exceed human abilities. If these futuristic cyborgs become a reality, they could compete with humans and replace them, thereby increasing the professional and social gap between humans and enhanced humans. Another ethical concern is related to the cyborg services availability for high-income customers, which could create a new social class that can buy the proposed superior services, and this could increase the equity gap too.

REFERENCES

Aggelidis, V. P., & Chatzoglou, P. D. (2009). Using a Modified Technology Acceptance Model in Hospitals. *International Journal of Medical Informatics*, *78*(2), 115–126. https://doi.org/10.1016/j.ijmedinf.2008.06.006

Ahadzadeh, A. S., Pahlevan Sharif, S., Ong, F. S., & Khong, K. W. (2015). Integrating Health Belief Model and Technology Acceptance Model: An Investigation of Health-Related Internet Use. *Journal of Medical Internet Research*, *17*(2), 1–27. https://doi.org/10.2196/jmir.3564

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alsharo, M., Alnsour, Y., & Alabdallah, M. (2018). How Habit Affects continuous Use: Evidence from Jordan’s National Health Information System. *Informatics for Health & Social Care*, 14. https://doi.org/10.1080/17538157.2018.1540423

Bawack, R. E., & Kamdjoug, J. R. K. (2018). Adequacy of UTAUT in Clinician Adoption of Health Information Systems in Developing Countries: The Case of Cameroon. *International Journal of Medical Informatics*, *109*, 15–22. https://doi.org/10.1016/j.ijmedinf.2017.10.016

Blut, M., Wünderlich, N. V, & Brock, C. (2018). Innovative Technologies in Branded-Service Encounters: How Robot Characteristics Affect Brand Trust and Experience. In *Thirty Ninth International Conference on Information Systems*. San Francisco. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1126&context=icis2018

Britton, L. M., & Semaan, B. (2017). Manifesting the Cyborg via Techno-Body Modification : From Human Computer Interaction to Integration. In *CSCW '17 Companion* (pp. 2499–2510). Portland, Oregon: ACM. https://doi.org/10.1145/3025453.3025629

Chang, M. Y., Pang, C., Michael Tarn, J., Liu, T. S., & Yen, D. C. (2015). Exploring User Acceptance of an E-hospital Service: An Empirical Study in Taiwan. *Computer Standards and Interfaces*, *38*, 35–43. https://doi.org/10.1016/j.csi.2014.08.004

Chen, S.-C., Liu, S.-C., Li, S.-H., & Yen, D. C. (2013). Understanding the Mediating Effects of Relationship Quality on Technology Acceptance: An Empirical Study of E-Appointment System. *Journal of Medical Systems*, *37:9981*, 1–13. https://doi.org/10.1007/s10916-013-9981-0

Chow, M., Chan, L., Lo, B., Chu, W. P., Chan, T., & Lai, Y. M. (2013). Exploring the Intention to Use a Clinical Imaging Portal for Enhancing Healthcare Education. *Nurse Education Today*, *33*(6), 655-662. https://doi.org/10.1016/j.nedt.2012.01.009

Christie, E., & Bloustien, G. (2010). I-cyborg: Disability, affect and public pedagogy. *Discourse: Studies in the Cultural Politics of Education*, *31*(4), 483-498. https://doi.org/10.1080/01596306.2010.504364

Chu, X., Lei, R., Liu, T., Li, L., Yang, C., & Feng, Y. (2018). An Empirical Study on the Intention to Use Online Medical Service. In *2018 15Th International Conference on Service Systems and Service Management (Icsssm)*. https://doi.org/10.1109/ICSSSM.2018.8464965

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. *Doctoral dissertation*. Massachusetts Institute of Technology.

Davis, F. D. (1989). Perceived Usefulness , Perceived Ease Of Use , And User Acceptance. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Destephe, M., Brandao, M., Kishi, T., Zecca, M., Hashimoto, K., & Takanishi, A. (2015). Walking in the Uncanny Valley: Importance of the Attractiveness on the Acceptance of a Robot as a Working Partner. *Frontiers in Psychology*, *6*, *204*. https://doi.org/10.3389/fpsyg.2015.00204

Dünnebeil, S., Sunyaev, A., Blohm, I., Leimeister, J. M., & Krcmar, H. (2012). Determinants of Physicians' Technology Acceptance for E-Health in Ambulatory Care. *International Journal of Medical Informatics*, *81*(11), 746–760. https://doi.org/10.1016/j.ijmedinf.2012.02.002

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Füller, J., Faullant, R., & Matzler, K. (2010). Triggers for Virtual Customer Integration in the Development of Medical Equipment - From a Manufacturer and a User's Perspective. *Industrial Marketing Management*, *39*, 1376–1383. https://doi.org/10.1016/j.indmarman.2010.04.003

Gao, Y., Li, H., & Luo, Y. (2015). An Empirical Study of Wearable Technology Acceptance in Healthcare. *Industrial Management and Data Systems*, *115*(9), 1704–1723. https://doi.org/10.1108/IMDS-03-2015-0087

Gauttier, S. (2018). 'I've got you under my skin' – The Role of Ethical Consideration in the (non-) Acceptance of Insideables in the Workplace. *Technology in Society*, *56*, 93–108. https://doi.org/10.1016/j.techsoc.2018.09.008

Greguric, I. (2014). Ethical issues of human enhancement technologies: Cyborg technology as the extension of human biology. *Journal of Information, Communication and Ethics in Society*, *12*(2), 133–148. https://doi.org/10.1108/JICES-10-2013-0040

Guo, X. T., Yuan, J. Q., Cao, X. F., & Chen, X. D. (2012). Understanding the acceptance of mobile health services: A service participants analysis. In *International Conference on Management Science and Engineering - Annual Conference Proceedings* (pp. 1868–1873). IEEE. https://doi.org/10.1109/ICMSE.2012.6414426

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Heisele, B., Serre, T., Pontil, M., Vetter, T., & Poggio, T. (2002). Categorization by Learning and Combining Object Parts. *Advances in Neural Information Processing Systems*, *14*(2), 1239–1245.

Hendrikx, H. C. A. A., Pippel, S., van de Wetering, R., & Batenburg, R. S. (2013). Expectations and Attitudes in eHealth: A Survey Among Patients of Dutch Private Healthcare Organizations. *International Journal of Healthcare Management*, *6*(4), 263–268. https://doi.org/10.1179/2047971913Y.0000000050

Hogle, L. F. (2005). Enhancement Technologies and the Body. *Annual Review of Anthropology*, *34*, 695–716. https://doi.org/10.1146/annurev.anthro.33.070203.144020

Hossain, A., Quaresma, R., & Rahman, H. (2019). Investigating Factors Influencing the Physicians' Adoption of Electronic Health Record (EHR) in Healthcare System of Bangladesh: An Empirical Study. *International Journal of Information Management*, *44*, 76–87. https://doi.org/10.1016/j.ijinfomgt.2018.09.016

Hsieh, P. (2014). Physicians ' Acceptance of Electronic Medical Records Exchange : An Extension of the Decomposed TPB Model with Institutional Trust and Perceived Risk. *International Journal of Medical Informatics*, *84*(1), 1–14. https://doi.org/10.1016/j.ijmedinf.2014.08.008

Kates, M., Ball, M. W., Patel, H. D., Gorin, M. a, Pierorazio, P. M., & Allaf, M. E. (2015). The Financial Impact of Robotic Technology for Partial and Radical Nephrectomy. *Journal of Endourology*, *29*(3), 6. https://doi.org/10.1089/end.2014.0559

Keikhosrokiani, P., Mustaffa, N., Zakaria, N., & Baharudin, A. S. (2018). User Behavioral Intention Toward Using Mobile Healthcare System. In *Consumer-Driven Technologies in Healthcare* (pp. 128–143). Pennsylvania, USA: Information Resources Management Association. https://doi.org/10.4018/978-1-5225-6198-9.ch022

Kijsanayotin, B., Pannarunothai, S., & Speedie, S. M. (2009). Factors Influencing Health Information Technology Adoption in Thailand's Community Health Centers: Applying the UTAUT Model. *International Journal of Medical Informatics*, *78*(6), 404–416. https://doi.org/10.1016/j.ijmedinf.2008.12.005

Koschate, M., Potter, R., Bremner, P., & Levine, M. (2016). Overcoming the Uncanny Valley: Displays of Emotions Reduce the Uncanniness of Humanlike Robots. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 359–365). Christchurch, New Zealand: IEEE.

Kostrica, D. (2018). Medical Approach of Transhumanism. *HUMANUM*, *28*(1), 67–74. Retrieved from http://www.humanum.org.pl/images/2018/humanum_28_1_2018.pdf#page=67

Kulviwat, S., Bruner, G. C., Kumar, A., Nasco, S. A., & Clark, T. (2007). Toward a Unified Theory of Consumer Acceptance Technology. *Psychology & Marketing*, *24*(12), 1059–1084. https://doi.org/10.1002/mar.20196

Lai, Y.-H. (2014). A study on the attitude of use the mobile clinic registration system in Taiwan. *International Journal of Computer and Information Technology*, *3*(4), 750–754.

Lee, Jaebeom, & Rho, M. J. (2013). Perception of Influencing Factors on Acceptance of Mobile Health Monitoring Service: A Comparison between Users and Non-users. *Healthcare Informatics Research*, *19*(3), 167–176. https://doi.org/10.4258/hir.2013.19.3.167

Lee, Joseph. (2016). Cochlear Implantation, Enhancements, Transhumanism and Posthumanism: Some Human Questions. *Science and Engineering Ethics*, *22*(1), 67–92. https://doi.org/10.1007/s11948-015-9640-6

Li, H., Wu, J., Gao, Y., & Shi, Y. (2016). Examining Individuals' Adoption of Healthcare Wearable Devices: An Empirical Study from Privacy Calculus Perspective. *International Journal of Medical Informatics*, *88*, 8–17. https://doi.org/10.1016/j.ijmedinf.2015.12.010

Lilley, S. (2013). *Transhumanism and Society: The Social Debate Over Human Enhancement*. Fairfield, CT, USA: Springer. https://doi.org/10.1007/978-94-007-4981-8

Matsui, D., Minato, T., Macdorman, K. F., & Ishiguro, H. (2018). Generating Natural Motion in an Android by Mapping Human Motion. In *Geminoid Studies* (pp. 57–73). Springer Singapore. https://doi.org/10.1007/978-981-10-8702-8

McColl, D., Louie, W. Y. G., & Nejat, G. (2013). Brian 2.1: A Socially Assistive Robot for the Elderly and Cognitively Impaired. *IEEE Robotics and Automation Magazine*, *20*(1), 74–83. https://doi.org/10.1109/MRA.2012.2229939

McGee, E. M., & Maguire, G. Q. (2007). Becoming borg to become immortal: Regulating brain implant technologies. *Cambridge Quarterly of Healthcare Ethics*, *16*(3), 291–302. https://doi.org/10.1017/S0963180107070326

Mori, M. (1970). The Uncanny Valley. *Energy*, *7*(4), 33–35.

Moro, C. (2018). *Learning Socially Assistive Robot Behaviors for Personalized Human-Robot Interaction*. *Master Thesis*. University of Toronto. Retrieved from http://hdl.handle.net/1807/82909

Moser, S. E., & Aiken, L. S. (2011). Cognitive and Emotional Factors Associated with Elective Breast Augmentation among Young Women. *Psychology & Health*, *26*(1), 41–60. https://doi.org/10.1080/08870440903207635

Mulken, S. van, André, E., & Müller, J. (1999). An Empirical Study on the Trustworthiness of Life-Like Interface Agents. In *The 8th International Conference on Human-Computer Interaction* (Vol. 2, pp. 152–156). Munich, Germany: DBLP.

Nasir, S., & Yurder, Y. (2015). Consumers' and Physicians' Perceptions about High Tech Wearable Health Products. *Procedia - Social and Behavioral Sciences*, *195*, 1261–1267. https://doi.org/10.1016/j.sbspro.2015.06.279

Oh, C., Lee, T., Kim, Y., Park, S., Kwon, S. bom, & Suh, B. (2017). Us vs. Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (pp. 2523–2534). Denver, CO, USA. https://doi.org/10.1145/3025453.3025539

Olarte-Pascual, C., Pelegrín-Borondo, J., & Reinares-Lara, E. (2015). Implants to increase innate capacities: Integrated vs. apocalyptic attitudes. Is there a new market? *Universia Business Review*, *2015*(48), 86–117. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84949637861&partnerID=40&md5=e412b5a12f21962f95d3abfe4bb18e04

Olarte, C., Pelegrín, J., & Reinares, E. (2017). Model of acceptance of a new type of beverage: Application to natural sparkling red wine. *Spanish Journal of Agricultural Research*, *15*(1), 1–11. https://doi.org/10.5424/sjar/2017151-10064

Pai, F. Y., & Huang, K. I. (2011). Applying the Technology Acceptance Model to the introduction of healthcare information systems. *Technological Forecasting and Social Change*, *78*(4), 650–660. https://doi.org/10.1016/j.techfore.2010.11.007

Pappas, I. O., Giannakos, M. N., & Chrissikopoulos, V. (2013). Do Privacy and Enjoyment Matter in Personalized Services? *International Journal of Digital Society*, *4*(1), 705–713. https://doi.org/10.20533/ijds.2040.2570.2013.0091

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2018). Does Ethical Judgment Determine the Decision to Become a Cyborg? *Journal of Business Ethics*, 1–13. https://doi.org/10.1007/s10551-018-3970-7

Pelegrín-Borondo, J., Juaneda-Ayensa, E., González-Menorca, L., & González-Menorca, C. (2015). Dimensions and basic emotions: A complementary approach to the emotions produced to tourists by the hotel. *Journal of Vacation Marketing*, *21*(4), 351–365. https://doi.org/10.1177/1356766715580869

Pelegrin-Borondo, J., Orito, Y., Fukuta, Y., Murata, K., Arias-Oliva, M., & Adams, A. A. (2017). From a Science Fiction to the Reality: Cyborg Ethics in Japan. *ORBIT Journal*, *1*(2), 1–15. https://doi.org/10.29297/orbit.v1i2.42

Pelegrín-Borondo, J., Reinares-Lara, E., & Olarte-Pascual, C. (2017). Assessing the acceptance of technological implants (the cyborg): Evidences and challenges. *Computers in Human Behavior*, *70*, 104–112. https://doi.org/10.1016/j.chb.2016.12.063

Pelegrín-Borondo, J., Reinares-Lara, E., Olarte-Pascual, C., & Garcia-Sierra, M. (2016). Assessing the moderating effect of the end user in consumer behavior: The acceptance of technological implants to increase innate human capacities. *Frontiers in Psychology*, *7:132*, 1–13. https://doi.org/10.3389/fpsyg.2016.00132

Phichitchaisopa, N., & Naenna, T. (2013). Factors Affecting the Adoption of Healthcare Information Technology. *EXCLI Journal*, *12*, 413–436. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4566918/

Reinares-Lara, E., Olarte-Pascual, C., & Pelegrín-Borondo, J. (2018). Do you Want to be a Cyborg? The Moderating Effect of Ethics on Neural Implant Acceptance. *Computers in Human Behavior*, *85*, 43–53. https://doi.org/10.1016/j.chb.2018.03.032

Reinares-Lara, E., Olarte-Pascual, C., Pelegrin-borondo, J., & Pino, G. (2016). Nanoimplants that Enhance Human Capabilities: A Cognitive-Affective Approach to Assess Individuals’ Acceptance of this Controversial Technology. *Psychology & Marketing*, *33*(9), 704–712. https://doi.org/10.1002/mar.20911

Romportl, J. (2015). *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. https://doi.org/10.1007/978-3-319-09668-1

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*, 617–640. https://doi.org/10.1016/S0167-4870(01)00059-9

Schermer, M. (2009). The Mind and the Machine. On the Conceptual and Moral Implications of Brain-Machine Interaction. *NanoEthics*, *3*, 217–230. https://doi.org/10.1007/s11569-009-0076-9

Schicktanz, S., Amelung, T., & Rieger, J. W. (2015). Qualitative assessment of patients' attitudes and expectations toward BCIs and implications for future technology development. *Frontiers in Systems Neuroscience*, *9:64*. https://doi.org/10.3389/fnsys.2015.00064

Schifter, D. E., & Ajzen, I. (1985). Intention, Perceived Control, and Weight Loss: An Application of the Theory of Planned Behavior. *Journal of Personality and Social Psychology*, *49*(3), 843–851. https://doi.org/10.1037/0022-3514.49.3.843

Sezgin, E., Özkan-Yildirim, S., & Yildirim, S. (2017). Investigation of Physicians' Awareness and Use of M-Health Apps: A Mixed Method Study. *Health Policy and Technology*, *6*(3), 251–267. https://doi.org/10.1016/j.hlpt.2017.07.007

Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43–50. https://doi.org/10.1016/j.cognition.2016.12.010

Sun, Y., Wang, N., Guo, X., & Peng, Z. (2013). Understanding the Acceptance of Mobile Health Services : A Comparison and integration of alternative models. *Journal of Electronic Commerce Research*, *14*(2), 183–200. Retrieved from http://web.csulb.edu/journals/jecr/issues/20132/paper4.pdf

Triviño, J. L. P. (2015). Equality of Access to Enhancement Technology in a Posthumanist Society. *Dilemata*, *7*(19), 53-63. Retrieved from https://www.dilemata.net/revista/index.php/dilemata/article/view/400

van der Heijden. (2004). User Acceptance of Hedonic Information Systems. *MIS Quarterly*, *28*(4), 695. https://doi.org/10.2307/25148660

Venkatesh, V. (2000). Determinants of Perceived Ease of Use : Integrating Control , Intrinsic Motivation , and Emotion into the Technology Acceptance Model. *Information System Research*, *11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology : Extending the Unified Theory od Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Warwick, K. (2016). Transhumanism: Some Practical Possibilities. *FIfF-Kommunikation. Zeitschrift Für Informatik Und Gesellschaft*, (2), 24–27. Retrieved from http://www.fiff.de/publikationen/fiff-kommunikation/fk-2016/fk-2016-2/fk-2016-2- content/fk-2-16-p24.pdf

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave New World: Service Robots in the Frontline. *Journal of Service Management*, *29*(5), 907–931. https://doi.org/10.1108/JOSM-04-2018-0119

Y, M. D. I., Reza, M., Meyliana, Widjaja, H. A. ., & Hidayanto, A. N. (2017). Acceptance of HIS Usage Level in Hospital with SEM-PLS as Analysis Methodology: Case Study of a Private Hospital in Indonesia. In

2016 International Conference on Information Management and Technology (ICIMTech) (pp. 112–117). Bandung, Indonesia: IEEE. https://doi.org/10.1109/ICIMTech.2016.7930313

Yang, H., Yu, J., Zo, H., & Choi, M. (2016). User Acceptance of Wearable Devices: An Extended Perspective of Perceived Value. *Telematics and Informatics*, *33*(2), 256–269. https://doi.org/10.1016/j.tele.2015.08.007

Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2009). Toward acceptable domestic robots: Applying insights from social psychology. *International Journal of Social Robotics*, *1*(1), 95–108. https://doi.org/10.1007/s12369-008-0006-y

ETHICAL REFLECTIONS IN A TRANSHUMANISM FRAMEWORK

Ferran Sánchez Margalef

University of Barcelona (Spain)

ferran.sanchez@ub.edu

ABSTRACT

We are at the beginning of a technological tsunami, known as the Digital Revolution, that will transform many of the spheres of human reality. In this context, transhumanism appears as a tendency capable to take humanity to its own transcendence. Taking into account that some corporations or institutions of great social and economic importance (e.g. NASA, Google or the University of the Singularity) are already starting to invest in projects that facilitate the arrival of a post-human world, it is essential for humanity to consider the challenges that this movement implies.

In this article, we take as a starting point the heterodox but consolidated tendencies that are grouped according to their position, for or against, when transcending humanity, namely transhumanists and bioconservatives, with the aim of propose some of the arguments that are today on the table. Hence, the first step to acquire the required consciousness to reach the understandings of the transhumanist phenomenon is to maintain a constructive, argued and peaceful debate. However, in a scenario of a non-agreement on the required consensus about the limits of transhumanism, there is a certain possibility that humanity will stop using technology as a means to put itself at the service of its logic.

To focus this brief communication about the transhumanists and bioconservatives arguments, we will first establish and contextualize this movement, as providing and commenting some of the motivations of both sides from several lectures. Then, from an axiological point of view, we will offer a critique of transhumanism and, finally, we will provide some conclusions.

KEYWORDS: ethic, cyborg, freedom, technology, transhumanism, transcendence.

1. INTRODUCTION TO THE TRANSHUMANIST COSMOVISION

Transhumanism (H+) throws out the idea of overcoming and transcending the human being. Although this phenomenon seems to be exclusively a product of a hyper-technological society, the reality is that we can find the roots of transhumanism in the glimpses of transcendence that our ancestors already had at the origins of civilization. We must not forget that funeral rituals constitute a milestone for humanity, since they imply the conception of death (a unique characteristic of our species) and the belief in an afterlife. Therefore, it is no coincidence that, already in Mesopotamia, we find stories such as the Epic of Gilgamesh, in which a mention of immortality is made: “There is a plant… like a boxthorn, whose thorns will prick your hand like a rose. If your hands reach that plant you will become a young man again. Hearing this, Gilgamesh opened a conduit and attached heavy stones to his feet. They dragged him down, to the Apsu they pulled him. He took the plant, though it pricked his hand, and cut the heavy stones from his feet, letting the waves throw him onto its shores” (Carnahan, 1998, p.50). Immortality, which is one of the key elements of the Sumerian legend, is the same dream that, more than four thousand years later, transhumanism promises to make a reality thanks to scientific progress. In this way, we see how the glimpse of eternity has always been part of the imagination, implicit in the human condition, as well as the will to translate those dreams into reality. However, all these dreams have been limited, until now, by the biological condition itself. Today, transhumanism, a product of the artificiality of civilization, will use technology to overcome biology and thus alter the course of nature.

Hence, it seems that the positions around the transhumanist debate are divided between the transhumanists and the bioconservatives. The former intend to abandon humanity to achieve, with the support of technology, a more perfect being (and apparently a better society). The latter position themselves against it and warn of the risks that certain actions may entail.

The human is not a definitive being. Actually, he never has been. The Darwinian theory of evolution explains how the little ape *Pliopithecus* has evolved into modern *Homo Sapiens*. In this way, Transhumanism states that the next evolutionary step, the one that will bring a new man, known as the *Homo Deus* according to Harari (Kiryat Atta, 1976), will also be the result of evolution, with the "small" change that this one will not be imposed by the biological logic, but will be in the hands of the technological rationality. If we think about it, the name *Homo Deus* is not presumptuous. Our ancestors prayed to the gods to take care of their crops after sowing or to provide health at birth to the child that mothers carried in their wombs (Harari, 2016). Today, we are already able to genetically manipulate seeds to obtain transgenic foods that resist pests better, produce more fruit and adapt to different climates and landscapes. Nowadays, man has even been able to create the first living machines (Kriegman, Blackiston, Levin & Bongard, 2020). One could say that, in some way, we are becoming Gods, since creation has ceased to be an exclusive property of the divinity to be shared with the human (or transhuman) subject: a being that will be able to create better beings and even to recreate himself. As we have been warning, overcoming the organic limits is essential to reach a post-human stage that escapes the biological frontiers to which the human condition is subject.

We should ask ourselves, in the event that the human being is transcended, what moral repercussions will it carry. That is, if we take into account that things are only good or bad in relation to the human being (Scheler, 2001), that is to say that any ethical assessment is subject to a human reference and that transhumanism tries to overcome the category of humanity, we can begin to elucidate in the complicated position that would remain the ethics in a transhumanist or post-humanist scenario. Many questions arise such as the following, to which we will briefly try to shed some light: "Who will not be tempted [in this new transhuman era] by the possibility of being always young and eternal? What if this depended on replacing our body and living eternally, within the networks of information, a virtual reality, as real as ours without the danger of dying? How many people will be willing to continue living even if only as a post body? And how many will die locked up in their obsolete body? […] Will this be one of the most controversial ideologies in the future? If so, will there really be full freedom to modify and transcend our body? But if this were to have an economic cost, would it worsen the social imbalance, where superior bodies and simple mortals would coexist? Will the freedom of choice proclaimed by transhumanists depend solely on the purchasing power of each person?" Córdoba, 2007, pp. 611-612).

2. THE MOST AMBITIOUS IDEA EVER CONCEIVED

Youth, eternity, superiority or improvement are only some of the terms that are usually together with the transhumanist discourse. There is no doubt that its projects are ambitious. If we analyze the word on a semantic level, Transhumanism is composed of *trans* (prefix of Latin root that means "beyond of" or "on the other side of"), *humanus* (also of Latin root that refers to the human species), and *ism* (suffix of Greek root that implies a doctrine, belief or vital posture). In a synthetic way, and in consequence of its etymological approach, we understand that the H+ is a current that outlines the overcoming of the human being (arriving to a post-human state once the human condition has been overcome).

However, the limits to overcome the human condition are not clear. Moreover, the lack of precision of Fereidoun M. Esfandiary or F.M. 2030 (Brussels, 1930 – New York, 2000), the first to use the concept of Transhumanism in an instructive way, when clarifying them in his book, has not helped the scientific community. The author states that it is a human in transition, but without clarifying exactly where the limits of this transition are, nor what can be considered as a transition (beyond some general characteristics such as the use of prostheses, plastic surgery, intensive use of telecommunications or a cosmopolitan profile without any religious beliefs and with a rejection towards traditional values) (F.M. 2030, 1989). In any case, which is highlighted is the emergence of values such as dynamism, fluidity and change, which become essential in the Digital Revolution and that necessarily follows the transhumanist discourse.

On the other hand, we must also pay attention to the reasons that accompany the popularity of Transhumanism, since they also prevail in the will of any being that wishes to improve or transcend himself to become, even, eternal. Overcoming the terrible idea of death, which resonates in the head of any human being, is therefore also one of the main goals of Transhumanism. Bostrom (Helsingborg, 1973), one of the theorists of this movement, uses the metaphor of the Tyrannical Dragon to refer, precisely, to the desire to overcome aging and, thus, to kill the death.

His reasoning is the following: assuming that aging is the cause that generates more deaths on the planet (therefore attacking human well-being), it must be a (moral) priority of humanity to face and defeat the Tyrant Dragon that devours people (Bostrom, 2005).

It must be borne in mind that this desire to improve oneself is a characteristic that has always been intrinsic with the human being. In this way, transhumanists will try to correlate their particular vision of (technological) improvement to the human spirit. Therefore, we can observe in Savulescu (Melbourne, 1963) that “if these [genetic] manipulations improve our ability to make rational and normative judgements, they further improve what is fundamentally human. Far from being against the human spirit, such improvements express the human spirit. To be human is to be better” (Savulescu, 2009, p.428).

These manipulations will be carried out thanks to the relationship between several fields of knowledge that have been, so far, compartmentalized between each other. In this way, Transhumanism, which is only possible by the interconnection of the propitiated digitalization and technological implementation in different areas of knowledge, such as the NTBI (Nanorobotics, Technology of information, Biotechnology and artificial Intelligence), will point out towards the future with grandiloquent promises awaiting the next scientific advances, which are going to be capable to make them a reality, until we reach Posthumanity.

Hence, one of its most characteristic features is that, while other worldview, sensibilities, phenomena or movements have been inspired in the past to build their speech, Transhumanism denies the past to venerate the future. According to the predictions, an advanced and superior specie is going to replace humans as known in the present-day. As stated by Lafontaine (Canada, 1970), a new being, and therefore a new species, perhaps still biological but with built-in devices and technological elements, seems to be the destiny of today’s society: “contemporary society, with its large contribution of technologies of the information and biotechnology, also has the hope of finally seeing the appearance of a new man, capable to adapt by his great flexibility to the whims of caprices of the communication flows” (Lafontaine, 2000).

The appearance of a being that has its origin in the cultural development, as a result of science and technology instead of the nature itself, has no precedent in the history of humanity. Without being able to make a comparison to keep a certain equivalence, it is necessary to comment that transcending the human being is going to represent, at least, a Copernican turn as great as the one that humanity took place during the Renaissance and culminated in the French Revolution. We are referring to the humanism that moved God from the center of the Cosmos to place the human in it (*anthropocentrism)*. If Humanism put the man in the epicenter of the Universe, Transhumanism will displace him from it to give place to a new being (probably a human turned into God) that will appear from the human digitalization. Not in vain, it is fair to recognize that, despite the first change required several centuries to be implemented, the immediacy that characterizes technological devices can provide that this can happen in a few years.

Although it cannot be said that there is a human being who has transcended humanity, it is no less true that the cyborg has ceased to be exclusively part of science fiction. In Western society, it can already be seen how "the increasing variety and availability of models of prosthesis/artifacts that can be built-in in a body, either for functional and/or aesthetic purposes, will progressively transform the human body into a complex sum of artifacts, with an increasingly extensive interface between the technological and the biological, between the cybernetic and the organic, like in the futuristic creatures known as cyborgs, created by science fiction writers" (Koval, 2006, p.13). The reality is that humanity already has cyborgs (at least on a terminological level), since Harbisson (London, 1984) has been recognized as such for the British state. Harbisson was born with a congenital eye disease that prevented him from distinguishing colors except for black and white. After years of effort and study, he has managed to develop a functional antenna, which is integrated into the occipital bone of the cranium. This device has a sensor capable to capture frequencies of light and transform them into frequencies of sound. Thus, Harbisson is able to hear colors and even to detect infrared and ultraviolet rays, which are imperceptible to any human (Ledesma, 2018).

Therefore, it does not seem unthinkable that, if technological hybridization continues, there will be more and more people who will be considered as cyborgs because of their capacities, which are unachievable by the only means of the human condition. If so, although it will not be measured in a specific period of time, it is possible that the organic will be gradually replaced by the cybernetic. However, according to the transhumanists, we should not look with nostalgia the probable disappearance of the human being, since we should embrace the possibility of enjoying a life that reaches quotas of greater perfection (although terms such as "perfection" or "improvement" are somewhat abstract). Furthermore, given the Kurweil's Law of Accelerating Returns (1999), which points out that technological development is exponential, the transhumanist revolution is revealed to be unstoppable (Kurzweil, 2015). This is also considered by Baylis (Montreal, 1962) and Robert (?): "The development and application to humans of the technology for genetic improvement is inevitable. They constitute the next and definitive step of the evolutionary process of our species. All resistance is condemned to failure". In this way, the cyborg is not only the future for humanity, but also represents the last opportunity for it to not be left out of the post-human world that is already under construction.

It is also necessary to point out where this transformation will take place, which is none other than the body itself. Transhumanist ideas have spread at a time when the body is no longer conceived as a sacred temple (as in Greco-Latin culture) or as a source of sin (as in the Judeo-Christian tradition). As pointed out by Farrero (Barcelona, 1980) and Vilanou (Barcelona, 1953) (2016), "the body is a political setting for insurrection and desecration, as evidenced by the different experiences currently being carried out that take the body to the extreme of tattooing or piercing". Thus, in postmodernity, there is no major force to prevent the body from being alienated, modified, or even replaced for the benefit of the subject himself. Hence, the human will cease to be human to reach post-humanity through the intervention and manipulation of the somatic. It is not possible, therefore, to embrace the benefits of Transhumanism without detaching oneself from the human condition. As F.M.2030 says, “if we want to extend each life far into future, we have to make radical changes. We cannot live for hundreds of years with these fragile limited bodies” (F.M.2030, 1989, p.201).

3. MORAL CRITICISM OF THE TRANSHUMANIST MOVEMENT

As aforementioned, several voices have arisen either to defend the postulates of the Transhumanism or to refute them. That is, the repercussion of this movement to the humanity is still unknown and, therefore, rejectable for several people. In this section, some of the challenges that are raised in an axiological level of H+, proposed by bioconservatives authors, are going to be discussed. F. Fukuyama described Transhumanism as the most alarming idea ever expected (Fukuyama, 2002), when considering it as a frontal attack to humanity (Fukuyama and Reina, 2002). The occidental society has a consensus about the Human Rights since there is an international acceptance of the fundamental premises such as the life dignity or the equality of lives.

Despite the diverse legislations of the different democratic cultures, which subject humanity to the rights, defend (to a greater or lesser extent) to abide these universal maxims, it must be borne in mind that the first fundamental right, indispensable to be able to exercise any other, is none other than the natural right to life. Taking into account that H+ directly modifies the human condition and life as understood today, it is also clear that it attacks the very dignity of the species. In addition, the breach of the right to a biological life is also the breakdown of the right to a proper and spontaneous identity resulting from a biological chance and from an extremely complex set of variables and conditioning factors. Trying to control and influence these variables implies directing a life, ergo violating its most intimate dignity: the freedom for each one to be what he or she must be.

Another argument at the axiological level, which revolves around identity, is given by Sandel (Minneapolis, 1953). He points out the pressure that will be placed on the future improved subjects, considering the expectation that we have foreseen in them, and the possible serious disappointments or even depressions that they will have if they are not up to the task (Sandel, 2015). Thus, we must contemplate the possibility that these improved future beings will rebel against the eugenic goals of their own designers, that these will not share the purpose of such improvement and, consequently, will not understand why they have been genetically manipulated. That being the case, we must keep in mind that a transhumanist individual can suffer serious disruptions of personal identity.

Moreover, as stated by Sandel, instead of continuing this commitment to improve humanity and eradicate any imperfection through the transhumanist channels, it would perhaps be more logical to use our efforts to create the conditions to enjoy a kinder world. In his own words, perhaps “Instead of using our new power to strengthen ‘the twisted shaft of humanity’, we should do everything in our power to create social and political conditions that are kinder to the gifts and limitations of imperfect human beings” (Sandel, 2015, pp.146-147). We must consider, then, that a transhuman future is a scenario without people with disabilities or functional diversity. Without entering into a complex debate, we did want to point out, at least, the same right of disabled people to decide their own future (in the same way as any other citizen) or over that of their future children. Would it be legitimate for the disease to be eradicated in order to improve the species? Does a person who is deaf (hereditary) not have the right to have a daughter, if they so choose, that will have the same disability?

The third argument that we want to comment is about equality. Even though the differences between humans are large and varied, they all share the same genetic condition. However, if H+ fractures with this equality, the contrast between the intelligent species inhabiting the planet will be emphasized. In other words, part of a privileged population will be able to access to the biotechnological benefits, while others will not.

Hence, one may be concerning about the relationship between both communities: the transhuman (*Homo Deus*) and the human (*Homo Sapiens*). Will those who have broken the biological condition and those who remain tied to their mortality compete for the same oppositions and in the same competitions? Transhumanist development inherently implies an imbalance at the social level produced by those people who begin to access this type of technology. Harari indicates that, although it is impossible to determine under which parameters this relationship will be established, what can be known with exactitude is the present-day relationship between human beings and other species endowed with less intelligence than them, as are the animals (Harari, 2016). The future of the human being in a transhuman or post-human context does not seem very hopeful because, if the transhuman being has a similar treatment to the human as the one maintained between this and the animals, the human future can be that of subordination and subjugation.

Furthermore, it is also appropriate to comment that human life is based on certain coordinates such as space, time, life or death, among others. Taking into account the revolution, at all levels, that Transhumanism implies (blurring the borders between space/time and life/death), it is very difficult to imagine that the same values that today serve as a reference for humanity continue being those that mark a transhumanist society. This fact brings a special unpredictability to the transhumanist movement that, in the worst case, can make it turn towards unsuspected parameters, perpetrating inconceivable tragedies. This is the reason why it can be dangerous to leave transhumanist experimentation in the exclusive hands of technicians or scientists, since, as we have been saying, the consequences of these are far-reaching for all of humanity. The same way is how Arendt (Hannover, 1906 – Nova York, 1975) understands it, when he warns that "the only question that arises is whether or not we want to use our scientific and technical knowledge in this sense, such a question cannot be decided by scientific means; it is a political problem of first order and, therefore, cannot be left to the decision of professional scientists or politicians" (Arendt, 1993, p.15). Thus, the possibility of democratizing decisions on technological developments is on the table. In this sense, Diéguez (Málaga, 1957) opens up the possibility of subjecting Transhumanism to an ethical framework affirming that "technological development can be controlled by means of an appropriate technological policy and by conditioning it to accepted values" (Diéguez, 2017, p.68).

Another of the theses concerns around freedom. As Panikkar suggests, we must understand technology as the science of control (Panikkar, 1991). Assuming that Transhumanism will integrate technology into the human biology, we must consider the possibility of encountering our freedom restricted. The hypothesis of a transhumanist dystopia has already been widely exploited in literature in works such as *New Brave World* (A. Huxley, 1932) or *1984* (G. Orwell, 1949), in which hyper-technological realities are constructed where subjects experience a limited freedom. Although in our society it is far from the worlds conceived by Orwell or Huxley, certainly the big technological companies (e.g. Apple, Samsung, Microsoft, IBM or Tencent Holdings) are starting to create programs that are capable of managing the data of millions of people, establishing new forms of domination. Here is an example if this that we are commenting: It was March 17, 2018 when journalists Cadwalladr (Tauton, 1969) and Graham-Harrison (London, ?) (2018) of *The Guardian* newspaper, made the following news public: "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach". This reveled the misappropriation of data (which would later be used in Donald Trump's election campaign to win the US presidency) by Cambridge Analytica, with information that would have been provided by Facebook.

With this being only one example, we must certainly take into account the possibility that, as human beings become more dependent on technological devices, these will diminish our freedom. Considering the reflections of Kant (Königsberg, 1724 – 1804) about the Illustration, freedom is defined as the overcoming of the man to the under-age (Kant, 2009) due to, precisely, the autonomy and freedom that have been acquired. Thus, it is appropriate to assume the paradox that Transhumanism can return the man to this under-age stage by removing the tools that have allowed him to think for himself. In the same line, Sandel rejects the transhumanist approach "because it manifests and promotes a certain attitude towards the world: an attitude of control and domination that does not recognize that gift character of human capacities and achievements, and forgets that freedom consists in a certain sense in a permanent negotiation with what has been received" (Sandel, 2015, p.137).

The last of the criticisms that we wanted to consider is about the irreversibility of transhumanist actions. We must not forget that there is no turning back in the biotechnological revolution that Transhumanism encloses, as it seeks to "dominate the territory of human 'natural nature' and of the entire biosphere, preserved so far in biological unity and cultural diversity, to transmute it from a radical and irreversible way into a biological-genetic-neural-factual diversification and into a paradoxical cultural uniformity" (Linares Salgado, 2018, p.86).

This is why prevention is a fundamental element in order to avoid future disasters, being necessary, also in relation to the argument that demands a democratization of technological advances, the "creation of conditions around which not only it is possible to control the freedom of choices that involve the modification or programming of the human beings, but also it is plausible that, if this is done, there will be measures of containment that allow us to become aware of what kind of programming we want for us" (Cardozo and Cabrera, 2014, p.86).

4. CONCLUSION

We would like to end this article by addressing two conclusions. Firstly, we would like to acknowledge that any change at a social level requires, previously, a change of awareness from the society. In this sense, the battle over technological hegemony has already begun and both the scientific discourse and human reality are gradually impregnated with the growing technological assimilation. It is not a minor matter that many of the spheres of human activity are already filled with applied sciences nor that our society venerate innovation, dynamism or consumerism, since these are the same values that will facilitate the arrival of Posthumanism.

Thus, it can be assumed that our behavior, our language or our reasoning are already being highly influenced by transhumanist postulates. A clear example can be seen in the concept itself and in its own antonym. The word *transhmanist*, chosen by the followers of this current, implies a series of positive connotations. On the other hand, if we refer to the antagonist word, *bioconservative*, the name evokes a certain perception of antiquity or something retrograde. It is necessary to highlight in order to get an idea as faithful as possible, that while transhumanists chose the name used by Huxley and recovered by Esfanidary, *bioconservative* is the alias, clearly derogatory, that transhumanists have imposed. Seeking new concepts on which to build discourses is therefore a prerequisite before starting the debate on ethical conditions; otherwise, battles will have been lost even before the start of the war. To conclude, then, we suggest another concept, *biovitalism*, on which to build a discourse and an alternative narrative to H+.

In an axiological sense, the morality on which law is based, depending on the ideal of what is right in accordance with a tradition, is undergoing an axiological transfiguration, consequence of the Digital Revolution, which facilitates the emergence of new values, such as efficiency, capability, adaptability or innovation. Moreover, in the same way that new values appeared due to the Industrial Revolution, such as obsolescence, the Digital Revolution and, later, the Transhumanism, will give birth to new axiological constructs, as long as humanity is present to provide them some value.

In addition, we would like to discuss the need to dissociate from the human progress what is strictly a technological and scientific advance. Although technology has improved the life of human beings on countless occasions, it has also led to some headaches. To give just one example, without scientific progress, climate change and its consequent environmental disasters (due to bad practices in the extraction of materials or the pollution of industries that manufacture the devices), would not be one of the biggest challenges to be globally tackled in the present-day. It is therefore necessary to denounce the falsehood that technological progress is positive regardless of whether it must be conditioned by ethical or democratic criteria in order to submit Transhumanism to other filters beyond economic or scientists.

REFERENCES

Arendt, H. (1993). La condición humana (Vol. 306). Barcelona: Paidós.

Baylis, F., & Robert, J. S. (2004). The inevitability of genetic enhancement technologies. *Bioethics*, *18*(1), 1-26.

Bostrom, N. (2005). The fable of the dragon tyrant. *Journal of Medical Ethics*, *31*(5), 273-277.

Cadwalladr, C., & Graham-Harrison, E. (2018). The Guardian. Revealed: 50 million Facebook profiles harvested for Cambridge Analyitica in major data breach, 17/3/2018. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election [Data de consulta: 24/10/2019]

Cardozo, J. J., & Cabrera, T. M. (2014). Transhumanismo: concepciones, alcances y tendencias. Análisis. Revista Colombiana de Humanidades, 46(84), 63-88.

Carnahan, W. (1998). The epic of Gilgamesh. Electronic Edition. Retrieved from https://uruk-warka.dk/Gilgamish/The%20Epic%20of%20Gilgamesh.pdf

Córdoba, S. (2007): *La representación del cuerpo futuro* [tesis doctoral], Madrid: Universidad Complutense de Madrid. Retrieved from http://biblioteca.ucm.es/tesis/bba/ucm-t29917.Pdf

Diéguez, A. (2017). Transhumanismo. *La búsqueda* del mejoramiento humano. Barcelona: Herder.

F.M. 2030. (1989). *Are You a Transhuman?* Nova York: Warner Books.

Farrero, J. G., Ortega, G. T., & Vilanou Torrano, C. (2016). El deporte europeo en la crisis del siglo XX. Un palimpsesto posmoderno. *Ars Brevis*, (22), 304-351.

Fukuyama, F., & Reina, P. (2002). *El fin del hombre: consecuencias de la revolución biotecnológica.* Barcelona: Ediciones B.

Harari, Y. N. (2016). *Homo Deus: breve historia del mañana*. Madrid: Debate.

Kant, E. (2009). ¿Qué es la Ilustración? *Foro de Educación*, *7*(11), 249-254.

Koval, S. (2006). Androides y posthumanos. La integración hombre-máquina, Diego Levis, sin número de publicación, 1-22.

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1853–1859.

Kurzweil. R. (2015). La singularidad està cerca. Cuando los humanos trascendamos la biologia. Madrid: LolaBooks.

Lafontaine, C. (2000). La cybernétique matrice du posthumanisme. Cités, 4, 59-71.

Ledesma, E. T. (2018). Construcción de una Tipología de las Formas Tecnológicas de Vida. El caso del cyborg Neil Harbisson. Congreso Internacional de Tecnología, Ciencia y Sociedad. Lisboa.

Linares Salgado, J. E. (2018). De la naturaleza a la tecnoespecie: La proyección antropotécnica de la condición humana. Contrastes: revista internacional de filosofía, 23(2), 77-95.

Panikkar, R. (1991) *El "tecnocentrisme". Algunes tesis sobre tecnologia*. Barcelona: La Llar del Llibre.

Panikkar, R. (1991). El tecnocentrisme, algunes tesis sobre la tecnologia, en la nova innocència. Barcelona: La llar del llibre.

Sandel, M. (2015). *Contra la perfección*. Barcelona: Marbot Ediciones.

Savulescu, J. (2009). Genetic interventions and the ethics of enhancement of human beings. *Readings in the Philosophy of Technology*, 417-430.

Scheler, M. (2001). *Ética*. Madrid: Caparrós Editores.

THE ETHICAL ASPECTS OF A "PSYCHOKINESIS MACHINE": AN EXPERIMENTAL SURVEY ON THE USE OF A BRAIN-MACHINE INTERFACE

**Yohko Orito, Tomonori Yamamoto, Hidenobu Sai, Kiyoshi Murata,
Yasunori Fukuta, Taichi Isobe, Masashi Hori**

Ehime University (Japan), Ehime University (Japan), Ehime University (Japan),
Meiji University (Japan), Meiji University (Japan), Health Sciences University of Hokkaido (Japan),
Waseda University (Japan)

orito.yohko.mm@ehime-u.ac.jp; yamamoto.tomonori.mh@ehime-u.ac.jp;
sai.hidenobu.mk@ehime-u.ac.jp; kmurata@meiji.ac.jp; yasufkt@meiji.ac.jp;
tisobe@hoku-iryo-u.ac.jp; horimasa@waseda.jp

ABSTRACT

A brain-machine interface (BMI), one of the emerging cyborg devices, processes signals acquired from a human brain and translate them into a meaningful output in accordance with a given purpose such as operating a machine remotely. In this respect, a BMI system can function as a psychokinesis mechanism. This system can widely be utilised for various purposes including for enhancing healthy people's intellectual and/or physical abilities. However, ethical and social issues concerning such technology use have not been fully examined. This study aims at investigating these issues. To attain the aim, the authors constructed a simple experimental environment where a non-invasive wearable BMI device was put on the head of a healthy person as a subject of the experiment to operate a robotic arm remotely or without touching it. The interview surveys were conducted with subjects before, during, and after the experiments, to investigate their attitudes to BMI usage, feelings of robotic arm operation and ethical awareness of brain signal collection by the BMI device. The results of the surveys revealed their attitudes to and ethical concern about the BMI use and suggested research agenda for the future.

KEYWORDS: brain-machine interface, cyborg, privacy, responsibility.

1. INTRODUCTION

Owing to the development and increasing use of cyborg devices such as smart glasses, smart watches, powered exoskeletons and RFID chips in various fields including of medicine, education, commerce and sports, the cyborgisation of human beings is being accelerated. The cyborg devices can extend human intellectual and physical abilities, thus they are expected to assist those with congenital and/or acquired disabilities. Among them, a brain-machine interface (BMI), or a brain-computer interface (BCI), has recently attracted attention. Most BMI systems consist of four sequential components: signal acquisition, feature extraction, feature translation and classification output (Rupp et al., 2014). Based on these components, a BMI enables communication between a human brain and external devices through sending signals from the brain to devices and vice versa using dedicated hardware and software. According to the results of a survey conducted by Nijboer et al. (2013), people involved in BCIs tend to consider that BCI systems 'measure signals from the central nervous system and "translate" those signals into output signals' (p.545). As their study suggests, BMI systems process signals acquired from a human brain and translate them into a meaningful output in accordance with given purposes such as remotely operating a machine or sending messages over a long distance. In this respect, a BMI system can function as a psychokinesis or telepathy machine.

Non-invasive wearable BMI devices have been used for medical or rehabilitation purposes. In this case, typically, patients' brain signals are collected by BMI hardware such as an electroencephalograph (EEG) and are processed by a dedicated BMI software application to operate devices remotely or in a non-contact manner just by setting their intention to do so. Thanks to such a BMI system, for example, those who could not move their bodies at will successfully worked as waiters in a coffee shop by remotely operating humanoid robots; during this process, the wearable BMI devices were non-invasively connected to their brains (Ory Labo). So far, BMI devices and systems have been proposed to be used for other purposes than medical one, such as for gaming (Nijholt et al, 2009; Nijholt , 2008) and marketing (Guger et al., 2014).

BMI use for a wider range of purposes may exert a substantial influence over individuals, organisations and society as a whole. However, ethical and social issues regarding the social penetration of BMI systems have not been fully discussed. The BMI research to date has tended to focus on the operability, functionality and/or usability of BMI devices or the effectiveness of the BMI system for medical treatment or rehabilitation. In addition, there are fundamental problems in predicting and evaluating the social risks or ethical issues relating to BMI usage, because such technology has not been used by healthy people in daily-life settings. Healthy people do not usually recognise the necessity for such technology, thus it is unlikely that the use of wearable or implantable BMI will be used by many of healthy people in the future. One way to investigate the ethical and social aspects of BMI usage in a wider context is to conduct an experimental survey of healthy people's using BMI devices.

To adopt this way, the authors constructed a simple experimental environment where a non-invasive wearable BMI device was used by a healthy person as a subject of the experiment. He/she was asked to operate a robotic arm connected to the BMI device, which was put on his/her head, without using any part of his/her body. Interviews with data subjects were conducted, before, during and after experiments, to examine their attitudes to BMI device usage, feelings of robotic arm operation and ethical awareness of brain signal collection by the BMI device. Interview questions were prepared based on the results of the authors' previous studies (Murata et al., 2019; Murata et al., 2018; Murata et al., 2017; Isobe, 2013).

2. ETHICAL QUESTIONS ON THE USE OF BMI

While most of existing studies on BMIs have focused on their clinical or rehabilitation use, some researchers have discussed the ethical issues surrounding BMI use (e.g. Kansaku, 2013; Schermer, 2009). They conducted questionnaire and interview surveys concerning ethical evaluation on the development and use of BMIs, most respondents to which were medical or rehabilitation professionals and patients using BMI devices (e.g. Gilbert, 2019; Nijboer et al., 2013; Isobe, 2013). However, only a few surveys of healthy people, who are neither experts nor professionals in relevant fields, concerning their attitudes to BMI usage have been conducted. Given that BMIs will soon be used in society for a broader range of purposes, the ethical issues and social risks caused by its usage should be examined in a proactive manner taking socio-cultural and economic contexts around BMIs into account. To conduct this examination, the following questions, for example, may be raised.

- If a BMI system malfunctions contrary to its users' intentions, who is responsible for the malfunction? How can we decide who are responsible people or organisations?

- Should users' brain signals collected by BMI systems be protected as sensitive personal information?
- Is it socially or legally acceptable that brain signals obtained from individuals while their using BMI systems are utilised for not-initially-intended purposes? For example, is it acceptable that a BMI system is used as a lie detector, and, if that's the case, under what conditions?
- What benefits and risks do exist when BMI are used in a specific context?

In the near future, it is expected that implantable BMIs or BMI brain chips will become available. In that case, people may be required to decide whether they implant the chip in their brains, evaluating risks and benefits associated with the implantation and continuous use of the chip. In this regard, the following questions are also raised.

- Under what conditions is the implantation of a BMI chip into the brain – a very complex and not-well-known organ, which deeply relates to human dignity – justified?
- Should an equal opportunity in BMI chip implantation be guaranteed for all? Should the disabled be prioritised? Is the difference of opportunity between the rich and the poor acceptable?
- Should the autonomy of an individual's decision to become a cyborg using an implantable BMI be respected? Is it acceptable that an individual is forced to implant a BMI chip to play his/her social or professional role?
- How is an individual's self-recognition and self-identity transformed when an implantable BMI device is embedded in his/her brain? Should the mental transformation be cured, and how?

These ethical and social questions regarding the use of BMI chips should proactively be addressed to predict and avoid any subsequent and future risk. Based on the interests and concerns described thus far, this study attempts to examine how individuals feel about their enhanced abilities acquired by using a non-invasive BMI device in an experimental laboratory setting, as the first step toward responding the interests and concerns.

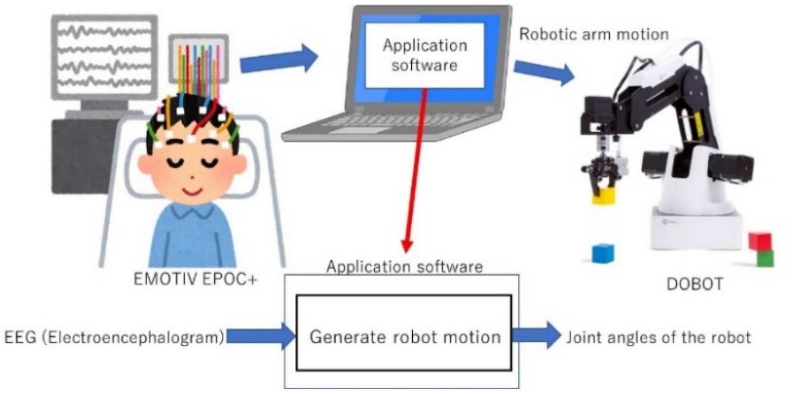
3. OVERVIEW OF THE BMI EXPERIMENT AND THE INTERVIEW SURVEY

3.1. Outline of the experiment and the interview

An experimental environment was designed and set up to allow subjects of the experiment to control a robotic arm using only their brain signals. As shown in Figure 1, a non-invasive wearable BMI/EEG device (EMOTIV EPOC+), a robotic arm (DOBOT) and the dedicated application software were located in an experimental laboratory.

At the beginning of an experiment, the EEG device was put on a subject's head, which enabled to measure his/her brain signals. He/she was then invited to join the training process, where his/her brain signal data were collected and recorded by the EEG system. Brain signal data in two kinds of mental states – a 'relaxed state' and an 'in-operation state' – were collected. To acquire the data of a subject in the in-operation state of mind, he/she was required to imagine that he/she was pushing a small box displayed on the computer screen toward the back. The two kinds of data – the relax state data and the in-operation state data – acquired in the training process were transmitted to and stored in the application software system installed in a personal computer, to which the EEC device is connected.

Figure 1. The experimental environment.



After completing the training process, a subject was asked to maintain his/her relax state of mind for a while, and then to intentionally move to the in-operation state with imagining he/she was pushing the robotic arm toward the back. When his/her brain signal patterns acquired at this time was similar to the in-operation state ones stored in the software system to a satisfactory extent, a robotic arm turned toward the back. During the experiment, a subject was required to conduct this several times. In addition, the experiment was designed so that a subject once experienced a preprogrammed robotic arm movement towards the near side when he/she asked to move to the in-operation state. This is intended for letting a subject experience malfunction of the BMI system.

Before, during and after the experiment, a subject was asked to grant semi-structured interviews prepared for it. Some interview questions were responded in written form at a later date. The interview sheet was designed so as to examine subjects' attitudes to and recognitions of their experience with the BMI system during the experiments, based on previous studies (e.g. Gilbert et al., 2019; Nijboer et al., 2013; Tamburrini, 2014; Isobe, 2013; Fukushi & Sakura, 2007). The interview questions pertained to: (a) privacy and personal data protection, (b) human autonomy and dignity, (c) identity development and personal transformation, (d) the acceptance of body extension in an individual and organisational context, (e) the workplace cyborgisation and (f) social responsibility and informed consent. The interview sheet was designed so that a subject could consider the benefits and risks associated with implantable BMI devices, though a non-invasive device was used in the experiment.

3.2. Survey participants

The surveys of five healthy undergraduate students, who majored in commercial science, were conducted in February 2020 at Meiji University, Tokyo. Their attributes, knowledge about a BMI, and expectation and anxiety about the experiments are shown in Table 1. Most subjects had known little about a BMI and all of them had positive feelings about the experiment rather than negative.

Each subject was informed of the purposes and methodology of this study and the contact information for enquiry in advance the experiment. In particular, based on the ethical policy adopted in this research, he/she was clearly notified that he/she had total say over whether to participate in this survey or not, and the experiment would never put him/her in a disadvantageous position. The expected risks entailed in the experiment and the measures to prevent them were also explained. After offering the full explanations about the study, each subject was required to sign a consent form when they determined to participate in the experiment and the interviews.

Table 1. Experimental subjects (n = 5).

ID	Age	Gender	Have you ever heard about a BMI or are you familiar with a BMI?	Expectation/anxiety about the experiment (Weak 0 – Strong 7)
1	21	Female	I have heard only a little about it before in class.	5/1
2	22	Male	No, not at all.	6/1
3	21	Male	No. I have no idea what I'll do during the experiment.	5/3
4	21	Male	Yes. I have seen people with disabilities who operated a robotic arm on a TV programme.	6/3
5	21	Male	No, I have never heard about BMI, although I have heard about brain tech.	7/1

4. THE RESULTS OF INTERVIEW SURVEY

4.1. The feelings about the operability of the BMI system

The observed ease or difficulty in operating the robotic arm and the speed of operation varied among subjects. During the experiment, they were asked how they felt about the operation of the robotic arm using a BMI device. Their responses are summarised in Table 2.

Table 2. Feelings about the operation of the robotic arm using a BMI.

Q: Did you have a sense that you operated the robotic arm during the experiment?	
1	The robotic arm actually turned when just casting my eyes. I had little sense of my operating it. I felt that the robotic arm autonomously turned when something came into my field of vision during the time of my intensively imagining that I was pushing the robotic arm.
2	Toward the end of the experiment, I felt that my operation of the robotic arm became easier, although this feeling was not necessarily solid. In the beginning, the image of pressing the robotic arm in my mind was weak. The best tip I can give other participants is that the robotic arm turns when you become calm and empty your mind.
3	I had the sense that I was actually pushing the robotic arm. However, when my will to push became weaker accidentally, the robotic arm moved, so I felt frustrated. This robotic arm move was observed between my attempts to say 'move!' in my head. So, there was definitely a time lag between my attempt and the movement of the robotic arm.
4	From the third or fourth attempt, I gained the sense that I was actually operating the robotic arm. I tried to see the scene of the robotic arm's move in my mind, which I saw at the first or second attempt. When I recalled the scene with the sound at that time, the robotic arm actually started to move. In the beginning, I vaguely imagined the move of the arm. Through repeated attempts, I became able to visualise the specific image of how I operate the robotic arm.
5	I did not have the sense that I operated the robotic arm at all. The robotic arm moved at an unexpected time. This is strange. Actually, the robotic arm moved at the moment when I thought it would not move and my mental concentration became weak. It might move with a delay. Maybe, the robotic arm moves when I don't have a strong intention to push it toward the back. Intuitively, the robotic arm likes a perverse person, because it moves against other's will.

The interviews after the experiment provided the authors with interesting findings: after multiple operations, Subjects 1, 2 and 3 felt that their physical states related to robotic arm movement.

> Subject 1: ‘I don’t know why the arm turned. The key to understand this may be my experience during the experiment that the arm didn’t move unless I didn’t make a motion of pushing something with my hands. If I had a lever or the like to operate the arm, I would have had a different feeling. But, in this experiment setting (where she could not operate the robotic arm physically), I still want to operate the arm with making that motion. I can’t operate without doing so. Maybe, this is because I am a dancer.’
>
> Subject 2: ‘The robotic arm did not turn when I kept my eyes open. It moved immediately as I closed my eyes. This may be a just coincidence, but it’s a fact that my attempts to push the arm in my head with closing my eyes worked.’
>
> Subject 3: ‘I think I could easily make the robotic arm in motion when my body wasn’t tensed up or was rather relaxed. It became easier for me to operate it in the latter half of the experiment in which I took myself less seriously.’

4.2. Subjects’ recognition of the intended malfunction of the BMI system

When the robotic arm moved in an unexpected way, how did they recognise this? Subjects’ feelings about such a move, which were talked about during the experiment, are summarised in Table 3. All subjects considered that the movement of the arm, which pretended the malfunction of the BMI system, was their faults or due to their failure.

Table 3. Feelings about unexpected moves of the robotic arm.

Q: Why do you consider the robotic arm moved in an unexpected way?	
1	It’s due to my changing the position of my hands. I didn’t do anything else in particular. I think that when it moved toward the near side, so did my hands.
2	I highly concentrated my mind on pushing the arm, after attempting to pull it lightly in my head. I can guess this caused its strange movement. But, I cannot understand why. I don’t think the robot made a mistake. I’m a little bit confused.
3	Maybe because I attempted to move the arm anyway – not to push it toward the back – in my head, the arm moved toward the near side. In the beginning of the experiment, I understood that the robotic arm could turn to any direction, and this was embedded in my memory. So, I might unconsciously consider that ‘to move the arm’ is to it anyway, and because of this, the arm turned toward the near side. My unconscious mind was read by the system, and consequently the arms turned toward the back and near side. I don’t think that the robot was out of order, but something was wrong with my attempt in my head. I don’t feel that there was some coding error.
4	I consciously stopped my attempt to move the robotic arm in my head – actually, I considered nothing then – but, the arm moved toward the different direction. My mental operation didn’t work well. I had several successful attempts, but the last one took a little time and maybe I became impatient or a little bit upset. I told myself ‘relax, relax’, but then the arm moved. I didn’t think anything in particular. Perhaps my impatience resulted in such a strange movement.
5	If several kinds of movements of the robotic arm were preprogrammed, it would move in accordance with the programs. But, I don’t have any evidence about this, and I think I’m completely wrong. If my brain signal patterns were steady, the machine arm would not malfunctioned at all. I think that my brain signal patterns were somewhat unstable, so my attempt to move the arm in my head was misinterpreted, leading to the strange move of the arm.

4.3. The recognitions of enhanced abilities enabled by the BMI system

Subjects’ recognition of their abilities enhanced by the BMI system, which was questioned about after the experiments, were mixed, as shown in Table 4.

Table 4. The recognition of enhanced ability.

Q: Do you think your ability was enhanced by using the robotic arm?	
1	I don’t think my ability was enhanced, at least in the sense that the arm was not part of my body. It was, I felt, rather like my buddy or a replacement of my hand outside my body. If the arm showed its functions better than my hand’s, I would feel my ability was enhanced. But, as long as the arm is outside my body, all I can do is that I operate it as intended. (How would you feel, if there were many robotic arms at home and they performed household chores instead of you?) In that case, also, my ability would not be enhanced, but I could just control the convenient tools.
2	I don’t think so, because I couldn’t really imagine the robotic arm’s move, though it moved. If I get used to operate it, or if it has similar traits and appearance to a human arm and has fingers to pick something up, I might think so. In that case, I might be able to feel it was my own arm. In reality, my physical ability was decreased, because I did not move the arm with my body.
3	I had a sense that my ability was enhanced a little. I’ll become better at operating the robotic arm, if I train myself a bit more. If we have robots at home and workplace which can be controlled by our mind, our capability will surely be increased. In addition, the environment surrounding robots plays a key role. If such robots become available to anyone, human beings will evolve to a new level. We’ll be able to do what we can’t do now.
4	The experiment brought a brand-new experience to me. I felt my ability was improved, and I was enhanced. If I can operate the robotic arm as if it were one of my limbs, I’ll surely feel that I become more able. I cannot imagine any opposite situation.
5	My ability was expanded, but not strengthened. Thanks to the robotic arm, my arm’s reach was expanded, but this is not the improvement of my ability. The arm is just a tool. If my ability to use this tool is better than others’, then my ability is improved. But, if I rely on it too much, my ability will be weakened because I’ll less engage in physical work.

4.4. The attitudes to the application of BMI technology

After the experiment, each subject was asked to answer the questions about possible application of BMI technology in daily life and the risks entailed in it. The results are shown in Table 5. While almost all of subjects considered the technology as the useful devices for their daily life, some of them mentioned the risks caused by malfunction of a BMI system.

Table 5. Attitudes to the application of BMI technology in daily life.

Q: In your daily life, for what kind of work or activity do you want BMI technology to be applied? Is there any risk or problem such application would entail?	
1	Simple tasks like housework.
2	I’d like to become able to remotely turn on and off an air conditioner and other appliances using a BMI. But, I’m concerned about malfunction of the system and brain damage it would give.
3	I hope smart home environment will be created. In a business setting, I want to finish simple, miscellaneous tasks (such as responses to emails) before commuting to work using a BMI . But, any trouble caused by the malfunction of the technology is problematic.
4	I would like to use a BMI to regulate my life such as maintaining good sleep habits. I’m afraid hacking into or malfunction of such a BMI system would bring serious danger such as an individual user’s falling asleep suddenly during the day.
5	Mail order of daily necessities and social networking service handling. Problems are malfunction of the BMI system, and so on.

After the experiment, they were asked to give written responses to the question about expected application fields of BMI technology, and the following fields were mentioned: construction industry, disaster support, medical field including disability aid, physical labour support, advanced intellectual activity support and entertainment including games. In addition, when subjects were asked about the situations in which an implantable BMI are used, almost all of respondents expressed their concern about malfunction of it. For example, Subject 4 and 5 pointed out the risky phenomenon regarding implantable BMI use as well as the benefit of it as follows.

> Subject 4: 'This should never be used with a war objective. I'm worried about the advent of pain-insensitive or excessively fearless soldiers by using this brain chip. In addition, the use of this implantable chip could bring about a socio-economic gap between the haves and have-nots. This can't function as a scholarship, though a brain chip to enhance an implantee's learning capability sounds a good idea.'
>
> Subject 5: 'I don't like to be implanted it, even if I can't deny the implantation. If it can be implanted and removed at will, it may be acceptable. I never want to be implanted it for safety reasons and given the risk of brain damage. If its use becomes widespread in society, I may want to use it. If I can't live without it, then I will use.'

4.5. The attitudes to personal data collections by a BMI device

Other questions asked after the experiment pertained to subjects' awareness of data collection by the BMI system. The answers to the questions are shown in Tables 6 and 7. As described in Table 7, a subject expressed his concern about misuse of those data, whereas other subjects thought that brain signal data were not sensitive.

Table 6. Attitudes to EEG data collection.

Q: How did you feel about the EEG's collecting data of you?	
1	No problem at all.
2	It was a little uncomfortable, because I had never put something like that (the EEG) on my head. I am interested in it, but I know nothing about any technological feature of it.
3	I don't feel any aversion. I think it is because I am ignorant about the technology. I wonder what can be found from my brain signal, and I think no one wants to look into my heart. There may be no relationship between one's thought and brain signal. It's no problem for me that my brain signals are read and used.
4	It's not uncomfortable for me at all. I don't want my mental states to be grasped as an image, but because I don't think the brain signal does not convey details of that image, it is OK. On the other hand, it's not pleased for me that my mental states are grasped in detail by others.
5	I don't recognise any risk. I hate to think that my thinking patterns are read by someone else. But, I don't feel bad about someone else's acquiring my detailed brain signals.

Table 7. Attitude to brain signal data collected by the EEG.

Q: Do you think that brain signal data the EEG collects need to be protected carefully as sensitive personal data? What do you think if your brain signal data are utilised for lie detection?	
1	It's not too bad for me that the brain signal data are used to read my emotions or for lie detection. I believe my brain signals represent what I think more accurately than my recognition of it. But, I don't feel that the data are my personal information.
2	From such data, I can learn about my own physical condition that others can't see. It is okay that I look at the data. Whether other people can see the data or not is decided on a case-by-case basis. If specialists look at my data, they will find out everything about me. But, anyone other than them can't do this, so the data are not sensitive. If I am an employee and my condition was evaluated by my employer based on my brain signals, they demonstrate my true condition. I can use the data to explain my condition rationally. In the case of a lie detector, the brain signals do not express everything. It is only one of many things that can be used to decide whether I'm a liar or not. Brain signal data are not sensitive from my personal viewpoint.

3	I agree that a certain protection is needed. Perhaps it is the same as when it comes to marketing. As long as my personal information is provided to researchers of brain signal, I hope my data is managed properly. It's a little bit sensitive, isn't it? A lie can be detected by how nervous you are – this is rather sensitive. But, compared with other kind of personal information, I don't think brain signals are sensitive.
4	I think brain signal data should be managed properly, because not all of us feel good about arbitrary usage of the data. If a lie is detected using the data in real time, such data usage will pose a problem for ordinary human communication. I think brain signal data is particularly important personal information, more than other kinds of one. If my brain signal data are handed to others and used to evaluate me, this is really terrible.
5	I think emotions are personal information that can't be hidden. Because I don't trust machines, I pay no mind to whether machines handle my information or not. Information gained from machine processing is not necessarily correct, and we can treat it as merely one of many. I would provide information if its price is appropriate. I think brain signal data are not sensitive because the information is constantly changing.

4.6. The recognitions of legal regulations for BMI usage

Finally, the question concerning respondents' recognition of the necessity of legal regulations for BMI use was asked. The outcomes are shown in Table 8. Except Subject 1, all subjects recognised the necessity of legal restriction on BMI usage.

Table 8. The necessity of legal restriction on BMI usage.

Q: Do you think legal restrictions on BMI devices usage is necessary?	
1	I don't think it should be regulated by law. In a company, better standards or rules can be set up. The rules don't need to contain a punitive clause. It's a company's job to set their own standard as to how to deal with the information they collect. They can certainly decide on it, because this relates to their productivity.
2	I think it is necessary. Training for using BMI devices in accordance with the regulation is absolutely necessary. If this is conducted well, we don't need to worry about any problematic use of them. If the devices are connected to appliances at home to control them, any malfunction of them can be prevented. If the devices are used for military purposes, their users have to be trained harder. It is necessary to establish a clear criterion in your mind of acceptable usage of the devices. I can't think of any justifiable reason for using them at work. It is better that the users have the right to claim information disclosure on the devices.
3	Regulations are necessary. Even if risks are explained, some people would be suspicious about it. (In terms of using the brain signal to manage employees) While this may be rational, I personally and instinctively feel uncomfortable about this. In this regard, I feel regulation may be necessary. (If a person works with such BMI devices, how do you think?) I would be envious of him/her. But, if it comes to a member of my family or someone close to me, I would be worried about whether they would be harmed by the device use.
4	Regulations are necessary. Such a device can be used by criminals to commit violence, murders and abduction. We need to know how to use it, and to have an established way to stop its functioning by the police. I agree that the police can use it, but obviously this is not the case when it comes to miscarriages of justice and crime committed by the police.
5	Some crisis awareness is necessary. I don't know what to think about the risks. Because the risks are invisible, we should be more cautious. There is a possibility that some people would face an irrational situation in which a decision is made based on brain signals, such as in an analysis of emotions. This is meaningful information for those in a customer-facing industry. Standardising human feeling leads to impoverishment, but it cannot deal with diversity of humanity.

5. DISCUSSIONS

Through the experiments and subsequent interview and questionnaire surveys regarding BMI devices and systems, the authors gained the following findings and insights.

- Operability of BMI: There was a wide range of variations in subjects' recognition of the operability of the BMI system. This may have depended on the accuracy of the BMI devices; therefore, improvements in experimental devices such as changes in default settings used by experimental systems should be reconsidered. On the other hand, the subjects seemed to develop original ideas regarding how the robotic arm could be moved by their brain signals; they subsequently attempted to explain their own interpretation, and some of them subjectively experienced physical synchronisation of their body move with the robotic arm's one. Similarities between the methods, which may be found with repetition of the experiments among the subjects, could be useful for understanding the deep relationships between physical movement and brain function and/or to examine how human recognitions and feelings are strongly related to physical body movements.
- Recognition of responsibilities to operate a BMI device: When the robotic arm moved in an unintended way, the common responses by all subjects was to assume that it was their faults or due to their failure to emit correct brain signals. These results may have been caused by the authors' explanation or by the experimental environment; therefore, there is a need to re-examine the experimental situation. Alternately, their attitudes may imply that the users of information systems or cyborg devices tend to have higher recipiency or to feel a responsibility for the misconduct. Furthermore, it may be difficult to consider various possibilities and reasons around the malfunction of the system when the default settings are given and explained.
- Self-enhancement, enhanced abilities: While the subjects recorded various opinions and feelings about their enhanced abilities enabled by BMI devices, there is the possibility that their feelings and awareness of the BMI devices may change when the experiments are repeated and they become able to operate the robotic arm more effectively. Therefore, it is necessary to evaluate their own awareness of their enhanced ability and self-consciousness with BMI devices in continuous manner.
- Risk awareness on development and usage of BMI: Throughout this study, it seems from the data that the subjects expressed their recognition regarding the use of BMI more concretely than they would have done if they had merely completed a questionnaire or interview survey without the experiment. In this regard, the experiment itself had an educational effect in terms of information ethics to analyse the social influence of emerging technologies, such as cyborg technologies. While the subjects were aware of the risks caused by malfunction or unintended movements of BMI devices and the necessity of regulations, they also had unarticulated anxiety about them. Furthermore, the invasion of privacy issues related to the collection of brain signal data and its analysis may have been difficult for the subjects to imagine. Then, it may be required that the interview sheets and question items should be revised to be easier to respond for the subjects.

6. CONCLUSIONS

As a first step in examining the ethical aspects of BMI usage, this exploratory survey conducted qualitative investigations of BMI usage of five data subjects. Through the investigations, subjects' personal experience of BMI usage and their expectations and ethical concerns about using BMI devices were investigated. While the number of subjects was small, these outcomes provided valuable insights for the development of more appropriate experimental environments and questionnaire sheets.

As these results and subject responses imply, the development and application of BMI devices face plenty of operational challenges and controversial ethical issues; the acceptance of body extension, awareness on protection of brain signal data, the necessity to restriction. Moreover, considering all subjects in this study are university students who have represented their opinions and feelings to a certain degree, if the experiment for the professionals and researchers conducted, they may emit more specific opinions on the survey. With many kinds of research, future challenges and ethical issues regarding BMI technologies should be addressed in an ongoing manner.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 17K03879 and 19K12528, the Kurata Grants subsidised by the Hitachi Global Foundation, and the Meiji University Grant-in-Aid for the international collaborative research project 'Cyborg Ethics'. We also appreciate Prof. Shizuka Suzuki, Dr. Yoshitaka Moritsugu, and all the participants in the experiments and researchers who support our study. It was certified that all procedures performed in the experiments of this study, which involved human participants, were in accordance with the ethical standards of the research ethics committee set up at the Faculty of Medicine, Ehime University (issued Jan.27.2020, No. 2001001).

REFERENCES

Gilbert, F., Cook, M., O'Brien, T., & Illes, J. (2019). Embodiment and estrangement: Results from a first-in-human 'Intelligent BCI' trial. *Science and engineering ethics,* 25(1), 83-96.

Guger, C., Brendan, Z. A., & Edinger, G. (2014). Emerging BCI opportunities from a market perspectives. In Glübler, G. and Hildt, E. (eds.), *Brain-Computer Interfaces in their ethical, social and cultural context* (pp. 85-98). Dordrecht: Springer .

Fukushi, T. & Sakura, O. (2007). Ethical implementation of research and development on Brain-Machine Interface. *Keisoku to Seigyo*, 46(10), 772-777, Retrieved from https://doi.org/10.11499/sicejl1962.46.772 (in Japanese)

Isobe, T. (2013). The perceptions of ELSI researchers to Brain-Machine Interface: Ethical & social issues and the relationship with society. *Journal of Information Studies*, (84), 47-63 (in Japanese).

Kansaku, K. (2013). EEG based BMI systems for practical application. *Japanese Journal of Cognitive Neuroscience*, 14(3), 185-192, Retrieved from https://doi.org/10.11253/ninchishinkeikagaku.14.185 (in Japanese).

Murata, K., Arias-Oliva, M., & Pelegrín-Borondo, J. (2019). Cross-cultural study about cyborg market acceptance: Japan versus Spain. *European Research on Management and Business Economics*, 25(3), 129-137.

Murata, K., Fukuta, Y., Orito,Y., Adams, A. A., Arias-Oliva, M. & Pelegrín-Borondo, J. (2018). Cyborg athletes or technodoping: How far can people become cyborgs to play sports? Presented at ETHICOMP 2018, 25 September 2018, Retrieved from https://www.researchgate.net/publication/327904976_Cyborg_Athletes_or_Technodoping_How_Far_Can_People_Become_Cyborgs_to_Play_Sports

Murata, K., Adams, A.A., Fukuta, Y., Orito, Y., Arias-Oliva, M., & Pelegrín-Borondo, J. (2017). From a science fiction to reality: Cyborg ethics in Japan. *Computers and Society*, 47(3), 72-85.

Nijboer, F., Clausen, J., Allison, B. Z., & Haselager, P. (2013). The Asilomar survey: Stakeholders' opinions on ethical issues related to Brain-computer Interfacing. *Neuroethics*, 6(3), 541-578.

Nijholt, A. (2008). BCI for games: A 'state of the art' survey. In *International Conference on Entertainment Computing* (pp.225-228). Berlin, Heidelberg: Springer,

Nijholt, A., Bos, D. P. O., & Reuderink, B. (2009). Turning shortcomings into challenges: Brain–computer interfaces for games. *Entertainment computing*, 1(2), 85-94.

Ory Labo, Retrieved from https://tinyurl.com/y4acthl8

Rupp, R., Kleih S.C., Leeb, R., del R. Millan J., Kübler A., & Müller-Putz G.R. (2014). Brain–Computer Interfaces and assistive technology. In Glübler, G. and Hildt, E. eds., *Brain-Computer Interfaces in their ethical, social and cultural context* (pp.7-38). Dordrecht: Springer .

Schermer, M. (2009). The mind and the machine. On the conceptual and moral implications of brain-machine interaction. *Nanoethics*, 3(3), 217-230.

Tamburrini, G. (2014), Philosophical reflections on Brain-Computer Interface, In Glübler, G. and Hildt, E. eds., *Brain-Computer Interfaces in their ethical, social and cultural context* (pp. 147-162). Dordrecht: Springer.

3. Educate for a Positive ICT Future

COMPUTER ETHICS IN BRICKS

Gosia Plotka, Bartosz Marcinkowski

De Montfort University (United Kingdom), University of Gdansk (Poland)

malgorzata.plotka@dmu.ac.uk; bartosz.marcinkowski@ug.edu.pl

ABSTRACT

It has been recognised that students of Information & Communication Technologies (ICT) tend to find their majors more of a challenge than their peers from any other course. That has a significant impact on their retention and engagement within the subject taught. What makes the content difficult to grasp is the fact that it is usually full of abstractive concepts that students tend to silo from those learnt on different modules, so they miss their chance to build an in-deep understanding of the topic. Hence, threshold concepts have been introduced to help students focus on what is vital and construct their understanding of the topic by organising troublesome content. Teaching social, ethical, and professional aspects of ICT additionally requires bringing up different perspectives to give learners the ability to discuss – and reflect critically. Therefore, the paper establishes a theoretical framework for incorporating ethical and professional values into curricula and raising the awareness of students regarding the relevance of such values for their sustainable ICT professional development. The students, by engaging in communicative learning, not only build up their self-esteem but also get a space to experiment with technologies they will soon be developing and discuss any concerns that may affect their development as professionals ready to enter the labour market.

KEYWORDS: Ethics; Lego Serious Play; Curriculum; ICT; Constructivism; Threshold concept.

1. INTRODUCTION

Modern businesses are increasingly demanded by society to comply with the standards of Corporate Social Responsibility (CSR) (Kim & Han, 2019; Patrignani & Kavathatzopoulos, 2016). Not only failing to meet these standards exposes an organization to several risks – such as contributing to environmental or corruption-related scandals, consumer boycott, or even state intervention – but also prevents the organization from taking advantage of certain opportunities. Such opportunities may include, but are not limited to, increasing brand recognition, attracting investors, increase workforce commitment, retaining consumer loyalty, or improving the bottom line. Over last few weeks, together with the outbreak of COVID-19, we observe a number of examples when companies and governments are taking advantage of the situation and putting privacy and data of civilians at risk (Stein, 2020).

One might argue that computer professionals are in a specific position in the context of this trend. On the one hand, the level of their access to corporate data often makes them vital links in keeping an eye on CSR of parent organizations. On the other, they often elude institutional CSR programs due to the high demand for IT professionals on both the freelancers and start-ups markets. Therefore, raising awareness regarding the potential impact of their decisions on individuals, society and environment among future computing professionals cannot be overestimated.

The title of this paper has been inspired by the movement called *ethics in bricks* that uses popular Lego bricks to disseminate and explain some of the ethical dilemmas and concepts on Social Media (Twitter, Facebook and Instagram). This paper aims at presenting first stages of R=T (i.e. research equals teaching) study on how to make some troublesome content popular and easier to understand and equip students in soft skills that are not the most natural one for those studying computer science. It is computer ethics that is among the topics covered by such content. The originality of the research lies in:

- identifying the needs of different stakeholders involved in the educational process – including students, teachers, Information & Communication Technologies (ICT) industry, society, and professional bodies like BCS, IEEE, or ACM (Tassone et al., 2018; Voskoglou & Buckley, 2012);
- putting the threshold concept, fundamental ideas, and transformative learning as lenses that helped the researchers to see the face value of explored phenomena to work;
- seeking the synergy of multiple methods – including Design Thinking (DT), i.e. a solution-based approach to finding what would-be users really need (Dam & Siang, 2018);
- adoption of the Lego Serious Play and reframing.

Following the Introduction, the contributors provide an overview of the problem and related literature studies in section 2. The third section outlines the research approach taken. Next, the theoretical framework for this research is elaborated. The framework aims at delivering a set of guidelines on how ethical and professional values should be incorporated into curricula and presented to students, so they see such values as must-have competencies for their sustainable development that meets today's and future job market needs. The findings so far are introduced in section 5, followed by the conclusions.

2. RESEARCH BACKGROUND

2.1. Why shall we bother?

Prof. Simon Rogerson (2010) – the founder and former director of the Centre for Computing and Social Responsibility at the De Montfort University in Leicester, the first centre of such type in Europe – naming benefits and disadvantages of technological innovation claims that people becoming addicted to digital amenities are inclined to unreflectively accept them without considering or not being aware of their impact. This assessment is still valid a decade later, as confirmed by Patrignani and Whitehouse (2018). The number of domains going through a digital transformation driven by the industry, business and governments in a physical, psychological and economic way being driven by the ICT industry, business and governments. Many of those have already been discussed on several occasions during the ETHICOMP conference series. All the impact that technology may have needs to be understood and properly addressed. Didactic, social, ethical and professional aspects of computing seem to be a must for those who are entering the profession so that they are equipped in the right skillset and knowledge to make relevant decisions in their workplaces. In his speech for Orkney College, University of the Highlands & Islands, Rogerson (2019) argues that one needs to look after the least in power, i.e. members of the society who have no capability to take care of themselves against the negative impact of technological development.

There is an expectation that computing-related courses ought to be accredited by professional bodies such as BCS to ensure that the students gain industry-standard training/skills and are prepared for employment upon graduation (Times Higher Education, 2017). Both honesty and ethicality are included on the list of skills that are highly demanded by the labour market (Chartered Management Institute, 2018; Lindley et al., 2013). Therefore, it seems to be vital to provide future computing professionals with the relevant information regarding their potential impact on individuals, society and environment (Tassone et al., 2018; Voskoglou & Buckley, 2012) on top of the knowledge on some more technical aspects of system development.

2.2. How to teach computer ethics?

Costa and Pawlak (2018) in their abstract submitted to ETHICOMP2018 summarise some previously expressed views on how practical computer ethics should look like. They bring up an assessment by Soraker (2010) who highlights that (1) the bulk of computer ethics-related literature is directed towards other computer ethicists; (2) is simply boring; (3) explore self-evident topics; (4) is irrelevant to the actual practice of software engineering. Another highlighted view comes from Connolly and Fedoruk (2014). They state that education in computer ethics is theoretically unsound and empirically under-supported. Moreover, ICT professionals need to explicitly understand the social contexts of computing – while faculty staff ought to put significantly less focus on ethical evaluation. Costa and Pawlak (2018) argue that despite the case studies, recent publications continue a strategy that features a lack of social context or people's behaviour/physiological response. Rogerson (2019), in turn, sees the value in encouraging students to debate real-life problems and try to look for both problems and solutions by discussing given topics. This approach needs to involve both rigour and justification that come from students' qualitative reflections.

The article written by Portela (2017) may come handy here. The authors describe their educational model based on Kolb's learning cycle and Gary et al.'s (2013) iterative teaching methodology. The latter integrates preparation, experimenting, reflecting and conceptualisation to evolve students' competencies and was validated in practice. Similarly, Portela's model combines:

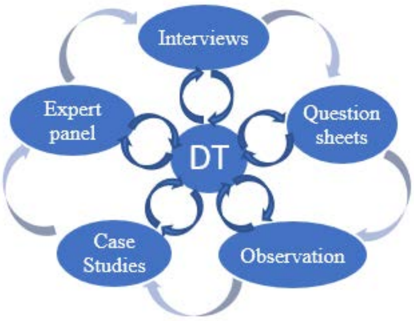
- preparatory exercise – reading of technical articles (that requires the verbal-linguistic intelligence) or watching relevant video materials (what in turn is related to the spatial-visual intelligence);
- followed by discussions of case studies to understand the state of the art (logical intelligence);
- problem-based learning – the execution of practical exercises (bodily-kinaesthetic intelligence);
- gamification that helps to contextualise problem (spatial-visual intelligence);
- and reflection (intrapersonal intelligence).

This approach not only involves practice – but also refers to multiple intelligences theory and ways how people learn as defined by Gardner (2011). Such a mix satisfies UDL (Universal Design for Learning) principles that involve providing flexible (1) study resources; (2) ways to learn; and (3) ways to demonstrate knowledge and makes this approach even more inclusive (Marcinkowski, Carroll-Mayer & Plotka, 2020). Rogerson (2019) adapts the contributions by Confucius (450 BC) and Aristotle (349 BC) to contemporary conditions by highlighting the value of learning ICT by doing and experimenting; members of academia simply allow computing completing their courses and joining other professionals without understanding technology, its impact on the society and taking full responsibility of their actions or inaction. On top of that, participative learning is only possible should colleges and universities move from tutor- to student-led teaching, where – as observed by Carruthers (1953) – lecturers make themselves "progressively unnecessary".

3. RESEARCH APPROACH

To build pedagogical capacity in computing for the benefit of the student and future computing, we sought an answer to the following question: *what is the best way to incorporate social, ethical and professional aspects into a computing curriculum*? In this project, a multi-method approach was used. Years of domain-related practice that featured a number of empirical cases allowed the authors of the paper to make the DT a centric component of the research design. DT, as a solution-based approach that allows testing different ideas, suits well the project helps to find out what users really need (Dam & Siang, 2018). The approach to data collection was inspired by inter-relationship cycle suggested by Rowe (2002) and Weyman (2007) as it is shown in Figure 1.

Figure 1. Design Thinking.



Source: self-elaboration based on Rowe (2002)

Employing a few techniques as a part of the multi-method approach enabled examining both the current theory and practical experiences as well as opinions (Weyman, 2007). Namely, capturing different perspectives and building a bigger picture helped, in a pragmatic manner, to better understand the context of the problem. Such an approach is consistent with a number of well-described examples (Cavallo & Ireland, 2014; Merali & Allen, 2011). Therefore, the overall project in some ways requires a thoughtful and holistic approach. Also, the DT method is open for creative techniques using LSP to generate (ideation phase) and synthesis data gives very good results.

4. THEORETICAL FRAMEWORK

Development of guidelines for the computing-related curriculum that includes social, ethical and professional aspects through hands-on activities required adapting some theoretical framework. The theoretical framework is a way that any researcher looks at the world that allows him/her to get inside the problem within the investigated context. In this study that lenses that helped the researchers see the face value of explored phenomena are a combination of a threshold concept, fundamental ideas and transformative learning.

4.1. Through practitioner community to a sense of belonging and security

Students of computing courses – as per their disadvantaged background, lack of maturity or fact that computing courses are perceived as "nerdy" – are likely to struggle with identity (Gordon, 2016). This impacts their confidence and engagement with the group/content taught as well as a will to remain part of their groups (courses). As explained by Lave and Wenger (1991) in their works on practitioner community, people, students and teachers must engage fully (physically, emotionally as well as with their relations and thinking) into active learning together. That social regulation of education not only reinforces their learning and understanding the subject. It also impacts their sense of belonging and sense of security through being recognised as valid members of the group throughout their progress from newcomers to advanced participants – which helps to negotiate their identities. Such depersonalisation is understood as a change of perception of the individual in respect to the group conditions becoming a part of the group when their reach the level of their social identity that enables them to place themselves within that group, engage (Hogg & Terry, 2000).

Davies (2006) also emphasises the importance of a community that shares the same way of thinking and practise across a learning process. By interacting with people, a learner may construct their perspective on the world in line with a community point of view – but without taking ownership for their own understanding of the subject (Wenger , 1999). Unless they make a conscious decision whether they want to be part of it or stay outside of the discourse (should their way of seeing the world did not correspond with the point of view of the community), they may be unable to see the world through their lenses. This approach makes a difference between students being only able to repeat the content and those who apply it successfully. People who acquire knowledge through real-world or realistic experience (learning and acting) get a better understanding of the meaning of the subject and the world. According to Brown, Collins and Doguid (1988), even should there be an initial decrease in understanding, it eventually enables opening perspective and results in learning to become a life-long process. Building a relationship within the discipline and/or community may impact students' way of thinking about their potential more than their actual performance. Causing or reinforcing well-being and mental health is more important to those who are first-generation students enrolled in higher education (Stebleton, Soria & Huesman Jr, 2014). However, a false sense of security may lead the higher-year computing students to build on their misconception of what they know and are able to do as a result of surface knowledge (Gary, 2015). As students find difficult to link together previously acquired knowledge Gary finds it challenging to quiz students on what they may know or not, and, subsequently, encourage to transfer into deeper learning. The authors of this paper based on their practice over years observed a similar phenomenon: students, especially the knowledgeable ones, tend to be afraid of the questions that do not come directly from the coursebook and are more likely to challenge the teachers regarding their approach. The students do not like any deviation from the way they were taught so far. Also, they like to know beforehand what questions are going to be there – so that they can prepare. They are reluctant to sail into uncharted waters because they may not succeed – expressing the fear of failure. Hence, one of the possible means to pave the way for depersonalisation could be by imposing certain rules – such as telling a coherent story behind a model, just as it takes place in Lego® Serious Play™ (LSP®) (Harn, 2018).

Why are social, ethical and professional aspects of computing any different from any subject discussed during Computer Science, Software Engineering, ICT and other similar courses? This kind of content expects students to employ lots of skills that are not usually required when completing other modules. Those skills include, but are not limited to, critical thinking, debating, evidencing their arguments and reflection. Any ethical aspects can be found contra-intuitive as they naturally allow for discussion where people may disagree about "what" to do rather than "why" to do it (Greene, 2015). Additionally, learning computer ethics requires participation in the group and approaching real-world problems – what, as previously mentioned, proves to be a significant challenge for computing students. Therefore, looking for a suitable learning tool is important to look for the one that helps properly to address that challenge.

4.2. Dealing with troublesome knowledge

The moment of "getting it" is like lighting bulb effect: students are now able to see what used to be invisible and read between the lines. Eckerdal et al. (2007) observed that effect moving from "līmen" (Latin equiv. for threshold) to lúmen (light, an opening) when learners leaving liminal space (liminality as a state of being in-between) after a while of getting stuck there or reaching an in-deep understanding of the subject. They dubbed it a sudden insight. Individuals, in line with principles of constructivism, are actively constructing their knowledge and transfer their understanding of the topic (Clancy, 2004; Eckerdal et al., 2006). However, being overloaded with new concepts, linking new and existing knowledge (moving from known to unfamiliar contexts) in silos from other related modules may lead to misconception. According to Eckerdal et al. (2006), isolating knowledge acquired across different modules does not help a student to transfer and a link between subjects the knowledge that could enable them to get an in-depth understanding of the topic. The way to overcome this problem could be by introducing threshold concepts – i.e. subsets of the core concepts in the discipline (Eckerdal, et al., 2006) that help to organise content taught and to focus on what is the most important. Nicola-Richmond, Pépin and Larkin (2018) stress that once a threshold concept is grasped, the world is changed forever. In their article, they discuss lighting bulb moments with the participants of their study run in the healthcare environment. They also observe that it is not so much about acquiring the skills and knowledge required by the curriculum but about the transition from student to professional mentality. Rogerson (2019) calls it an experimental journey of maturity: the ability to truly understand and confidently apply the knowledge into practice. In result, students not only master technical knowledge (*what*?) but do it in a way (*how*?) so they make a positive impact as well. To really understand something, it is expected from learners' not just to memorise facts and understand how to apply the rules, but rather to actively look for an opportunity to construct their knowledge based on the experience of learning with others – as only this guarantee being exposed to different perspectives (Walker, 2013). This comes very useful in teaching social, ethical and professional aspects of computing, where the learning process ought to take into account familiarising with such rules as code of ethics and conducts, applying those in an analysed situation, but also discussing different scenarios and challenging status quo. It seems to be obvious that in order to be able to see things in a different way one needs to be exposed to different perspectives. Suggested by Gary et al. (2013) and promoted by Portela (2017), the iterative teaching methodology sets an order of delivering content in a way that process starts with a theoretical introduction of the topic and follows through group discussion of possible options, practical application, and – finally – reflection.

This research uses the threshold concept, one of the educational principles to make teaching computer ethics more effective. Threshold concepts introduced by Mayer and Land (2003) enable a new way of thinking about a phenomenon, thus enhancing the students' ability to master their subjects (Advance HE, 2015). Identification of threshold concepts may start with pinpointing a list of core ideas – like, proposed by Schwill (1994), fundamental ideas. It is worth to bring back here the vertical character of fundamental ideas that implies teaching the same aspects at different levels, with making them more and more complex by adding additional details as we progress. It can be equated to growth within a community from adept/newcomer (outsider, or a person who just joined) to a full member (insider).

Irvine and Carmichael (2009) explain that since threshold concepts depend on the context, they quickly become a point of focus to build a shared understanding of ideas covered within a practitioner or expert community – what makes them very useful in professional learning. Participation (dialogical) metaphor, unlike acquisition (monological) metaphor, requires building knowledge as a part of the community that shares the space and object of their development. That can, of course, be a practitioner community. Wegner (2011) describes it in terms of a group of people who, intentionally or incidentally, work hand in hand on a collective goal. Such a community could be an Indian tribe, rock band or students trying to understand a subject taught – as originally introduced by Lave and Wenger (1991) in their learning model. To ensure that this collective, in fact, forms a practitioner community there must be: (1) a shared interest; (2) members' cooperation while performing tasks; (3) practice involved. Concept of newcomers joining a group of professionals to learn and becoming part of the community through legitimate peripheral participation was introduced by Lave and Wenger (1991) and explored by Allen (2005). The former authors, similarly to Clouder (2005), remark usefulness of the threshold concept in confronting students with range aspects of dilemmas coming from the subject learnt. These threshold concepts could be therefore what Nissani (1995) calls distinctive components – that describe specific ways of approaching a problem for each discipline and include its important elements. It is particularly important when it comes to creating new knowledge within interdisciplinary research.

Taking different perspectives, for instance during discussing the topic with colleagues within the practitioner community, provides principles for interpreting (Mezirow, 1990). Action bereave of reflection becomes habitual while reflective action may lead to further reflection (critical reflection) on the process – such as learning. To develop critical reflection, one needs to allow their presuppositions to be challenged and revised. Therefore, a reflective thinker makes an informed decision based on collected evidence to support his/her judgment. The ability to critically see own action and inaction is crucial in understanding social, ethical and professional aspects of computing.

4.3. Lego ® Serious Play^{TM} and reframing

As the threshold concept is often described as troublesome, knowledge teaching requires an approach that helps students with engaging effectively in their own learning process (Barton & James, 2017). Namely, using metaphors and applying active and creative methods such as Lego Serious Play and reframing seems to produce a very positive outcome. Lego® Serious Play® (LSP) is a method introduced in the ‘90s of the XX century by Kristiensen, Victor and Roos. It was popularised (as an open-source) in 2010 by the Lego Group to support communication and problem-solving. LSP can be used as an alternative tool to ideate (brainstorm) and conceptualise the outcomes – for example during meetings or focus groups. In LSP®, bricks become mediators that enable participants to answer even the questions that seem to be very abstract at the beginning through the story (Barton & James, 2017).

It has been recognised that cognitive processes – such as learning and memory – are highly influenced by the way people use their bodies to interact with the physical world (Gauntlett, 2010). For example, “talking and thinking with hands” by employing social constructivism principles and facilitating physical interaction with the properties of the problem in a more natural way (Vallée-Tourangeau & Vallée-Tourangeau, 2016), proved to be a powerful way of overcoming some barriers with expressing an opinion and reflecting on own work or discussed topic (Executive Discovery, 2014). Together with depersonalisation, thinking with hands makes a way to develop perspective thinking that (1) helps to embrace a diversity of learning needs; and (2) enhances the sense of security through the narrative process and reframing personal experience (Harn, 2018).

Reframing is another cognitive technique that could be successfully used to assist students in building up their confidence in their skills. It breaks a problem down into layers by going through questions: what → who → when → where → why. The approach shows similarity to the one presented in Kipling’s poem *Six honest serving men* – although the latter introduces an additional question: “how”. As observed by Reeve (2014) during her work with Art & Design students, this approach helps to think of study direction, the context of the problem as well as generate and synthesis new ideas. It also enables focusing on a single aspect at the time, reframing way of thinking about the subject (James & Brookfield, 2014), and, by going through different frames, visualise thought process (English, 2011). Once frames are completed, students get a better understanding of the complexity of their topic (Reeve, 2014). In turn, when they read back, the order is reversed and starts with “why” for a better storyline.

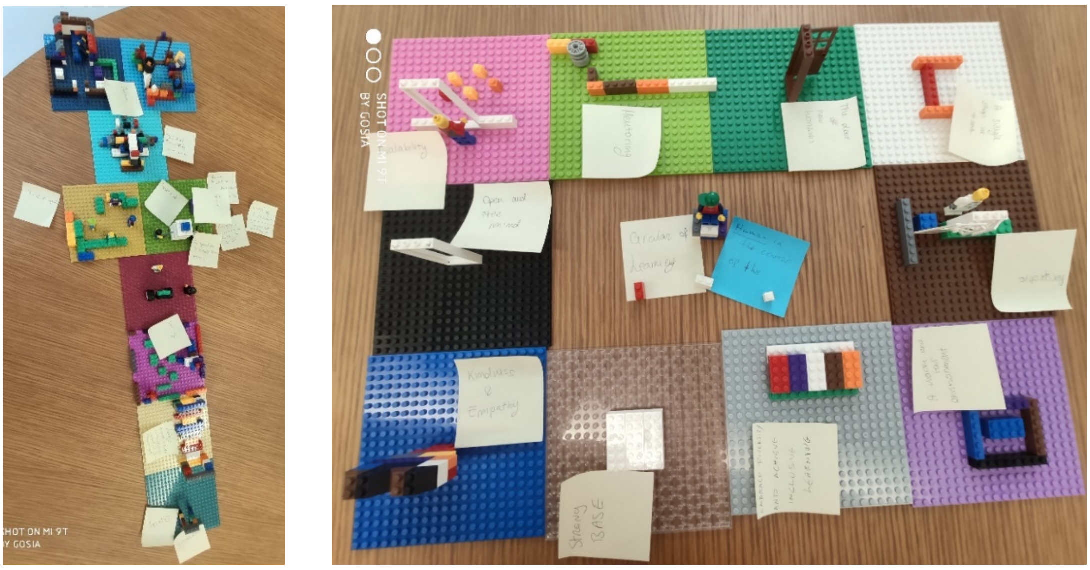
Both LSP and reframing enable people to reflect on their experiences and rethink them, ensuring their sense of security at the same time. Reframing, as a way to look at the discussed situation from a different angle, enables changing or adjusting further move towards the innovative problem-solving. According to Anderson (2019), storytelling constitutes a powerful communication tool. A story is a way to show the other person perspective. By listening or sharing a story, people develop empathy and build a relationship with each other (Snow & Lazauskas, 2018).

5. PRELIMINARY RESULTS

As per data collected in accordance with the adopted research approach, the authors of this paper learnt that the proposed iterative approach helps a student to achieve a lighting bulb effect. The flexibility of study resources, ways of learning and demonstrating knowledge in our experience makes this approach truly UDL-friendly by addressing different learning needs and preferences. Not every student experienced the moment of revelation at the same point of time, or when performing the same type of activity. Some got it during more traditional activities – e.g. while answering questions set up by tutor. Others through practical activities, like building LEGO models and elaborating stories behind them, or during a reflective group discussion over their shared model. Talking through some concepts in a classroom environment helps to eliminate extremes and behaviours that are not accepted by the group. Some of the students still struggled. That said, as this approach helps them confront a sense of security and open themselves to different perspectives and needs, the number of students who may feel left behind decreases. It is the discussion on copyrights jointly with researchers that may serve as an example to back this mechanics up. Is it socially justified to download content that is not meant for open use without paying for that? Despite some legal environments that allow it for private use exclusively, the group opted for the position that individuals presenting pro-downloading arguments were just seeking to provide a rationale and to justify their theft.

One occasion when authors had a chance to apply their approach was the 90-minute-long *ACM celebration of women in computing* workshop (including methodological introduction and a hands-on activity performed by participants) in Rome, 2019. During the practical part, the participants divided into two groups of nine were asked how to build a positive ICT future and an inclusive society in the Information Age. Even though they initially found question far too abstract, they ultimately managed to come up with their own model and communicate its story to the rest of the team. And then, jointly as a group, they combined the components into a shared idea (Figure 2). Constructing models was followed up by the discussion on what is important in becoming an ICT professional with a human-centred approach.

Figure 2. ICT student pathway towards professionalism and a human-centred cycle of learning.



This event help contributors to check first hand in practice how LSP can (1) help people deal even with initially very abstractive idea, (2) bring together on the same page people from different backgrounds making creating a group out of them within a relatively short time. That was also observed by the teachers who participated in the workshop. One of them declared to buy LEGO to use in their classroom with their students.

6. CONCLUSIONS

As both the research background and the preliminary outcomes of the study demonstrate, there is a need to create an environment where they can explore their course subjects and have an opportunity to practice what they have learnt. It is especially noticeable in the computing domain – where students seem more likely to be disengaged and fail. Knowledge transfer should be based on the practitioner community, where students can explore metaphors and discuss abstraction – as one of the leading concepts in computing. At the same time, a sense of security ought to be ensured, so students were not afraid to explore given topics. Building a practitioner community aids in teaching regular computing modules, and may help in teaching social, ethical and professional aspects of computing on top of that. As the latter might be considered highly subjective – probably more than any other subject within this domain – it is crucial that students engaged with the topic and came across perspectives different than their own. Students should become a part of a community where they are exposed to experimentation, involved in discussions, argumentation, practical tasks etc. In our humble opinion, it is the optimal way to build up critical thinking – by being a part of the community, practitioners' community in particular, and continuously engaged.

ACKNOWLEDGEMENTS

The authors of the current paper would like to extend their sincere thanks to Dorota Filipczuk and Ruth Lennon. Without their contribution, the workshop in Rome would not have a shape it ultimately had.

REFERENCES

Advance HE (2015). *Threshold Concepts*. Retrieved from https://www.heacademy.ac.uk/knowledge-hub/threshold-concepts

Allen, V. (2005). A reflection on Teaching Law to Business Students. In *Proceedings of the Society for Research into Higher Education Conference.* Edinburgh: University of Edinburgh.

Anderson, C. (2019). *Storytelling Is a Powerful Communication Tool – Here's How to Use It, from TED*. Retrieved from https://ideas.ted.com/storytelling-is-a-powerful-communication-tool-heres-how-to-use-it-from-ted

Barton, G., & James, A. (2017). Threshold Concepts, LEGO Serious Play® and Whole Systems Thinking: Towards a Combined Methodology. *Practice and Evidence of Scholarship of Teaching and Learning in Higher Education*, 12(2), 249-271.

Brown, J., Collins, A., & Duguid, P. (1988). *Situated Cognition and the Culture of Learning* [Report No. 6886]. Cambridge, MA: Bolt Beranek and Newman Inc.

Carruthers, T. J. (1953). Discipline as a Means of Development. *The Phi Delta Kappan*, 35(3), 137-139.

Cavallo, A., & Ireland, V. (2014). Preparing for Complex Interdependent Risks: A System of Systems Approach to Building Disaster Resilience. *International Journal of Disaster Risk Reduction*, 9, 181-193.

Chartered Management Institute (2018). *All University Students Must Gain Leadership Skills, Says New Employability Report.* London: Chartered Management Institute.

Clancy, M. (2004). Misconceptions and Attitudes that Interfere with Learning to Program. In S. Fincher, & M. Petre (Eds.), *Computer Science Education Research* (pp. 85-100). London: RoutledgeFalmer.

Clouder, L. (2005). Caring As a 'Threshold Concept': Transforming Students in Higher Education into Health (Care) Professionals. *Teaching in Higher Education*, 10(4), 505-517.

Connolly, R., & Fedoruk, A. (2014). Why Computing Needs to Go Beyond Good and Evil Impacts. *Proceedings of ETHICOMP 2014. Liberty and Security in an Age of ICTs*. Paris: The University of Pierre and Marie Curie.

Costa, G., & Pawlak, P. (2018). *Practical Computer Ethics – An Unsolved Puzzle! Creating, Changing, and Coalescing Ways of Life with Technologies*. Warsaw: Polish-Japanese Academy of Information Technology.

Dam, R., & Siang, T. (2018). *What Is Design Thinking and Why Is It So Popular?* Aarhus: Interaction Design Foundation.

Davies, P. (2006). Threshold Concepts: How Can We Recognise Them? In J. H. R. Meyer, & R. Land (Eds.), *Overcoming Barriers to Student Understanding: Threshold Concepts and Troublesome Knowledge* (pp. 94-108). Abingdon: Routledge.

Eckerdal, A., McCartney, R., Moström, J., Ratcliffe, M., Sanders, K., & Zander, C. (2006). Putting Threshold Concepts into Context in Computer Science Education. *ACM SIGCSE Bulletin*, 38(3), 103-107.

Eckerdal, A., McCartney, R., Moström, J., Sanders, K., Thomas, L., & Zander, C. (2007). From Limen to Lumen: Computing Students in Liminal Spaces. In *Proceedings of the Third International Workshop on Computing Education Research* (pp. 123-132). New York, NY: ACM.

English, F. (2011). *Student Writing and Genre: Reconfiguring Academic Knowledge.* London: Bloomsbury Academic.

Executive Discovery. (2014). *The Science of LEGO® SERIOUS PLAY™.* Retrieved from https://thinkjarcollective.com/wp-content/uploads/2014/09/the-science-of-lego-serious-play.pdf

Gardner, H. (2011). *Frames of Mind: The Theory of Multiple Intelligences, 3rd Edition.* New York, NY: Basic Books.

Gary, K. (2015). Project-Based Learning. *Computer*, 48(9), 98-100.

Gary, K., Lindquist, T., Bansal, S., & Ghazarian, A. (2013). A Project Spine for Software Engineering Curricular Design. *Proceedings of 26th Conference on Software Engineering Education and Training* (pp. 299-303). San Francisco: IEEE.

Gauntlett, D. (2010). *Introduction to LEGO® SERIOUS PLAY®.* Retrieved from https://davidgauntlett.com/wp-content/uploads/2013/04/LEGO_SERIOUS_PLAY_OpenSource_14mb.pdf

Gordon, N. A. (2016). *Issues in Retention and Attainment in Computer Science*. Retrieved from https://documents.advance-he.ac.uk/download/file/4652

Greene, J. (2015). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics. *The Law & Ethics of Human Rights*, 9(2), 141-172.

Harn, P. L. (2018). LEGO®-Based Clinical Intervention with LEGO®SERIOUS PLAY® and Six Bricks for Emotional Regulation and Cognitional Reconstruction. *Examines in Physical Medicine & Rehabilitation*, 1(3), 1-3.

Hogg, M., & Terry, D. (2000). Social Identity and Self-Categorization Processes in Organizational Contexts. *Academy of Management Review*, 25(1), 121-140.

Irvine, N., & Carmichael, P. (2009). Threshold Concepts: A Point of Focus for Practitioner Research. *Active Learning in Higher Education*, 10(2), 103-119.

James, A., & Brookfield, S. (2014). *Engaging Imagination: Helping Students Become Creative and Reflective Thinkers.* Hoboken, NJ: John Wiley & Sons.

Kim, H., & Han, J. (2019). Do Employees in a “Good” Company Comply Better with Information Security Policy? A corporate social responsibility perspective. *Information Technology & People*, 32(4), 858-875.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge: Cambridge University Press.

Lindley, D., Aynsley, B., Driver, M., Godfrey, R., Hart, R., Heinrich, G., Unhelkar, B., & Wilkinson, K. (2013). Educating for Professionalism in ICT: Is Learning Ethics Professional Development? In J. Weckert, & R. Lucas (Eds.), *Professionalism in the Information and Communication Technology Industry* (pp. 211-232). Canberra: ANU Press.

Marcinkowski, B., Carroll-Mayer, M., & Plotka, M. A. (2020). Non-Attendance Factors – Can e-Learning Be Considered a Disincentive. *Information Technologies and Learning Tools*.

Meyer, J. H. F., & Land, R. (2003). Threshold Concepts and Troublesome Knowledge: Linkages to Ways of Thinking and Practising within the Disciplines. In C. Rust (Ed.), *Improving Student Learning: Theory and Practice – 10 Years On* (pp. 412-424). Oxford: Oxford Brookes University.

Merali, Y., & Allen, P. (2011). Complexity and Systems Thinking. In P. Allen, S. Maguire, & B. McKelvey (Eds.), *The Sage Handbook of Complexity and Management.* London: SAGE Publications Ltd.

Mezirow, J. (1990). How Critical Reflection Triggers Transformative Learning. *Fostering Critical Reflection in Adulthood*, 1(20), 1-6.

Nicola-Richmond, K., Pépin, G., & Larkin, H. (2018). Once You Get the Threshold Concepts the World Is Changed Forever: The Exploration of Threshold Concepts to Promote Work-Ready Occupational Therapy Graduates. *International Journal of Practice-Based Learning in Health and Social Care*, 6(1), 1-17.

Nissani, M. (1995). Fruits, Salads, and Smoothies: A Working Definition of Interdisciplinarity. *The Journal of Educational Thought (JET)/Revue de La Pensée Educative*, 121-128.

Patrignani, N., & Kavathatzopoulos, I. (2016). Cloud Computing: The Ultimate Step Towards the Virtual Enterprise? *ACM SIGCAS Computers and Society*, 45(3), 68-72.

Patrignani, N., & Whitehouse, D. (2018). Applying Slow Tech in Real Life. In N. Patrignani, & D. Whitehouse (Eds.), *In Slow Tech and ICT* (pp. 113-127). Cham: Palgrave Macmillan.

Portela, C. (2017). *Modelo Iterativo Para o Ensino de Engenharia de Software* [PhD Thesis]. Recife: Universidade Federal de Pernambuco.

Reeve, J. (2014). How Can Adopting the Materials and Environment of the Studio Engage Art & Design Students More Deeply with Research and Writing? An Investigation into the Reframing Research Technique. *Journal of Writing in Creative Practice*, 7(2), 267-281.

Rogerson, S. (2010). Ethics of Emerging Technologies. Retrieved from https://www.youtube.com/watch?v=enRARJEuBVk#

Rogerson, S. (2019). Teaching Computer Ethics. Retrieved from https://www.youtube.com/watch?v=sZWRj6G67x4

Rowe, R. (2002). A Multi-Methodological Approach to Emergency Call Handling in the Metropolitan Police Service. In G. Ragsdell, D. West, & J. Wilby (Eds.), *Systems Theory and Practice in the Knowledge Age* (pp. 79-86). New York: Springer Science+Business Media, LLC.

Schwill, A. (1994). Fundamental Ideas of Computer Science. *Bulletin European Association for Theoretical Computer Science*, 53, 274-295.

Snow, S., & Lazauskas, J. (2018). *The Storytelling Edge: How to Transform Your Business, Stop Screaming into the Void, and Make People Love You.* Hoboken, NJ: John Wiley & Sons.

Soraker, J. (2010). *Designing a Computer Ethics Course from Scratch*. Retrieved from http://www.soraker.com/designing-a-computer-ethics-course-from-scratch

Stebleton, M., Soria, K., & Huesman Jr, R. (2014). First-Generation Students' Sense of Belonging, Mental Health, and Use of Counseling Services at Public Research Universities. *Journal of College Counseling*, 17(1), 6-20.

Stein, S. (2020). *How to Restore Data Privacy After the Coronavirus Pandemic*. Retrieved from https://www.weforum.org/agenda/2020/03/restore-data-privacy-after-coronavirus-pandemic

Tassone, V. C., O'Mahony, C., McKenna, E., Eppink, H. J., & Wals, A. E. (2018). (Re-) Designing Higher Education Curricula in Times of Systemic Dysfunction: A Responsible Research and Innovation Perspective. *Higher Education*, 76(2), 337-352.

Times Higher Education (2017). *Computer Science – De Montfort University*. Retrieved from https://www.timeshighereducation.com/world-university-rankings/de-montfort-university/courses/computer-science

Vallée-Tourangeau, G., & Vallée-Tourangeau, F. (2016). Why the Best Problem-Solvers Think with Their Hands, As Well As Their Heads. *The Conversation*.

Voskoglou, M. G., & Buckley, S. (2012). Problem Solving and Computers in a Learning Environment. *Egyptian Computer Science Journal*, 36(4), 28-46.

Walker, G. (2013). A Cognitive Approach to Threshold Concepts. *Higher Education*, 65(2), 247-263.

Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity.* Cambridge: Cambridge University Press.

Wenger, E. (2011). *Communities of Practice: A Brief Introduction.* Cambridge: Cambridge University Press.

Weyman, T. R. (2007). *Spatial Information Sharing for Better Regional Decision Making* [PhD Thesis]. Penrith: University of Western Sydney.

EDUCATIONAL GAMES FOR CHILDREN WITH DOWN SYNDROME

Katerina Zdravkova

University Ss. Cyril and Methodius, Faculty of Computer Science and Engineering (N. Macedonia)

katerina.zdravkova@finki.ukim.mk

ABSTRACT

In the last 10 years, the incidence of Down syndrome increased worldwide. In order to improve the quality of life of these children, and to increase their life expectancy, many systematic measures have been undertaken. Inclusive education, which embraces educational, social and emotional practises, based on well-structured instruction, interventions and support in the classroom is definitely one of them. In parallel with the in-class activities, educational software stimulates the inclusion. This paper presents the recommendations how to create such educational applications together with the pilot study intended to develop literacy skills, basic mathematical competencies and memory. A small Android application was created and presented to children attending the recently open Day Care Centre for Down syndrome (DCCDS) in Skopje. The enthusiasm and interest to use the application is the greatest motivation to carry on with the study, to create more ambitious applications and after an approval by the experts and parents to offer them to the all the children in the country. The application can be easily adapted to all languages, making it available to much wider community.

KEYWORDS: Basic learning skills, Down syndrome, Educational games, Inclusive education, Mobile and tablet applications for children with disabilities, Preparation for independent life.

1. INTRODUCTION

Regardless of the considerably improved prenatal detection, and the good prenatal diagnostics of fetal DNA, including the highly sensitive Down syndrome specific non-invasive screening, the incidence of this congenital anomaly increased worldwide. Down syndrome is a chromosomal disorder, which causes: phenotypic characteristics; physical growth and nonverbal cognitive delays; mild to moderate intellectual disability; adaptive behaviour problems; and evidence of adult dementia (Chapman & Hesketh, 2000).

People with Down syndrome deserve the same opportunities and care as others, which results in increased life expectancy and better quality of life. This can be achieved by constant parental care and support, monitoring of the mental and physical conditions, medical therapies, and consistent community support (Reid, 2018).

Inclusive education proved to be the best way to provide educational, social and emotional benefits starting from very early childhood (Felix, 2017). The pilot study of the use of emerging computer technologies proved the improvement of the effectiveness of reading and writing therapies in children with Down syndrome (Gilmore et al., 2003). Moreover, it changed the attitudes towards this disability and improved the interaction with children with Down syndrome (Campbell et al., 2003). If well designed and implemented, specially created educational applications can significantly facilitate the process of inclusive education, enhancing the cognitive and learning skills of these vulnerable children.

The paper continues with a brief overview of the cognitive and neuropsychological profile of these children, which determine the features of the dedicated educational software for them. The cognitive and neuropsychological profiles of the children with Down syndrome, and the recommended features are briefly presented in Section 2. The creation of the pilot project implementing the recommended features is presented in Section 3. Each part of the application is introduced and illustrated in more details, preceded by an explanation of the virtual tutor. Section 4 is dedicated to children feedback. The paper concludes with the suggestions how to increase inclusive education for children with Down syndrome and with the intended future work within the project.

2. FEATURES OF EDUATIONAL SOFTWARE INTENDED FOR DOWN SYNDROME

Similarly to most children from Generation Z, children with Down syndrome have become familiar with the computer technology since their early childhood, particularly to tablets and mobile phones (Feng, 2010). Therefore, the potential of various applications, including educational software can be very important for their development and education, enabling them to stretch the skills and abilities. Children with Down syndrome are usually gifted for one-type skills: language, math, strategic thought or physical coordination. They typically manifest a deficit of attention, thus they are not capable of comprehending longer or more complex rules (Mason, 2015). Children with Down syndrome are not patient to wait for the application to download or to process the following steps (Skotko, 2005). They also need instant rewards for each successful outcome. Furthermore, it was noticed that children with Down syndrome have significant vision deficit and anomalies in colour discrimination (Krinsky-McHale, 2014), and a lack of control of muscles stiffness affecting their motor skills (Vicari, 2006).

These cognitive and neuropsychological profiles, amplified with the guidelines for supporting children with disabilities (Encarnação, 2018) and the recommendations of the specialists from DCCDS resulted in the following conceptual design criteria:

1. Intuitive gameplay with easy navigation and few, simple functionalities accessible by clicking over a perceptive icon, which is active throughout the whole image;
2. Clear interface with bright colours, clear contours, realistic and simple images, and without anthropomorphic features or facial expressions (Lee, 2018);
3. Adjustable progression pace, based on the performance of the Down syndrome child;
4. Virtual tutor who announces the game, and responds with an appropriate facial and voice expression (Herring, 2017);
5. Simple and unambiguous instructions, which are repeated whenever an image is touched;
6. Substituted single and double finger gestures by two touches: from the source place to the target (Landowska, 2018);

7. Learners are not capable of reading, so the instructions should be spoken or presented with the sign language;
8. Quick download and very short waiting time to advance from beginning to end;
9. Free of charge.

All the eight conceptual design criteria were carefully followed while creating the whole application. It is currently installed on 10 tablets, which are a donation to Day Care Centre supplied by a successful fundraising event. The stable version, which will include the modification initiated after observing the children's feedback, reactions provided by their parents, and particularly the specialists from the Day Care Centre will be offered free of charge for all the children interested to use it.

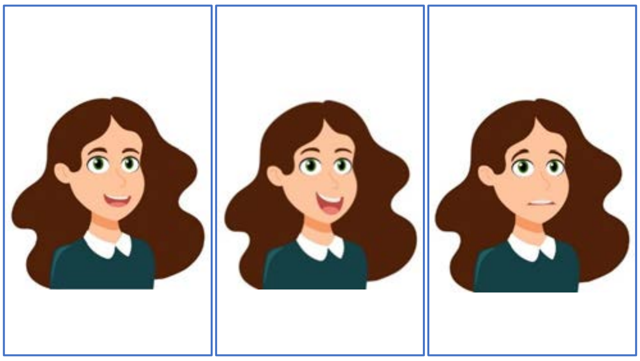
3. CREATED APPLICATION

The application consists of three integral parts: developing literacy skills, developing basic mathematical competencies and practising memory. The screens have a white background, few images and intuitive navigation, thus they fulfil the first two design criteria from the previous section. To give learners an opportunity to set their own pace, and to enable progress, all three parts have several levels, starting from the simplest and ending with the most advanced. They are compatible with the third design criteria.

The application was created using App Inventor 2 (App Inventor, 2020), a very nice and user-friendly development environment, which supports several operating systems. App Inventor 2 is a cutting age educational tool initiated by Google and maintained by the Massachusetts Institute of Technology. It was created in accordance to modern constructivist learning theories (Giordano & Maiorana, 2014). The decision to select the Android operating system was made due to its predominant share on the local smartphone market.

Each part of the application has its own virtual teacher, two young female teachers responsible for basic literacy skills and memory practice, and a young male teacher for mathematical skills (Fig. 1). Tutors have a full and deep voice and a perfect pronunciation. They introduce the task, the levels, speak out the names of the touched objects or pronounces the two navigation icons (the arrows in the upper left corner on Fig 2, and on the lower left corner on Fig 4.).

Figure 1. Three tutor's moods: instructional, happy and sad.



Source: Designs created by Ana Zdravkova

Most of the prospective learners are not capable of reading the instructions, the corresponding virtual teacher slowly speaks them, first by announcing the goal of the level, and then by introducing the task. If the learner accurately performs the task, tutor's face smiles and says a randomly picked congratulation with a happy voice. If the learner has failed, tutor's face becomes sad. After three wrong attempts, sad faced tutor suggests to repeat the task with a calm voice. After five consecutive mistakes, the advice is to go back to the previous level or to ask for help.

Whenever the learner successfully performs the task, one of the congratulations is loudly spoken. They are randomly picked out from the following list: amazing, bravo, compliments, great job, excellent, and well done. After finishing a whole task, the congratulations become stronger and personalized: you are a genius, you are gorgeous, and you are remarkable. At the end of the level, the learner is awarded with a badge (Fig 3.). If the learner has not been successful, the virtual teacher politely suggests to repeat the task. Whenever the learner persistently gives wrong answers, the suggestion is to ask the parents, guardians, or learning assistants for help.

The presence of a virtual tutor, which reacts after each successful or unproductive activity fulfils the fourth and the fifth conceptual design recommendation. Bearing in mind that the majority of the children with Down syndrome are not capable of reading, all the messages are spoken, which is in accordance with the seventh design criterion. In the near future, it is intended to combine the application with the avatars from the (Joksimoski, Chorbev, Zdravkova, Mihajlov, 2015).

During first testing of the pilot application, it was noticed that almost all of the children were not capable of implementing the drag and drop gestures. To avoid this obstacle, they were replaced by two separate activities, touching over the source place, and then touching of the target place. Implementing this approach, the sixth design criterion is also achieved.

Finally, all the images are extremely simple, and they are presented using the vector graphics. In total, the size of the whole application doesn't exceed 1MB. This constraint is much lower that the App Inventor 2 maximum size limit, enabling very quick download, as recommended by the eight design criterion. The following three subsections describe the three integral parts of the application in more details.

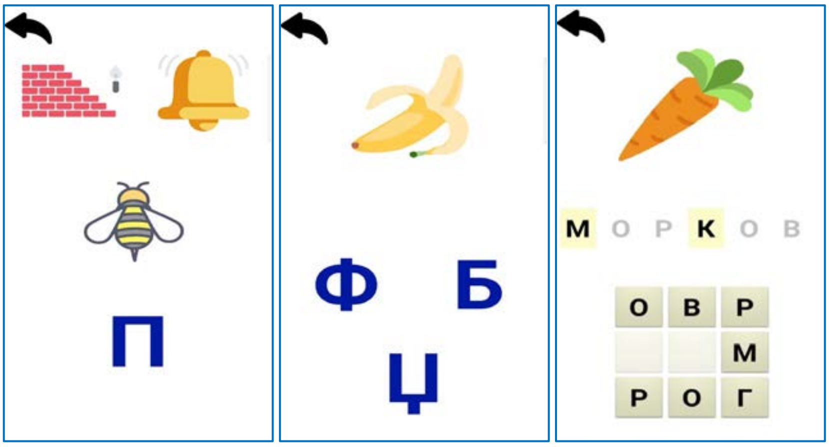
3.1. Developing literacy skills

The part intended for developing basic literacy starts with the introduction of the upper case and lower case letters, which are presented below the image that starts with that character. For each letter, at least three different images exist in the pool. For example, under the letter A a, the images of a plane (авион = avion), a car (автомобил = avtomobil), and a pineapple (ананас = ananas) are presented one by one. For the letter Ж ж, the corresponding images are a frog (жаба = zhaba), a turtle (желка = zhelka), and a giraffe (жирафа = zhirafa). After the presentation of the images for one letter, children are given an option to draw the uppercase letter. It can't be checked by App Inventor 2, so the success can't be verified.

The successful recognition of the letters is the task of the three games (Figure 2):

- Finding the corresponding image that starts with a presented letter,
- Finding the corresponding initial letter of the presented image
- Spelling a word presented as an image from distributed letters

Figure 2. Intermediate level of finding the correct word starting with an initial letter, the correct initial letter of the presented image and spelling a word with two syllabi.



Source: Mobile application for learning the alphabet for children with Down syndrome, developed by Iva Mihajlovska, B.Sc.

In the simplest level of the first game, two images are presented under the letter, the image that starts with it, and an image with a different initial letter. The intermediate level has three images, one of which is the correct one, and the most advanced level has four images.

The reverse game starts with two uppercase letters offered for one image, continues with three letters, and ends with four. Similarly to previous game, only one letter corresponds to that image.

In the third game, the letters of a simple word are randomly presented on the bottom of the screen. By clicking and positioning them at the right place in the middle of the screen, they make the word which is presented as an image, and after the successful spelling, it is spoken by the virtual teacher. The simple level consists of words with one syllable, the intermediate has two syllabi, while the most advanced level has more than three syllabi. In order to stimulate the recognition of the letters, the set of prospective letters contains characters that don't exist in the word.

3.2. Developing basic math competencies

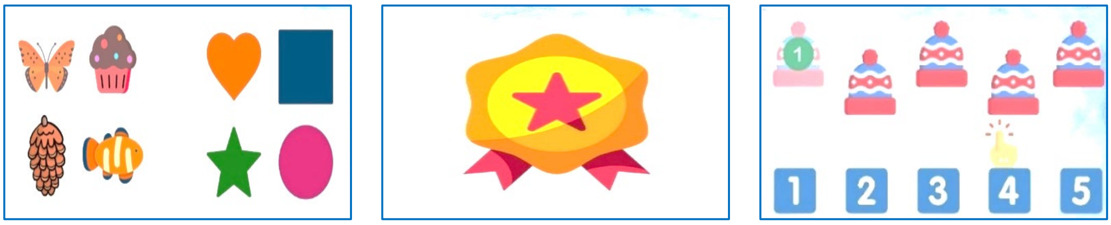
The order of the traditional concepts for developing basic mathematical competencies is the following: forms, relations, numbers, and measures. The first goal is to teach the learner to determine the exact form among these 3D forms: sphere, box, cylinder, and the 2D forms: square, triangle, circle and rectangle. The relations usually include: up – down; over – below; in front of – beside – between; inside – outside – over; left – right; identical – different – similar; big – bigger – the biggest; bigger – smaller; wide - narrow; high – low, fat – thin; deep – shallow; and same quantity – less – more. The introduction of numbers is accompanied by the relations less – more – the same. The measures start with the length, the weight, the time and currencies.

Each of these concepts is usually performed as an in-class activity, where learners use toys or special tools that enable them to practically resolve a task. The use of the tangible objects enable them to implement problem solving method based on trials and errors (Newell et al., 1958), which is more convincing and effective than a computer game. Therefore, the games intended for developing the basic mathematical skills in our application covered the amounts of objects and forms (Figure 3). The simplest level comprises the values up to 3, the intermediate up to 5, and the most advanced, up to 10. After the training session, where children see how to count the objects and the forms, they are invited to perform the same task. Depending on the level, the list with the squares containing the numbers up to 3, 5 or 10 is presented in the lower part of the screen. To improve the clarity of the interface, the orientation of the screen in this game is landscape.

This game introduced the rewarding with badges, and the demonstration of the correct answer (Figure 3, middle screen). A badge is obtained after 5 consecutive correct answers. After next 5 correct answers, a new badge is awarded, and the screen with all collected badges so far appears on the screen, together with a spoken personalized congratulation. However, not all attempts will be successful. If the child is not capable of getting the correct answer within two attempts, the application turns into a training mode, showing steadily the answer. Then, the child has an opportunity to repeat the same task individually. If even after the demonstration the success is not achieved, virtual teachers suggest to ask for help from someone.

These two concepts: rewarding with badges, and demonstration of the task solutions will be soon added to the games responsible for developing literacy skills, and practicing memory.

Figure 3. Introductory screen, the badge, collected and explanation of the correct answer.

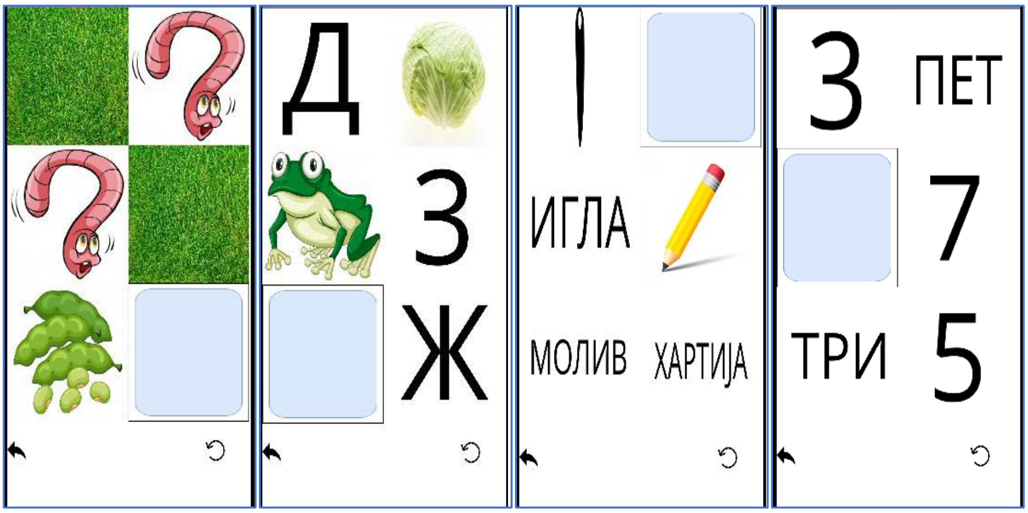


Source: Mobile application for developing math skills for children with Down syndrome, developed by Marija Krsteska, B.Sc.

3.3. Memory game

The memory game reinforces the skills gained in the previous two games, uniting the letters, the objects, the forms, and the numbers. For that purpose, four smaller units are created: coupling pairs of equal images; coupling the initial letter with the image; coupling the written word corresponding to the image; and coupling the numbers with the corresponding written word. Similarly to previous two games, three levels are established. The simplest level couples two pairs, the intermediate three, and the most advanced four pairs. In a near future, it is intended to extend the game with units responsible for coupling pairs of equal numbers and words, including the words representing the numbers, as well as coupling several identical objects with the value expressed with numbers.

Figure 4. Memory game, intermediate level: coupling equal images, the initial letter with an image, the written name with an image, and numbers with a word.



Source: Mobile application for practicing memory of children with Down syndrome, developed by Davor Trifunov, B.Sc.

This game was presented and tested in the Day Care Centre for Down syndrome in Skopje during September 2019. The feedback of the game is presented in the next section.

In February 2020, Davor Trifunov, one of the collaborators in the project, organized a very successful fundraising event, and with the support of ANHOCH (https://www.anhoch.com/), 10 tablets were provided for the Day Care Centre. The pilot application was downloaded on each of the tablets as currently the only content. At that time, the amount of children attending the Centre tripled, and the range of the children increased. It was a great opportunity to make a more methodical assessment and collect crucial information for its evolution. Unfortunately, due to the Covid-19 pandemic restrictions, the Centre was closed in the beginning of March. Ii will remain inactive until the end of this academic year. Therefore, the next version of the application will be presented next September, and the new feedback will be available later on.

4. FEEDBACK

Seven young boys and two girls aging from 15 to 19 and their parents were the first evaluators of the application. The game was installed on one tablet and demonstrated to every child individually. The age and the basic reading skills enabled them to successfully play the memory game. The whole event was touching for everyone. The kids were noticeably amused and attracted, except one girl, who was too shy. She listened the tutor with great attention and observed how the others played.

The most experienced boy comprehended the game immediately and asked to play the first. After trying all the options several times, he generously let others play. He manifested his frustration from the absence of an immediate congratulation after each successful coupling by lifting the speaker to hear the greeting.

Other five kids explored him, tried the game and managed to play it independently. The most extrovert boy succeeded after several trials and errors, and then tried to download the game from Google Play. Two kids, a boy and a girl created a strategy to first open all the tiles, and then couple them.

Two boys were not competent with the written words, one couldn't even discover the initial characters. They turned to the easier level of the game of own accord and were not enthusiastic to play it again.

During the second visit, all the kids, except the shy girl, activated the game and played it more competently, including the boys with lower literacy skills.

5. CONCLUSIONS

The ultimate goal of Day Care Centre for Down syndrome in Skopje is to prepare the kids for an independent life. They started making own meals under a full supervision of DCCDS staff and organized a cocktail with self-made bread and snacks. The next stage is to purchase the ingredients and start cooking according to a written recipe. To achieve this goal, their literacy and understanding of quantities should increase significantly. According to DCCDS staff and their parents, the educational game will be of a great use.

The major challenge is the indifference and the anxiety of some kids. Hopefully, they are very confident in using the smart phones. Before launching it on Google Play, the application will be polished and upgraded with new contents suggested by the specialists from DCCDS. As a consequence, those kids who were shy to show their incompetence or who were not interested to use it will be able to experience it with the support by their family members.

The educational game is in Macedonian only. It can easily be adapted to other languages, making it available to wider community.

ACKNOWLEDGEMENTS

This work is part of the project "Contribution to inclusive education of children with Down syndrome", which is partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University.

REFERENCES

App Inventor (2020), Retrieved from https://appinventor.mit.edu/explore/content/what-app-inventor

Appleton, M., Buckley, S., & MacDonald, J. (2002). The early reading skills of preschoolers with Down syndrome and their typically developing peers–findings from recent research. *Down syndrome news and update*, *2*(1), 9-10.

Campbell, J., Gilmore, L., & Cuskelly, M. (2003). Changing student teachers' attitudes towards disability and inclusion. *Journal of Intellectual and Developmental Disability*, *28*(4), 369-379.

Chapman, R. S., & Hesketh, L. J. (2000). Behavioral phenotype of individuals with Down syndrome. *Mental retardation and developmental disabilities research reviews*, *6*(2), 84-95.

Ekstein, S., Glick, B., Weill, M., Kay, B., & Berger, I. (2011). Down syndrome and attention-deficit/hyperactivity disorder (ADHD). *Journal of child neurology*, *26*(10), 1290-1295.

Encarnação, P., Ray-Kaeser, S., & Bianquin, N., 2018. *Guidelines for supporting children with disabilities' play: Methodologies, tools, and contexts*. De Gruyter Open.

Felix, V., Mena, L., Ostos, R., & Maestre, G. (2017). A pilot study of the use of emerging computer technologies to improve the effectiveness of reading and writing therapies in children with Down syndrome. *British Journal of Educational Technology*, *48*(2), 611-624.

Feng, J., Lazar, J., Kumin, L., & Ozok, A. (2010). Computer usage by children with Down syndrome: Challenges and future research. *ACM Transactions on Accessible Computing (TACCESS)*, *2*(3), 13.

Gilmore, L., Campbell, J., & Cuskelly, M. (2003). Developmental expectations, personality stereotypes, and attitudes towards inclusive education: Community and teacher views of Down syndrome. *International Journal of Disability, Development and Education*, *50*(1), 65-76.

Giordano, D., & Maiorana, F. (2014, April). Use of cutting edge educational tools for an initial programming course. In *2014 IEEE Global Engineering Education Conference (EDUCON)* (pp. 556-563). IEEE.

Herring, P., Kear, K., Sheehy, K., & Jones, R., 2017. A virtual tutor for children with autism. *Journal of Enabling Technologies*, *11*(1), 19-27.

Hodapp, R. M., Ly, T. M., Fidler, D. J., & Ricci, L. A. (2001). Less stress, more rewarding: Parenting children with Down syndrome. *Parenting: Science and practice*, *1*(4), 317-337.

Joksimoski, B., Chorbev, I., Zdravkova, K., & Mihajlov, D. (2015, October). Toward 3D Avatar Visualization of Macedonian Sign Language. In *International Conference on ICT Innovations* (pp. 195-203). Springer, Cham.

Krinsky-McHale, S. J., Silverman, W., Gordon, J., Devenny, D. A., Oley, N., & Abramov, I. (2014). Vision deficits in adults with Down syndrome. *Journal of Applied Research in Intellectual Disabilities*, *27*(3), 247-263.

Landowska, A., 2018. 8 Which digital games are appropriate for our children?. In *Guidelines for supporting children with disabilities' play*, 85-97.

Lee, J.M., Baek, J., & Ju, D.Y. (2018). Anthropomorphic Design: Emotional Perception for Deformable Object. *Frontiers in psychology*, *9*, 1829.

Mason, G.M., Spanó, G. & Edgin, J. (2015). Symptoms of attention-deficit/hyperactivity disorder in Down syndrome: effects of the dopamine receptor D4 gene. *American journal on intellectual and developmental disabilities*, *120*(1), 58-71.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological review*, *65*(3), 151.

Reid, W. H., Balis, G. U., Wicoff, J. S., & Tomasovic, J. J. (2018). *The treatment of psychiatric disorders*. Routledge.

Skotko, B. (2005). Mothers of children with Down syndrome reflect on their postnatal support. *Pediatrics*, *115*(1), 64-77.

Taylor-Phillips, S., Freeman, K., Geppert, J., Agbebiyi, A., Uthman, O. A., Madan, J., ... & Clarke, A. (2016). Accuracy of non-invasive prenatal testing using cell-free DNA for detection of Down, Edwards and Patau syndromes: a systematic review and meta-analysis. *BMJ open*, *6*(1), e010002.

Vicari, S. (2006). Motor development and neuropsychological patterns in persons with Down syndrome. *Behavior genetics*, *36*(3), 355-364.

IMPACT OF EDUCATE IN A SERVICE LEARNING PROJECT. OPENING UP VALUES AND SOCIAL GOOD IN HIGHER EDUCATION

Ana María Lara-Palma, Montserrat Santamaría-Vázquez, Juan Hilario Ortiz-Huerta

Universidad de Burgos (Spain)

amlara@ubu.es; msvazquez@ubu.es; jhortiz@ubu.es

ABSTRACT

In the academic course 19-20, the University of Burgos has launched a call for Service Learning Projects (SLP) with the aim of reinforce the academic skills of the students endorsing a societal transformation. These projects nonetheless pursue very worthwhile goals: to educate students at higher education not only in cognitive aspects, but also in personal growth. Service Learning is a groundbreaking and appeal methodology focused in acquire knowledge by doing community service work. This paper describes a real case scenario composed by students of the University of Burgos from two different careers at the University of Burgos (Degree in Occupational Therapy and Degree in Management Engineering). (N=23 3th, 4th grade students); all have contributed to resolve different challenges by using skills, competences and knowledge of their corresponding disciplines (health and engineering), more specifically, to design and manufacture low cost tools for helping disable people. The contribution adds a picture of how to achieve cognitive competences (technical), social competences (consciousness), ethical competences (compassion) and professional competences (productive versatility), among others. This novel scenario serve for a purpose: opening up values and social good in Higher Education.

KEYWORDS: Service Learning Projects, Innovative Education, Higher Education, Cognitive Competences, Social Competences, Ethical Competences, Professional Competences.

1. INTRODUCTION

The rate at which universities have been assimilating proposals in their educational environments has been constant. Since the first meeting at Praga in May 2001, efforts drive on getting improvements in specific and transversal competences of the students. Nowadays, eighteen years later, it is still present the way of doing innovation at the universities, and, the sustainable human development concept has been included within the topics and guides. Quoting Brotóns, (2009), "to improve the quality of teaching is mandatory to create real learning situations: with new innovative tasks, thinking in a positive ICT future and with acquisition, transfer and updating knowledge processes". Folgueiras, Luna and Puig (2011, pp.159) point out learning by using Service Learning tasks enhance students to "take part directly with those who are supporting, adapting to their needing's and facing up a realistic circumstances, really different from the classroom lectures and environments".

In order to reinforce the theoretical framework of the research, next it is highlight three approaches to the matter; firstly, the concepts of learning theories, and the ethical, social structure of SLP projects. Secondly, novel disciplines such as Design Thinking and its correlation with SLP Projects, and, lastly, an analysis of the impact of STEM education system (science, technology, engineering and mathematics) in our real case scenario.

2. LEARNING THEORIES, ETHICAL AND SOCIAL STRUCTURE FOR SLP PROJECTS

Nowadays, the development of medicine and the advances achieved in rehabilitation processes, generate the need to pay attention to the classic bioethical principles: non-maleficence, beneficence, justice and autonomy. These principles are rational criteria that enable conflicts resolution, although the difficulties of implementation lies in decision-making.

The principle of non-maleficence refers to avoid harming, recklessness and negligence. In the rehabilitation process, specialists must be cautious in decision-making and consider all the technical aspects and consequences of their prescription. Therefore, professionals must consider previous scientific research and adapt it to the needing of their users.

Beneficence is the principle in which healthcare professionals have been educated, therefore diagnostic and therapeutic procedures must be safe and effective. This principle also refers to doing what is good, not only to the user but to society as a whole. Keep in mind that good is a subjective concept, so healthcare professional have to have the necessary mechanisms to know, act and respect the user’s necessities.

The principle of justice is about being equitable and fair with the distribution of health resources; that is, users who have the same rehabilitation needs should receive the same services and resources in terms of quantity and quality; and users with higher needs, higher services and resources. This principle guarantees that all users deserve a decent and fair distribution of all health resources.

The principle of autonomy is defined as the user's ability to make their own decisions related to their illness. This implies that the user has to know the consequences of their actions; therefore, health professionals must communicate reciprocally with the user about all the information, recommendations and alternatives for the user about their rehabilitation processes.

The applicability and knowledge of bioethical principles are necessary for quality professional practice. All healthcare professionals must provide their users with efficient, equitable, fair and adequate care in order to reintegrate them into the changing society. This implies the application of all bioethical principles without any hierarchical order, i.e., the principles are all equally important and all they are mandatory. However, clinical practice generates situations where the negotiation among them is required. These situations are real and arise with some frequency in professional practice, so every health professional must have a thorough training in bioethics.

The learning of bioethical principles must be included together with the rest of the subjects of the different health disciplines. All students acquire essential knowledge and skills for their profession, but they must also know the social, cultural and ethical environment before the different situations that may arise in the exercise of their profession (Vera, 2017). Therefore, the learning of bioethics facilitate the student a thinking of the bioethical principles in different situations, and not be limited to the teaching of theories of bioethical attitude.

Bioethics, therefore, is a transversal academic discipline that provides students with skills that allow them to handle the conflicts of values that may appear in clinical practice (Garzón & Zárate, 2015). Current education emphasizes the training of students in competencies, understanding competence as a whole of knowledge (knowing and understanding), skills (knowing how to act) and human attitudes (knowing how to do) that allow excellent clinical practice appropriate to the bioethical context. Bioethics learning encompasses all kinds of skills, both transversal and specific to each discipline and profession, since bioethical principles must be present in all situations.

Learning bioethical principles requires that academic subjects allow the student to become an informant, counsellor and collaborator for users when prescribing a rehabilitation process. Therefore, in many colleges, students receive two types of training; the first relates to the ethical codes of their professions and, the second, focuses on the formation of ethical reasoning. The ultimate goal of bioethical learning is to train professionals to be able to act according to their structured ethical reasoning.

In order to achieve this final goal of bioethical learning, one of the most appropriate ways to teach bioethics is giving the student the chance to see real-life situations where ethical dilemmas can occur, always with the possibility of giving students an accompaniment by teachers. A pedagogical framework where the protagonists of learning are students is "learning and services" (SLP), in which students, guided by a teacher, detect a need in society, develop a project, carry it out and evaluate it.

SLP methodology arises from the continuous search for pedagogical methods that encourage motivational, practical and dynamic learning (Zayas, González Pérez & Gracia, 2018). The SLP is considered an educational proposal that combines learning processes and community services, in which participants learn by working on real needs of the environment in order to improve it (Uruñuela, 2018). That is, it establishes an active relationship between theory and practice, giving the student the opportunity to learn while contributing to society. It is a pedagogical method that integrates the benefits of experimental learning and community service.

SLP is considered a dynamic educational method, proposed to meet educational objectives, including changing the role of instructor to facilitator by teachers. Students have the most active role in their learning by providing a context where they can learn ethics and social responsibility and teach interdependence and partnership within society; due to this, SLP is one of the most appropriate methods to teach students bioethics. Bioethics, like ethics, depends on the cultural, historical and social environment; SLP Projects offers the opportunity to understand those environments that guide how society thinks and behaves (Ventres, 2017). SLP Projects have a structure for preparation, reflection and evaluation of bioethical principles that provide a great opportunity for students to integrate experiences that favour their personal and professional development.

For all the aforementioned, it is appropriate to use the SLP methodology so that students understand the importance of ethical principles (non-maleficence, beneficence, justice and autonomy) that guide their professional practice when establishing rehabilitation processes, since real experimentation of cases creates different situations that improves this learning and allows society to obtain some benefit since the SLP Projects facilitates it.

3. DESIGN THINKING AND ITS CORRELATION WITH SLP PROJECTS

Design Thinking (DTh), refers to a methodology that aims to create innovative solutions to real problems; quoting Tim Brown (2010), DTh is a "person-centred innovation". DTh methodology focus on innovation within interdisciplinary work, including different professional profiles in teams. According to the Hasso-Platter-Institut (HPI, 2020), DTh process is composed by six phases: Understand, Observation, Defining the point of view, Ideation, Prototyping and Test. The team goes through the different phases not necessarily in order of appearance.

The first phase (understand) consists of acquiring the basic knowledge about the problems that lack the potential users of the service to innovate. The second (observation), empathizes with the users and connect with their necessities; the third (definition the point of view) seeks to create the profile of the typical user; in the next phase (ideation), teams must generate all the possible ideas, preferentially without filters and go to the next phase to make real prototypes. Finally, it is about testing the prototypes, not only in laboratory conditions, but also with the users in situ.

According this concept, Steinbeck (2011), talks about the Design Thinking as an innovative pedagogical strategy, which aims to provide students analytics and creative competences. Using DTh methodology in the university context, the author identifies four key points: teams with students from different branches of knowledge, teachers from diverse disciplines, relation with the industrial sector and, workspaces where different teams can work at the same time (mobile furniture, variety of technologies, etc.). In addition, motivate the students and help to achieve the competences with the schedule of the different milestones along the process, such as training and consolidation of the teams, deliver prototypes or the final presentation of the developed device in a Design Exhibition.

There are previous experiences about the application of DTh in the training of health sciences professional (Lori & Reed, 2019; Falcao, Savoy & Markey, 2020). These experiences underline the need to train these professionals within multidisciplinary teams, as well as develop their creativity.

3.1. Analysis of the Design Thinking Model and SLP Experience

The comparison between the Design Thinking Model and the SLP proposal "With you I am Capable" has been carried out at the University of Burgos from two different perspectives: (1) from the DTh as a pedagogical strategy and (2) from the DTh as an own methodology for create new products.

Related to the first perspective (DTh as pedagogical strategy), is important to highlight that the four key points are achieved, although the one related to facilitate common workspaces need to be improved. Due to the milestone of the project, meet the needs of specific disabilities, diversity in the composition of the teams of students and teachers is relevant, it is, Health Science and Engineering.

Moreover, the projects are developed in some organizations; the university contact with them, sign the collaboration agreement, stablish the contact person between the organization and the institution), look for the practical case, and act as a coach providing required knowledge about the user and the necessities to resolve. Regarding the workplace, there is a specific area, called UBUmaker, what is a digital room. Students can fabric their products when prototypes are ready.

Summarizing these framework reviews, Table 1 relate key DTh points according to Steinbeck (2011) and de SLP Projects.

Table 1. Comparison between DTh as pedagogical strategy and SLP project.

DTh Key points as pedagogical strategy	**SLP Project**	**Comments**
Students from different branches of Knowledge	Yes	Two degrees: Occupational Therapy and Management Engineering
Diversity disciplines (Teachers)	Yes	Two degrees: Occupational Therapy and Management Engineering
Contacts with the industrial sector	Yes	Through participant organizations
Different Teams work in a specific Workspace	Yes, with improvements	Try to allocate fixed workspaces to favour interaction since the beginning
Training and consolidation of the teams	Yes, with improvements	To stablish a specific plan for set up and consolidate the teams
Schedule milestone delivery	Yes, with improvements	To provide timeline from the beginning of the project
Design Exhibition	Yes	The products were exposed along two weeks in an exhibition

Source: own elaboration

Continuing with the DTh analysis as pedagogical strategy is necessary to mention not only the key points, but also check with the milestones defined by Steinbeck (2011). Accordingly, the training of the teams and its subsequent consolidation was done naturally as the project progresses; however, DTh as pedagogical strategy implies a faster consolidation, what will need to be improved in the next experiences. Likewise, it is identified as a future improvement, the necessity of a better planification of the milestone delivery in the early stage. Finally, the devices were presented through a conference supported by the local council. The participant teams showed their final devices, which after the presentation were also exposed for two weeks, allowing citizens to learn about the innovations made by students and at the same time, highlight the connection of the university with society.

Before starting with the comparative analysis of the DTh as own methodology to create new products, we underline that all students received a specific training about DTh methodology. Table 2 summarize the actions carried out within the SLP Project in comparison with the mentioned methodology.

Checking how the six phases defined by HPI are applied, phases 1 and 2, *Understand* and *Observation*, were done practically at the same time, therefore, students had at least two meetings in order to identify and evaluate their user's necessities and abilities. This evaluation was done through observation, but also through specific Occupational Therapy (OT) assessments; thus, it was made a fusion between DTh methodology and the own process of OT. The aim of this fusion was to develop a factible device. The third phase (*Defining the point of view*) in not applicable to this project, because DTh is typically applied for groups or collectives and SLP project conducted is focused on specific human being with functional diversity.

The *Ideation,* 4th phase, was carried out within each group, with the team members themselves who leaked the best ideas that passed to the *Prototyping* phase; This penultimate phase, had more support from teachers and external collaborators, advising students on better design options, selected materials and feasibility.

Table 2 Comparison between DTh as design methodology and SLP Project

Phases	Project SLP	Comments
Understand	Visit centres and users	Aim: To identify necessities
Observation	Evaluation according OT process	Aim: to evaluate abilities
Defining the point of view	-	It is design for specific individual, not for a collective
Ideation	Intragroup	-
Prototyping	Intragroup with support from teachers and collaborators	-
Test	Intragroup and with the particular user	Tiny changes (parameterization and materials)

Source: Own elaboration

Finally, the *Test* phase needs to join the *Prototyping* phase, following the DTh concept model (HPI, 2020). As long as the prototypes are created, are being tested, first by the students and later, by the final user; all products need to be improved with tiny changes (parameterization and or materials).

4. METHODOLOGY

According the methodology, the scheme of work in the SLP Projects has consisted of four stages: the first one for organizing the students in teams and groups of work; the second one, for acquiring conceptual and practical knowledge by doing learning workshops; the third one, for developing the projects (students use a novel Technological Center of the University of Burgos equipped with new and recent IT technologies such as 3D Printers, electronic devices, machines and material for design and manufacture the support products, and, the fourth one, supervisors have developed a survey for collecting data of satisfaction by using a rubric based on Campo (2015). The expectation about the effects of the use of new learning methodologies (Service Learning Projects) for Higher Education and its repercussion on cognitive competences, social competences, ethical competences and professional competences are the basis of the Hypothesis.

4.1. Empirical Study

In order to complete the study of the SLP Projects at the University of Burgos, participants have answered a survey based on a Rubric developed by Campo (2015). The questionnaire consists in 13 questions (Campo, 2015) and 2 more questions (own elaboration), which matched the structural areas of SLP Projects, such as the approach of learning, competences, level of participation, evaluation, transdisciplinarity, impact and social projection, professional field, resources and multiculturalism. The complete questionnaire can be found at page 15 of the publication[1], but a sample of the rubric used is included in Table 3. All are qualitative questions can be transformed to numerical values quite straightforwardly; by assigning a score number ranging from 1 (leftmost option) to 4 (rightmost option), depending on the answer selected. The students completed this survey in an anonymous way in order to prevent unintended data recollection and to encourage the student to answer in the most honest possible way. For the quantitative analysis, the mathematical tool used is SPSS v24.

Table 3. Rubric (summarize) for evaluating SLP Projects (survey).

Parameter	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Learning Approach	SLP Projects are based on memorize concepts and fulfilment institutional requirements with no possible attitude changes.	SLP Projects develop learning that aims to make students change their way of seeing the world, be creators of their reality and encourage metacognition.	SLP Projects develop learning that aims to make students change their way of seeing the world, be creators of their reality. In addition, there is a specific space for reflection on PhD studies.	SLP Projects propose learning strategies that are based on students' interest in maximize understanding and satisfy their curiosity. There are specific spaces for this.
…	…	…	…	…

Source: own translation from Campo (2015, p. 15)

5. ANALISYS OF RESULTS

Due to the novelty of the implementation of the LSP Projects at the UBU, in the current state of our study findings come from a relatively low quantity of students (23 students). Nevertheless, this study is an ongoing research that will try to reach a much greater number of students in the next academic years in order to obtain more statistically significant conclusions.

In the sample, the 52.2% of the students are from the engineering field while the 47.8% come from the health discipline. Figure 1 shows the results of the survey. Valuing the percentages, the most relevant findings are the next:

1. Learning Approach: students of both disciplines agree with the idea that by developing these projects they have changed the way of seeing the world, being creators of their reality and encouraging metacognition.
2. Competences: engineering students have improved their transversal skills, such as autonomy, creativity, critical thinking, personal initiative and sensitivity. Health students have improved their specific proffesional skills.
3. Level of involvement: 75% of the engineering participants consider participation has been projective (determine the project, objectives, design, planning, implementation and evaluation); 17% of them consider participation has been purely consultative. As far as health students, 36%, value the participation as proyective, 36% metaparticipation and a mere 27% purely consultative. Summarizing results, three-quarters of the

[1] https://doi.org/10.1344/ridas2015.1.6

students value the item participation as proyective and metaparticipative and, the others, purely consultative.

4. Evaluation of the project: all students perceive the project is checked by all participants (institutions, associations and tutors). Nevertheless, the 27% of health students do not notice the service offered to the community.

5. Academic monitoring: 50% of engineering students consider there is an academic follow-up, 50% perceive there is follow-up between the entities and the training institution. Regarding the health students, all sample perceive the follow-up is evaluated between entities and institution.

6. Transdisciplinarity: more than the three-quarters of the students (both disciplines) notice all them work on the same challenges with the need to complement each other.

7. Social impact and projection: 60% of engineering sample realize they work on real and proximate needs and influence society. The 33% point out with these projects is feasible to provide tools to the community when the work is completed (empowering it). Thus, health students (91%) value these experiences can be addressed beyond their execution.

8. Work in nets: all students feel they work in collaboration agreenment projects to build a common work, and, the 35% out of them consider it is possible to exchange reflections and improvements in regular meetings.

9. Proffesional field: 17% of engineering students notice these projects contribute to open a vision of the proffesional field with greater emphasis on knowledge generation (any health student observe it). 42% of the engineering students and a 27% of health students value the projects developed similar to their disciplines. By contrast, 73% of health sample and a 42% of engineering sample see the possibility to open up new proffesional challenges within the community and social implication; still, add the possitive view of working with different disciplines.

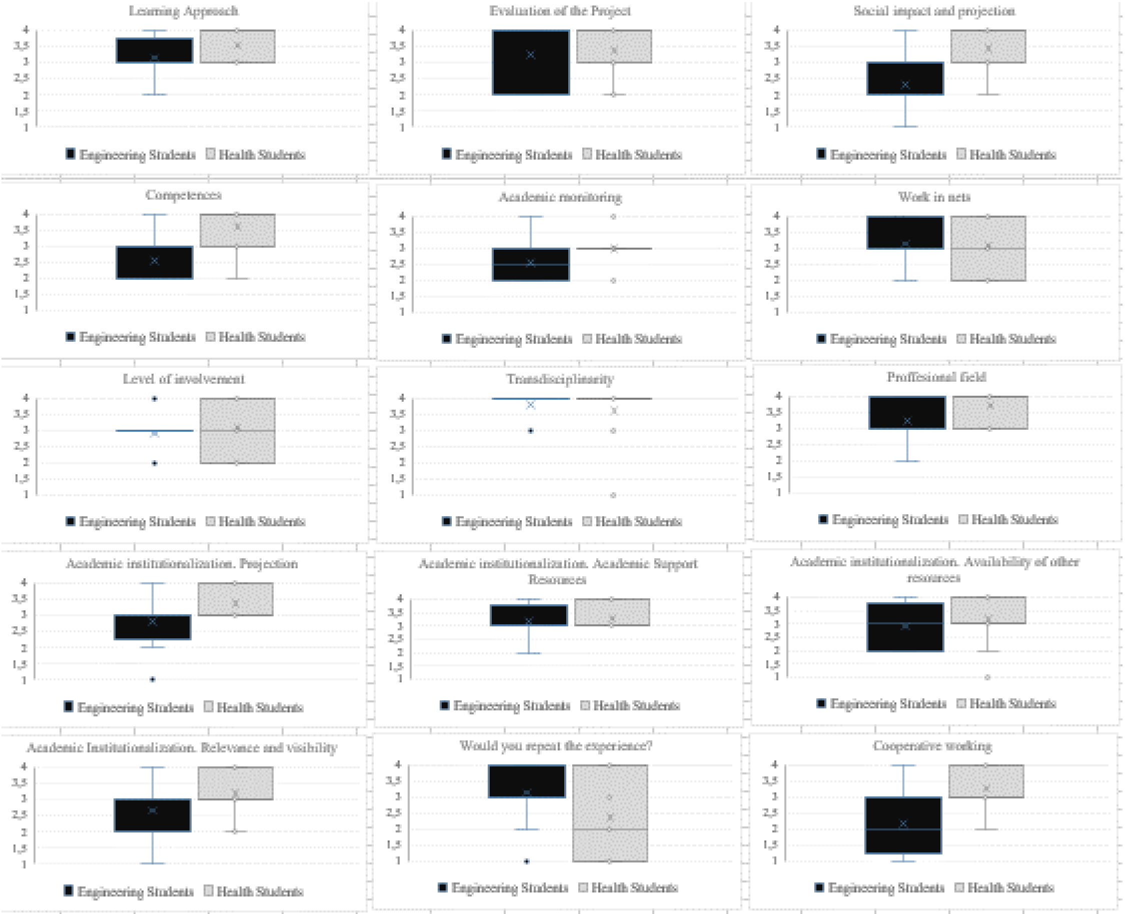
10. Academic institutionalization. Projection: the whole sample (both disciplines) notice these experience help the promotion of service learn work and point out the importance of doing sistematically.

11. Acadmic institucionalization. Academic support resources: 75% of the students value the projects are located in any structure of the university (subjects). They point out a major support and recognition.

12. Academic institutionalization. Availability of other resources: 33% of engineering students and a 9% of health students recognize the flexible groups of work and open schedules to go ahead with the activities. Approximately the same percentage of students in both disciplines (40%) consider that auhtorizations, agreements and conventions are facilitaded. Only 25% of engineering students, compared to 45% of health students have seen contacts facilitated for project networking.

13. Academic institutionalization. Relevance and visibility: 83% of engineering students and 63% of health students see some institutional recognition and celebration, but it is not institutionalized nor systematized. The most opposite perception is whether these projects favour social recognition through awards and grants, an opinion expressed by 8% of engineering students compared to 36% of health students.

Figure 1. Survey results SLP Projects.



Source: self-elaboration based on Campo (2015)

6. CONCLUSION

The benefit of this research is the contribution for a universal benefit: the impact of educate university students in SLP Projects. This line of work has raised the possibility to opening up values and social good in higher education.

Results in our study inferred there is evidence that student from both disciplines value the experience because the learning approach, level of involvement, transdiciplinarity and collaborative work; as consequence, positive gradient in their academic marks. Specifically speaking, engineering students have improved their transversal skills, such as autonomy, creativity, critical thinking, personal initiative and sensitivity. Health students have improved their concrete proffesional skills.

The future line of work is based on enhance the items with lower percentages, such as, follow-up intervention from all entities involved, impact on society and the possibility to open up new proffesional challenges within the community and social implication.

Overall, in accordance with university responsibility on educate in excellence and values, the proposal of SLP Projects in higher education enable students, professors, stakeholders and society for running a cohesive consortium.

ACKNOWLEDGEMENTS

We gratefully thank all entities involved in the LSP Projects (APACE BURGOS, COCEMFE CyL, PUERTA DEL PARRAL, OTRI, UBUmaker, Unidad de Cultura Científica and Unidad de Cooperación y Acción Solidaria) and survey participants of the University of Burgos for their collaboration and support.

REFERENCES

Brotóns-Cano, R., Lara-Palma, A. M., Stuart, K., Karpe, J., Faeskorn-Woyke, H. & Poler, R. (2009): Competitive Universities need to Internationalize Learning: Perspectives from three European Universities. *Journal of Industrial Engineering and Management*, 2(1), 299-318. Retrieved from http://www.jiem.org/index.php/jiem/article/view/89

Brown, T. (2010). IDEO *Why Design thinking*? *Approach.* Design thinking is a process for creative problem solving. Retrieved from https://www.ideou.com/pages/design-thinking

Campo-Cano, L. (2015): Una rúbrica para evaluar y mejorar los proyectos de aprendizaje servicio en la Universidad. *Revista Iberoamericana de Aprendizaje Servicio* (RIDAS), 1, 91-111.

Casado-Muñoz, R., Greca I., Tricio-Gómez, V., Collado-Fernández, M. & Lara-Palma, A. M. (2014): Impacto de un Plan de Acción Tutorial Universitario. Resultados Académicos, Implicación y Satisfacción. Revista de la Red Estatal de Docencia Universitaria (REDU) 12(4), 323-340. Retrieved from http://red-u.net/redu/index.php/REDU/article/view/587.

Falcao, M, Savoy, J., Markey M. (2020): Teaching cross-cultural design thinking for healthcare. *The Breast*, 50, 1-10.

Folgueiras Bertomeu, P., Luna González, E., Puig Latorre, G. (2011): Service Learning: study of the degree of satisfaction of university students. *Revista de Educación*, 362, 159-185

Garzón Díaz, F., Zárate, B. (2015): El Aprendizaje de la Bioética Basado en Problemas (ABBP): un nuevo enfoque pedagógico*. Acta bioethica, 21(1), 19-28.*

*Hasso-Platt*er-Institut (HPI) (2020) *What is Design Thinking?* Retrieved el 18 de February 2020, de HPI Academy, Education for professionals. Website:https://hpi-academy.de/en/design-thinking/what-is-design-thinking.html

Lara-Palma, A. M. & Collado-Fernández, M., Tricio-Gómez, V. (2010): Improving Education: A Knowledge Transfer Map Proposal for University Tutorship. 5th International Conference of Education, Research and Innovation. 6043-6050. Retrieved from https://library.iated.org/view/LARAPALMA2012IMP

Lara-Palma, A. M. & Giacinto, R. (2014): Improving the Effectiveness of Virtual Teams: Tackling Knowledge Management and Knowledge Sharing. A Real Case Scenario. *Journal of Social Sciences* (COES&RJ-JSS), 4(1), 626-634. Centre of Excellence for Scientific & Research Journalism, COES&RJ LLC.

Lara-Palma, A. M., Cámara-Nebreda, J. M., & Vicente-Domingo E. M. (2015): Impact level within the English Friendly Program in Virtual Degrees at the University of Burgos. A Real Case Scenario of Transnational Education. *Advances in Intelligent Systems and Computing*, 369, 525-532. Springer International Publishing Switzerland. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84946779842&partnerID=MN8TOARS

Lorenzo, C. & Lorenzo, E. (2020): Opening Up Higher Education: An E-Learning Program on Service-Learning for University Students. *Advances in Intelligent Systems and Computing*, 963, 27-38. AHFE International Conference on Human Factors in Training, Education, and Learning Sciences

Lori, A & Reed, A. (2019) The Power of Design Thinking in Medical Education. *Academic radiology*, 26(10) 1417-1420.

Maquilón Sánchez, J. J., Alonso Roque, J. I. (et. al.) (2014): Experiencias de innovación y formación en educación.

Steinbeck, R. (2011): El "Design Thinking" como estrategia de creatividad a distancia. *Comunica*r, 19(37), 27-35. https://doi.org/10.3916/C37-2011-02-02

Torres-Coronas, T., Arias-Oliva, M., Yáñez-Luna, J. C., Lara-Palma, A. M. (2015): Virtual Teams in Higher Education: A Review of Factors Affecting Creative Performance. *Advances in Intelligent Systems and Computing*, 369, 629-637. Springer International Publishing Switzerland. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84946771368&partnerID=MN8TOARS

Uruñueña, P. M. (2018): La metodología del aprendizaje servicio. Narcea.

Ventres, WB. (2017): Intentional Exploration on International Service Learning Trips: Three Questions for Global Health. *Annals of Global Health*, 83(3-4), 584-589.

Vera Carrasco, O (2017): La enseñanza de la ética y bioética en las facultades de medicina. *Revista médica la Paz*. 23(1), 52-59.

Zayas Latorre, B., González Pérez, V., Gracia Calandín, J. (2018): La Dimensión Ética y Ciudadana del Aprendizaje Servicio: Una apuesta por su institucionalización en la Educación Superior. *Revista Complutense de Educación*, 30(1), 1-15.

OVERCOMING BARRIERS TO INCLUDING ETHICS AND SOCIAL RESPONSIBILITY IN COMPUTING COURSES

Colleen Greer, Marty J. Wolf

Bemidji State University (USA)

Colleen.Greer@bemidjistate.edu; Marty.Wolf@bemidjistate.edu

ABSTRACT

In this paper we describe qualitative work that identified barriers to incorporating ethics and social responsibility in computing curricula in public teaching universities in the US. Three themes emerged: competent creation, expertise, and deference to authority. There is evidence that faculty see the incorporation of ethics and social issues into the curriculum in a systematic manner as important, yet they consistently point to the need for efficiencies, concerns regarding their own expertise, and prioritization of need.

KEYWORDS: computing ethics, teaching computing ethics, integrating computing ethics, social responsibility in computing-

1. THE ISSUE

A recent focus of concern in ethical and social responsibility in the digital environment is the way in which privacy has been breached and human behavior has been monitored, recorded, and sold. Increased surveillance and the ability to engage in information data mining without consent has been the topic of multiple studies (e.g., Altaweel, 2015; Libbert, 2015; Mengwei Xu et al., 2017; Zuboff, 2019). These developments have contributed to an awareness among many of the need to better educate computing students regarding ethics and social responsibility. In the United States faculty at some research institutions and private liberal arts colleges have started to explore what it means to increase understanding and awareness, however there is little evidence that these activities have been replicated at public colleges and universities where faculty have high teaching loads (Burton et al. 2018; Groz et al. 2019; Saltz, et al. 2019; Shaer & Peck, 2018). Until recently there has been limited focus on cross-disciplinary approaches and incorporating philosophical interpretations of ethical frameworks into computing (Burton et al., 2017). Yet, while some have used philosophical understandings of ethics in computing, there has been less effort directed at incorporating social science approaches that move beyond an emphasis on the individual and utilitarian interpretations toward the intersubjective and articulations of “the good.”

Recently Wolf (2016) argued for the need for enhanced interdisciplinary attention to ethical matters and social responsibility in computing. This attention is well-warranted in the current political, social, and economic landscape where humanity is working and living, and in the academy where we seek to educate for character, career, and future leadership. The ACM Code of Ethics and Professional Conduct (2018) raises deeper awareness of what it means to be a computing professional. It suggests the computing profession and the computing professional must be astute, insightful participants in the social fabric who demonstrate a public understanding of the broader social environment. Increasingly, individuals find themselves in an environment where they are at a crossroads regarding what it means to be responsible professionals and socially engaged actors who have the capacity to ask challenging questions of their social worlds, the broader environment, and themselves.

Our project builds on these contemporary issues and challenges by exploring perceptions of computer science department chairs and faculty at target institutions surrounding notions of "ethical and socially responsible computing" and "collaboration." To lay the groundwork, we highlight the social issues and point to calls for enhanced ethical and social understanding by scholars from multiple disciplines. Next, we survey the literature related to incorporating ethical and social responsibility into the computer science curriculum. Then we outline the methods we used to gather information from current computer science faculty. Finally, we share the results from our study, documenting the perceptions of computer science department chairs and faculty at target institutions regarding notions of "ethical and socially responsible computing" and the overall design of a process whereby curriculum modules on these topics are created for delivery in typical computer science courses.

2. ETHICS & SOCIAL RESPONSIBILITY AND CS INSTRUCTION

Over the last fifteen years we have seen an unprecedented expansion in enhanced information extraction and analysis techniques to create big data and significant advances in artificial intelligence. Gary T. Marx (1998) in an early analysis of this expansion encouraged careful consideration of how to inform and how to protect those who will be indirectly and directly impacted. Yet as news organizations and multiple scholars have noted, since 2004 Google, and subsequently Facebook, have found ways to track, extract, and sell the human experience of millions (Zuboff, 2019). Bauman & Lyon (2013:135-136) specifically note that, "… the road to submission to an offer leads through the elimination of choice … The willing, nay enthusiastic cooperation of the manipulated is the paramount resource deployed by the synopticons of consumer markets." We are living in an era where differentiation, categorization, and manipulation of traits is framed as a moral good, yet the collection that is sold is not a human with moral obligations and agency (Bauman & Lyon, 2013; Bauman, 2013). Publications on surveillance often focus on governmental investigations and background patterns of surveillance, yet the current flow of information is from multiple locations and establishes privileges and detriments for peoples (Ullrich, 2018; Young, 2017). Societal values are used and disarmed through the establishment of filters that leverage control (Helles & Flyverbom, 2019). Responses to multiple breeches of trust have frequently been after-the-fact correctives, leading to significant questions about how consumers, industry experts, and educators should respond and engage in proactive attempts to incorporate an ethical focus.

Ethics discussions within computer science have taken multiple forms over the years, and the merits of a focus on professional competence is frequently positioned as counter to responsibility to society and the benefit of humanity. Configured as competent creation, Stieb (2008:226) expounds on the "unnecessary and unfortunate 'add on'" of benefit to humanity in professional discussions and professional codes. He articulates the rights of professionals to focus on individual needs, with a narrow emphasis on ensuring product quality and the immediate safety of the user. The central issue of ethics and responsibility is perceived as either too expansive to consider, too complicated to master, or put down to the fact that it is impossible to articulate the “good.” It appears that Stieb (2008, 2009, 2011) is arguing for a form of negative freedom that does not directly acknowledge the social arena within which the acts of engineering and computing are occurring, thus missing, in particular, the cultural and structural realities of neoliberalism and its consequences (Pendenza & Lamattina, 2019). While other approaches that expand the discourse encourage self-awareness through a virtue ethics lens or encourage a way of looking at the world that enhances the sense of self-identity and personal responsibility, they still emphasize the individual first through a social identity connection that does not directly interpret or structure an understanding of the “public good.” Miller (2008) in a debate with the ideas of competent creation encourages computing and engineering professionals to not abandon the concept of the “public good,” yet there is not a direct expansion of what this might entail. Gotterbarn (1995, 2001), has, however, consistently encouraged a recognition of the impact of computing on humanity. As he does so, he points to the responsibilities that computing professionals need to consider as they do their work, the social environment within which they are acting, and the implications of what it is they are doing. This is certainly distinct from competent creation, yet, the social remains ill-defined and questions about how to define the public good, what it means to engage in social responsibility, and how particular goals and outcomes should be established are left open.

Certainly, these conversations occur among professionals as they relate to the products they create, as well as among educators. As Stahl et al. (2016) note, professionals in computing have been exposed through their education to the ethical standards of associated professional bodies (e.g., ACM, BCS). Yet, understanding how to define and prioritize ethical needs and the relevant issues that are at stake has not been part of that education. This leaves professionals in the field—including practitioners and educators--to establish, for themselves and their students, what ethical and social considerations might mean for those new to the profession. Certainly, the complex process of “becoming” includes a premise under neoliberalism that it is through your labor that you will develop a self-awareness. In essence it is labor itself that will stand as the social (Farrugia, 2019:1098-99). To not accomplish is to potentially lose your sense of being. For current computing professionals, including computer science faculty, who are functioning within a social-economic-political environment where efficiency is primary, and for whom instruction has typically been narrow and specialized, reaching beyond the confines of existing understandings of the individual and competence requires additional resources for self and profession. It is not clear that there has been sufficient education on the differences and overlap among various ethical frameworks and interpretations of social responsibility to inform future instruction.

There have, however, been attempts to create an ethics education framework. Over the years a variety of approaches have incorporated ethics and social responsibility into the computer science curriculum. There are two primary methods. The first involves a standalone course in computing ethics and social responsibility. The second incorporates ethics and social responsibility modules into most courses in the computer science curriculum.

There are numerous textbooks that are designed to support the standalone approach. For example, Brinkman and Sanders (2013), Tavani (2015), and Quinn (2017). Others such a Burton, Goldsmith, and Mattei (2018) describe a course that uses science fiction to drive the pedagogy of a computing ethics course. Like many who teach computing ethics they are interested in seeing students understand computing as something more than competent creation of computing artifacts: “Teachers and leaders in the field have a responsibility to drive the discussion about the effects of their own work and the work of their students" (2018:57). Urman and Blumenthal (2018) focus on engaging students in a different way. Their computing ethics course is one that promotes the common good. Moore (2020) takes a different approach and calls for the incorporation of politics into computing ethics education. Moore does acknowledge a practical challenge of this approach: "instructors in computer science departments might not be interested, available, or capable of teaching such classes" (2020:421). Henderson (2019) describes a course focused on Data Ethics and Privacy that is a "fifth-level module" that is available for master's students as well as undergraduates who are in their final year.

There is widespread concern that the standalone computing ethics course model presents several structural challenges. When the course is not taught by a computer science faculty member, it sends a message to students that it is somehow less important than their "real" computer science courses. This negative message is exacerbated when the course does not have a "CompSci" prefix. Even when the course has that prefix and is taught by a computer science faculty member, the fact that it is a senior-level course sends the message that the consideration of ethical and social impacts of computing is something that is done late in the "computing process" and after the "difficult technical work" is nearing completion. To address these concerns, many are exploring ways to incorporate ethics modules in a variety of computer science courses. Grosz et al. (2019) describe a project in which computing ethics modules are developed for delivery in a variety of computer science courses at Harvard. Saltz et al. (2019) describe how they integrate ethics into a variety of machine-learning courses. Skirpan et al. (2018) developed modules where students are expected to incorporate the social and ethical lessons into the projects they do in a variety of computer science courses.

There are many potential variations of the two extremes suggested above, including hybrid models that include both approaches. Regardless of model, there is the bootstrapping problem Moore (2020) acknowledged when it comes to incorporating politics into computer science. It takes special expertise to teach computing ethics, use science fiction as a pedagogy, develop students' understanding of the common good, and importantly, to consider the "ethical and social aspects" of a project from conception to completion. All computer science faculty, but especially those at teaching-focused undergraduate institutions, need preparation that enables them to effectively develop within their students these essential professional abilities. It may not be easy. Henderson notes that he took a year's leave to further his "understanding of law as a mechanism for regulating and enabling ethical behaviour" (2019). Further, many computer science faculty may feel ill-equipped to evaluate the sorts of work that tends to come along with the teaching and learning of ethics and social responsibility.

The incorporation of ethics, and if it occurs, social responsibility, is not consistent across levels of delivery within the computer science curriculum at the undergraduate level. Consequently, students often find themselves exploring the ethical questions from a utilitarian perspective, from an individualistic understanding of consequences for individuals who make up the collective, or for themselves as professionals.

3. METHODOLOGY

In order to develop a deeper understanding of how ethics and social responsibility are understood and when they are used by computer science faculty, we developed a multi-tiered approach that is based in critical methodology and is designed to both interpret and engage. First, we identified state, public universities that had computer science, humanities, and social science programs/departments. Each identified department had between 3 and 65 faculty. Second, we designed interview and survey instruments to collect data from the selected departments (Babbie, 2016). We received Institutional Review Board approval for all documents and research processes.

Our research/intervention process occurred in three phases. In phase one we designed and carried out a qualitative interview study with department chairs of the identified computer science, humanities, and social science departments. We sent emails to the selected department chairs announcing our project and requesting their participation in an interview. Interested participants signed and returned an informed consent document prior to the interview. Individual interviews occurred via an audio-only Zoom meeting, and we sent a debriefing document after the interview. We designed the questions to identify existing understandings of academics and academic practices related to the infusion of ethics and social responsibility into the computer science curriculum, and computing concerns into humanities and social science curriculum. We also asked about faculty collaboration and how those collaborations are supported by colleagues, the department, and by the university. In addition, we sought information on whether and how work in their department is aligned with strategic or master academic planning at their institution. We used four questions, with additional probes, during both the computer science and the humanities and social science interviews.

In phase two of the project we distributed a short survey instrument to all faculty in departments where we had engaged in interviews with department chairs. Fifteen survey questions distributed to computer science faculty explored the extent to which ethics and social responsibility were perceived as important to instruction, research, and service, the types of pedagogical approaches used, which courses incorporated ethics, what assessments were used to understand skill development, the level of confidence faculty had in their knowledge of social and ethical issues, and their pattern of collaboration. We also asked some basic demographic questions regarding years of experience and rank. We created and distributed seventeen survey questions to humanities and social science faculty. Similar questions regarding pedagogical approaches, assessment, collaboration and demographics were asked of them We also included questions to identify the extent to which they address digital communication, information technology, or social media in their research and instruction, and their comfort level associated with this delivery. Finally, for phase three, we identified, via online data collection and through direct contacts, faculty interested in participating in workshops on ethics and social responsibility.

For purposes of this paper we analyzed results from the interviews with computer science department chairs and the survey results from the computer science faculty in the associated departments. Our findings are based on an analysis of the five interviews conducted with CS department chairs and the survey responses received from CS faculty. As of this writing, of the 150 surveys distributed, 16 have been completed. Our emphasis is therefore on a qualitative interpretation. Berg (2009) argues that it is through qualitative processes that we can better understand the interpretive processes of humans, their levels of awareness, and their thought patterns. Our first step was to engage in open and focused coding of the five transcripts of the interviews (Berg 2009, Katz 1983). Competent creation, expertise, and deference to authority emerged from a review of the codes we identified. We also identified several important elements related to pedagogical approaches and interest. Descriptive information from the surveys conducted with faculty in the associated departments supports various aspects of these themes, and those results will be used to point to the significance of the details shared by department chairs.

4. THE FINDINGS

The interest in and challenge of identifying and incorporating ethics and social responsibility in the curriculum, instruction, and research of our selected computer science faculty was apparent in a variety of ways. While there is professed interest, there are frequent questions about placement of ethics. There are also descriptions without a direct understanding of whether the placement of ethics is effective or, at points, necessary. Each of these aspects will be explored under the themes outlined below.

4.1. Competent Creation

Department chairs addressed ethics in a variety of ways during the interviews, but the various points at which they invoked aspects of competent creation in their responses are most notable. Creation of a reliable product and the efficient means of instructing are clear priorities. Ensuring that students are independent, successful, and meet the outcomes associated with particular courses and the overall program of study are central concerns, and this focus is evident in the way in which they describe faculty responsibilities, resource acquisition, resource use, and development (e.g., grants).

Descriptions of instruction typically brought to the forefront their understanding of the necessity of efficient instruction patterns, and how meeting particular outcomes, often set by industry, were essential. As one chair indicates, [we]

> "give our students a lot of really relevant up to date experience on how this material is being used right this minute in industry." M3B
>
> "[T]he curriculum really focuses on the tech, the skillsets of actually doing the programming, setting up servers, configuring servers, just to get the job done and get the product delivered." M3C

Additionally, expectations that the program of study will support economic direction and student needs is clear from the following quote:

> "We place really well our students to local companies in [our state] so that's one way for us to contribute to the economics of the states." M1R

Even as they touted the ability of faculty, they focused on the practical over and against the ethical and social issues that might be addressed in the classroom. For example,

> "our faculty are very good at covering the necessary theoretical aspects, but they've also got a lot of practical experience" M3A
>
> "this course needs to prepare students for the next course and the instructors in the next course are complaining that students are showing up and they're not prepared and

> so there's a lot of pressure to use the limited time available to make sure students know how to program and get them ready for the next class" M3I

A discussion of ethics under this framework highlights applied understandings and does not ask questions about the social implications of computing. As the chairperson stated, in courses:

> "you talk about everything from security ethics to programming ethics to intellectual property ethics to the ethics of application." M2B

An emphasis on ensuring appropriate end-products was highlighted, with a focus on dealing with understanding the rules surrounding research development and intellectual property:

> "we do have to get a little into technology transfer and intellectual property, because in some cases we are… and increasingly so, we are developing products for clients." M2H

The pressures of instruction, research, and product development often lead faculty toward a prioritization of time and delivery that follows patterns of neoliberal concern with product use and safety and efficiency of development for clients. Even when there is an emphasis on ethics, it is more directly related to the rules of use and how the transfer of a product is handled. When emphasis goes beyond this level, instruction of the ethical and social responsibility issues is placed differently within the curriculum or is outsourced to another department for support. Competent creation is then revisited through the voicing of interest in a curriculum that meets industry needs, through an emphasis on efficient use of resources, or through a deep concern about faculty competence to deliver on ethical and social concerns. It is this emphasis on expertise that carries with it multiple challenges related to overcoming competent creation.

4.2. Expertise

Responses from participants regarding how ethics and social responsibility is delivered brings up faculty concerns related to instructing in areas that are outside of their direct training. While a few faculty have more comprehensive training, many do not. As one department chair noted,

> "having some background in philosophical ethics is a real help. And not all of the faculty have a strong background in philosophy." M2E

Faculty also recognized that aligning expertise with lower division courses where interest in ethics is significant is important:

> "But really in that liberal arts course, in that one non-major course, putting our best, most experienced faculty in ethics is the way we like to go." M2G

In order to ensure successful learning, there have been some efforts to encourage faculty without expertise in ethics to engage in professional development, and certainly as it impacts instruction and collaboration, it is perceived as important for faculty. For example:

> “Others have picked up the course with say a summer to prepare it and have done some directed readings that are led by the experienced faculty.” M2F

Yet, department chairs acknowledge that it is challenging to deliver on ethics and social issues associated with computing, when most of the faculty lack that expertise. As the following quotes point out, the emphasis on specialization in training for a faculty position in computer science, along with understandings of the boundaries of the discipline, preclude the possibility of particular types of discussions and instruction.

> “You're not going to bump into a lot of people who would be qualified to teach both a course in computer science and say, a course in psychology.” M3T
>
> “We, meaning those faculty teaching those courses, did not feel prepared and competent to be teaching ethics, carrying the weight of the ethics teaching responsibility.” M4D
>
> “it's more the expertise or background preparation than it is the time per se. … They're hired to be a machine learning or a database or whatever expert.” M4H
>
> “we're looking at about six of our forty faculty, and then the faculty that now teach the one course, which is in essence the ethics course or the professionalism and practice, that's two of us.” M4G

On a faculty of forty only two focus on ethics delivery at any level. Faculty have adapted to a particular understanding of delivery and emphasize its functionality. However, as they leave ethics delivery to a small number of individuals in their department, or as they emphasize delivery of ethics at the lower division or within an application, they marginalize it and present ethics and social issues to students as either specific concerns (i.e., applications), end point concerns (i.e., when you find particular cases within a setting), or areas that are outside their purview. This implies that it is also outside the students’ purview. If they “run into” an issue it will be considered, but from what angle and with what tools? Modeling expertise as it is currently structured, implies that while ethics and social responsibility are important, they are not central to the life of the discipline. In essence, functionality means “does it work?”

In large measure, the expertise interpretation present in these discussions also carries a sense of deference to authority, a deference to the training previously received as well as a deference to other forms of authority that exist within and adjacent to instruction. Looking at those patterns of deference provides additional insight into how computer science faculty understand expectations.

4.3. Deference to Authority

Public universities exist within the authority structure of the state and adhere to laws and policies that are externally driven and internally interpreted. Rules and regulations apply to multiple aspects of instruction and course delivery, and review of practice occurs by the state, regional accrediting bodies, and accreditors specific to the programs (Bowen & Tobin 2015). *A priori* to any state or local structure and rules, disciplines themselves internally seek validation from their trajectory of knowledge creation and application. It is a form of authority that exists within the language, method, and substance of the discipline, as well as in and through those deemed “worthy” to speak with and for the discipline itself (Bowen & Tobin, 2015; Hallett, 2007; Kaufman-Osborn, 2017; Keith, 1994). Identifying levels of authority, and who can speak for what under what circumstances can be a topic of debate. There is a form of deference present in the understanding of who has expertise related to which knowledge that serves both in support of and a barrier to knowledge creation and delivery. Yet, for the department chairs that were interviewed other forms of deference to authority were often predominant. Sometimes interestingly so:

> “We've got a lot of accreditors breathing down our neck” M3H.

Another noted,

> “because we're ABET, we tend to follow a traditional, create the objectives, create where we're going to measure that and what ranges we're going to look for in determining whether it's effective.” M4F

Deferring to program accreditors both provides support for the delivery of the program and for the attainment of resources that programs perceive they need. Importantly accreditors also establish a mechanism for where and whether discussions of ethics and social responsibility are incorporated into the program of study. For many, program accreditation is a sign of success. Without it, resource discussions and understanding of program goals and outcomes may become more difficult.

Interest in program accreditation needs can also be identified in discussions about when and how a faculty person may decide to change what they deliver. Certainly, faculty are responsible for structuring their courses and syllabi, however, they are also responsible for delivering on departmentally understood curricular outcomes that support accreditation. When asked about whether faculty might insert more discussions of ethics and social responsibility into their courses, a department chair responded:

> “faculty would be willing to work in elements so long as they're comfortable that the existing core curriculum items that they're responsible for and they're going to get dinged for if they don't cover, if there's room to squeeze it in. If they've got the authorization from the department that this is the deployment level responsibility” M3J

> “If that's a department level decision, then I think they would follow it. That would give them the cover that they would need, and they wouldn't be nervous that they were going to get dinged for stealing time from something else that they thought was more important.” M3M

In these quotes we see both concern about what a faculty person will be downgraded for not delivering and a sense that a faculty person needs “cover” in order to feel comfortable adjusting instruction. This suggests that the department has a set understanding and that “deployment” of the curriculum is not in the total purview of the faculty person. Departmental authority to frame, set, and implement appears to be officially set, and deference to that set understanding of the substance of a course is essential to faculty success.

Interpretations of authority are also evident related to university structure, procedures, and processes. When asked about the feasibility of more cross-disciplinary instruction and research, a chair responded:

> “there is a gigantic administrative barrier in the form of the faculty union.” M3S

The implication is that there are significant challenges depending on faculty/administrative structure related to setting up team instruction.

The chairperson, identifying the limits to their own administrative authority in scheduling team-teaching, stated:

> “if I were to take say, an instructor with a PhD in psychology and appoint them to be the instructor of record for a senior level course in the computer science department. I'm not sure I could do that.” M3U

Even with accreditation and internal departmental control over curriculum, faculty, in this case department chairs, do not feel that they have the authority to make significant delivery adjustments that cross disciplinary boundaries or reach into the community at large. The same department chair noted:

> “but there's a lot of red tape. I know they ran into a bunch of red tape problems because you have to be careful that you're not setting things up where students appear to be competing with outside industry.” M3Q

The caution related to process and procedure and the impact it has on making significant change to curriculum or delivery is also evident when chairs were asked about department involvement in institutional strategic planning or master academic planning. One participant indicated:

> [The] “University just finished the strategic plan and then [the] college actually had a committee that they developed a stage and then the departments are kind of asked to revise or revisit their strategic plan. … So it's a top down process that's, so we are meeting the university missions and the goals as we develop our own and strategic plan.” M5P

The hierarchical nature of public universities is exemplified in this quote, and in the comments regarding course content, faculty delivery of that content, and accreditation. Perceptions related to change includes references to faculty being “nervous” about “stealing” from something considered more important, and a need for “cover” to allow for any significant incorporation of ideas that are outside of curriculum delivery toward product development. The implication is that the trajectory of the department, its curriculum, and its students are set, and that adjustments that do not follow the line of sequenced development are suspect. Perhaps the real struggle is with balancing understandings of the hierarchical university, the accrediting bodies, and the discipline itself. What is most significant here? Where is the time to process? Faculty find themselves challenged by a need to balance various aspects of course delivery, research, and service.

5. SHIFTING THE FOCUS—WORKSHOPS FOR CRITICAL CONSCIOUSNESS

A review of the department chair interviews and responses to the faculty survey underscores the relevance of the identified themes. Responses to the survey demonstrate an understanding of the necessity of instruction in responsible computing and the current placement of ethics and social responsibility in the curriculum. Of the sixteen faculty who responded to the question, "As a faculty person in higher education, to what extent do you see responsible computing (i.e., ethics and social responsibility in the design and implementation of information technology and social media) as important to your instruction?," twelve indicated that it was important and four that it was somewhat important. When these response patterns are compared to responses to the survey question, "Considering the standard CS courses identified below, to what extent do you address ethics or ethical and/or social issues in each course?," we note some discrepancies. Response options for this question were: to a great extent, to some extent, not at all, or does not apply. Of the twelve participants who responded to this question, nine indicated "does not apply" for Artificial Intelligence courses and Database courses. Eight chose "does not apply" to Computer Graphics Courses and seven chose "does not apply" to Web Programming. On the other hand, other courses were identified as including responsible computing at a high or fairly high level. The courses that were identified as addressing ethics and social issues the most were Computer Science 0, Networking, and Thesis.

These responses are consistent with department chair comments regarding responsible computing placement in the curriculum and raise questions about how additional instruction in responsible computing could be infused into the curriculum. Perhaps, given how faculty currently handle the curriculum, the feasibility of incorporating responsible computing is not viable due to elements of time, expertise, and expectations surrounding curriculum sequencing. Or perhaps delivery of topics such as AI and databases is done mechanistically and portends conversation about the inclusion of substance and consequences of ethics in instruction. Nevertheless, these data and the interviews with department chairs confirm the presence and absence of particular types of substantive conversation regarding responsible computing at public state universities and highlights where professional development may be possible.

To assist faculty who are interested in infusing ethics more systematically across the curriculum, we have designed a workshop that will help faculty reach across disciplines. Each workshop involves computer science and humanities or social science faculty who come together for five to six hours. Rather than having the workshop designed around a case analysis or a direct lecture framework, we engage faculty in a staged process to address interests, capacities, and modalities. The pair and share process encourages the development of empathy regarding individual interpretation of responsible computing, ethics, and social elements. The model also demonstrates how engaged dialogue can build toward deeper reflection and can act as the groundwork for pedagogical practice. We model designing a teaching module using a dialogical process that demonstrates stages of development. At the end of this step, participants have an awareness of not only what a module may contain, but how alternative viewpoints are used to approach a standard computer science topic. Using elements from critical curriculum studies and interpretations of virtue ethics, we encourage both personal and professional self-awareness and development through this workshop (Au 2018, Stovall, 2011).

6. CONCLUSION

The responses to the interviews and the supplemental data from the survey demonstrates how particular interpretations of expertise and delivery are embedded within academic cultural and structural dynamics. The themes of competent creation, expertise, and deference to authority capture the essence of the strengths and challenges faculty face in meeting the expectations of departments, universities, and industry. While many of these faculty see the incorporation of ethics and social issues into the curriculum in a systematic manner as important, and perhaps even wish for more, they consistently point to the need for efficiencies, concerns regarding their own expertise, and prioritization of need. Since our social, professional, and digital environments are now deeply entwined, it is time to help faculty enhance their capacity to embed ethics and critical awareness into the culture of their teaching and research.

ACKNOWLEDGEMENTS

This project is supported by the Responsible Computer Science Challenge, funded by Omidyar Network, Mozilla, Schmidt Futures, and Craig Newmark Philanthropies.

REFERENCES

Au, Wayne (2018). *A Marxist Education: Learning to Change the World.* Chicago: Haymarket Books.

Altaweel, I., Good, N. & Hoofnagle, C.J. (2015). Web privacy census. *Technology Sciences*, December 15, 2015. Retrieved from https://techscience.org/a/2015121502

Babbie, E. (2016). *The practice of social research* (14th ed.). Boston, MA: Cengage Learning.

Bauman, Z. (2013). *Does the richness of the few benefit us all?* Cambridge: Polity Press.

Bauman, Z., & Lyon, D. (2013). *Liquid surveillance: A conversation.* Cambridge: Polity Press.

Berg, B. (2009). *Qualitative research methods for the social sciences.* Boston: Allyn & Bacon.

Bowen, W.G. & Tobin, E.M. (2015). *Locus of authority: The evolution of faculty roles in the governance of higher education.* Princeton, NJ: Princeton University Press.

Brinkman, B. & Sanders, A.F. (2013). *Ethics in a computing culture*. Boston, MA: Cengage Learning.

Burton, E., Goldsmith, J., & Mattei, N. (2018). How to teach computer ethics through science fiction. *Communications of the ACM,* 61(8), https://doi.org/10.1145/3154485

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattel, N. & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI Magazine*, 38(2), 22-34.

Farrugia, D. (2019). Class and the post-Fordist work ethic: Subjects of passion and subjects of achievement in the work society. *The Sociological Review*, 67(5), 1086-1101.

Gotterbarn, D. (1995). The moral responsibility of software developers: Three levels of professional software engineering. *The Journal of Information Ethics*, 4(1), 54-64.

Gotterbarn, D. (2001). Informatics and professional responsibility. *Sci Eng Ethics,* 7, 221-230.

Grosz, B.J., Grant, D.G., Vredenburgh, V., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM,* 62(8), 54-61.

Hallett, T. (2007). Between deference and Distinction: Interaction ritual through symbolic power in an educational institution. *Social Psychological Quarterly*, 70(2), 148-171.

Helles, R. & Flyverbom, M. (2019). Meshes of surveillance, prediction, and infrastructure: On the cultural and commercial consequences of digital platforms. *Surveillance & Society*, 17(1/2), 34-39.

Henderson, T. (2019). Teaching data ethics. Proceedings of the 3rd Conference on Computing Education Practices. New York, NY: ACM.

Katz, J. (1983). *A theory of qualitative methodology: The social system of analytic fieldwork.* In Robert M. Emerson (Ed.), Contemporary field research: A collection of readings (pp. 127-148). Prospect Heights, IL: Waveland.

Kaufman-Osborn, T. (2017-03). Disenchanted professionals: The politics of faculty governance in the neoliberal academy. *Perspectives on Politics*, 15(1), 100-115.

Keith, B. (1994). The institutional structure of eminence: Alignment of prestige among intra-university academic departments. *Sociological Focus*, 27(4), 363.

Libbert, T. (2015). Exposing the invisible web: An analysis of third-party http requests on 1 million websites. *International Journal of Communication,* 9(18), 3544-3561.

Marx, G.T. (1998). An ethics for the new surveillance. *Information Society*, 14(3), 171-185.

Mengwei, X., Ma,Y., Liu, X., Liu, Y., & Lin, F.X. (2017). AppHolmes: Detecting and characterizing app collusion among third-party android markets. In *Proceedings of the 26th International Conference on World Wide Web*. https://doi.org/10.1145/3038912.3052645

Miller, K.W. (2008). Critiquing a critique: A comment on “A critique of positive responsibility in computing.” *Sci Eng Ethics*, 14, 245-249.

Moore, J. (2020). Towards a more representative politics in the ethics of computer science. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’20)*. New York, NY:ACM. https://doi.org/10.1145/3351095.3372854

Pendenza, M. & Lamattina, V. (2019). Rethinking self-responsibility: An alternative vision to the neoliberal concept of freedom. *American Behavioral Scientist*, 63(1), 100-115.

Quinn, M. (2017). *Ethics for the information age*, 7th ed. Boston, MA: Pearson.

Saltz,J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Deward, N., & Beard, N. (2019). Integrating ethics within machine-learning courses. *ACM Trans. Comput. Educ.,* (19)4, Article 32, 26 pages. https://doi.org/10.1145/3341164

Shaer, O., & Peck, E. (2018). Teaching pervasive computing in Liberal Arts colleges. *IEEE Pervasive Computing*, 17(3), 64-69. https://doi.org/10.1109/MPRV.2018.03367736

Stahl, B.C., Timmermans, J., Millelstadt, B.D. (2016). The ethics of computing: A survey of the computing oriented literature. *ACM computing surveys*, 48(4), 1-40.

Stieb, J.A. (2008). A critique of positive responsibility in computing. *Sci Eng Ethics*, 14, 219-233.

Stieb, J.A. (2009). Response to commentators of “A critique of positive responsibility.” *Sci Eng Ethics,* 15, 11-18.

Stieb, J.A. (2011). Understanding engineering professionalism: A reflection on the rights of engineers. *Sci Eng Ethics*, 17, 149-169.

Stovall, P. (2011). Professional virtue and professional self-awareness: A case study in engineering ethics. *Sci Eng Ethics*, 17, 109-132.

Tavani, H. (2015) *Ethics and technology: Controversies, questions, and strategies for ethical computing*, 5th ed. Hoboken, NJ: Wiley.

The ACM Code of Ethics and Professional Conduct (2018). https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct

Ullrich, P. & Knopp, P. (2018). Protesters’ reaction to video surveillance of demonstrations: Countermoves, security culture and the spiral of surveillance and counter-surveillance. *Surveillance & Society*, 16(2), 183-202.

Urman, J.E. & Blumenthal, R. (2018). An undergraduate ethics course for promoting common good computing: a progress report. *J. Comput. Sci. Coll.* 34(2) 39–45.

Young, S. (2017). Slipping through the cracks: Background investigation after Snowden. *Surveillance & Society,* 15(1), 123-136.

Wolf, M.J. (2016). The ACM Code of Ethics: A Call to Action. *Communications of the ACM,* 59(12), 6. https://doi.org/10.1145/3012934

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York, NY: Hachette Book Group.

PROJECT MANAGEMENT: EXPERIENTIAL LEARNING PEDAGOGY

Shalini Kesar, James Pollard

Southern Utah University (USA)

kesar@suu.edu; jamespollard1@suu.edu

ABSTRACT

This paper reflects on collaborative research to create an experiential learning pedagogy for a project management class. The motivation for this research was based on the authors belief that core skills such as teamwork, communication, professionalism and ethics are part of experiential learning pedagogy, which prepares the students to deal with challenges faced in today's technology related businesses. This class is taken by computer science and information system students prior to graduating. The National Society for Experiential Education's framework was used while developing the curriculum. The theoretical framework is also used to reflect on the past and present's outcomes of the project as well to provide guidelines on curriculum robustness for an experiential learning project management class. The findings and lesson learned will contribute towards pedagogy where instructors intend to create an experiential learning platform for their project management students.

KEYWORDS: Project management, Information systems, Experiential learning, DATIM, NSEE.

1. INTRODUCTION

This paper reflects on collaborative research to create an experiential learning pedagogy for a project management (capstone) class. The motivation for this research was based on the authors belief that core skills such as teamwork, communication, professionalism and ethics are part of experiential learning pedagogy, which prepares the students to deal with challenges faced in today's technology related businesses. This class is taken by computer science and information system students prior to graduating. Prior to taking this class, the students are required to take foundation classes in computer science and information systems courses. In addition to the foundation classes, the students also have taken classes that introduce them to topics linked with understanding the interface between computer software and hardware including processor architecture, computer arithmetic, instruction set architecture, and assembly language. They also learn about and conduct projects in higher-level languages, computer performance analysis, basic concepts of pipeline, introduction to memory management, Computer IO, and disk storage systems. Both computer science and information systems students are required to take programming language courses in C, C+ and Java Script. They are also required to take a class in database management systems that include: database processing, data modelling, database, database design, development, implementation, alternative modelling approaches, and implementation of current DBMS tools and SQL.

The Computer Science and Information Systems (CSIS) capstone project's gives senior-year students an opportunity to manage a major information systems development/enhancement project, in which they apply what they have learned in various other courses to a single project. The emphasis is on enterprise-level project management. This course's learning outcomes include: 1) Providing senior-year students an opportunity to manage a major information systems development or enhancement project, which is proposed by the instructor or the students themselves. Students will utilize various technical skills that they have learned from various courses in a single project. Project management techniques will be emphasized in class; 2) Developing students' abilities to initiate, analyse, evaluate and manage an IS project in preparation for making informed decisions as a future IT project manager; 3) Developing students' ability to distinguish among opinions, facts, and inferences; to identify underlying or implicit assumptions; to make informed judgments; and to solve problems by applying evaluative standards when working with an IS project; 4) Providing students an understanding of the challenges of different project stages, and will develop skills to understand and handle a variety of project management challenges; and 5) Developing a comprehensive understanding of implementing and managing an Information Systems project. The number of students ranged from ten to fifteen in each class annually. The class duration was fifteen weeks and was conducted in the last semester before the students graduated.

2. BACKGROUND

This section describes the National Society for Experiential Education (NSEE) framework and the Forest Service application, Design and Analysis Toolkit for Inventory and Monitoring (DATIM) project which has been a collaborate project for capstone classes since 2015. The National Society for Experiential Education's framework was used while developing the curriculum. The theoretical framework was also used to reflect on the past and present's outcomes of the project as well to provide guidelines on curriculum robustness for an experiential learning project management class.

2.1. NSEE framework in context of project management class

The National Society for Experiential Education (NSEE) framework the varied roles and responsibilities represented in the field of experiential education. Founded in 1971. The members of NSEE advocate for the use of experiential learning throughout the educational system; to disseminate principles of best practices and innovations in the field; to encourage the development of research and theory related to experiential learning; to support the growth and leadership of experiential educators; and to create partnerships with the community. Since the founding of the Society, the Board of Directors, staff, and membership have been governed by policies and practices that guide ethical actions, relationships, and decisions. The distinctive purposes and conditions of experiential learning demand that all those involved in the process of learning through experience are held to the highest standards of mutual respect and responsibility, and that ethical behaviour is understood and practiced at every level of the learning process. Eight Principles of Good Practice for All Experiential Learning Activities include: Intention; Preparedness and Planning; Authenticity; Reflection; Orientation and Training; Monitoring and Continuous Improvement; Assessment and Evaluation; and Acknowledgment. The NSEE points out the importance of organizational partnership. As mentioned earlier, the instructor encouraged guest lectures from different departments, including English, communication, and law. This provided background information and learning skill about technical writing, communication and presentation skill and relate the importance of these skills in the context and working environment in which they will be exposed to after graduation.

In the first principle, intention, it is important that demonstrates the purposefulness that enables experience to become knowledge. It is goes beyond more than just outlining the goals and objectives, and activities that define the experience. This directly leads to the next principle of preparedness and planning. Participants must ensure that they enter the experience with sufficient foundation to support a successful experience. This early stage requires to carefully align the identified intentions with the goals, objectives and activities to be flexible enough to allow for adaptations as the experience unfolds. Authenticity principal must have a real-world business environment that will be useful and meaningful in reference to context selected. This is followed by the reflection, that helps to transform simple experience to a learning experience. These activities designed can create knowledge that the learner can internalized. Furthermore, it allows the instructor to reflect on the assumptions and hypotheses about the outcomes of decisions and actions taken, along the outcomes against past learning and future implications. This reflective process is integral to all phases of experiential learning, from identifying intention and choosing the experience, to considering preconceptions and observing how they change as the experience unfolds. Orientation and Training principle within the NSEE framework adds value of the experience to be accessible to both the learner and the learning facilitator(s), and other parties who are part of the experiential learning. Monitoring and continuous improvement ensures that the experience, as it is in process, continues to provide the richest learning possible, while allowing responsibility and accountability. The feedback will contribute towards the contentious improvement for an experiential learning experience. Assessment and evaluation processes should be systematically documented with regard to initial intentions and quality outcomes. Assessment is a method to not only develop and but refine the initial goals and quality objectives identified. Finally, acknowledgment principal is a recognition process of learning and its impact on all the parties involved throughout the experience. This principle is a form of a celebration of learning and helps provide closure and sustainability to the experience.

2.2. DATIM application

The context of the project was the Forest Service application, Design and Analysis Toolkit for Inventory and Monitoring (DATIM). This project is partially funded by the USDA Forest Service, Forest Inventory Analysis. This paper specially reflects on the four years of DATIM project's structure, results, and lessons learned. The goal was to facilitate an experiential learning environment to provide student's flexibility to identify their milestones within the project scope. Based on the feedback from both students and client, every year the project and pedagogy style was modified. Experiential learning topics such as ethics were included as part of learning outcomes to help prepare computing students to meet global challenges that they may face in real business settings.

The United States Forest Service (USFS) is an agency of the United Sates Department of Agriculture that administers the nation's 154 national forests and 20 national grasslands, which encompass 193 million acres. Major divisions of the agency include the National Forest System, State and Private Forestry, Business Operations, and the Research and Development branch. The application, Design and Analysis Toolkit for Inventory and Monitoring (DATIM) is being developed for Inventory and Analysis (FIA), a research branch of the USFS. The DATIM project is a collaborative effort between the National Forest System (NFS) and USFS Research & Development (R&D), Forest Inventory and Analysis (FIA), and Ecosystem Management Coordination (EMC) staff. The co-author, a Research Fellow at Southern Utah University is the main collaborator with the development team of DATIM. He has been part of the DATIM capstone project for the last four years and has worked closely with the primary author as a client for the student' projects.

3. EXPERIENTIAL LEARNING PROJECT MANAGEMENT

This section reflects on the NSEE framework and project management for information system and computer science students. The capstone class was comprised of senior undergraduate Computer Science (CS) and Information Systems (IS) students. The DATIM project was part of the project management between 2016-2020. The projects involved teamwork and lasted fifteen weeks. The aim of the capstone curriculum was to foster a teaching environment to: 1) include interdisciplinary partnership among university departments; 2) cultivate local industry alliances; 3) encourage students' analysis and synthesis of skills and knowledge in a real business set-ting project.

3.1. Experiential Pedagogy

In 2016, the first year of the capstone class, twenty students who participated in DATIM application project included both computer science (5) and information systems (15) students. The student diversity included: three Caucasian females, seven male international students, and remaining male Caucasian students. In 2017, eighteen information systems students participated, which included zero females. Whereas, in 2018, there were six students, of which one was a female student. In 2019, there were eleven students, all male students. In 2020, six students were part of DATIM project with one female student. While planning the lessons, different types of assessments were designed. The evaluation process of each assignment varied, for example, report writing, class presentation, public presentation. The design for assessment incorporated core skills evaluations. These were systematically designed by keeping in mind the initial intentions. A Few months prior to the class, all students who had registered for the course were emailed details and context of the DATIM project. Meetings were held with students interested in the project to discuss the planning phase for their individual projects. On the first day of class, the DATIM project was sub-divided into smaller, individual projects. The students were provided an opportunity to ask questions. It was interesting to see how varied the point of view of the students was. Some students were more concerned with the timeline and the usefulness of such projects that required working with business. Over the five years, different projects were developed which were associated DATIM. Having said that, the project management pedagogy was designed with the NSEE framework, providing an annual communality to the overall capstone class design. The planning phase also included asking students the same set of questions at the beginning and at the end of the semester.

By the end of fifteen weeks, the same students who questioned the intention of the purposefulness and the experience gained in the classroom were appreciative of the exposure they received to a real business setting. The intention was to examine students' experience of the project. The goals of individual projects was to prepare students for the working life, making them familiar with the work place by practicing their skills on real-world business setting. The

Client (co-author) and his team visited the class nearly every week to provide the students feedback during project development. They were required to update their weekly report with their experience, questions for the clients and progress of their projects. The Learning Management System, Canvas was used to post resources as well as students' findings. The weekly report included students' weekly goal as a team individual task, challenge encountered, and how did they overcome the challenges as both a group and as individuals. While designing the project management, the next two principles, authenticity and reflection, included formal and informal feedback, and assignments. Based on the feedback received from students, peers and the client, every year, the class curriculum as well DATIM application project was modified. Both the client and the instructor felt it was important that the students understand the background of DATIM application and its purpose. Training and orientation both in person and online tutorials were given to the students each year. Reviewing progress of students, feedback and outcomes of the project, helped the instructor to modify the curriculum each year. The rubric for assessments, presentation and group report were carefully designed for an experiential learning classroom environment. Given that reflection is also an essential tool for adjusting the experience and measuring outcomes, students were required to present progress report three times in the semester.

3.2. Outcomes of experiential pedagogy

The table below summarizes the main outcomes during the years 2016- 2020 capstone project linked with the DATIM application.

Table 1. Capstone Project Outcomes.

Year	Sub-projects	Outcomes
2016	Programming, Training Webinar, Functionality of DATIM	Template of Webinar Training, Guidelines Programming report in Silverlight, 1000+ tests ran on DATIM
2017	Google Analytics, Database Expansion on DTIM & ATIM, Functionality of DATIM	Report of Google Analytics on DATIM, 2,700+ tests ran on DATIM, Training session for next class.
2018	Cybersecurity & FIA, Section 508 Complaint testing (1st time).	3 local conference presentations, Security report for SUU FIA team, 508 testing report, Recommendations for PEN testing.
2019	Section 508 Compliant Testing, Gender Neutrality, International FS sites Comparison.	2 local conference presentations, Detailed report on 508 Section 508 Compliant solutions, Resource of 45 articles on Gender Neutrality, Report to compared DATIM to 6 international countries for design.
2020	Find solutions to automate the issues identified in Section 508 Complaint	Solutions & resources for 9 issues Posters accepted for local conference.

4. DISCUSSION AND LESSON LEARNED

Every year the pedagogy was modified based on previous year's findings. Students were exposed to a real business setting where they appreciated the depth of responsibility, accountability and skills needed to work in a team. Students enjoyed the freedom to design their own sub-projects within the context of the larger DATIM project. This freedom to create subprojects within a large complex application development fostered a spirit of collaboration and team work. Subsequently, it also provided a means to create a sense of work ethics and professionalism as can be found within a professional team. Literature suggest that such skills are crucial when preparing students facing global challenges in the work field. In their paper, Leidig and Lange (2012) highlight lessons learned from one hundred projects over the course of ten years. Although their focus was on community-based non-profit organizations, they provide useful insight about information systems capstone. It was also found designing pedagogy with NSEE framework created an informed learning context that foster students' growth and actualization of potential, achieve academic and civic goals, and reflect excellence in curriculum design and quality. Quality and design of the curriculum of the capstone class was validated by the same students were hired by the businesses because of their technical and soft skills. It could be argued that the clients had found an opportunity to develop the skills in the classroom catering for their businesses. These skills give the graduates the ability to work in an ever-changing environment with individual and group challenges. Developing experiential pedagogy also posed some challenges. For example, some students resisted this style of learning and wanted structured assignments. During planning, guest lectures, examples of successful and unsuccessful projects, importance of soft skills in project management were incorporated into the curriculum. At the end of the semester, it was interesting to note that some of the same students who resisted working with the DATIM project actually felt enriched and said it added "value" to their experience.

Overall students appreciated the pedagogy style of teaching and believed that that experiencing core skills added value to the class. The intention of choosing DATIM application project was to provide an exposure to the class of the different complexities and business settings. It was interesting that this pedagogical approach initially did not spark as much interest as the authors wanted but as the weeks passed, lectures and real case study scenarios related to the project made it interesting for them. Consequently, the pedagogy designed did provide an enriching learning style different with flexibility, opportunity to work as teams, and enhance their leadership skills. Subsequently, students were able to enhance their abilities to initiate, analyse, evaluate and manage an IS project in preparation for making informed decisions as a future IT project manager. Working in a diverse team environment proved beneficial to the students as they began to appreciate and understand ethical and professional code of conduct when handling different phases of project management challenges.

5. CONCLUSION

Pedagogy for a CS and IS capstone class should be an experiential educational that includes core skills such teaching ethics and professionalism. This will prepare the students to face the global challenges in today's technology-based businesses. The NSEE framework outline principles that to incorporate various phases as part of the experiential learning platform. The framework used in this paper to design capstone project management class where student's feedback and suggestions, collaborative work, modified pedagogy to prepare students with skills that will help them when facing challenges in computing field. Pedagogy also emphasize team work, communication skills, leadership skills, critical thinking solving problems and finding alternative solutions. To conclude, the DATIM application project provided unique opportunities for CSIS students to gain leadership skills, understand the different aspects of project management, and gain a real-business setting experience in a experiential classroom.

REFERENCES

Kesar, M., (2015). Including Teaching Ethics into Pedagogy: Preparing Information Systems Students to Meet Global Challenges of Real Business Settings, S. Kesar. ACM SIGCAS Computers and Society - Special Issue on Ethicomp, 45 (3).

Leidig, P. M. and Lange, D. K. (2012). Lessons Learned From A Decade Of Using Community-Based Non-Profit Organizations In Information Systems Capstone Projects. Proceedings of the Information Systems Educators Conference ISSN, 1435.

National Society for Experiential Education. (1998) Eight Principals of Good Practices for ALL Experiential Learning Activities. Retrieved from https://www.nsee.org/8-principles

START A REVOLUTION IN YOUR HEAD!
THE REBIRTH OF ICT ETHICS EDUCATION

Simon Rogerson

Centre for Computing & Social Responsibility, De Montfort University (UK)

srog@dmu.ac.uk

ABSTRACT

This paper is a viewpoint rather than grounded in research. It questions some of the established ICT norms and traditions which exist both in industry and academia. The aim is to review current ICT ethics educational strategy and suggest a repositioning which aligns with the concept of computing by everyone for everyone. Professional bodies, in their current role, have little influence on 97 percent of global software developers whose ethical code and attitude to social responsibility comes from elsewhere. There needs to be a radical change in how the ethical and social responsibility dimension of ICT is included in education of the whole population rather than focusing on the elitist computing professional community. It is against this backdrop that this paper explores new avenues for widening education, both formal and informal, to all those who may become involved in computing. The discussion concludes by laying out a new pathway for ICT ethics education which embraces people of all ages and all walks of life.

KEYWORDS: Thought Experiment, STEM, STEAM, Poetry, ICT Ethics, Oral History.

1. INTRODUCTION

Computing is no longer the sole domain of professionals, educated and trained through traditional routes to service public and private sector organisations under paid contracts. Computing is now by everyone for everyone with the advent of economically accessible hardware, a multitude of software tools and the Internet (Rogerson, 2019a). The IDC survey of 2018 found that there were, worldwide, 18,000,000 professional software developers and 4,300,000 additional hobbyists. The combined membership of leading professional bodies, ACM, ACS, BCS and IFIP represents only 3.09 per cent of that global total. The youngest app developer at Apple's Worldwide Developers Conference in June 2019 was Ayush Kumar aged 10 who started coding when he was 4 years old (Graham, 2019). He is not alone, 15 year old, Tanmay Bakshi, who is the world's youngest IBM Watson Developer, started software development when he was 5 years old (Param, 2018). These facts suggest that professional bodies, in their current role, have little influence on 97 percent of global software developers whose ethical code and attitude to social responsibility comes from elsewhere.

It is now over a year since the launch of the new code of ethics for the ACM. At the last ETHICOMP conference much time was devoted to discussing the Code and the part it would play in moving ICT ethics forward. The code has spawned the ACM Integrity Project: Promoting Ethics in the Profession (https://ethics.acm.org/integrity-project/). The aim of this 2-year project of the ACM Committee on Professional Ethics is to promote ethics in the profession though modern media: YouTube videos, podcasts, social media, and streaming video. The use of modern media should certainly appeal to post millennials and offers a new approach to engage with future generations of computer scientists. Unfortunately, there has been little exposure of this project in, for example, the ACM's flagship publication, the Communications of the ACM. However, that same publication ran as its cover story in the August 2019 edition "Embedded EthiCS: integrating ethics across CS education" (Grosz et al, 2019). This is a paper about Harvard reinventing the wheel of computer ethics education which has a long and comprehensive history stretching back to the 1980s (see, for example, Aiken, 1983, Johnson, 1985, and Miller, 1988). It offers little new insight, does not link to a 40 year history and experience (for example see Pecorino and Maner, 1985; Martin, Huff, Gotterbarn and Miller, 1996; and Bynum and Rogerson 2004), nor does it appear to connect with the Integrity Project.

Within industry and government, the compliance culture has taken a firm hold and so strangles the opportunity for dialogue and analysis of complex multi-faceted socio-ethical issues related to ICT. Superficial compliance is dangerously unethical and must be challenged vigorously in a technologically-dependent world. The timeframes for ICT development and ICT regulation and governance are, and will always be, misaligned. By the time some control mechanism is agreed, the technology will have moved on several generations and thus what has been agreed is likely to be ineffective. Currently, this seems to be the case with the governance of Artificial Intelligence, as there are so many opinions and vested interests causing protracted debate whilst AI marches onwards. Thus, it is paramount to imbue strategists, developers, operators and users with practical ICT ethics. In this way ethical computing has a chance of becoming the norm. Traditional approaches of professional bodies seem ineffective in a society which is moving rapidly towards complete dependency on technology

It is this landscape which makes the ETHICOMP 2020 theme, *Paradigm shifts in ICT* Ethics, so relevant. It is time to change. In the spirit of Kuhn (1962) we need a paradigm shift in ICT Ethics to address the societal challenges in the not-so-smart society of today. He suggests that scientific progress of any discipline has three phases: pre-paradigm phase, a normal phase and a revolution phase. Progress occurs when a revolution takes place after a dormant normal period and the community moves ahead to a paradigm shift. Given the ongoing frequent occurrence of ICT disasters, it seems ICT ethics education in its current dormant normal phase is in need of revolution. There needs to be a radical change in how the ethical and social responsibility dimension of ICT is included in education of the whole population rather than focusing on the elitist computing professional community. It is against this backdrop that this viewpoint explores four new avenues for widening education, both formal and informal, to all those who may become involved in computing. These avenues are: science and technology museums, history, thought experiments, and poetry. Such avenues also offer greater awareness to the public at large and align with Burton et al (2018), who use science fiction to teach ICT ethics, rather than Harvard's unimaginative, traditional approach already discussed.

2. AVENUE ONE: SCIENCE AND TECHNOLOGY MUSEUMS

An innovative interactive facility, Ethical Technology could be rolled out across the global network of science and technology museums and activity centres. It would be a programme for children and adults of all ages. It would be the catalyst for public awareness and public voice, schools' cross curricular activities, higher education research, teaching and learning, and new meaningful purpose for professional bodies.

Ethical Technology comprises four elements: *If by chance*, *Story time, Worldwide watch* and *Out and about* which combine to form a compound view of the ethical dimension of computing technologies and the ramifications for general population.

The first element, *If by chance*, is an opportunity to see giants of computing and literature in a different light through a set of hypothetical conversations between contemporary pairs. The conversation is on an ethical technology topic which is relevant to the expertise, experience and thinking of the pair of individuals. For example, Charles Babbage and Ada Lovelace discuss the increasing use of moral algorithms in everyday technology. The moral algorithm is used to embed ethical decision making into, for example, driverless cars. In a second example, Isaac Asimov and an intelligent robot discuss the relationship between human and android. Global laws and legal frameworks provide the scaffolding for a civilised society. How do *The Three Laws of Robotics* and android rights impact on a human society?

The second element, *Story time*, is a series of case scenarios which contain at least one ethical dilemma. Every story is based on either a real event or created by combining existing technologies in a societally-damaging manner. A typical example is this hypothetical story, *The Data Shadow*, about personal data which resides on the internet. It has its foundation in things which have happened (Rogerson,2017). It raises serious questions about whether we should be more aware of the risks associated with data shadows and whether there are things organisations and individuals could do to reduce such risks. The person engaging with the story is given a range of options to choose from to resolve the dilemma. The likely outcome of a chosen option is displayed. An infograph displays the totals of each option chosen with a summary of the likely outcomes for each.

The third element, *Worldwide watch,* is a repository of ethical and unethical technology occurrences. This element is an interactive blog which tracks worldwide media to collect stories of unethical and ethical technology. People will be able to give their opinions via a simple Likert scale for ethics. This could be part of a web-based offering where people could also add comments. In this way a rich dynamic record of ethical technology issues could be captured and retained. Again total opinions are captured using an infograph. This element offers an online link to be established across the participating community.

The fourth and final element, *Out and about*, is the outreach element of Ethical Technology. Much of the technology which pervades living space is ethically and socially sensitive. Such commonplace technology becomes invisible and unsurprising. *Out and about* aims to increase public awareness by providing multimedia information about technology and the associated ethical and societal sensitivity. Access could be, for example, through a QR reader on a smart phone. This element links people to Ethical Technology in their living space. Computer-based technology in everyday use in public spaces is used to illustrate the associated ethical and social issues. In this way public awareness is increased resulting in greater community questioning and calls for justified accountability. The applications such as ATMs, CCTV, and contactless payment systems would be fitted with QR codes which provide links to multimedia information about the ethical and social dimensions. Using this element, virtual assistants, such as Alexa and Siri, which masquerade as pseudo-friends, could be ethically contextualised.

3. AVENUE TWO: LEARNING FROM HISTORY

Deborah Johnson (1997, p61) wrote “The ethical issues surrounding computers are new species of generic moral problems. This is as true when it comes to online communication as it is in any other area of computing. The generic problems involve privacy, property, drawing the line between individual freedom and (public and private) authority, respect, ascribing responsibility, and so on. When activities are mediated or implemented by computers, they have new features. The issues have a new twist that make them unusual, even though the core issue is not.” She has been proved correct and consequently there is much to be learnt from the history of computers through, for example trade journal archives and the Communications of the ACM archive. However, in the context of ICT ethics the annals of ETHICOMP provide a particularly rich resource from which to learn. These historical records are important because history forces both scholar and practitioner “to lift their heads beyond the lab bench or the clipboard and realize the greater social, economic, and racial contexts in which their [work] plays out. It gives them a sensitivity that only the humanities can teach.” (Dubcovsky, 2014). The themes of the first conference, ETHICOMP 95, were *Ethical Development* (The use of development methodologies and the consideration of ethical dilemmas, user education and professionalism); *Ethical Technology* (Advances in technologies and the ethical issues they are likely to raise when applied to business and social problems.); and .*Ethical Applications* (Developing ethical strategies which allow technology to be exploited in an ethically acceptable way.). This remains a relevant landscape and so illustrates that reflecting on the past can help in addressing the ICT ethical challenges of today and the future.

In its 25 years history ETHICOMP has evolved from a fledgling conference to become a multi-generational global community. Therefore the value of ETHICOMP is not simply the conference themes and associated papers but it is its community of people and their thoughts and observations over the passage of time. Such narrative is often forgotten or overlooked. Oral history addresses this oversight because it is a way of gathering, recording, and preserving a diverse range of personal experiences that generally are not well documented in written sources (Dalton, 2017). It enhances reflective thinking in both typical and non-typical educational settings (Gazi and Nakou, 2015). For this reason oral histories began being collected at ETHICOMP 2018. To date 18 have been collected involving around 25 people belonging to the ETHICOMP community. Summaries are being prepared so that easy access can be facilitated.

Currently there exists a comprehensive record of ETHICOMP (held by this author) which maps the complete history of the ETHICOMP conference series from the kernel of the idea through to the latest conference However, this collection is inaccessible and the potential value for future generations of scholars, practitioners and observers is lost. This archive comprises conference themes and calls; programmes; abstracts, proceedings, posters and flyers; pictures; videos; oral histories; and spin-off activities and miscellaneous materials. This could be housed on a purpose-built website to establish an interactive repository which not only holds the archive but also had a blog and social media facility which enables visitors to add content and make comment. This interactive repository using a chronological taxonomy would offer practical ethical insight to all those involved in computing. The Chronological Taxonomy is novel and potentially valuable across all empirical research disciplines (Rogerson, 2018). It is a two-dimensional method of ordering; the first dimension is Chronology which focuses on time and the second dimension is Taxonomy which focuses on classification. A typical classification could be based in IFIP’s technical committees and working groups structure. The Chronological Taxonomy is a powerful tool which can be used to structure and analyse the interactive repository through using concatenated keys to sort the data thus making issues, trends and patterns more visible. As the data set held in the repository expands and perhaps automatic feeds are established to harvest advances and issues, the use of Big Data analytics might be needed.

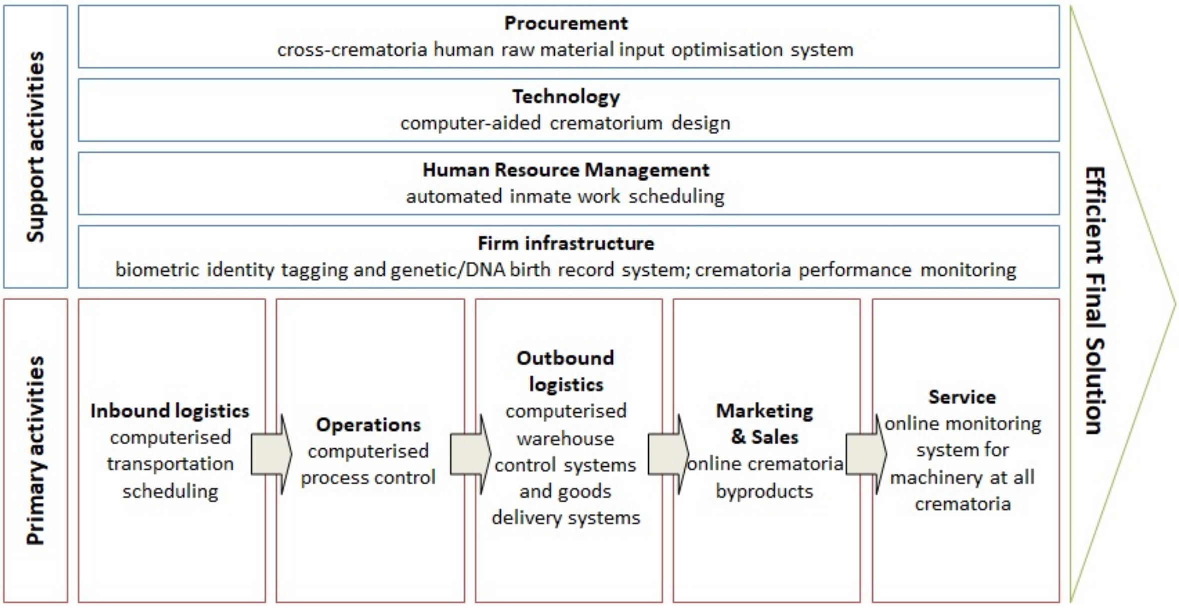
4. AVENUE THREE: THOUGHT EXPERIMENTS

Brown and Fehige (2014) explain that *thought experiments* are used to investigate the nature of things through one’s imagination. Usually they are communicated through narratives and accompanying diagrams. Brown and Fehige state that, “Thought experiments should be distinguished from thinking about experiments, from merely imagining any experiments to be conducted outside the imagination, and from psychological experiments with thoughts. They should also be distinguished from counterfactual reasoning in general, …". This approach can be used to explore the possible dangers of dual use of technological advances that could occur in the absence of effective ethical scrutiny. Rogerson (2020a) uses two thought experiment instruments to acquire new knowledge about the dangers of Free and Open-Source Software (FOSS) components which could have dual usage in the context of a fictitious system, *Open Genocide*. It is an investigation which cannot use empirical data as this would require the actual and immoral construction of a system of annihilation.

These thought experiments are grounded in the Holocaust enacted by the Nazis. By the end of the war some 6 million Jews and many millions of Poles, gypsies, prisoners of war, homosexuals, mentally and physically handicapped individuals, and Jehovah's Witnesses had been murdered. The historical account of human suffering is sickeningly shocking but alongside this is the realisation of evil brilliance, not mindless thuggery, that orchestrated the *Final Solution* (a Nazi euphemism for the plan to exterminate the Jews of Europe).

On a bitterly cold day in February 2006 the author (SR) was quietly standing looking at the building which housed the gas chamber at Auschwitz. He reflected on the evil brilliance which had facilitated the Final Solution. He wondered what might have happened if the computer technology of 2006 had been available to the Nazis. It was a consideration which resonated with *Would you sell a computer to Hitler* by Nadel & Wiener (1977). On returning home he completed the first thought experiment which was subsequently published (Rogerson 2006). Technological Determinism argues that technology is the force which shapes society. Computing power would therefore be a major force in activating the Final Solution. Value Chain Analysis (Porter, 1985) is one way to consider the impact of this force. Indeed Porter and Miller (1985, p151) wrote, "Information technology is permeating the value chain at every point, transforming the way value activities are performed and the nature of the linkages among them. … information technology has acquired strategic significance and is different from the many other technologies businesses use." Many computer application systems that existed in 2006 and which were proven and accepted could have been used to realise the Final Solution. This is illustrated in Figure 1.

Figure 1. Indicative examples across the *Final Solution* Value Chain.


Support activities
Procurement
cross-crematoria human raw material input optimisation system
Technology
computer-aided crematorium design
Human Resource Management
automated inmate work scheduling
Firm infrastructure
biometric identity tagging and genetic/DNA birth record system; crematoria performance monitoring
Primary activities
Inbound logistics
computerised transportation scheduling
Operations
computerised process control
Outbound logistics
computerised warehouse control systems and goods delivery systems
Marketing & Sales
online crematoria byproducts
Service
online monitoring system for machinery at all crematoria
Efficient Final Solution

In 2018, 12 years later, the second thought experiment was undertaken. Technology has evolved at a seemingly increasing pace. Indeed, "In the not-too-distant future with the cloud, big data, and maybe 80%-90% of the world's population online and connected the scope for systems of oppression seems limitless. Consequently, we must consider and counter what oppressive regimes of tomorrow's world could and might do in their drive to subjugate humankind." (Rogerson, 2015, p4) Imagine a world where all software is available as free open source. It might seem improbable but it is possible. If it were so, the wherewithal to exploit every technological advance for any cause, albeit good or bad, would exist. The scene is set for an *Open Genocide* system brutally to further the cause of an extreme faction at the expense of the world at large and "destroy in whole or in part, a national, ethnical, racial or religious group" (United Nations, 1948, Article 2). *Open Genocide* would comprise seven components:

1. Identify: Systematically review the whole geographic region to identify every person within the targeted group as well as every sympathiser of this group.
2. Detain: Organise the detention of all identified persons in distributed holding pounds. Each detainee is appropriately tagged.
3. Deport: Manage the distribution and redistribution of detainees to work compounds and prisons.
4. Use: Select detainees exhibiting work value for allocation to appropriate tasks.
5. Dispose: Remove to disposal units all valueless or dead detainees thereby freeing up space in prisons and work compounds.
6. Recycle: Collect, sort, recycle and market all seized assets. Produce and market detainee by-products.
7. Broadcast: Devise plausible propaganda for local and international audiences and communicate widely.

The pervasive nature of current computing technology facilitates all components of *Open Genocide*. A cursory inspection of two open source portals, SourceForge and The Black Duck Open Hub, reveals many useful items for the construction of *Open Genocide*.

These thought experiments might be shocking to many readers. That is their intention. It seems that if the Holocaust had occurred in our technologically-advanced modern world there is a very good chance that it would have completely succeeded. If ever there was an example to convince computing professionals, as custodians of the most powerful technology yet devised, of their responsibilities and obligations to humankind, this is it.

5. AVENUE FOUR: POETRY

There is value of linking the arts with the sciences in the delivery of ICT ethics education and poetry can serve as the vehicle (Rogerson, 2020b). Over 500 years ago, Leonardo da Vinci wrote that poetry is painting that is felt rather than seen. Such sentiment is echoed in "We can never learn too much about a poem, but always we come back to the work itself, for it exists not only as an historical object and the product of a particular mind and vision, but also in its own right, as an enduring work of art." (Anon, 1980). Poetry challenges us to think beyond the obvious and reflect on what has been, what is and what might be. Poetry can reboot the way in which social impact education is delivered to technologists. According to Rule et al (2004) incorporating poetry in science and technology teaching expands the curriculum beyond subject knowledge and process skills. They argue that images and metaphors in poems can clarify and intensify meaning. A poem has many layers and such richness can promote enlightenment and understanding. Poems can provide meaningful context. This is imperative in ICT education and awareness at all levels for all people as the social impact of technological advances is ever increasing. In partnership, computer science and liberal arts educators could offer an exciting new perspective through poetry as an instrument of presentation and discussion as well as in creative exercises for students.

Consider this example of using haikus. For readers who are unfamiliar with haikus, Trumbull (2003) explains that a haiku takes a three line format of 5-7-5 syllables known as a *kigo* and uses a cutting technique called *kire* to divide the verse into two parts for contrast or comparison. He argues that the more radical the verse the better the haiku. This is certainly the case when considering the broader issues surrounding the development and use of ICT.

Warning of technological advances without careful consideration beyond the technology is the focus of *Machine – the final chapter* (Rogerson, 2019b). Lack of proper checks and balances means that advances in computer technology from its inception by pioneers such as Babbage and Lovelace, through to the forefronts of artificial intelligence giving rise to ensuing disaster, potentially moving towards Armageddon as laid out in these three haikus.

Machine – the final chapter

Computer
Bits, bytes, ones, zeros
So Charles and Ada conceive -
IT’s Pandora’s box

Robot
Man and beast replaced
Same task over and over -
Objective carnage

AI
Boolean bible
Artificial ignorance -
Logical ending

Armageddon!

Poetry can be the key to unlock the door so the room can be explored. In the ICT setting this is important because often the most challenging ethical dilemmas are the least obvious. Different perspectives, for example, the poetical lens, can provide greater visibility.

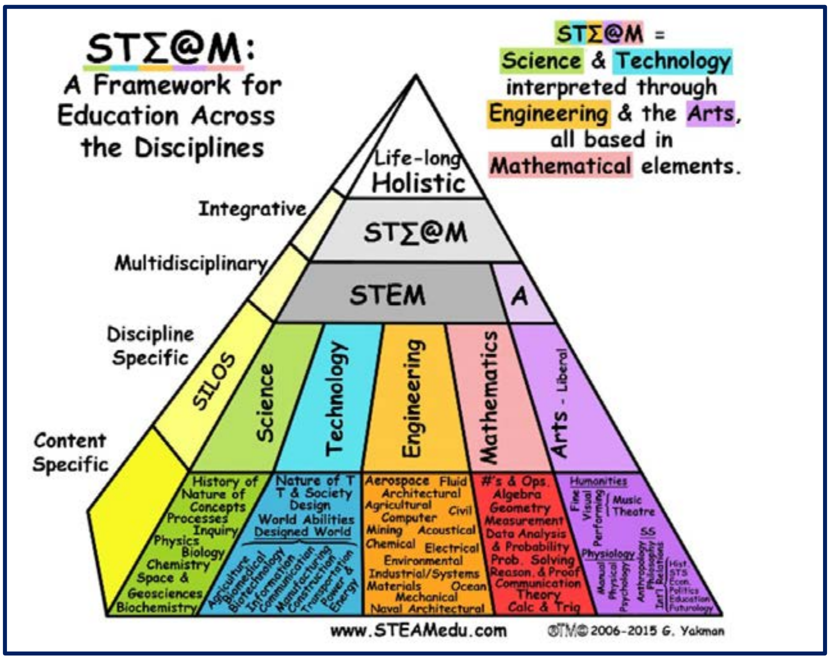
6. BEYOND STEM BOTH FORMALLY AND INFORMALLY

ETHICOMP 2020’s overarching theme of “Paradigm Shifts in ICT Ethics: Societal Challenges in the Smart Society” encourages its community to think beyond the obvious and traditional, and re-evaluate what should be done to ensure computing by everyone for everyone is ethically and societally acceptable. The four avenues discussed illustrate the way in which a paradigm shift might take place but for this to happen there needs to be an overarching framework to provide necessary scaffolding. A modification of the STEM model offers such a framework.

STEM has its roots in the US National Science Foundation and refers to teaching and learning in the fields of science, technology, engineering, and mathematics. It typically includes educational activities across all levels from pre-school to post-doctorate in both formal and informal settings. It requires the abandonment of top down approaches with teachers willing to talk to each other and to believe that interactions between subjects will result in enhanced learning opportunities (Williams, 2011). However, the problem with STEM is that it sustains the principle that there are two separate fundamental cultures; the scientific and the humanistic. This can restrict reflection and innovation. As Yakman (2008, p19) states, “Trends have also shown many of the branches of the arts being more and more marginalized …. this is a tragedy, as it eliminates many primary ways for students to obtain contextual understanding.” Indeed members of the ICT ethics community often encounter opposition from those who subscribe to this principle thus hampering the quest for ethical technology by design rather than by accident.

In 2006, Yakman conceived a model which blends STEM and the arts in a way which addresses this shortcoming. The STEAM Pyramid, as it is named and shown in Figure 2, aims, “…to correlate the subject areas to one another and the business and social development worlds … [and] … to create a matrix by which researchers, professionals, and educators could share information to keep education as up to date as possible while still having a basis in methodologies” (STEAM Pyramid History at http://steamedu.com/pyramidhistory/). Watson & Watson (2013, p3) explain that “The arts contribute to STEM education by exposing students to a different way of seeing the world. Students learn through different pedagogical modalities engaging their other interests. By applying the STEM disciplines, combined with real-world experience, students become more comfortable in both worlds.” In this holistic view the single discipline silos are augmented by a blended approach which better reflects the real world.

Figure 2. The STEAM approach as conceived by G. Yakman of STEAM Education in 2006.


STΣ@M:
A Framework for Education Across the Disciplines
STΣ@M = Science & Technology interpreted through Engineering & the Arts, all based in Mathematical elements.
Life-long
Holistic
Integrative
STΣ@M
Multidisciplinary
STEM
A
Discipline Specific
SILOS
Science
Technology
Engineering
Mathematics
Arts - Liberal
Content Specific
History of Nature of Concepts Processes Inquiry
Physics Biology Chemistry Space & Geosciences Biochemistry
Nature of T T & Society Design World Abilities Designed World
Agriculture Biomedical Biotechnology Information Communication Manufacturing Construction Transportation Power & Energy
Aerospace Fluid Architectural Agricultural Civil Computer Mining Acoustical Chemical Electrical Environmental Industrial/Systems Materials Ocean Mechanical Naval Architectural
#'s & Ops. Algebra Geometry Measurement Data Analysis & Probability Prob. Solving Reason. & Proof Communication Theory Calc & Trig
Humanities
Fine Visual Performing Music Theatre
Physiology
Manual Physical Psychology
SS
Anthropology Philosophy Int'l Relations
Hist. STS Econ. Politics Education Futurology
www.STEAMedu.com
©TM® 2006-2015 G. Yakman

STEAM is becoming increasingly important in ICT-related education and awareness. For example, recently published research discussed the combined use of robots and theatre for STEAM education across the science, art, and education communities (Barnes et al, 2020), whereas Song (2020) describes developing STEAM game content for infant learning. Finally, Ong et al (2020) investigate the effects of creative drama on situational interest, career interest, and science-related attitudes of science majors and non-science majors. It has been argued that computing is by everyone for everyone. Therefore, the ethical conduct of us all influences the acceptability of ICT and we all have the responsibility to challenge the unethical elements in ICT from inception through to implementation and use. Clearly, some more than others will have greater in-depth knowledge and experience of different facets. However, it is the population as a whole which has a complete view. This will include (by way of illustration) the hesitant user of a computerised public service who is the victim of poor system design as well as the junior software engineer who is bullied into unethical yet commercially valuable action. ICT ethics education and awareness must provide the tools and confidence to enable everyone to act responsibly and ethically.

The ICT ethics framework must cater for educational requirements of all. This is a paradigm shift. Yakman (2019) explains that "A" in STEAM represents the Liberal Arts which "include the ethics, ideals and emotional and physical expression grouped into overlapping categories of Humanities, Physiology and Social Studies (SS). In this way, STEAM formally adds in the subject area 'silos' of Language Arts, Social Studies, Music, Fine Arts, and PE." STEAM therefore appears to offer the basis of an ICT ethics framework. This author has been involved in ICT ethics education since 1994. This experience has led to the conclusion that for ICT ethics education programmes to succeed in making a difference there has to be persistent high visibility of these programmes and associated content. Ethics and Responsibility are the keywords. If STEAM is to be used it should be done so in an explicit ethics and responsibility landscape. Therefore to ensure ongoing visibility it could become STEAM-ER which would provide the fuel, impetus and environment to ensure that actions and outcomes in the technological world are more likely to be societally positive rather than societally negative.

Consider this example of how a STEAM-ER ICT ethics activity might work in practice. The example involves the poetry avenue discussed earlier. A cross-curricular project could be established for 11-18 year olds. The preparatory work, in various classes and activities, would focus on the uses and challenges of ICT in various settings. Poetry writing would also be covered both implicitly and explicitly. The culmination of this project would be a poetry writing exercise where pupils would be encouraged to consider the positive and negative effects of ICT, choosing a particular theme to be the subject of a poem. All poems would then be displayed in a public exhibition with pupils having the opportunity to engage with visitors to discuss their work. Those benefitting from this project would include: staff through positive interdisciplinary work relating to ICT ethics; pupils through increased understanding of the social impact issues which surround ICT, as well as practising a range of communication skills; and the general public through increased implicit understanding of ICT ethics.

The STEAM-ER proposal parallels the Responsible Research and Innovation (RRI) advances in recent years. For example, Stahl and Coeckelbergh (2016) argue that traditional approaches to ethics and risk analysis need to be modified to include reflection, dialogue and experiment which explicitly links to innovation practices and contexts of use. STEAM-ER offers a new approach to ICT ethics education which embraces a spectrum of disciplines in an integrated fashion. It could herald the rebirth of ICT ethics education and awareness which aligns more appropriately with a technologically-dependant world of the present and the future.

7. CONCLUSION

This paper has discussed the current shortcoming in ICT ethics education because of the ongoing focus on ICT students who aspire to enter the profession. Four new avenues, by way of illustration, have been outlined which offer novel informal and formal educational experiences. The ICT ethics education framework has been outlined which embraces people of all ages and all walks of life. It is time to start a revolution in your head which will culminate in ethical computing by everyone for everyone. We have to accept and adjust to the fact that we are all technologist to a lesser or greater degree. How we educate our future generations must reflect this change to ensure ICT is societally beneficial. This paper attempts to act as a catalyst for a much-needed paradigm shift in our thinking and application of ICT ethics education, one which heralds a rebirth.

REFERENCES

Aiken, R. M. (1983). Reflections on teaching computer ethics. *ACM SIGCSE Bulletin*, *15*(3), 8-12.

Anon (1980). How to enjoy a poem. In Cook, C. (Ed.) (1980). *Pears Cyclopaedia*. 89th edition, Book Club Associates, London. M3.

Barnes, J., FakhrHosseini, S.M., Vasey, E., Park, C.H. & Jeon, M. (2020). Child-Robot Theater: Engaging Elementary Students in Informal STEAM Education Using Robots. *IEEE Pervasive Computing*, 19(1), 10-21.

Brown, J. R. & Fehige, Y. (2014). Thought Experiments. In Zalta, E. N. (Ed.) (2017) *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition). Retrieved from https://plato.stanford.edu/entries/thought-experiment/

Bynum, T. W. & Rogerson,S. (Eds.) (2004). *Computer ethics and professional responsibility*. Blackwell Publishing.

Burton, E., Goldsmith, J. & Mattei, N. (2018). How to teach computer ethics through science fiction. *Communications of the ACM*, *61*(8), 54-64.

Dalton, S. (2017). What are oral histories and why are they important? 9 August. Retrieved from https://womenslibrary.org.uk/2017/08/09/what-are-oral-histories-and-why-are-they-important/

Dubcovsky, A. (2014). To understand science, study history. *Chronicle of Higher Education*, *60*(24). Retrieved from https://www.chronicle.com/article/To-Understand-Science-Study/144947

Gazi, A. & Nakou, I. (2015). Oral history in museums and education: Where do we stand today. *Museumedu*, 2(Nov), 13-30.

Graham, J. (2019). WWDC 2019: Meet Apple’s youngest app developer, Ayush. *USA Today*, 5 June.

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A. & Waldo, J. (2019). Embedded EthiCS: integrating ethics across CS education. *Communications of the ACM*, *62*(8), 54-61.

Johnson, D. G. (1985). *Computer ethics*. Englewood Cliffs (NJ).

Johnson, D. G. (1997). Ethics online. *Communications of the ACM*, *40*(1), 60-65.

Kuhn, T. (1962). *The structure of scientific revolutions*. University of Chicago Press. Chicago.

Martin, C. D., Huff, C., Gotterbarn, D. & Miller, K. (1996). Implementing a tenth strand in the CS curriculum. *Communications of the ACM*, *39*(12), 75-84.

Miller, K. (1988). Integrating computer ethics into the computer science curriculum. *Computer Science Education*, *1*(1), 37-52.

Nadel, L. & Wiener, H. (1977). Would you sell a computer to Hitler. *Computer Decisions*, 28, .22-27.

Ong, K. J., Chou, Y. C., Yang, D. Y. and Lin, C. C. (2020). Creative Drama in Science Education: The Effects on Situational Interest, Career Interest, and Science-Related Attitudes of Science Majors and Non-Science Majors. *EURASIA Journal of Mathematics, Science and Technology Education*, *16*, 4-21.

Param, S. (2018). Tanmay Bakshi: The Youngest IBM Watson Developer in the World. TechGig, 5 June.

Pecorino, P. A. & Maner, W. (1985). A proposal for a course on computer ethics. *Metaphilosophy*, *16*(4), 327-337.

Porter, M. E. & Millar, V. E. (1985). How information gives you competitive advantage. *Harvard Business Review*. 63(4), 149-160.

Porter, M. E. (1985). *Competitive advantage: creating and sustaining superior performance*. The Free Press, New York.

Rogerson, S. (2006). ETHIcol – A lesson from Auschwitz. *IMIS Journal*. Vol 16(2).

Rogerson, S. (2015). *The ETHICOMP Odyssey: 1995 to 2015*. Self-published on www.researchgate.net, 12 September DOI: 10.13140/RG.2.1.2660.1444.

Rogerson, S. (2017). The data shadow. *ACM SIGCAS Computers and Society*, *47*(1), 8-11.

Rogerson, S. (2018). Towards a Chronological Taxonomy of Tourism Technology: an Ethical Perspective. *ETHICOMP 2018*.

Rogerson, S. (2019a). Computing by everyone for everyone. *Journal of Information, Communication and Ethics in Society*. 17(4), 373-374. Translated into Japanese for inclusion in Murata, K. and Orito, Y (Eds.) Introduction to information ethics. Minerva Shobo, Kyoto, forthcoming.

Rogerson, S. (2019b). Machine - the final chapter. posted on *PoetrySoup*. 2 October. Retrieved from http://www.poetrysoup.com/poem/machine_the_final_chapter_1185554

Rogerson, S. (2020a). The dangers of dual use technology: a thought experiment exposé. In preparation.

Rogerson, S. (2020b). Poetical potentials: the value of poems in social impact education. *ACM Inroads*, *11*(1), 30-32.

Rule, A. C., Carnicelli, L. A. & Kane, S. S. (2004).Using poetry to teach about minerals in earth science class. *Journal of Geoscience Education*, 52(1), 10-14.

Song, M. Y. (2020). Design and Implementation of STEAM Game Contents for infant Learning Education using Gyroscope Sensor. *Journal of the Korea Society of Computer and Information*, *25*(1), 93-99.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, *86*, 152-161.

Trumbull, C. (2003). An Analysis of Haiku in 12-dimensional Space. *Paper for HSA_Meeting*.

United Nations (1948). *Convention on the Prevention and Punishment of the Crime of Genocide*, 9 December.

Watson, A. D. & Watson, G. H. (2013). Transitioning STEM to STEAM: Reformation of engineering education. *Journal for Quality and Participation*, *36*(3), 1-5.

Williams, J. (2011). STEM education: Proceed with caution. *Design and Technology Education: An International Journal*, *16*(1), 26-35.

Yakman, G., 2008, February. STEAM education: An overview of creating a model of integrative education. In *Pupils' Attitudes Towards Technology (PATT-19) Conference: Research on Technology, Innovation, Design & Engineering Teaching, Salt Lake City, Utah, USA*.

Yakman, G. (2019) STEAM- An Educational Framework to Relate Things To Each Other And Reality. *K12Digest*, December 12.

Note: This is a revision of the original paper to reflect the feedback provided by STEAM Education which is gratefully accepted. STEAM Education is a registered trademark.

4. Internet Speech Problems - Responsibility and Governance of Social Media Platforms

INTERNET SPEECH PROBLEMS – RESPONSIBILITY AND GOVERNANCE OF SOCIAL MEDIA PLATFORMS

Adriana Belgodere Rivera, Fabiana Piñeda Naredo

Interamerican University, School of Law (Puerto Rico)

Adriana.belgodere@lex.inter.edu; Fabiana.pineda@lex.inter.edu

ABSTRACT

The rise of fake news in our overly technological era has had a snowball effect due to social media. With extremely fast sharing and spreading of data, misinformation is bound to get tangled in the Internet's feeds. With tons of information, ideas and thoughts being poured into social media platforms every second, restrictions and censorship are almost impossible to avoid. However, with free speech on the line, regulation to prevent fake news is an uphill battle. In the debate of social media regulation, it remains unclear who should assume this responsibility. This research explores the complex and delicate issues that exists for fake news regulation through private actors. Through statistics, jurisprudence and more, the authors aim to find a clear look at the implications that go into the dissolution of fake news through social media governance.

KEYWORDS: fake news, free speech, social media, first amendment.

1. INTRODUCTION

Should the Government regulate what we say on social media? There's been a long and exhausting debate about this, and still no action has been taken by the government. Social media is currently regulated in a limited way by private actors and is largely immune from government regulation. For news creators and consumers, uncertainty is a fact and changes are endless, resulting in a confusion effect. On one hand it can be a helpful tool and on the other, it can be destructive and harmful. During the past decade, social media platforms have gained fame globally. Images, videos, podcasts, texts and innovations of all kinds have been generated, which can be broadcasted and shared, this includes fake news. Certainly, social media is a vehicle for social change.

One key benefit of social media is definitely how it has enhanced access to information. Accessing news about any given topic is just a click or tap away. Typing in a word, phrase or specific question into the most used search engine, Google, automatically generates millions of options that provide the knowledge the user requested. Individuals have created a sense of trust on the Internet, to the extent that we rely on it every day for, basically, anything. A poll created by Gallup reveals that today 40% of adults in the United States say they trust the accuracy of the news and information found on the Internet. Back in 1998, that percentage was at 25%. An even more significant increase is the amount of people that use the Internet to get information and news. The poll results show a 12% for 1998, while in 2019 64% of U.S. adults use this method when seeking information (Brenan, 2019).

2. WHAT IS FAKE NEWS?

On the Internet, nothing is what it seems. With a click of a button, you can find anything online. Although this is a great tool, at the same time we risk receiving wrong information or how it's commonly known as ‘fake news'. The term is defined by Cambridge Dictionary as “false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke”.

This phrase got even more famous after the 2016 U.S. presidential election. Donald Trump used it as a shield when media outlets ran stories that affected his image and campaign. This has created a false idea about what exactly ‘fake news’ is (Day and Weatherby, 2019).

> The possibility that false news stories on sites such as Twitter and Facebook impacted how Americans viewed national politics subsequently drove at least one major social media site to announce that it is now employing programs to fact-check stories on its platform and will flag those that do not meet certain press standards with warnings about their accuracy (VanLandingham, 2017, p. 12).

Digital platforms have created a whole new reading practice that has changed the processes by which people often interpret news and informational articles. A lot of the information we find online is not reliable, and although we may believe its true, many times it’s not. “Truthful information can be difficult to ascertain but can most likely be found on the majority of national and local news profiles. Major news publications have the burden to ensure the information they release is truthful and accurate” (Riddle, 2017). Doing the contrary, they might place themselves in a legal conflict, most likely to be defamation claims (Walters, 2018). This is why there is a need to identify the digital expression required to address the challenges caused by “fake news”.

> A poll conducted by Monmouth University reported that three out of four Americans believe that the media routinely report fake news, while a Gallup/Knight Foundation study found that 42 percent of Republicans consider any news stories that cast a political group or politician in a negative light to be fake news (Kirtley, n. d., para. 5)

A study from the Pew Research Center, states that Americans rate fake news as a problem bigger than racism, climate change, or terrorism and they blame political leaders for this. But they believe that journalists should be the ones fixing this problem (Mitchell, et. al, 2019). It seems that news sources in the United States have become subjected to a “Trump filter” that categorizes their credibility and journalistic skills into pro-Trump and anti-Trump.

> In April 2018, more than 170 television stations owned by conservative-leaning Sinclair Broadcast Group were ordered to use local anchors to produce a scripted “must-run” commentary decrying fake news. Responding to criticism from others in the industry that the segment was itself fake news intended to deceive viewers, Trump tweeted that “The Fake News Networks, those that knowingly have a sick and biased AGENDA, are worried about the competition and quality of Sinclair Broadcast.” (Kirtley, b, para. 6)

This constant labeling of the press by a political leader could lead to major repercussions like the total downfall of journalists, news and media channels. Not to mention, the President’s behavior towards the nation's press is viewed throughout the entire world. "Trump's words provide authoritarian leaders in countries such as Kenya, Venezuela, and the Philippines the ammunition to suppress opposition media, even as they spread fake video clips and stories through paid commentators and bots." (Kirtley, c, para. 8).

> Government can control and manipulate the flow of information about itself and its actors, so any determination of truth or falsity that fails to recognize the fundamental and coextensive right of the citizen to criticize without fear of sanctions or retribution—what Justice Brennan called "the central meaning of the First Amendment"—is flawed. A free and independent press, not a single leader or a government-run "Truth Tribunal," is the best means to ensure an informed citizenry, and to hold institutions and individuals to account. And that's not fake news. (Kirtley, d, para. 27).

2.1. The Problem with Fake News

The real problem behind fake news is not the amount of fake stories online, in fact the number of fake news stories is a small one. The actual severity of the problem is that these fake news stories reach more people than the real and factual stories. This causes people to abruptly act on misinformation, which translates into shares, likes and comments.

> In the ten months leading up to the 2016 presidential election, the top twenty fake news stories on Facebook had over nine million comments, reactions, and shares whereas articles from mainstream media saw a decline in comments, reactions and shares from 12 million to 7.3 million --fake news was shared more than real news. This sharing was not limited to average Facebook users. Television news hosts reported fake news stories, and then-President-elect Trump and his son shared other fake news stories on social media (Savino, 2017, p. 1101).

Parallel trends were seen on Twitter. In a research, the dissemination of true and fake news was verified on Twitter between 2006-2017. About 126,000 "tweets" were shared by 3 million people more than 4.5 million times. The findings included that "fake news" are more novel and inspired emotion of fear, disgust and surprise. "Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information" (Vosoughi, et. al, 2018).

These stories were shared and spread more than stories in the top news channels and pages, thus the phrase 'fake news' is often connected with digital platforms.

> Regardless of what 'fake news' actually means, it is typically tied up with anxieties about the democratic ramifications of the shift from consuming news from broadcast television and newspapers to consuming news on social platforms ... Thus, platforms including Facebook and Twitter have been heavily criticized for their role in spreading, facilitating, and even encouraging 'fake news' (Marwick, 2018, p. 476).

Search engines and social media give access to a worldwide audience and they give news creators access to extensive audiences. Therefore, consumers acquire an unlimited range of content on digital platforms and can also become producers, allowing them to express themselves. However, the harms can also be significant. This new era has crashed the pre-digital business model for news producers.

> Between 2011 and 2015, Australian newspaper and magazine publishers lost $1.5 billion and $349 million respectively in print advertising revenue, while gaining only $54 million and $44 million in digital (as noted by this inquiry's Issues Paper). By 2016, three quarters of the total Australian online advertising spend went to Google and Facebook. And since the US presidential election of 2016, the issue of fake news – and the ongoing dismissal by some public figures of unsympathetic coverage as 'fake news' – continues to challenge the credibility of journalism and news media (2018a, b).

These related platforms through which "fake news" can be disseminated have changed the reading practices of individuals. Nowadays, individuals are less likely to obtain news and information directly from news sources, instead they rely more in social media. "78% of users see news when they are using Facebook for other reasons. While only 34% of users subscribe to a news media source on social media" (Matsa and Mitchell, 2014).

Our legal system has remedies to manage other types of false statement claims against individuals. In libel or defamation claims there's an individual affected by the statement made about him/her by another individual or legal person. "The trouble with fighting back against fake news is it's hard to know who you're fighting against" (Gillin, 2020). Jayne Clemens, Senior Associate at Michel mores and Jacob Dean, Barrister at 5RB Chambers, explain the difference:

> Fake news and libelous material are both false. In the case of libelous publications, a complainant can sue for damages if they're able to demonstrate how the published material has caused, or is likely to cause, them serious harm. Fake news may well cause no harm at all, particularly if no one believes it. In short: libel is fake news, but fake news is not necessarily libelous (Clemens and Dean, 2019).

Another type of remedy provided for falsehood claims is intentional infliction of emotional distress (IIED). This "is a common law tort that is regularly alleged against fake news publishers under state law." (Klein and Wueller, 2019). IIED takes place when one person's intentional extreme behavior of one person provokes another individual's severe emotional distress. But, IIED claims require a stricter analysis of the statements. In order for a claim to proceed, these statements must be "so outrageous in character, and so extreme in degree, as to go beyond all possible bounds of decency, and to be regarded as atrocious, and utterly intolerable in a civilized community." (Klein, 2019a, b).

However, fake news poses an even bigger problem. How do we differentiate fake news from opinions? In this technological era, stories are shared thousands of times within seconds. "Libel suits are intended to provide compensation to those whose reputations have been harmed as a result of false statements made with actual malice." (Kirtley, d). But when it comes to fake news, how do we prove an actual damage or harm?

3. WHAT THE U.S. JUSTICE SYSTEM SAYS

Courts have seen a variety of claims regarding defamation, libel and other falsehood-related issues. "In the United States, truth is an absolute defense to libel and slander claims. Likewise, pursuant to First Amendment free speech protections, each defamation plaintiff must prove that defamatory statements were published with the requisite intent, which varies depending on the plaintiff's level of public prominence." (Klein, 2019a, b, c).

In *New York Times v. Sullivan* (1964), the Supreme Court faced for the first time "the extent to which the constitutional protections for speech and press limit a State's power to award damages in a libel action brought by a public official against critics of his official conduct." Then concluded that "[t]he Constitution accords citizens and press an unconditional freedom to criticize official conduct" (New York Times co. v. Sullivan, 1964). The Court also made clear an exception through which a public official can prevail if he/she proves that the statement was made with actual malice. This does not apply to private individuals.

A few years later, in *F.C.C. v. Pacifica Foundation* (1978), the Court held that:

> The fact that society may find speech offensive is not a sufficient reason for suppressing it. Indeed, if it is the speaker's opinion that gives offense, that consequence is a reason for according it constitutional protection. For it is a central tenet of the First Amendment that the government must remain neutral in the marketplace of ideas.

In *Snyder v. Phelps*, the Supreme Court faced whether First Amendment protect protesters at a funeral from liability for intentionally inflicting emotional distress on the family of the deceased. Justice Samuel Alito argued: "[o]ur profound national commitment to free and open debate is not a license for the vicious verbal assault that occurred in this case." (Snyder v. Phelps, 2011)

In a 2012 Supreme Court case, *United States v. Alvarez* (2012), the federal Stolen Valor Act of 2005 was invalidated. This statute criminalized false representation by individuals as having military awards. The Court held that interest in truthful speech was not sufficient to sustain the criminal statute.

> Some legal scholars describe the Alvarez ruling as delineating a "constitutional right to lie." While the FTC and Attorneys General have broad discretion to aggressively pursue unfair and deceptive trade practices claims against fake news publishers, defendants in other cases have had increasing success in raising First Amendment defenses to criminal and regulatory claims involving restrictions on false speech (United States v. Alvarez, 2012).

However, U.S. courts have not yet decided which standard applies when talking about 'serious intent' online. There is one less strict standard were courts "require the government to prove only that the defendant knowingly made a statement that 'was not the result of mistake, duress, or coercion' and that a 'reasonable person' would regard as threatening." (Larking and Richardson 2014). And another, which is stricter, were "courts analyze whether the speaker knew his speech was likely to be perceived by a reasonable person as threatening and was intended to be threatening." (Williams 2019).

> In a recent high-profile case, an actual photo- graph of Anas Modamani (a Syrian refugee living in Germany) taking a selfie with German Chancellor Angela Merkel was transformed into a fake news publication. Mr. Modamani's selfie photo was placed alongside photos of three other men, with the German headline "Homeless Man Set Alight in Berlin. Merkel Took a Selfie with One of the Perpetrators." After the false image

> began circulating on Facebook, Mr. Modamani sought an injunction from a German court that would have required Facebook to block its reproduction and circulation. On March 7, 2017, the court denied the injunction, ruling that Facebook had not manipulated the content itself and, therefore, could not be held legally responsible (Klein, 2019).

4. FAKE NEWS REGULATION: WHAT'S BEING DONE

The Internet has grown and evolved to be such a powerful tool that the need for some type of control or limit is logical. The amount of information accessible through search engines is unimaginable. Just as stated in previous sections, there are endless possibilities when it comes to navigating the web. Or are there?

4.1. Private Response: Social Media Giants' Role

Private companies have gained enormous amount of power and control. So much, that you may even label it as censoring. For example, "[t]he policies of Google, a company that has emerged in recent years as the clear leader among Internet search engines and is responsible for an enormous share of the nation's access to content online, represent a glaring example of corporate abuse of regulatory power." (Dickerson, 2009)

> In recent weeks, social media platforms like Facebook, Twitter, and YouTube have banned hate groups and controversial figures such as Louis Farrakhan of the Nation of Islam, Alex Jones of Infowars, and others. This resulted in a chorus of criticism from politicians (across the ideological spectrum), pundits, and the general public. The Trump administration even launched a website to allow users who have been suspended or banned from social media platforms to voice their complaints about political bias. But do social media sites have a legal obligation to allow equal access to all viewpoints? Do they violate the First Amendment if they exclude controversial speakers from their platform? Should the government step in to take corrective action? The answer to all these questions is a resounding no. The First Amendment applies to government actors. It means the government cannot punish you for speech it disapproves of. But social media platforms are private companies. Whether privately run platforms should censor speech is a separate issue ripe for debate. But there should be no debate as to whether the First Amendment bars Facebook, Twitter, or YouTube from restricting speech: No government, no First Amendment claim. (Ortner 2019)

In recent years, we've seen social media company executives like Mark Zuckerberg, be challenged in different aspects, but always regarding the policies of the platform. Quite possibly, creators of what have become a communication staple did not see this coming during the first stages. But, since these platforms have come to replace public squares, additional control and regulation is necessary.

> With public controversy over so-called "fake news" and hate speech swirling around them, leading internet companies are now being forced to confront their roles in the digital ecosystem: at birth, these companies were simply technology platforms; over the

> years, they have grown into brokers of content and truth on a global scale. (Open Mic 2017)

According to technologist and codirector of the Civic Signals project at the National Conference on Citizenship, Eli Pariser, social media platforms are quite similar to actual physical spaces. He uses a comparison between LinkedIn and Twitter to highlight the importance of structure and rules of these spaces. Pariser states that on LinkedIn, users will only see appropriate and professional content. Whereas on Twitter, it's the total opposite. (Pariser, n.d.). Through this perspective, social media platforms create the norms for users to follow. In this way, they can control what's expected of the users and consequently, what will develop as the platform's culture.

Another platform that has taken action to fight fake news is Snapchat. Since 2017, it "requires publications to fact-check articles for accuracy, not publish misleading or deceptive links, and not impersonate or claim to be a person or organization with the intention to confuse or misleads others." (Mejia, 2017).

4.2. Relevant Statutes

In order to regulate cyberspace, the United States has implemented different laws regarding the Internet. Among them are the following statutes: (1) Communications Decency Act of 1996 (CDA), (2) Child Online Protection Act (COPA), (3) Electronic Communications Privacy Act (ECPA), (4) Computer Fraud And Abuse Act (CFAA), and (5) Cyber Intelligence Sharing And Protection Act (CISPA).

On the other hand, Singapore has created legislation specifically against fake news. This law came into effect in October and "provides for prosecutions of individuals, who can face fines of up to 50,000 SGD (over $36,000), and, or, up to five years in prison." (Griffiths, 2019) It also provides sanctions of up to 1 million SGD or approximately $735,000 for companies who are found guilty of publishing fake news. However, concerns have arisen due to the possible effect on free speech.

Other countries who have taken a step towards the regulation of fake news are Russia, France and Germany. Although these governments claim the need to avoid the dissemination of misinformation, human rights advocates fear that the purpose of the legislation is to suppress political oppositions. (Ungku, 2019).

4.3. Freedom of Speech

Freedom of speech is a right guaranteed by the U.S. Constitution. In its First Amendment, it states that "[c]ongress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press, or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances."

Courts in the United States have seen a great amount of cases arguing the extent of this constitutional right. "Ruling unanimously in Reno v. ACLU, the Court declared the Internet to be a free speech zone, deserving of at least as much First Amendment protection as that afforded to books, newspapers and magazines." (ACLU, n.d.) Through innumerable cases regarding different types falsehood and/or tort claims, the "Court finds speech unprotected only when it does not contribute to the exchange of ideas as evidenced by external indicia of harm resulting from speech or from actions that are independently harmful, such as threats or lies."(Wells, 2010)

Although citizens have a constitutional right to speak and express themselves, this right is not unlimited. There's a fine line when it comes to falsehood claims and freedom of speech. In *United States v. Alvarez,* the Court expressed that "[t]he threat of criminal prosecution for making a false statement can inhibit the speaker from making true statements, thereby "chilling" a kind of speech that lies at the First Amendment's heart."

Nowadays, whenever a person feels like sharing, questioning or criticizing a particular topic, he or she can tap, type or upload to social media. "Many of the potential uses of social media go hand-in-hand with the freedoms that the Supreme Court has made clear are at the core of the First Amendment's protection." (Hitz, n.d.)

5. CONCLUSION

Articles found via social media can lead the reader to misinterpret the context, structure, style and voice of the news. Due to the popularity of social networks, the discovery of information is being transformed from an individual to a social endeavor where normally users are not objective while using these platforms. (Nikolov, et. al.) This will completely change the way people interact with the articles, and how they will discover information and news. Recent examples of "fake news" show the openness and disposition of the user with whether they can be manipulated by others or not, being directly proportional.

Educating the public about the harms of fake news is not enough to eradicate its effect. Regulating speech on social media is a difficult and delicate task for the United States government. Since Free Speech is guaranteed by the First Amendment, it forces the government to be extra cautious when regulating such areas.

> Ultimately, no algorithm alone can stop a moving target like fake news, which succeeds because it seeks to blend in like a chameleon with legitimate news stories. However, people and technology working together in creative ways can help limit the impact of fake news. The CDA silently allows all these methods to develop in a natural manner without a constant threat of litigation. (Walters, 2018).

No algorithm can actually stop "fake news" because it is perfectly suited to the fragmented news scenery, where "clickbait" has been linked to the rapid spread of misinformation online. (Chen, and Rubin, 2018). In an effort to avoid tainting the constitutional rights of citizens, it is necessary for the private sector to take hold of this much needed regulation. Since technology alone will not suffice, human intervention is essential for an effective system to work. Social media platforms need to reevaluate their algorithms based on certain shared characteristics that establish what could potentially be a "fake news" story. The algorithms could have a source fact-checking tool to track its origin and thus, flag questionable pieces. (Baron and Crootof, 2017).

Occasionally, something shared by friends on social media can be taken for granted and obtain validity, although it can be a "fake news" story. In order for citizens to distinguish factual news sources from fake news, an accreditation system should be created for these platforms. A professional organization should be established for the creation of codes of conduct on the Internet. This organization would grant accreditation to the different news sources. Based on this system, accredited news sources would then be held liable if their content is proven to be fake news. Social media platforms can create a sanction system when the accredited sources are flagged as fake news.

A similar approach has been suggested for implementation in Argentina. The legislation would create a commission for the verification of fake news in order to prevent false information spreading during national election campaigns.” In order to identify false information, the Commission for the Verification of Fake News would verify the content by comparing it with user comments, checking complaints about the data, and reviewing excessive viralization, among other evidence. (Rodríguez-Ferrand, 2019).

When creating this regulation, the First Amendment rights of citizens should be upheld while controlling the excess of falsehood. It will not become a means of silencing people, but a measure of regulating what is fake and what is true. As social media companies take hold of the regulation, users should become more aware of the type of information they’re receiving on a daily basis. This type of control can become a steppingstone in the development of fake news regulation.

REFERENCES

Baron, S., & Crootof, R. (2017, October 21). Fighting Fake News – Workshop Report. Retrieved from https://law.yale.edu/fighting-fake-news-workshop-report

Brenan, M. (2019). In U.S., 40% Trust Internet News Accuracy, Up 15 Points. Retrieved from https://news.gallup.com/poll/260492/trust-internet-news-accuracy-points.aspx

Chen, Y., & Rubin, V. (2018). Perceptions of Clickbait: A Q - Methodology Approach. Proceedings of the Annual Conference of CAIS / Actes Du Congrès Annuel De LACSI. http://doi.org/10.29173/cais1046

Clemens, J., & Dean, J. (2019). Fake News or Libel?. *CEO Review*. Retrieved 20 November 2019, from https://www.ceo-review.com/2019-fake-news-or-libel

Day, T. and Weatherby, D. (2019). Shackled Speech: How President Trump's Treatment of the Press and the Citizen-Critic Undermines the Central Meaning of the First Amendment. *Lewis & Clark Law Review*, 23(311).

Dickerson, N. (2009) What Makes the Internet So Special? And Why, Where, How, and by Whom Should Its Content Be Regulated?, *Hous. L. Rev. 46(61), 90.*

"Fake News," Hate Speech & Freedom of Expression: Corporate Responsibility in An Age of Alternative Facts (2017) Retrieved from http://fakenews.openmic.org [Accessed 10 Nov. 2019].

FAKE NEWS | meaning in the Cambridge English Dictionary. Retrieved 20 November 2019, from https://dictionary.cambridge.org/dictionary/english/fake-news

F.C.C. v. Pacifica Found., 438 U.S. 726, 745, 98 S. Ct. 3026, 3038, 57 L. Ed. 2d 1073 (1978).

Gillin, J. (2020). Fact-Checking Fake News Reveals How Hard it is to Kill Pervasive 'Nasty Weed' online. Retrieved from https://www.politifact.com/punditfact/article/2017/jan/27/fact-checking-fake-news-reveals-how-hard-it-kill-p/

Griffiths, J. (2019, October 2). Singapore 'fake news' law comes into force, offenders face fines and prison time. Retrieved from https://edition.cnn.com/2019/10/02/asia/singapore-fake-news-internet-censorship-intl-hnk/index.html

Hitz, J. (n.d) Removing Disfavored Faces from Facebook: The Freedom of Speech Implications of Banning Sex Offenders from Social Media. Retreived from http://ilj.law.indiana.edu/articles/13-Hitz.pdf

Klein, D., & Wueller, J. (2019). Fake News: A Legal Perspective. *Journal Of Internet Law*, *20*.

Kirtley, J. Getting to the Truth: Fake News, Libel Laws, and "Enemies of the American People". Retrieved from https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/the-ongoing-challenge-to-define-free-speech/getting-to-the-truth/

Marwick, A. (2018). Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technical Review*, *2*, 474-512.

Matsa, K., & Mitchell, A. (2014). 8 Takeaways About Social Media and News. *Pew Research Centre*.

Mejia, Z. (2017, January 23). Snapchat wants to make fake news on its platform disappear, too. Retrieved from https://qz.com/892774/snapchat-quietly-updates-its-guidelines-to-prevent-fake-news-on-its-discover-platform/

Mitchell, A., Gottfried, J., Stocking, G., Walker, M., & Fedeli, S. (2019). Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed. Retrieved from https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

New York Times Co. v. Sullivan, 376 U.S. 254, 84 S. Ct. 710, 11 L. Ed. 2d 686 (1964).

Nikolov, D., Oliveira, D. F., Flammini, A., & Menczer, F. (2015). Measuring online social bubbles. *PeerJ Computer Science, 1*, 1–14. doi: 10.7717/peerj-cs.38

Ortner, D. (2019, June 17). Government regulations for Social Media Companies that Censor Political Speech? No thanks. Retrieved from https://pacificlegal.org/government-regulations-for-social-media-companies-that-censor-political-speech-no-thanks/

Paul Larkin & Jordan Richardson, True Threats and the Limits of First Amendment Protection, THE HERITAGE FOUND. (Dec. 8, 2014), http://www.heritage.org/the-constitution/report/true-threats-and-the-limits-first-amendment-protection

Pariser E, (n.d.). "What obligation do social media platforms have to the greater good?". Retrieved from https://www.ted.com/talks/eli_pariser_what_obligation_do_social_media_platforms_have_to_the_greater_good/transcript?language=en

Riddle, J. (2017). All Too Easy: Spreading Information Through Social Media - The Arkansas Journal of Social Change and Public Service. Retrieved from https://ualr.edu/socialchange/2017/03/01/blog-riddle-social-media/

Rodríguez-Ferrand, G. (2019). Initiatives to Counter Fake News in Selected Countries. *The Law Library of Congress*. Retrieved from https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf

Savino, E. (2017). Fake News: No One is Liable, and That is a Problem. *Buffalo Law Review*, *65*(1101).

Snyder v. Phelps. (2011). Oyez. Retreived from https://www.oyez.org/cases/2010/09-751

Technology and Liberty: Internet Free Speech. (n.d.). Retrieved from https://www.aclu.org/other/technology-and-liberty-internet-free-speech

Ungku, F. Factbox: Factbox: 'Fake News' laws around the world. (2019, April 2). Retrieved from https://www.reuters.com/article/us-singapore-politics-fakenews-factbox-idUSKCN1RE0XN

U.S. Const. amend. IUnited States v. Alvarez, 567 U.S. 709, 132 S. Ct. 2537, 183 L. Ed. 2d 574 (2012)

VanLandingham, R. (2017). Jailing the Twitter Bird: Social Media, Material Support to Terrorism, and Muzzling the Modern Press. *Cardozo Law Review*,*39*(1).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science Magazine*, *359*(6380).

Walters, R. M. (2018). How to Tell A Fake: Fighting Back Against Fake News on the Front Lines of Social Media. Texas Review of Law & Politics, 23(1), 111–170.

Wells, C. (2010) Regulating Offensiveness: Snyder v. Phelps, Emotion, and the First Amendment, 1 Cal. L. Rev. Circuit 71, 79

Williams, A. (2019). You Want to Tweet About It but You Probably Can't: How Social Media Platforms Flagrantly Violate the First Amendment. *Rutgers Computer & Tech. Law Journal*, *45*.

PROBLEMS WITH PROBLEMATIC SPEECH ON SOCIAL MEDIA

William Fleischman, Leah Rosenbloom

Villanova University (USA), The Workshop School (USA)

william.fleischman@villanova.edu; leah.rosenbloom@workshopschool.org

ABSTRACT

In this paper, we consider some of the tensions and conflicts between freedom of speech on the Internet, and other public goods and individual rights. The dimensions of the problem include: Threats of physical violence to individuals; threats directed at groups defined by ethnic, national, religious, sexual or gender identity, or political orientation; abusive, harassing, and/or hateful speech; incitement to self-harm; doxing; social exclusion; and dissemination of false information. Since initiating this study, we have also come to see an additional dimension, the importance of which we were slow to recognize. This is a pattern of misleading, self-contradictory, content-free, and deceptive speech on the part of spokespersons for one of the dominant social media platforms – Facebook. We make provisional suggestions for discouraging the actions of troll armies and for applying more vigorous measures of transparency in regard to political advertising on social media.

KEYWORDS: freedom of speech, violent and abusive speech, internet, social media.

1. INTRODUCTION

The popularization of the Internet promised a radical democratization of communication: Everyone can be a publisher, cost of entry is low, and access is available to anyone connected to the Internet. But early on, prescient individuals understood that “cheap speech,” in Eugene Volokh’s pungent phrase, carried other implications not all of which are entirely conducive to the dissemination of reliable information or the reasoned discourse of the marketplace of ideas.

As Tim Wu points out in “Is the First Amendment Obsolete?” (Wu, 2017), the assumption that the most serious threats to freedom of speech come principally from governmental actors is no longer entirely valid. Direct censorship, either in the form of government action or by content filters and human content monitors employed by social media platforms, can now be supplemented or supplanted by the actions of privately constituted troll armies or bands of individuals and/or robots programmed to drown out disfavoured speech. These means are at the disposal of powerful private interests and loosely organized partisan groups.

In this paper, we consider some of the tensions and conflicts between freedom of speech on the Internet, and other public goods and individual rights. We argue that the widest scope should be afforded individuals’ right to free expression, but believe that social media platforms should be held to certain standards of responsibility for preventing or redressing harms resulting from speech on these platforms.

2. THE TROUBLED AND VIOLENT TERRAIN…

"We've got a speech problem on the Internet!" is an observation that covers a lot of ground. The dimensions of the problem include: Threats of physical violence to individuals; threats directed at groups defined by ethnic, national, religious, sexual or gender identity, or political orientation; abusive, harassing, and/or hateful speech; incitement to self-harm; doxing; social exclusion; and dissemination of false information.

2.1. Troll Armies

Gamergate (Wikipedia, 2019) is a well-known example of an online hate mob. Lately, troll armies have figured prominently in polarized political discourse. Tim Wu (2017) cites two examples: David French, a writer associated with the conservative *National Review*, and Rosa Brooks, a professor of law at Georgetown University, both targets of online mobs for criticism of the current U.S. president.

The rhetoric was murderous and hateful in both instances – Nazi imagery and the face of his daughter in the gas chamber in the case of French (French, 2016), and extremely violent misogynistic language directed at Brooks. (Brooks, 2017)

2.2. Reverse Censorship and Flooding

Another technique used by governments to marginalize dissident speech involves mobilizing a volume of opposing information to drown out inconvenient speech or distort the informational environment to render the speech dubious and unimportant. An important variant of reverse censorship, used in political advertisement, floods public discourse with patently false information or "fake news." It is widely understood that in 2016 targeted political advertisements disseminating false information were instrumental in both the U.K. Brexit referendum (Cadwalladr, 2019) and the U.S. Presidential election. (Lapowsky, 2018)

3. … NONETHELESS THERE ARE GOOD REASONS TO PROTECT EVEN EXTREME SPEECH…

In spite of these examples, there are important reasons to favor protecting even extreme speech on the web.

3.1. Free Speech Protects Protest Movements

Online speech is a cornerstone of modern social change. Organizers rely heavily on social media and online communication to disseminate information and coordinate action. The first widely-studied instance of online organizing was the Arab Spring, during which Egyptian and Tunisian dissidents used Facebook, Twitter, and blogs to discuss and promote revolutionary ideas before taking to the streets (Howard et al., 2011). More recently, University students in China relied heavily on social media to share information and encourage participation in Hong Kong's Umbrella Movement (Lee et al., 2016). Social media was also a vital part of the Euromaidan uprising in Ukraine (Bohdanova, 2014), and the native environmental movement in Standing Rock, North Dakota (Johnson, 2017). Governments recognize the potential of social media to amplify political discontent, which is why they censor and block posts, platforms, and sometimes even the entire Internet.

Social media is appealing to activists for its immediacy and accessibility. Allowing governments to regulate content on social media may introduce leeway for them to further silence activists and destabilize revolutionary movements.

3.2. Governments Stifle Speech to Protect Private Interests

Systemic regulation and censorship of online speech often goes hand in hand with other socially repressive tactics, for instance the incarceration and “re-education” of political dissidents in China (Human Rights Watch, 201?). While the most extreme examples are dictatorial governments that overtly censor and crush opposing voices, democratic governments also monitor and undermine dissent, especially when private financial interests are involved.

Among the most influential private sector interests is Big Oil. Global financial interest in the acquisition and distribution of oil has been a key driver of worldwide surveillance and censorship, even in countries with robust free speech protections. In the UK, counter-terrorism police labeled the non-violent environmental group Extinction Rebellion alongside neo-Nazis as an “extremist ideology” (Dodd & Grierson, 2020). Under the current policies of most social media platforms, a “terrorist” designation is grounds for immediate permanent dismissal from the platform. Any individual users found in support of “terrorism” would be similarly censored or dismissed.

During the Dakota Access Pipeline protests of 2016, independent media collective Unicorn Riot reported the disproportionate censorship and arrest of social media journalists, including Facebook’s (purportedly accidental) removal of a protest livestream for violating community standards (Unicorn Riot, 2016). Facebook has a designed “Law Enforcement Online Request System” that law enforcement can use to make requests for content forfeiture and removal. The specific nature and frequency of Facebook’s compliance with those requests is not public knowledge.

3.3. “Dangerous” Speech Is Defined by People in Power

It is natural for governments to want to minimize speech that encourages violence on its citizens, or otherwise undermines national interests. There is a delicate line, however, between censorship to maintain citizens’ health wellbeing, and censorship to maintain governments’ power and authority. Governments have been known to leverage the public’s need for security to censor and oppress opposition, and this is unlikely to change in the digital age. It is necessary to protect free speech rights not just for activists and dissidents, but for all people.

Defending free speech for everyone is not easy, but it is vital in order to maintain free speech for those who need it most. American Civil Liberties Union director Anthony Romero emphasizes the importance of defending neo-Nazis’ right to peaceably assemble: “We simply never want [the] government to be in a position to favor or disfavor particular viewpoints. And the fact is, government officials…are more apt to suppress the speech of individuals or groups who disagree with government decisions” (Romero, 2017). For over half its history, the United States government was more likely to agree with the KKK than the NAACP.

4. … BUT NOT NECESSARILY WITHOUT ANY LIMITS

First Amendment protections are intended to restrain the power of government to interfere with the freedom of expression of individuals. However, speech online occurs through an internet service provider or a social media platform which inherit First Amendment protections. As private enterprises they can and do set rules that should apply to their users. Often these rules are vague and inconsistently applied. In particular, when speech policies come into conflict with the profit motive of the platform, these policies frequently evaporate. The result is often flagrantly abusive, obscene, threatening or deceptive speech. We propose adoption of stricter and clearer standards and procedures to limit the harms associated with three aspects of speech on social media – troll armies, incitement, and dissemination of political advertisements containing verifiably false content.

We believe that social media platforms should hold themselves to standards and policies that are consistently applied and promote more responsible speech. Reddit provides an example that shows this is possible. (Marantz, 2016)

4.1. The Case of Troll Armies

We propose that in cases like those cited, where troll armies coordinate hateful speech and threats against an individual, the social media platform that facilitates such an attack take action against members of the mob. The principle here is that those claiming the right to speak violently and abusively under the doctrine of freedom of expression are, in fact, acting to attempt to silence another individual, and therefore, perversely curtailing the very same right of that individual.

These situations should be relatively easy to document, once the attacked individual registers a complaint with the social media platform, since they consist of N --> 1 more or less synchronized messages (multiple sources, one target). We recognize the limitations of algorithmic detection or wholesale human moderation of every instance of hateful and threatening speech. By contrast, the cases to which we refer are not so frequent that human inspection would be impossibly difficult. Naturally, this requires nuanced consideration of the multiple messages, some of which may be reasoned arguments expressed in strong language and should be differentiated from those that simply spew hate in language and (photoshopped) images.

4.2. Incitement

In the United States, legal theory governing cases involving incitement is not entirely satisfactory. The current standard for determining whether speech constitutes illegal incitement comes from the U.S. Supreme Court opinion in the 1969 case, *Brandenburg v. Ohio*. “There, the Court held that advocacy of violence is protected unless it ‘is directed to inciting or producing imminent lawless action and is likely to incite or produce such action.’” (Pew, 2015) The *Brandenburg* precedent has been cited on numerous occasions but there has been a certain difficulty. Conflicting interpretations of the word “imminent” have given rise to inconsistency in interpretation. According to Pew (2015), there is a consensus forming around the interpretation that “imminent” refers to “a matter of several days.”

However, the well-known case, *Planned Parenthood of Columbia/Willamette, Inc. v. Am. Coalition of Life Activists*, illustrates another difficulty in application of *Brandenburg*. This case stemmed from a 1995 action “in which antiabortionists uploaded approximately 200 more physicians’ names to a website, again including their photographs and addresses. Some of the physicians’ names were crossed out, others were in grey font, and the rest were in black font. The following legend accompanied the files: ‘Black font (working); Greyed-out Name (wounded); Strikethrough (fatality).’ In other words, the website recorded murders and other violent attacks against the abortion doctors. The names of the three doctors who had been murdered from 1993 to 1994 were struck through. Several physicians featured on the website, terrified for their lives, brought suit.” (Pew, 2015)

The original decision by a three-judge panel of the 9th Circuit held that the contested speech was protected in view of the lack of time frame indicating an “imminent” threat. However, the decision was reversed by the 9th Circuit sitting *en banc*. Although the court held the speech was protected under *Brandenburg*, it found another basis for declaring it unprotected. Under the “true threats doctrine,” since so many of the physicians identified on the website had already been killed or injured, the court held that no one posting the photographs, names, and addresses could believe otherwise than that those who were targeted by the website would live in fear that they might be the next target of an assailant in real life. (Pew, 2015)

The difficulty of “drawing lines” in cases bordering on incitement is apparent. On the other hand, living in a world in which the protracted state of anguish caused by the flood of hatred that engulfed David French and his family is “normal” seems deeply unsatisfactory. It seems that the only recourse in situations of this sort consists of protective reactions by the affected individual(s) such as blocking those responsible for virulently hateful attacks, and avoiding those sites where active participation results in further abuse. But how is this consistent with the idea of the web as the modern incarnation of the marketplace of ideas and reasoned discourse?

These difficulties are the result of the protections provided by the First Amendment (at least in the context of the U.S. Constitution) against government censorship of speech by individual citizens. But the situation is significantly altered when the speech occurs on a social media platform whose ownership is in private hands and whose owners have the freedom to set rules promoting a marketplace of ideas rather than a marketplace of murderous invective. Considered from this standpoint, we encounter some truly puzzling stories.

For one particularly flagrant example, there is the following: “A journalist on Monday tweeted, without naming them directly, that “they” need to be “killed” before “they kill us.” Although interpretations of the tweet may differ from person to person, many saw it as a tweet advocating violence against members of a particular community – Muslims – and called it genocidal in its intent. Many also reported the tweet as well as the account as abusive or advocating violence. However, Twittter doesn’t find anything wrong with the tweet and replied to many saying that the tweet advocating murder of people doesn’t violate its rules. (India Today Tech, 2018)

The full text of the tweet is as follows: “They killed us in Trains, Hijacked our Planes, held us Hostage in Hotels, Forced us to flee #Kashmir, & now Killing us for holding the Tricolor on #RepublicDay.

> Truth is We Live in Fear, NOT They.
>
> NO more. Always Carry Lethal Weapons. KILL them before they KILL us.
>
> #MondayMotivation”

> “The tweet was made from a profile that is verified and it is possible that because of the popularity of the account, Twitter decided that exhortations to kill people was probably alright to tweet from this particular account.” (India Today Tech, 2018)

Apparently, incitement to genocide is permitted if your Twitter profile is verified and popular.

In this context, it seems appropriate to quote the language pertaining to incitement articulated in Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR):

> “While the right to freedom of expression is fundamental, it is not absolute. A State may, exceptionally, limit the right under Article 19(3) of the ICCPR, provided that the limitation is:
>
> - Provided for by law, so any law or regulation must be formulated with sufficient precision to enable individuals to regulate their conduct accordingly;
> - In pursuit of a legitimate aim, listed exhaustively as: respect of the rights or reputations of others; or the protection of national security or of public order (ordre public), or of public health or morals; or
> - Necessary in a democratic society, requiring the State to demonstrate in a specific and individualised fashion the precise nature of the threat, and the necessity and proportionality of the specific action taken, in particular by establishing a direct and immediate connection between the expression and the threat.

Article 20(2) of the ICCPR obliges States to prohibit by law ‘any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence’.” (Article 19. 2018)

In particular, with regard to incitement:

- “Incitement. Prohibitions should only focus on the advocacy of discriminatory hatred that constitutes incitement to hostility, discrimination, or violence, rather than the advocacy of hatred without regard to its tendency to incite action by the audience against a protected group.
- Six-part threshold test. To assist in judicial assessments of whether a speaker intends and is capable of having the effect of inciting their audience to violent or discriminatory action through the advocacy of discriminatory hatred, six factors should be considered:
- Context: the expression should be considered within the political, economic, and social context prevalent at the time it was communicated, for example the existence or history of conflict, existence or history of institutionalised discrimination, the legal framework, and the media landscape;
- Identity of the speaker: the position of the speaker as it relates to their authority or influence over their audience, in particular if they are a politician, public official, religious or community leader;

- Intent of the speaker to engage in advocacy to hatred; intent to target a protected group on the basis of a protected characteristic, and knowledge that their conduct will likely incite the audience to discrimination, hostility, or violence;
- Content of the expression: what was said, including the form and the style of the expression, and what the audience understood by this;
- Extent and magnitude of the expression: the public nature of the expression, the means of the expression, and the intensity or magnitude of the expression in terms of its frequency or volume; and
- Likelihood of harm occurring, including its imminence: there must be a reasonable probability of discrimination, hostility, or violence occurring as a direct consequence of the incitement." (Article 19. 2018)

Our contention is that, with regard to incitement, social media platforms should have policies of self-regulation that at least meet the standard articulated in the ICCPR as described above. There is no rationalization, other than a shameless addiction to value-insensitive economic gain, that can justify labeling a message that calls for mass murder "in compliance with the rules of permissible expression" on a social media platform.

There is something profoundly discordant about the fact that, deliberating a case involving speech that was clearly intended to incite violence against identified individuals without specifying a time frame that would trigger application of the Brandenburg precedent, the 9th Circuit Court was able to find a basis for ruling the speech unprotected, whereas social media platforms which, as private entities, are under no obligation to adjudicate the nicety of determining the threat "imminent," cannot bring themselves to act in a conservative fashion when faced by a speech act of similarly explosive and violent intent.

Of course, we understand that if it is a speech act with the potential to generate numerous "clicks" and contribute significantly to the platform's bottom line, all bets are off.

4.3. The Case of False Political Advertising

Our other proposal has to do with political speech - specifically political ads that circulate false or discredited information. Facebook is currently involved in such a dispute. We don't want to say that these things should be outlawed - there's plenty of history, going back to the election of 1800 in our country, of scurrilous political speech (McCullough, 2001). But it is troubling that ads on Facebook and other platforms appear and then disappear without any trace so there is no possibility of auditing them or providing public scrutiny. They are particularly pernicious because they are targeted to people identified as susceptible through analysis of their Facebook profiles.

Carol Cadwalladr (2019) has documented how this occurred in the Brexit referendum and how Facebook has stonewalled any serious attempt to investigate the sources of funding and means of targeting these false and vanishing ads. Facebook executives have been notably oblivious and evasive about such advertising. (Lee, 2018)

Our proposal is to force the social media platform to keep publicly accessible, auditable records of political ads so that they can be scrutinized and rebutted in the same way that's possible with ads on other media. We are not alone in thinking that this is a reasonable measure for curbing the most egregious excesses of dishonesty in political advertising.

U.S. Senator Amy Klobuchar of Minnesota introduced a bill in 2017 bill – S. 1989, The Honest Ads Act – which appears to have stimulated pre-emptive action on the part of several social media platforms. At the present moment, Twitter has announced a complete ban on political ads.

Although Facebook claims to have set up an archive of the description we favor there are reasons to think it will fall short of the promise of promoting greater transparency in political advertising on social media. For one thing, Facebook continues to resist any voluntary action to remove ads that contain verifiably false content. In addition, we have the discouraging episode related in a recent article "On Dec. 10, [2019], just two days before the United Kingdom went to the polls, some 74,000 political advertisements vanished from Facebook's Ad Library, a website that serves as an archive of political and issue ads run on the platform. For a while, what the company described as a "bug" wiped 40% of all political Facebook ads in the UK from the public record." In fact, this was just one of a litany of disturbing failures that undercut the usefulness of the archive. (Smith, 2020)

5. FIRST YOU SAY YOU DO, AND THEN YOU DON't, THen YOU SAY YOU WILL…

Long ago, James Moor foresaw that computers would offer new capabilities and choices for action; these possibilities would, in turn, require new policies or call into question the adequacy of existing policies for ethical conduct in deployment of these new choices. He predicted further that the attempt to remedy an existing policy vacuum might bring us face to face with an underlying conceptual vacuum. (Moor, 1985)

This is an apt description of the problem of characterizing social media platforms in regard to the dissemination of news. Should these platforms be seen as neutral technology companies that have simply built an ingenious set of tools for passing along content including entertainment, artistic creation and news, or have they, in fact, evolved to play a role in the sphere of public information comparable to traditional media companies without having assumed any of the traditional responsibilities and ethical norms that define the legal and social expectations of journalism? What should we call them? What do they call themselves?

In the words of a song made famous by Ella Fitzgerald and Louis Armstrong,

"First, you say, you do
And then you don't
And then you say, you will
And then you won't
You're undecided now
So what are you gonna do?" (Genius, 2020)

If we are permitted to paraphrase somewhat facetiously: "First you say you are, and then you're not. And then you say you're not and then you are." This seems to be the stance of Facebook, Google, Twitter, and all the major social media platforms when confronted with the question as to whether they are tech companies or publishers.

Except that it has nothing to do with being undecided. It is, rather, a matter of exploiting what Shoshana Zuboff, echoing James Moor, has characterized as "a lag in social evolution" in the face of the rapid build-out of social media platform capabilities that "outrun public understanding and the eventual development of law and regulation that it produces." (Zuboff, 2015) Many observers representing very different perspectives on the spectrum of political affiliation and belief (Levin, 2018; Dougherty, 2019; Shaw, 2019) have noted the disparity between the public posture of Facebook as a self-identified tech platform and the contrary representations it makes in court filings where it seeks the protection accorded to publishers concerning decisions made about "what not to publish." (Levin, 2018)

We believe that regulating social media, based on the substantial advertising revenues realized from activity properly described as that of a publisher of news would, on the one hand, cut through the conceptual ambiguity cynically and opportunistically exploited by platforms like Facebook and, on the other hand, provide a basis for requiring that such media platforms put "their houses in order" by means of consistently applied journalistic oversight in regard to hate speech and incitement, the publication of which they permit.

The standard should be "You may publish anything the law allows as long as you refrain from monetizing it. If, however, you wish to derive streams of revenue from advertising associated with provocative speech, then apply the standards of good journalistic practice and responsibility to the publication thereof."

REFERENCES

Article 19. (2018) Responding to 'hate speech': Comparative overview of six EU countries. *Article 19*. Retrieved from http://europeanjournalists.org/mediaagainsthate/wp-content/uploads/2018/02/Final-compilation-off-regional-research-digital.pdf

Bohdanova, T. (2014). Unexpected revolution: the role of social media in Ukraine's Maidan uprising. *European View*, vol. 13, pp. 133-142. http://doi.org/10.1007/s12290-014-0296-4

Brooks, R. (2017). And then the Breitbart lynch mob came for me, *Foreign Policy*. Retrieved from https://foreignpolicy.com/2017/02/06/and-then-the-breitbart-lynch-mob-came-for-me-bannon-trolls-trump/

Cadwalladr, C. (2019). Facebook's role in Brexit and the threat to democracy. Retrieved from https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy?language=en

Dodd, V. & Grierson, J. (2020, 10 January). Terrorism police list Extinction Rebellion as terrorist ideology. *The Guardian*. Retrieved from https://www.theguardian.com/uk-news/2020/jan/10/xr-extinction-rebellion-listed-extremist-ideology-police-prevent-scheme-guidance

Dougherty, M. (2019, 25 June). What is Facebook? National Review. Retrieved from https://www.nationalreview.com/2019/06/facebook-legal-status-platform-publisher/

Feiner, L. (2020, 14 February). Facebook won't count influencer posts like Bloomberg's memes as political ads. Retrieved from https://www.cnbc.com/2020/02/14/facebook-wont-archive-sponsored-political-ads.html

French, D. (2016). The price I've paid for opposing Trump, *National Review*. Retrieved from https://www.nationalreview.com/2016/10/donald-trump-alt-right-internet-abuse-never-trump-movement/

Genius (2020). Undecided Lyrics, *Genius.com*. Retrieved from https://genius.com/Ella-fitzgerald-and-louis-armstrong-undecided-lyrics

Howard, P., Duffy, A., Freelon, D., Hussain, M., Mari, W. & Mazaid, M. (2011). Opening closed regimes. *Project on Information Technology and Political Islam*. Working paper 2011.1. Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/117568/2011_Howard-DuffyFreelon-Hussain-Mari-Mazaid_PITPI.pdf?sequence=1

Human Rights Watch (2018, 26 February). China: Big Data fuels crackdown in minority region. Retrieved from https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region#

India Today Tech (2018, 30 January). Tweet calling for killing of Muslims in India doesn't violate rules, says Twitter. Retrieved from https://www.indiatoday.in/technology/news/story/ tweet-calling-for-killing-of-muslims-in-india-doesn-t-violate-rules-says-twitter-1156743-2018-01-29

Johnson, H. (2017). #NoDAPL: social media, empowerment, and civic participation at Standing Rock. *Library Trends*, vol. 66 no. 2, pp. 155-175. Retrieved from https://muse.jhu.edu/article/686889/pdf

Kelly, J., Blood, D. & O Murchu, C. (2019, 10 December). Facebook under fire as political ads vanish from archive. *Financial Times*. Retrieved from https://www.ft.com/content/e6fb805e-1b78-11ea-97df-cc63de1d73f4

Lapowsky, I. (2018). Mark Zuckerberg speaks out on Cambridge Analytica Scandal, *Wired*. Retrieved from https://www.wired.com/story/mark-zuckerberg-statement-cambridge-analytica/

Lee, D. (2018). Mark Zuckerberg, missing in inaction, *BBC News*. Retrieved from https://www.bbc.com/news/technology-46231284

Lee, F. Chen, H-T. & Chan, M. (2016). Social media use and students' participation in a large-scale protest campaign: the case of Hong Kong's Umbrella Movement. *Telematics and Informatics*, vol. 34, pp. 457-469. Retrieved from https://www.researchgate.net/profile/Hsuan_Ting_Chen/publication/306075497_Social_Media_Use_and_University_Students'_Participation_in_a_Large-scale_Protest_Campaign_The_Case_of_Hong_Kong's_Umbrella_Movement/links/59e0313045851537160163a2/Social-Media-Use-and-University-Students -Participation-in-a-Large-scale-Protest-Campaign-The-Case-of-Hong-Kongs-Umbrella-Movement.pdf

Levin, S. (2018, 3 July). Is Facebook a publisher? In public it says no, but in court it says yes, *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/jul/02/facebook-mark-zuckerberg-platform-publisher-lawsuit

Marantz, A. (2018). Reddit and the struggle to detoxify the internet, *New Yorker*. Retrieved from https://www.newyorker.com/magazine/2018/03/19/reddit-and-the-struggle-to-detoxify-the-internet

Moor, J. (1985). What Is Computer Ethics?, Metaphilosophy, vol. 16, no. 4, pp. 266-275.

McCullough, D. (2001, at 545ff). *John Adams*, Simon & Schuster, New York

Nelson, J. (2017). Should Google and Facebook be considered media companies? Retrieved from https://rossdawson.com/blog/google-facebook-considered-media-companies/

Patterson, B. (2017). Police Spied on New York Black Lives Matter Group, Internal Police Documents Show, retrieved from https://www.motherjones.com/crime-justice/2017/10/police-spied-on-new-york-black-lives-matter-group-internal-police-documents-show/

Pew, B. (2015). How to Incite Crime with Words: Clarifying Brandenburg's Incitement Test with Speech Act Theory, *BYU Law Review*, vol. 2015, issue 4, no. 8. Retrieved from https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=2996&context=lawreview

Romero, A. (2017, 15 August). Equality, justice and the First Amendment. Retrieved from https://www.aclu.org/blog/free-speech/equality-justice-and-first-amendment

Shaw, C. (2019, 20 September). Facebook admits in court filing that it is a publisher, opening itself up to libel suits. *The New American*. Retrieved from https://www.thenewamerican.com/tech/computers/item/33469-facebook-admits-in-court-filing-that-it-is-a-publisher-opening-itself-up-to-libel-suits

Smith, R. (2020, 14 January). The UK election showed just how unreliable Facebook's security system for elections really is. *BuzzFeed News*. Retrieved from https://www.buzzfeednews.com/article/rorysmith/the-uk-election-showed-just-how-unreliable-facebooks

Unicorn Riot (2016, 20 September). Statement on recent censorship of #NoDAPL coverage. Retrieved from https://unicornriot.ninja/2016/statement-recent-censorship-nodapl-coverage/

Wikipedia (2019). Gamergate controversy. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Gamergate_controversy

Wu, T. (2017) Is the First Amendment obsolete? Knight Foundation First Amendment Institute. Retrieved from https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete

Zuboff, S. (2015), Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, vol. 30, pp. 75-89.

SRI LANKAN POLITICS AND SOCIAL MEDIA PARTICIPATION
A CASE STUDY OF THE PRESIDENTIAL ELECTION 2019

Chintha Kaluarachchi, Ruwan Nagahawatta, Matthew Warren

Deakin University (Australia)

c.kaluarachchi@deakin.edu.au; rnagahawatta@deakin.edu.au; matthew.warren@deakin.edu.au

ABSTRACT

Social media has been a recent phenomenon which impact all parts of the society. The aim of this study is to investigate the key themes of the posts that dominated social media landscape and user generated interactions related to those posts during the Sri Lankan presidential election 2019. The paper has shown that social media has the ability to generate discussion and debate. The most popular FB posting was to promote particular presidential candidates, and it may possibly a guided influence. The most interacted themes were “Social fragmentation and reduce voter’s loyalty theme” and “economic justice”. When we analyse the user interactions we can see both guided and freely evolving interactions related to the Sri Lankan presidential election 2019.

The authors have shown that Facebook did have guided and freely evolving influences on the Sri Lankan presidential election of 2019. Findings of the case study concluded that there is a significant impact on politics campaign and level of user’s interaction of social media. Further, it was established that the misuse of social media has becoming a major challenge for future free and fair elections. Therefore, necessitates the need for a national social media policy that focuses on election as key stakeholders in the registered political parties.

KEYWORDS: Social media, Political Campaigns, Voters, awareness, candidates, democratic.

1. INTRODUCTION

The Internet has become a part of the life of many people around the world (Kritzinger & Solms, 2010). “There is no argument whatsoever that the proliferation of devices and information are empowering. Technology is today far more democratically available than it was yesterday and less than it will be tomorrow” (Geer, 2015). The Internet evolvement in Sri Lanka is remarkable and most of the Internet related latest technologies were introduced to Sri Lanka sometime even before the other countries in the region (Abeysekara et al., 2012). Both the government and the corporate sectors of Sri Lanka have also incorporated the cyberspace into their operations. Thus, operations of the government and private sector institutions heavily rely on computers and the Internet. However, there are many threats and risks incorporated with the Internet (Riem, 2001). Furthermore, the Internet has led to criminal activities due to private information on it (De Joode, 2011). Hence, there is a risk of misusing and compromising personal data on the Internet (Tierney, 2018).

Social media is an outcome of the Internet platform on which individuals and groups can share their ideas, interests, and views with others. This media is more popular with the development of Internet technology and the attractiveness of web-based applications, enabling many-to-many communication and online sharing. Aral et al., (2013) claim that social media is “fundamentally changing the way we communicate, collaborate, consume, and create”. There is no a common definition for social media. However, Kaplan and Haenlein, (2010) defined social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 which allows the creation and exchange of user-generated content”. Social media comprises the usage of online websites such as Facebook (FB), YouTube, Twitter, LinkedIn and others which are used to reach their community base and to denote a variety of web-based tools that purportedly encourage communication. Similarly it enabled individuals to have access to variety of information sources facilitated by other consumers’ experiences and recommendations (Senecal & Nantel 2004). According to Hampton et al (2011) an average internet user has got over 669 social connections. Facebook statistics shows that they have more than 1.4 billion active users visited the social network on a daily basis and their annual revenue amounted to close to $ 40.65 billion, the majority of which was generated via advertising in 2017. Experts estimated that over 2.95 billion people would have access social networks regularly in 2020.

Then again, social media influences traditional media and has become an alternative to traditional media (Piechota, 2011). Negussie & Ketema (2014) argued that FB is a media for freedom of dialogue and consent to people from different ethnicities, religions, and backgrounds to directly share information without any restrictions. Also, FB election campaign is recognized as a facilitative mode to access political information in several ways. Through FB group activists and ordinary citizens could voice their opposition to the government when denied democracy and suppress their views and voices (Hanson et al., 2010). For example, The USA presidential campaign in 2008 was the first to use the world of YouTube, My Space, FB, and political blogging Internet based for such purposes. By 2010, 22% of Internet users have been using social media network for political activity (Bekafigo et al., 2013). Further, many studies argued that social media could influence people by changing their perception, and attitudes and promote people to think differently. From a political party perspective, social media provides a cost-effective medium to reach-out to large number of users (voters), it provides a rich two way engagement with users (voters) and by its nature creates interaction. Social media also offers a business benefits for political parties, by using social media they could engage with many more users (voters) rather than traditional media, so it means their investment in social media could give greater returns (Warren, 2018).

Oppositely, online political campaigns distorted pluralistic debates and limited voters’ access to the truthful information and ability to make informed decisions compared with ground campaigns. Coordinated dissemination of false and demeaning information presented in digital platforms including traditional media (EU election observation mission in Sri Lanka Presidential election, 2019). Most of the political parties and their candidates use cross-platform electioneering tactics online, with official party pages adjoining third-party sites that frequently served to discredit the opposing candidates. Also these social media political campaigns have been used to social fragmentation and reduce voter’s loyalty towards democratic political parties and candidates. This tread is rising rapidly aiming at variety of targets. These campaigns target personal lifestyle values to engage with variety of cases such as human rights, racist violence, economic justice, environmental protection (Bennett, 2012).

Another key aspect of the use of social media by political parties is that it allows them to influence voters and the way that could vote, this is also known as information operations. Information operations also known as influence operations, includes the collection of tactical information about an adversary as well as the dissemination of propaganda.e.g. fake news in pursuit of a competitive advantage over an opponent. (Waltzman, 2017). In a Sri Lankan context, the influence of social media on Sri Lankan politics has brought new dangers. According to the Prime minister Ranil Wickramasinghe

“Sri Lanka continues to face ‘New dangers’ posed by hate speech, fake news” (Maldives Independent, 2019).

The aim of this study is to investigate “Does the Facebook influence the Sri Lanka Presidential election and the key themes of the posts that dominated social media landscape and user generated interactions related to those posts during the Sri Lankan presidential election 2019.” Also we aimed to examine what shape the influence takes, whether the influence is guided, or evolving freely related to the “Responsibility and Governance” of social media platforms”.

2. OVERVIEW OF SRI LANKA

Sri Lanka is an island located in the Indian Ocean and has a population of around 22 million. Sri Lanka is a country with a strong background of traditional and conservative way of life. The way of life is influenced by many factors: its history of civil war, its Buddhist heritage, the influence from South India, influences from the colonization of the Portuguese, the Dutch and the British. The society living in bigger cities is relatively more open-minded than the ones in the smaller cities. Communal cooperation and harmony rather than conflict and violence are the predominant features of Sri Lankan society.

According to the census 2012, the total population is 21.94 million in Sri Lanka and consists of majority 70.1% Sinhalese, 12,6% Tamils, 9.7% Muslims and remainder are other ethnic groups. The majority of the population is Buddhist at 67 % and the balance consists of Islamists, Hindus, and Christians. It is estimated that around 80% of the total population lives in the rural and semi-urban areas. Sinhala, Tamil, and English are the major languages with 92% of the population that can converse in Sinhala, while 81% can read and write that language. 15% of the population can converse in English while 19% can read and write it. Literacy is fundamentally important to the ability of the user to access information. In Sri Lanka Digital Literacy is 38.7% and IT, Literacy is 28.3%. Mobile phone subscriptions per 100 inhabitants are 103.16, a fixed telephone subscription per 100 inhabitants is 12.49 and broadband subscriptions per 100 inhabitants is 10.45. Household computer ownership is 23.5% while, email usage is at 11%, and Internet usage stands at 21.3%. There is a major difference in ICT readiness among the Rural, Urban and Estate sectors in Sri Lanka. Further, amongst the over 2.3 million users of social media, over 60% of them are male. It is projected that typically they spend around 34 minutes a day on social media (DCS Sri Lanka, 2017).

3. LEGAL ENVIRONMENT AND REGULATORY BODIES IN SRI LANKA

While the right to freedom of expression, speech, and publishing is guaranteed under Article 14(1) (a) of Sri Lanka’s constitution, it is subject to numerous restrictions related to the protection of national security, racial and religious harmony, public order and morality. The Human Right Commissioner of Sri Lanka stated “The Commission recognizes the critical necessity to protect freedom of expression and the right to information as guaranteed by the Constitution of Sri Lanka and Sri Lanka's international human rights obligations”. Further, the Human Right Commission of Sri Lanka (HRCSL) had reiterated the vital need to take legal action against those who were using social media to propagate communal hatred and incite sectarian violence, under applicable laws, in particular under the International Covenant on Civil and Political Rights (ICCPR) Act No. 56 of 2007. Both Sri Lankan laws and English law have common landscapes with the digital media.

There are several legislations passed by the Sri Lanka parliament, namely, Computer Crimes Act (No. 24 of 2007), Payment Devices Frauds Act (No.30 of 2006), Information and Communication

Technology Act (No.27 of 2003) and Electronic Transactions Act (No. 19 of 2006). In addition, the technological framework for electronic signatures and authentication technologies and certificate authority was established in September 2013. The Telecommunications Regulatory Commission (TRC) was established under the Sri Lanka Telecommunications (Amendment) Act, No. 27 of 1996. As the national regulatory agency for telecommunications, the TRC's mandate is to ensure and protect the interests of the public, provision of effective telecommunications and maintain effective competition between commercial telecommunications enterprises.

As a consequence, Sri Lanka Ministry of Mass Media and Information has declared national policy for media freedom and right to access information, to safeguard the right of all citizens to express their views via any media and to receive, provide and gather information required for the proper functioning of society; ensure that the media would not in any manner harm Sri Lanka's National identity and would prevent any person or community from being subject to contempt, insult, disgrace or hate by the media; to facilitate and ensure to all the Sri Lankan citizens the right of access to information.

4. CASE STUDY OF SRI LANKAN PRESIDENTIAL ELECTION, 2019

Future of a country is decided by its voters. Sri Lankan nation has elected a new president in a landmark vote to overcome the challenges they posed due to sluggish economy, increasing political polarisation and security challenges such as Easter bomber attach which killed over 260 people and wounded hundreds more. This was the first Presidential election in Sri Lanka where sitting president, prime minister or opposition leader was not contesting for President. Gotabaya Rajapaksha won this election by defeating record of 35 candidates from across the political spectrum and elected as the 8th president of Sri Lanka.

Overall the presidential election was largely violence-free and a peaceful election which is well managed by the election department. However, the peaceful campaign on the ground contrasted by the few incidents related to the divisive rhetoric, hate speech and disinformation in traditional and social media (EU election observation mission in Sri Lanka Presidential election, 2019). According to the Statement by the European Union election observation mission (EU EOM) "2019 presidential election was largely free of violence and technically well-managed, but that unregulated campaign spending, abuse of state resources and media bias affected the level playing field".

The highest-profile candidates were Gotabaya Rajapaksa and Sajith Premadasa, who both attracted huge crowds at their ground rallies. Other than that they use traditional media, with a heavy presence in paid advertising in television and the print media such as newspapers. In addition they used cross-platform electioneering tactics online, with official party pages adjoining third-party sites that frequently served to discredit the opposing candidates. The volume of hostile commentary and interactions were higher on these third party sites compared to their official sites. However there was a significant gap between the two main candidates and other contestants in the election in terms of the resource allocation. The campaigns done by Janatha Vimukthi Peramuna (JVP) party of Anura Kumara Dissanayaka and National People's party of Mahesh Senanayake were less prominent compared to other two main candidates. The remaining contestants were hardly visible on the traditional media or social media space. The bias coverage by public and private media and unavailability of campaign financial laws created this huge gap in the playing field. (EU election observation mission in Sri Lanka Presidential election, 2019).

Facebook was the main contributor which shaped political narratives and electoral agenda in the social media space. Below table 1 shows the statistics related to the top 4 candidates and their Facebook participation to the 2019 presidential election. According to the table 1 both Gotabaya Rajapaksa and

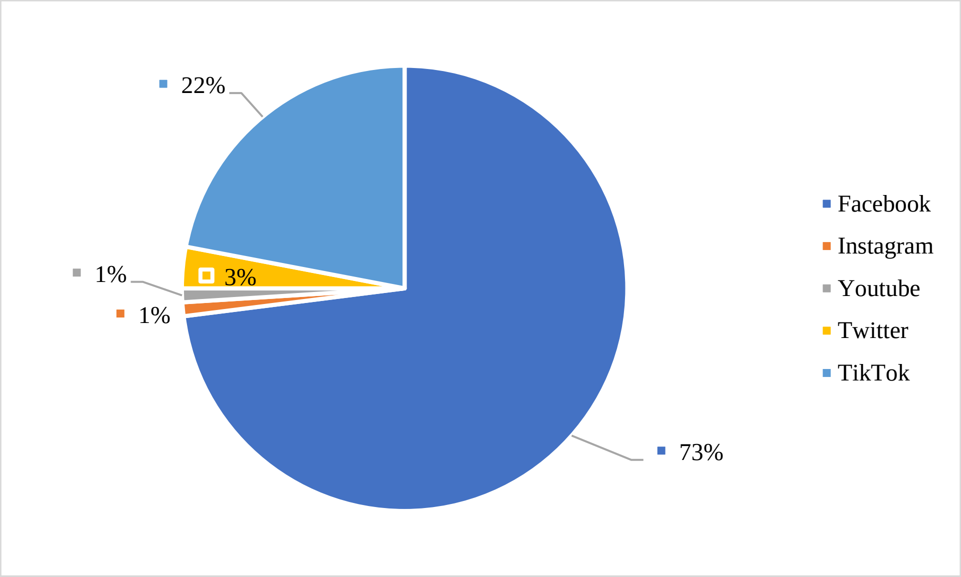
Sajith Premadasa have attracted highest number of followers and attractions to their posts as they did in their ground campaigns. According to the EU election observation mission (2019) The SLPP's online campaign was the most spending online campaign compared to the rivals. (EU election observation mission in Sri Lanka Presidential election, 2019). However we can see that New Democratic Front party has got huge number of average followers (320567) while SLPP has got highest number of average interactions (441). Both of these parties (SLPP and NDF) social media participation seems higher compared to the rivals.

Table 1. Official Facebook pages and their interactions - 2019 presidential election.

Candidate	Party	Official Facebook Page	Number of Post	Average Followers	Average Interactions
Gotabhaya Rajapaksa	Sri Lanka Podujana Peramuna (SLPP)	https://www.facebook.com/PodujanaParty	464	52713	441
Sajith Premadasa	New Democratic Front (NDF)	https://www.facebook.com/UNPofficialpage/	495	320567	391
Anura Kumara Dissanayaka	National People's Power	https://www.facebook.com/nppsrilanka	218	46872	267
Mahesh Senanayake	National People's Party National Peoples Movement	https://www.facebook.com/nationalpeoplesmovement/	209	41233	208

Overall misuse of media created a massive impact on voters' access to the factual information, which affected their ability to making fully informed decision. Coordinated dissemination of outright false and demeaning information presented in both traditional media and online platforms, however Facebook was leading this. Figure 1 shows that most of the social media misuse complains reported against the Facebook with 73%. Also most of the Facebook sites did not adherence to the campaign silence rules.

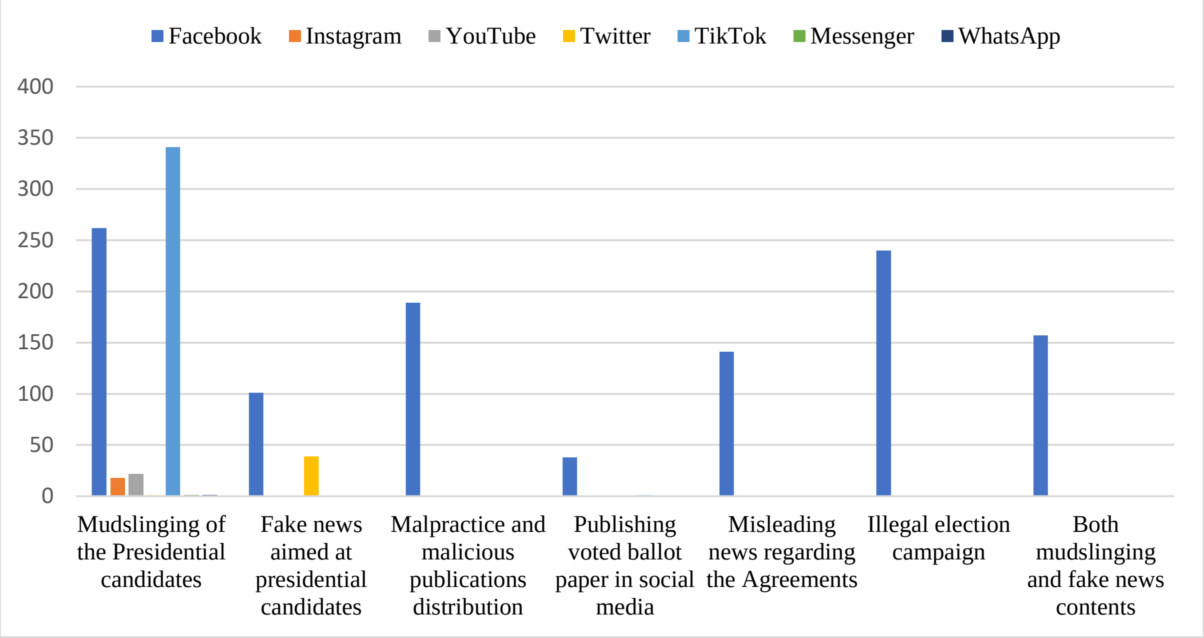
Figure 1. Misuse of social media by type - 2019 presidential election.



Source: ITSSL Social Media Monitoring Report (2019)

According to the social media monitoring report by information technology society (ITSSL) they have got 1593 complains related to the misuse of social media networks. In addition they have observed 240 incidents against campaign silence rules. Below figure 2 shows statistics related to the misuse of social media during the 2019 presidential election.

Figure 2. Misuse of social media for 2019 presidential election.



Source: ITSSL Social Media Monitoring Report (2019)

650 complains related to the mudslinging of the Presidential candidates and about 190 complains of malpractice and malicious publications received. Other than that, there have been 147 complaints of false news exchanges aimed at presidential candidates, and 142 complaints regarding the Millennium Challenge Agreement (MCC). There have also been complaints regarding eight fake social media accounts created using the names of presidential candidates. There are 157 complaints belonging to both mudslinging and fake news (ITSSL Social Media Monitoring Report, 2019).

However, in order to instigate the social media's influence on the Sri Lankan Presidential election 2019 and to examine what shape the influence takes, we performed an independent study and this study collected data from Facebook using CrowdTangle related to the Sri Lankan presidential election 2019. The research question related to the study is "Does the Facebook influence the Sri Lanka Presidential election 2019 and to examine what shape the influence takes, whether the influence is guided, or evolving freely related to the "Responsibility and Governance" of social media platforms". In order to answer the research questions we analysed key themes of the posts that dominated social media landscape and user generated interactions related to those posts during the Sri Lankan presidential election 2019."

The study focussed on collecting data from the four most popular open FB political groups namely "Ape Rata", "JVP Balakotuwa", "Ekayayata kola patata", "Sri AV TV Network". These have been selected to perform the analysis and 175 posts had been selected using purposive sampling technique from October 20-27 in 2019, to analyse using the key themes by the researchers.

Table 2 presents the key themes of the posts and viewers' interactions during October 2019 related to the Sri Lanka presidential election.

Table 2. Type of posts used by the selected FB public groups and followers interactions related to the 2019 presidential election.

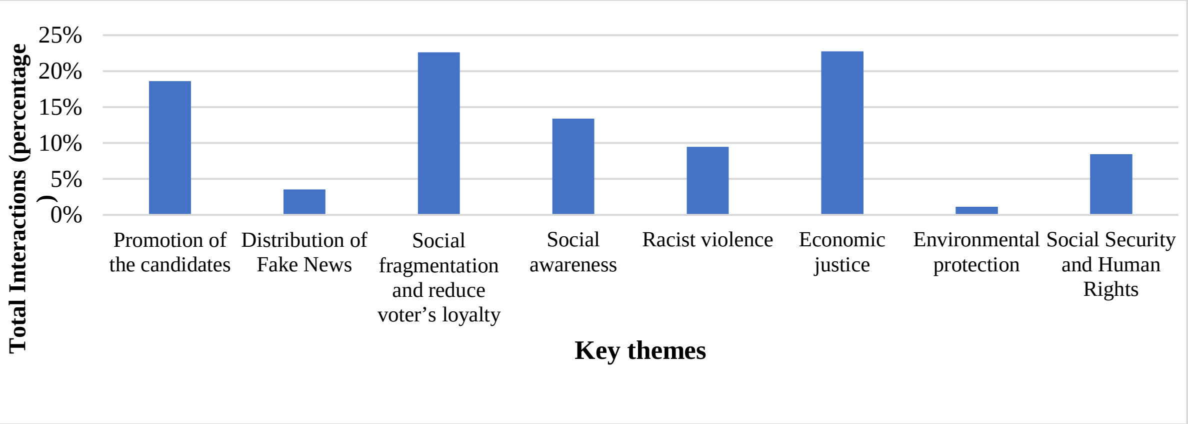
Key Themes	Total Post (Frequency)	Percentage (%)	Number of reactions	Number of Shares	Number of Comments	Total Interactions (Count)*	Total Interactions (%)*	Average Followers per Theme**
Promotion of the candidates	45	26%	6061	4059	817	10937	19%	324814
Distribution of Fake News (false and demeaning information)	16	9%	1041	906	129	2076	4%	144933
Social fragmentation and reduce voter's loyalty	29	17%	5135	7409	747	13291	23%	258282
Social Awareness	33	19%	4506	3114	259	7879	13%	426079
Racist violence	10	6%	2983	2233	365	5581	9%	362531
Economic justice	19	11%	9246	3493	644	13383	23%	659540
Environmental protection	6	3%	339	306	40	685	1%	215091
Social Security and Human Rights	17	10%	3177	1376	427	4980	8%	312524
Total	175	100%	32488	22896	3428	58812	100%	

*Total Interactions: The sum of Reactions, Shares and Comments related to the each theme.
**Average Followers: The sum of Page Likes, Instagram followers, Twitter followers, or sub edit subscribers for all of the matching results.

Table 2 shows that the most of the posts shared related to the following themes: Promotion of the candidate (26%), Social fragmentation and reduce voter's loyalty (17%) and Social Awareness (19%) while Environmental protection (3%) theme is the least post shared theme.

Then we looked at the data to determine the social interactions related to the themes of the posts.

Figure 3. Distribution of social interactions by theme related to the 2019 presidential election.



*Total Interactions: The sum of Reactions, Shares and Comments related to the each theme.

Figure 3 explore the user generated interactions by theme. Total interactions were vary according to the each theme the majority of followers interacted to the 'Social fragmentation & reduce voter's loyalty' and 'Economic justice' themes with each 23% interactions. Nevertheless Environmental protection theme has got least number of social interaction (1%) as shown.

The next stage was to analysis the distribution of Interaction by type related to the 2019 presidential election and the posts, this is shown in Figure 4.

Figure 4: Distribution of viewer's interactions by type related to the 2019 presidential election.

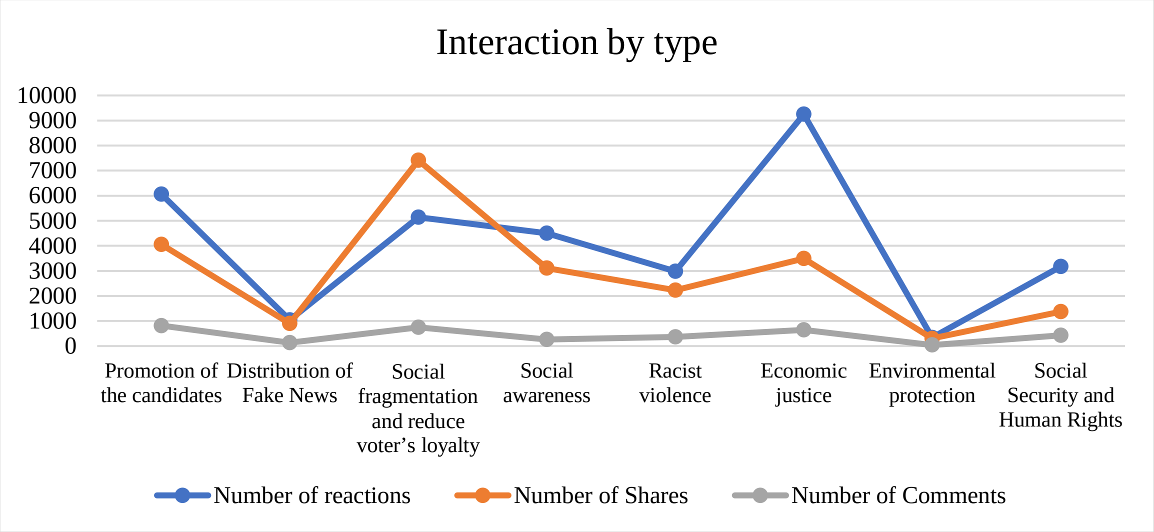


Figure 4 explore the viewer generated interactions by type. It's interesting to see that the most of the viewers reacted to 'Economic Justice' theme while most shared posts related to the 'Social fragmentation and reduce voter's loyalty' theme.

5. DISCUSSION

The analysis revealed what the major themes were in the Sri Lankan presidential election 2019. The analysis identified that the major themes 'Promotion of the candidates' (26%); 'Distribution of Fake News' (9%); 'Social fragmentation and reduce voter's loyalty' (17%); 'Social awareness' (19%); 'Racist violence' (6%); 'Economic justice' (11%); 'Environmental protection' (3%) and 'Social Security and Human Rights' (10%). What was of interest was that number one theme was the 'Promotion of particular candidates' many of the posts were promoting the main candidates, Gotabaya Rajapaksha and Sajith Premadasa were popular. The highest-profile candidates were Gotabaya Rajapaksha and Sajith Premadasa and they had used a sophisticated and heavy social media campaigns (EU election observation mission; 2019) and that seems to influence social media posts related to Promotion of the candidates theme. Another interesting outcome was that 'Distribution of Fake News' that only reflected 9% of the themes in the posts, the authors had expected this figure to much higher.

Another important outcome of the analysis was the disclosure of social interactions related to the themes in the posts. Mostly interacted themes by the followers were Social fragmentation and reduce voter's loyalty (23%) closely followed by Economic justice (23%). According to the EU election observation mission in Sri Lanka Presidential election (2019) most of the candidates used cross-platform electioneering tactics online, with official party pages adjoining third-party sites that frequently served to discredit the rival. This may lead to influence posts and social interactions related to the "Social fragmentation and reduce voter's loyalty theme mostly". This may possibly a guided influence by third party sites operated by the political parties. Also Sri Lankan nation has struggled by the challenges they posed due to sluggish economy, increasing political polarisation and security challenges. Because of that National security and Economic Justice were a prominent themes in the election campaigns. From our analysis also we can confirm that the economic justice theme was a prominent theme in terms of the posts and social interactions.

6. CONCLUSION

The paper has shown that social medial has the ability to generate discussion and debate, the authors showed that the most popular FB posting was to promote particular presidential candidates, and it may possibly a guided influence. The most interacted themes were "Social fragmentation and reduce voter's loyalty theme" and "economic justice". When we analyse the user interactions we can see both guided and freely evolving interactions related to the Sri Lankan presidential election 2019. Also issues such as Fake News were not a major issue. The authors have shown that Facebook did have guided and freely evolving influences on the Sri Lankan presidential election of 2019.

Findings of the case study concluded that there is a significant impact on politics campaign and level of user's interaction of social media. Further, it was established that the misuse of social media has becoming a major challenge for future free and fair elections. Therefore, necessitates the need for a national social media policy that focuses on election as key stakeholders in the registered political parties.

REFERENCES

Abeysekara, E.R. D., Liyanarachchi, M., Wijesinghe, W.S., Jayarathne, N., Wijethunga, M.T.N., Perera, M. (2012). Cyber Terrorism; is Sri Lanka Ready, General Sir John Kotelawala Defence University, Sri Lanka.

Aral, S., Dellarocas, C., & Godes, D., (2013). Introduction to the special issue – social media and business transformation: a framework for research, Information Systems Research, 24 (1), 3-13.

Bandarage, A., (2018). Voiding 'religious' violence in Sri Lanka, Retrieved March 30, 2019, from http://www.atimes.com/avoiding-religious-violence-sri-lanka/

Bekafigo M.A., Cohen D.T., Gainous J. and Wagner K.M. (2013). State Parties 2.0: Facebook, Campaigns, and Elections. The International Journal of Technology, Knowledge, and Society, 9(1), 99-112.

Bennett, W. L. (2012). The Personalization of Politics: Political Identity, Social Media, and Changing Patterns of Participation. The ANNALS of the American Academy of Political and Social Science, 644(1), 20–39. https://doi.org/10.1177/0002716212451428

De Joode, A. (2011). Effective corporate security and cybercrime. Network Security, 2011(9), 16-18.

Department of Census and Statistics Sri Lanka (DCS Sri Lanka), Computer Literacy Statistics – 2017 (First six months), (2017). Retrieved March 12, 2019, from http://www.statistics.gov.lk/education/ComputerLiteracy/ComputerLiteracy-2017Q1-Q2-final.pdf

Geer, D., 2015, 'Six key areas of investment for the science of cyber security', The Futurist, no. 1, p. 10, Retrieved May 4, 2019, from http://ezproxy.deakin.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.428176483&authtype=sso&custid=deakin&site=eds-live&scope=site.

Goodman, J., Wennerstrom, A., & Springgate, B. F. (2011). Participatory and social media to engage youth: from the Obama campaign to public health practice. Ethnicity & disease, 21(3 Suppl 1), S1-99.

Hanson G., Haridakis P.M., Cunningham A.W., Sharma R. and Ponder J.D. (2010). The 2008 Presidential Campaign: Political Cynicism in the Age of Facebook, MySpace, and YouTube. Journal of mass communication and society, 13(5), 584-607.

ITSSL Social Media Monitoring Report Presidential Election 16 November 2019 Sri Lanka. (n.d.).

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

Kritzinger, E., & Padayachee, K. (2013, September). Engendering an e-safety awareness culture within the South African context. In 2013 Africon (pp. 1-5). IEEE.

Lanka, Gunaratna, G., Sri Lanka News Paper by LankaPage.com (LLC) - Latest Hot News from Sri. "Sri Lanka: Muslim shops attacked in Buddhist-Muslim clash in Sri Lanka\'s East, security strengthened in Ampara", Retrieved November 10, 2018, from http://www.colombopage.com.

Meti V, Khandoba PK, Guru MC (2015) Social Media for Political Mobilization in India: A Study. J Mass Communicat Journalism 5:275. doi:10.4172/2165-7912.1000275

Maldives Independent, (2019). Sri Lanka continues to face ‘new dangers’ posed by hate speech, fake news – PM, Retrieved October 4, 2019, from http://www.adaderana.lk/news/57499/sri-lanka-continues-to-face-new-dangers-posed-by-hate-speech-fake-news-pm

Negussie N., and Ketema G., (2014). Relationship between FB Practice and Academic Performance of University Students, Asian Journal of Humanities and Social Sciences (AJHSS), 2(2), 31-37.

Piechota, G., (2011). Application of social media in political communication of local leaders in election processes (on the example of Facebook’s use by mayors of voivodship cities in Poland in the 2010 election campaign).

Riem, A., (2001). Cybercrimes of the 21st Century. Computer Fraud & Security, 2001(4), 12-15.

Statista. 2018. Facebook: global daily active users 2018 | Statistic. [ONLINE] Available at: https://www.statista.com/statistics/346167/facebook-global-dau/. [Accessed 1 July 2018].

Statement by the Spokesperson following the 2019 Presidential Elections in Sri Lanka. (2019, November 18). Retrieved from https://eeas.europa.eu/headquarters/headquarters-homepage/70576/statement-spokesperson-following-2019-presidential-elections-sri-lanka_en

Tierney, M. (2018). # TerroristFinancing: An Examination of Terrorism Financing via the Internet. International Journal of Cyber Warfare and Terrorism (IJCWT), 8(1), 1-11.

Thivanka, M., (2016). Young People and Social Media Addiction’, UNDP, Retrieved March 12, 2019, from http://www.lk.undp.org/content/srilanka/en/home/Blog/2016/10/21/Young-People-and-Social-Media-Addiction-.html

Thuseethan, S., and Vasanthapriyan, S., (2015). Social Media as a New Trend in Sri Lankan Digital Journalism: A Study. IUP Journal of Information Technology, 11(1).

Waltzman, R (2017), The Weaponization of Information, RAND. https://www.rand.org/content/dam/rand/pubs/testimonies/CT400/CT473/RAND_CT473.pdf, accessed 10/11/18.

Warren, M. (2018). Political Cyber Operations, Conference Proceedings of Australian Cyber Warfare Conference 2018, ISBN 978-0-6484570-0-8.

5. Management of Cybercrime: Where to From Here?

A FLOATING CONJECTURE:
IDENTIFICATION THROUGH FACIAL RECOGNITION

Wade Robison

Rochester Institute of Technology (United States)

wlrgsh@rit.edu

ABSTRACT

Using facial recognition as evidence in trials is part of a larger pattern of using suspect marks of identification to pinpoint those responsible for crimes. Those accused of crimes have been convicted on the basis of bite marks, hair samples, and fingerprints, and though those in law enforcement would no doubt want a mode of identification that ensures that those accused are in fact those who committed the crime being prosecuted, facial recognition technology fails to add any certainty to the modes of identification that are unfortunately now used and fail to sort out the guilty from the innocent.

We might well think that technological advances will eventually allow us to find us a sure proof mode of identification, but facial recognition technology is nowhere near the level of certainty of identification that we need to prove someone guilty beyond a reasonable doubt and there are serious doubts that it ever will be.

KEYWORDS: facial recognition, features comparisons, fingerprints, prosecution.

1. INTRODUCTION

Facial recognition technology is now being used by law enforcement and by prosecutors to identify and help convict criminal suspects. It is standard now whenever there is a crime to look at what the security cameras captured. They seem to be ubiquitous, and it is a rare article or broadcast on a criminal investigation in the United States that does not give a nod to how helpful a security camera was to identifying a suspect and, indeed, recording the criminal act. There is little doubt that they have been a huge help for law enforcement and that it is likely that we will see more and more cameras deployed throughout our city's streets, in businesses, and in homes.

That technology combined with the tracking information available via cell phones creates an alarming capacity on the part of governments to know exactly where someone is and what they are doing—and the subsequent concern about constant surveillance and control. That concern will no doubt take a back seat to the demands of law enforcement for its use in identifying possible perpetrators. It is too useful for those in law enforcement to be persuaded that a potential alarming capacity should curtail the deployment of surveillance cameras. Such cameras promise to be even more useful as the technology evolves. So deployment is going to continue, with greater and greater capacities to monitor citizens' activities.

That is disconcerting, but even more disconcerting is the use of facial recognition as evidence by prosecutors. Its use in criminal trials in the United States is part of a larger pattern of using questionable rules of skill that tell us how to identify a suspect based on fingerprints, bite marks, and other supposedly identifying information.

We will look at the rules of skill that are at issue in what is called features comparisons before examining how such features are used, and with what success, in prosecuting cases. We will then turn to facial recognition to examine its advantages and disadvantages in prosecution.

2. RULES OF SKILL

A rule of skill tells us how to achieve a particular end: to bake a cake, do such-and-such; to buttress a girder, do so-and-so. They are the tools of the trade, so to speak, for any profession, and they display a wide variety of functions. There are rules that tell us what things are—what symptoms go with which disease, what crosshair signatures are, what shape and position goes with which human organ, and on and on. There are rules that tell us how to do something—how to extract a tooth, how to use a Japanese saw, how to use Matlab, and on and on. There are rules that prescribe the procedures to follow—in writing a valid will or ensuring a fair trial, in minimizing the risks of infection, and on and on.

We are all familiar with rules of skill. We learn them early on as we learn to count or correct our pronunciation so we can be understood. We know as well what happens when we fail to follow the relevant rules. We open ourselves to criticism and to failure. We learn early on that games must be played in certain ways and not others, for instance, and we learn as well the limits of rules in ensuring that individuals do what they ought to do. Cakes do not always turn out the way they should. Girders are sometimes not properly buttressed. Surgeons amputate the wrong limb. Police officers stop a vehicle merely because the driver is African-American (Wang, 2017).

Rules of skill set a sequenced, coherent normative order to what we do. What matters for making fudge, for instance, is that add vanilla after we have melted in the chocolate for fudge, not before, and that we pay attention only to what is required for making fudge. Scratching one’s head while thinking about where the chocolate might be may occur while making fudge, but it not part of what it is to make fudge. What is required to make fudge is a coherent series of steps, and anything else that may occur while traversing those steps is not relevant. That is why the rule is normative: it tells us what we ought to do to make fudge and, in doing that, tells us what we ought to ignore as not part of the sequenced, coherent, normative order

Rules of skill are no different in that way than, say, the rules of logic. Valid argument forms, for instance, tell us how we ought to reason deductively. When we provide someone with the form for modus ponens—if p, then q and p, therefore q—we are providing them with a sequence of steps that they ought to follow if they are not to risk reasoning from truth to falsity. The same is true for any rule of calculation. That is why tellers at the grocery store cannot just give us any old handful of change. They are constrained by the rules that tell us how we ought to add and subtract. If they fail to give us the correct change, we may properly tell them that they have made a mistake, done something they ought not to do. We are telling them, effectively, that they are not doing what reason tells all of us we ought to do. When we subtract 76 cents from a dollar, we do not, with good reason, get 21 cents. The norm we have failed to satisfy is one of reason

3. FLOATING CONJECTURE

Some rules of skill are floating conjectures, without the sorts of evidential backing needed to make them reliable. Anyone who reads mysteries or watches crime shows knows that central to a crime's solution is what can be found at the crime scene—'DNA, hair, latent fingerprints, firearms and spent ammunition, toolmarks and bitemarks, shoeprints and tire tracks, and handwriting' (Report, 2016). Detectives hunt for samples at the scene that can then be compared to samples from a suspect, and they remind everyone not to touch anything at the scene so that when they dust for fingerprints, for instance, their findings will not have been contaminated. They are hunting for fingerprints or hair or something else left by whoever committed the scene crime.

Experts compare the features of what is found at the crime scene with the features of the relevant sample from a suspect, and if there is a match, they have significant evidence that the suspect is the criminal. The relevant rule of skill will vary depending upon what feature is being examined, but the general formula is the same for all features: if this sample looks like that sample, they are from the same person.

We can already see, from the vagueness of that formulation, how easy it must be for errors to enter into any identification. 'Looks like' requires someone to do the looking, and so one source of error is that the person doing the looking may make mistakes. Another source of error concerns in what way or ways the items being inspected look alike—and unalike. My siblings and I have a family resemblance and so look alike in certain ways, but not in others. What feature or features should count or count the most—the shape of our ears, their position relevant to our skulls, our noses, the shape of our nostrils, our projecting or receding chins, or what? What is compared with what requires a judgment based on evidence of which features, if any, are telling. And so, clearly, another potential source of error is the misidentification of what ought to count when comparing samples, and then yet another is the judgment that two samples are identical in regard to what has been judged to be the telling feature.

We can get a sense of how problematic the relevant rules of skill are by examining the track record of identifications for bite marks and hairs and fingerprints. We have a standard we can use in DNA.

DNA analysts don't tell jurors that a suspect is a match. Instead, they use percentages. Because we know the frequency with which specific DNA markers are distributed across the population, analysts can calculate the odds that anyone other than the suspect was the source of the DNA in question (Balko, 2020).

We have a basis for comparison with DNA. Since we know of any specific DNA marker how many there are in the population, we can tell how probable it is that one DNA marker is like another.

But we should emphasize that DNA is not the gold standard we may think it is. For one thing, it is not foolproof. A man who received a bone marrow transplant ended up with the DNA of the donor. He had both his own and his donor's DNA, and, somewhat to his chagrin, we must assume, 'all of the DNA in his semen belonged to his donor' (Murphy, 2019). We do not have any idea how often that happens, but once is enough to make DNA testing less than the gold standard.

There are also problems with how the testing is done. Those doing it can make mistakes, obviously, sending the police on the sort of fool's errand the German police engaged in for sixteen years after finding 'traces of identical female DNA…at 40 crime scenes across southern

Germany and Austria,' including six murders (DNA, 2009). The police used Q-tips that they had purchased from a store rather than sanitized ones, and the Q-tips had been contaminated by a woman working at a Q-tip factory in Bavaria.

So using DNA to tie a particular suspect to a crime scene is not without its problems (Otterman, 2019). It is, however, determinative enough that we can assess the validity and reliability of comparing hair and bite marks, for example, by determining if using DNA gives us the results we got in previous cases comparing other features from the crime scene.

4. HAIR AND BITE MARK

Santae Tribble was 17 when he was arrested for murder and convicted based on a comparison of his hair with hair found at the scene of the crime. As he put it, the experts said that the sample 'matched my hair in all microscope characteristics.' As the prosecutor said in summing up the evidence,'There is one chance…in 10 million that it could [be] someone else's hair' (Hsu, 2012a). Later analysis showed that of the thirteen hairs in question, nine were from one person, three from different individuals, and one from a dog. None belonged to Tribble (Oliver, 2017). He was freed (Hsu, 2012b) and then exonerated (Hsu, 2012c), but he spent 26 years in jail because of the mistaken judgment that his hair matched the samples found at the crime scene.

'Such is the true state of hair microscopy,' the lawyer representing Tribble said, that '[t]wo FBI-trained analysts, James Hilverda and Harold Deadman, could not even distinguish human hairs from canine hairs.' Researchers showed in 1974 that 'visual comparisons are so subjective that different analysts can reach different conclusions about the same hair. The FBI acknowledged in 1984 that such analysis cannot positively determine that a hair found at a crime scene belongs to one particular person' (Hsu, 2012a).

In 2012, the FBI and Department of Justice began a review of over 3000 'criminal cases involving microscopic hair analysis.' They found that 'that FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where that testimony was used to inculpate a defendant at trial' (Report 2016). So 19 out of every 20 defendants were falsely incriminated by FBI experts. It is difficult to imagine a less reliable way to determine if someone has committed a crime. Flipping a coin would give better than a 95% failure rate

Another example of a floating conjecture in forensic science concerns bite marks. Keith Harward 'narrowly escaped the death penalty,' but spent 33 years in prison after being convicted of rape and murder on the basis of six forensic dentists testifying that the bite marks on the rape victim's legs were his. DNA evidence showed that he was innocent and that a fellow sailor, Jerry Crotty, was responsible. Harward is one of at least 25 individuals 'to have been wrongfully convicted or indicted based at least in part on bite mark evidence' (Innocence, 2019). He is now free, but he says to those who tell him he is a free man, 'I will never be free of this…I spent more than half my life in prison behind the opinions and expert egos of two odontologists

Harward noted that there was 'a death-penalty case in Pennsylvania where the judge is going to allow bite-mark evidence' (Oliver, 2017). Indeed, 'bite-mark analysis…has yet to be disallowed by any courtroom in the country' (Balko, 2020)

The 2016 Report to the President pointed out that a '2010 study of experimentally created bitemarks…found that skin deformation distorts bitemarks so substantially and so variably that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter.' In fact, evidence 'showed a disturbing lack of consistency in the way that forensic odontologists go about analyzing bitemarks, including even on deciding whether there was sufficient evidence to determine whether a photographed bitemark was a human bitemark' (Report, 2016). That bite mark evidence still finds its way into court cases is a sad commentary on the failure of American judicial system to come to grips with such forensic floaters.

5. FINGERPRINT

On March 11, 2004, ten bombs killed 192 passengers on trains in Madrid and injured more than 1400, according to initial reports (Sciolino, 2004). The Spanish authorities found a fingerprint on a bag of detonators and forwarded it to the FBI to see if it could find a match in its database. The FBI's Integrated Automated Fingerprint Identification System (IAFIS) 'generated a list of 20 candidate prints.' None was a perfect match, but IAFIS also lists close matches, and one belonged to Brandon Mayfield, a lawyer in Oregon. The FBI 'immediately opened an intensive investigation of Mayfield, including 24-hour surveillance…and physical searches' of his law office and residence. When news somehow broke that an American was a suspect in the bombing, the FBI detained Mayfield on May 6th because they were 'absolutely confident' that Mayfield's fingerprint was on the detonator bags. They kept him in solitary confinement 'for up to 22 hours per day' (Office, 2006)

The fingerprint from Spain was examined by a fingerprint specialist in the FBI who verified it as belonging to Mayfield. That judgment was confirmed by a second FBI fingerprint specialist and by the fingerprint unit chief, all of whom agreed it was Mayfield's. That decision was confirmed by a court-appointed specialist (Office, 2006). Four fingerprint experts fingered Mayfield, as it were.

The defense attorney's own expert confirmed the judgment of the FBI experts and later said, 'No time before in history have there ever been two fingerprints with fifteen minutiae that were not the same person' (Bharara, 2020). So there was good reason for the FBI's confidence.

The Spanish authorities identified the person whose fingerprint was on the bag of detonators, and it was not Mayfield. As it turned out, further analysis of the fingerprints showed that Mayfield's was not identical to the one found in Spain, but what is of importance here is that specialists in fingerprint identification judged that it was and that they had absolute confidence in their judgment. The Mayfield case is a dramatic example of why such judgments cannot be relied upon and should not be relied on, especially in criminal cases where the stakes are high. We must have proof beyond a reasonable doubt, and the Mayfield case puts in doubt reliance on fingerprints comparisons. The case has become a classic example of how misidentification of a sample can mislead investigators, taking them off the scent of the perpetrator onto the scent of an innocent person who can be badly harmed by the mistake.

6. RELIABILITY

As it turns out, feature comparisons are not very reliable at all. The 2016 Report to the President on forensic science stated,

Reviews by the National Institute of Justice and others have found that DNA testing during the course of investigations has cleared tens of thousands of suspects and that DNA-based re-examination of past cases has led so far to the exonerations of 342 defendants (Report, 2016).

The failure of such feature comparisons as hair samples and fingerprints is illustrated by the number of exonerations each year as old cases are reexamined. 'More than 150 men and women were exonerated in 2018,' having 'spent more than 1,600 years in prison' for crimes they did not commit. The Innocence Project exonerated more than 350 individuals, and in 45% of the cases, those individuals were convicted because of a failure of feature comparisons combined with misleading testimony from experts who ensured juries and judges that they were sure within a 'reasonable degree of scientific certainty.' But the 'experts…used exaggerated statistical claims to bolster unscientific assertions.' That is a phrase that a jury is likely to believe gives great weight to the evidence but has no scientific validity (Innocence, 2019).

The 2016 Report quotes a judge about testimony from an expert that 'markings on certain bullets were unique to a gun recovered from a defendant's apartment':

As matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual's foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.

The Report adds,

> In science, assertions that a metrological method is more accurate than has been empirically demonstrated are rightly regarded as mere speculation, not valid conclusions that merit credence (Report, 2016).
>
> The need for evidence and testimony based on evidence is nicely put by U.S. District Judge John Potter, in 'an early case on the use of DNA analysis,' U.S. v. Yee (1991):
>
> Without the probability assessment, the jury does not know what to make of the fact that the patterns match: the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa (Report, 2016).

That, in a nutshell, is the problem with the comparison of features: 'There is no way to calculate a margin for error.' Unlike DNA testing, where we know how probable it is that one marker is like another because we know of any specific DNA marker how many there are in the population, comparing a hair found at the scene of a crime to one of a suspect can at best exclude it—if, say, the one is blond and other one black. Depending on how many features of a hair sample are compared, it may not exclude many at all.

'[T]he FBI agent testified at trial that the hair from the stocking matched Tribble's "in all microscopic characteristics",' (Hsu, 2012c), but the FBI expert, Hilverda, 'recorded in his lab notes that he had measured only three characteristics of the hair…—it was black, it was a human head hair, and it was from an African American' (Hsu, 2012a). We can presume that under a microscope more than three characteristics are discernible and that countless African Americans have black hair. So the FBI agent's testimony was misleading, to say the least, and Tribble spent 26 years in jail for being an African American.

Here we know that the three characteristics are hardly unique, but no matter how many features are found, we would have no idea how many hairs in the world share those features. The hairs may even match in all discernible ways, but with no idea how many different hairs of different individuals match, we have no idea whether the match is unique or only to be expected since millions could match.

The same is true for any comparison. We cannot know we have a unique match with marks on shell casings, or bite marks, or pry marks on a door because there is no way of knowing how many different guns or teeth or crowbars might, under the right conditions, produce identical marks (Balko, 2020).

We have floaters in forensic science. Because of them, some individuals were executed. Floaters can have grievous consequences, and those professionals who testified to their valid application in particular cases were wrong.

7. DEGREES OF CONFIDENC

It is no surprise that people can be confident about something or find something plausible or even obvious when the facts do not warrant confidence. We all have beliefs which range from the implausible to certain, and to assess them, we must rely not on how we feel about them, but on what the facts support. A feeling that a belief is certain is no guarantee it is true. If we were to construct an argument for the FBI experts' judgment that Tribble's hair was found at the scene of the crime, it would include the following premises regarding the degree of confidence the FBI experts had in their judgment implicating Tribble:

- There is one chance in 10 million that the hair is not Tribble's.
- We have only been mistaken 19 times out of 20 in making such judgments.
- So we experts are absolutely confident the hair is Tribble's.

We have terms of criticism for beliefs, and the one most relevant here concerns the degree of likelihood that the belief is true. We ought to be more or less confident in our beliefs in accordance with the quality of our evidence, and in this case we ought to lack any confidence at all.

A birder trying to identify a particular warbler will follow the usual methodology, making a judgment based on the bird's size, flight pattern, song, and other distinctive characteristics. The birder ought to be more or less confident depending upon how many identifying marks are discernible and how easily they can be discerned. 'It's a Palm Warbler' is a quite different judgment than 'Well, could be a Palm Warbler,' and they mark how many identifying marks the birder was able to discern and with what degree of certainty. Does the bird have a distinctive yellow eyebrow? A chestnut-colored crown (Sibley, 2000)? Catching a glimpse of something chestnut-colored is very different from being able to observe the bird for some period of time.

The methodology for identifying birds is not perfect. Experts can use it and still make mistakes. But when used correctly by a competent birder, the success rate is significantly higher than 5%. A 95% failure rate tells us that the methodology is unreliable and that having a second and third expert check another's judgment using the same methodology will not provide us with any more evidence for the truth of the belief.

If the methodology is faulty, it does not matter how experienced an expert may be, or how many experts chime in. An unreliable methodology will lead to unreliable results. As the President's Report of 2016 put it, Without appropriate estimates of accuracy [and error rates], an examiner's statement that two samples are similar–or even indistinguishable–is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.

As the Report notes,

> Nothing–not training, personal experience, nor professional practices–can substitute for adequate empirical demonstration of accuracy (Report, 2016).

The rules of skill that supposedly gave credence to 'expert' testimony are all recipes for mistakes. In comparing Mayfield's fingerprint with the fingerprint from the bag of detonators, FBI fingerprint specialists found ten points of similarity, and the defense's expert found fifteen. 'Points' is a technical term here. They occur where individual ridges end or split, and the similarities were 'the relative location of the points, the orientation of the ridges coming into the points, and the number of intervening ridges between the points.' The Office of Inspector General's Review of the FBI's Handling of the Brandon Mayfield Case points out that there is no research on how frequently such similar constellations of points occur in different individuals, but that 'anecdotal reports suggest that this degree of similarity…is an extremely unusual circumstance' (Office, 2019).

The bottom line, however, is that the experts were relying on a rule of skill that told them that if there are so many points of comparison between two fingerprints, they can have 'absolute confidence' that the two were made by the same person when they have no way to gauge a margin of error. We have no idea how often such a constellation of points occurs among all the fingers in the world, and without that information, we can only use a particular constellation of points as a way of excluding some possible suspects. We cannot pinpoint a suspect because we have no idea how many others share the relevant constellation. So the rule of skill the experts used is a floater, a recipe that provides no justification for any confidence at all in its outcome.

The other floaters are no better supported. They all depend on rules of skill that tell supposed experts that if they have such-and-such a configuration in two samples—of markings on a bullet, of the impression of teeth marks, of the details of hair—they can be absolutely confident that the samples came from the same firearm, or the same mouth, or the same head of hair. Such confidence is not responsive to reality, but reflects an unwarranted judgment about the reliability of a faulty rule of skill (National Research Council, 2009)

It is not just experts who make mistaken judgments about feature similarities. Eye-witness identifications are standard and are remarkably unreliable. A witness or the victim to a mugging gets a glance at someone's face and then identifies the defendant when asked by a prosecutor in the courtroom to point out the person responsible for the crime, but 'inaccurate eyewitness identifications…were introduced as evidence in over 70 percent of the more than 360 cases that the Innocence Project… proved were wrongful convictions' (Rakoff, 2019). Amateurs are no better than experts, that is.

8. FACIAL RECOGNITION

The use of facial recognition technology only adds to the floaters, with additional problems. The history of floating conjectures—fingerprints, bite marks, and so on—brings out most of the problems that plague using facial recognition for identifying and prosecuting suspects

The main problems are the ones that plagued the FBI when it misidentified Brandon Mayfield. His fingerprint was not a perfect match, but one of twenty that the FBI algorithm picked out from its massive data base as closely similar. The algorithm did not get a direct hit, but twenty possibilities, that is, and the failure of the algorithm to provide an exact match meant that judgment calls were necessary to determine which of the twenty, if any, was the most likely match. But without knowing how many individuals have fingerprints that fall within that range of possibilities, there is no way of knowing that any one individual has been properly identified. The most that can be said is that those individuals whose fingerprints are not within that range are not suspects.

This summary of the main problems captures the essence of what is wrong with using facial identification: the failure to zero in on exactly one person by comparing features requires judgment calls with no way of knowing the likelihood of one's getting it right. But this summary also obscures just how difficult a facial features judgment is.

We know that for fingerprints the likelihood of finding an exact match is exceedingly small. No matter how detailed the FBI's sample may be, it is being compared to a sample from a crime scene. It is highly unlikely that the two prints are going to match exactly. Smudging, a lighter touch here and a heavier touch there, a twist as one lets go, degradation through exposure to contaminants—all sorts of things can get in the way of a clear print. The consequence is that there are always gaps that need to be filled, and where there are gaps requiring judgment calls on the part of those doing the examination, we have an opening, a gap, that is, for mistakes

Facial identification also faces that issue, so to speak. No two facial images of a person are any more likely to be identical in all respects than two fingerprints of a person. What feature or features should count or count the most? That is the first decision to be made, and it is not at all clear what to choose to minimize mis-identification. A head turned slightly away, the beginnings of a smile, an irritated expression, a new hairstyle that covers, or uncovers, parts of one's face—all sorts of things can get in the way of a perfect match

What counts cannot be the whole face, presumably, because any two images are almost certainly going to vary because of the angles from which they are taken or the direction a person is facing or the quality of the image itself. What counts cannot be anything that changes when one's facial expression changes. Just imagine how different a person's face can look when the person is smiling, frowning, angry, grinning from ear to ear, disgusted, sad, pouting, eye-rolling, surprised, and so on.

Whatever is chosen as the telling feature or features must be constant, invariable despite different camera angles, different positions of the person's head, or different emotional expressions. What is telling must not alter no matter what other differences there are. Human faces do share some relatively constant characteristics. If you want to draw a face, you need to start with an oval, divide it vertically and horizontally in half, place the eyes on the horizontal line on each side of the vertical line, and so on with the nose and lips and ears and other features taking their usual places. But although those features are relatively constant, they are not always so, and in any event, they are too general to allow anyone to zero in on any one face using those constants. It is unclear what other feature or features would provide the detail and constancy needed to make an identification. It is unclear, that is, how we are to complete the first step of determining what is telling.

In any event, however that first determination is made, we have another problem. Whatever the telling feature may be, it is not going to distinguish between identical twins or, presumably, doppelgängers. Identical twins look alike, obviously, and doppelgängers look enough alike that, in the best—or the worst of cases for an investigator or prosecutor—they may as well be identical twins. We know of cases where identical twins were separated at birth, neither knowing of the other, only to discover one another years later (Paparella et al., 2018). What we do not know is how many identical twins have been separated and have never found out that they were identical twins. The number affects how likely it is that an identification based on facial recognition is correct, but without that information, we cannot be at all sure what the likelihood is that we have a match.

And there are more than a few individuals who look alike without being so much alike we would call them doppelgängers. I have been mistaken countless times for the actor in the Halloween series (sans costume), once by a flight attendant who said, 'Oh!' when she saw me, asked why I was sitting in the very back of the plane, told me she would make sure I was not disturbed when I said, 'For peace and quiet,' and then, as I left, having presumably looking up my name, commiserated with me for not being famous. I was not quick thinking enough to tell her that I always travel under my real name

So even if a determination can be made of what counts as a tell, we are no better off than we were with the other floating conjectures. We can only exclude some and not pinpoint any particular person. That problem ought to be sufficient to rule out facial recognition as definitive to prove guilt beyond a shadow of a doubt. As we quoted above on the reliability of bullet markings, 'This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.

But facial recognition faces other problems. One I have not mentioned before, but will occur regardless of what features are being compared—police misconduct. With Photoshop and other editing software, it is easy to manipulate images to fill in the gaps that will occur when comparing one image with another. We already have evidence that police investigators have doctored photos to increase the likelihood of a hit and so the chances of an arrest. 'Some investigators,' it has been reported, 'edited the photos in hopes of revealing more matches, including swapping out facial features, blurring or combining parts of photos and pasting in images of other people's lips or eyes' (Harwell, 2019a)

Facial recognition software faces yet another problem. It is biassed. The algorithms used misidentify asians and blacks far more often than whites, females far more often than males, and native Americans most of all. 'Middle-aged white men generally benefitted from the highest accuracy rates.'

The National Institute of Standards and Technology, the federal laboratory known as NIST that develops standards for new technology, found 'empirical evidence' that most of the facial-recognition algorithms exhibit 'demographic differentials' that can worsen their accuracy based on a person's age, gender or race (Harwell, 2019b).

We have a situation rather like the one we faced when airbags were first introduced. The engineers chose as their norm, the one best protected by the exploding airbag, five-foot-nine males weighing 170 pounds, a choice that best protected the 50th percentile of men and the 95th percentile of women. It is not a far reach to wonder whether the sex of those designing the airbags mattered (Robison, 2016)

Just so, it is not a far reach to wonder about the race and sex of those designing facial recognition software. The problem is similar to the soap dispenser that fails to recognize any hands but those of whites (Hale, 2017). Clearly those who designed it failed to test it across the range of diverse hands that it would need to recognize.

The point is that the algorithms experts use to help fill in the gaps are a function of the vast amount of data now available to be mined for accurate assessments, but that data captures our biasses as well. 'In 2015, for example, the Google Photos app was caught labeling African-Americans as "gorillas"' (Metz, 2019). Such mistakes can be corrected, but that misses the point. They can permeate the algorithms used in facial recognition, and until we find such mistakes, we have no idea how frequently they occur and how they can bias the results. We are in the same position as those who use fingerprints, bite marks, bullet markings and other features. Without knowing how often we will find the same bullet markings in all the guns in the world, the best we can do is to exclude some, leaving however large an unknown number suspect. Just so, because we have no idea how often such mistakes in algorithms occur, the best we can do is exclude some individuals, perhaps leaving an impossibly large number in the pool of suspects.

9. CONCLUSION

The bottom line is that facial recognition cannot be any more definitive in establishing guilt or innocence than fingerprints or any other feature. The rule of skill experts must use regarding facial recognition floats, without the sorts of evidential backing needed to make it reliable. It shares the two failings we identified for the other features being compared:

1. We do not know how many share the particular features that are taken to be telling. We do not know, for instance, how many individuals have fifteen or ten identical points in a fingerprint or how many have such-and-such a distance between the centers of their eyes.
2. The gaps that are found between any two samples must be filled by the judgment of an expert, but without any knowledge of how many individuals share the original configuration, how large a pool there is, that is, no expert, no matter how experienced, can provide a knowledgeable judgment about any particular individual or, indeed, even a judgment of the likelihood of a particular person being the person to be charged or convicted.

So facial recognition is not some new and wonderfully different and effective technique for identifying and prosecuting individuals. The advantage of providing a history of features comparisons and the ways in which they have been helpful and harmful is that we can see that comparing facial features is more of the same, with all the problems we already canvassed and more.

That is not to diminish its value. Law enforcement can and no doubt will use it to rule out a relatively large class of individuals in any specific case. The images are good enough to allow for a great deal of discrimination, and that can be crucial to those trying to track down a suspect. But it will have no value for prosecutors. A supposed match would only show that a defendant is one of an indeterminate number of individuals with relevantly similar features.

REFERENCES

Balko, Radley. (2020, February 28). A D.C. judge issues a much-needed opinion on “junk science.” The Washington Post.

Bharara, Preet. (2020). Doing Justice. New York: Vintage Books.

'DNA bungle' haunts German police. Retrieved from http://news.bbc.co.uk/2/hi/europe/7966641.stm.

Hale, Tom. (2017, August 17). This Viral Video Of A Racist Soap Dispenser Reveals A Much, Much Bigger Problem. IFLScience. Retrieved from https://www.iflscience.com/technology/this-racist-soap-dispenser-reveals-why-diversity-in-tech-is-muchneeded/.

Harwell, Drew (2019a, May 16). Police have used celebrity lookalikes, distorted images to boost facial-recognition results, research finds. The Washington Post.

Harwell, Drew (2019b, December 19). Federal study confirms racial bias of many facial recognition systems, casts doubt on their expanding use. The Washington Post.

Hsu, Spencer S. (2012a, April 16). Convicted defendants left uninformed of forensic flaws found by Justice Dept. The Washington Post.

Harwell, Drew (2012b, May 16). Santae Tribble’s 1980 murder conviction overturned by D.C. judge. The Washington Post.

Harwell, Drew (2012c, December 14). Santae Tribble cleared in 1978 murder based on DNA hair test. The Washington Post.

The Innocence Project: Keith Allen Harward (2019). Retrieved from https://www.innocenceproject.org/cases/keith-allen-harward/.

Metz, Cade (2019, November 11). We Teach A.I. Systems Everything, Including Our Biases. The New York Times.

Murphy, Heather. (2019, December 7). When a DNA Test Says You’re a Younger Man, Who Lives 5,000 Miles Away. New York Times.

National Research Council of the National Academies. (2009). Strengthening Forensic Science in the United States: A Path Forward. Washington D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/12589.html

Office of Inspector General (2006). A Review of the FBI’s Handling of the Brandon Mayfield Case. Retrieved from https://oig.justice.gov/special/s0601/exec.pdf.

Oliver, John (2017, October 1. Forensic Science: Last Week Tonight. Retrieved from https://www.youtube.com/watch?v=ScmJvmzDcG0; from 13:21 to 14:40.

Otterman, Sharon (2019, April 23). She Was Fired After Raising Questions About a DNA Test. Now She’s Getting $1 Million. The New York Times.

Paparella, Andrew et al. (2018, March 9). Twins make astonishing discovery that they were separated shortly after birth and then part of a secret study. Retrieved from https://abcnews.go.com/US/twins-make-astonishing-discovery-separated-birth-part-secret/story?id=53593943.

Rakoff, Jed S. (2019, April 18). Our Lying Eyes. New York Review of Books.

Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity in Feature-Comparison Methods. (2016). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Robison, Wade. (2016). Ethics Within Engineering, An Introduction. London: Bloomsbury Academic Publishing.

Sciolino, Elaine. (2004, March 12). BOMBINGS IN MADRID: THE ATTACK: 10 Bombs Shatter Trains in Madrid, Killing 192. The New York Times.

Sibley, David Allen. (2000). The Sibley Guide to Birds. New York: Alfred A. Knopf.

Wang, Amy B. (2017, July 13). Video shows police trying to explain why they pulled over a Florida state attorney. The Washington Post.

CYBERSECURITY AND SMART CITIES

Shalini Kesar

Southern Utah University (USA)

kesar@suu.edu

ABSTRACT

This paper outlines some challenges and suggestion to manage and minimize cybersecurity breach within smart cities. It also reviews various research and recommendations proposed to minimize, manage, and mitigate cyber breaches within the smart cities. According to definition, a smart city is designation given to a city that incorporates Information and Communication Technologies (ICT) to enhance the quality and performance of urban services such as energy, transportation and utilities in order to reduce resource consumption, wastage and overall costs. Various reports, worldwide spending on cybersecurity is going to reach $133.7 billion in 2022. Furthermore, smart cities are complex ecosystem where city infrastructure (public and private entities, people, processes, and devices) are constantly interacting with each other and with technology also.

KEYWORDS: Smart cities, cyber breaches, ecosystem.

1. INTRODUCTION

With increase in technology use, the misuse of it is also on the rise. According to Verizon report (2019): worldwide spending on cybersecurity is going to reach $133.7 billion in 2022; approximately 68% of business leaders feel their cybersecurity risks are increasing; data breaches exposed 4.1 billion records in the first half of 2019; 71% of breaches were financially motivated and 25% were motivated by espionage; and 52% of breaches featured hacking; 28% involved malware; and 32–33% included phishing or social engineering.

Cybersecurity, by definition refers to a set of techniques used to protect the integrity of an organization's security architecture and safeguard its data against attack, damage or unauthorized access. Verizon report (2019) commented that smart cities are increasingly under attack by a variety of threats. Further, "these include sophisticated cyberattacks on critical infrastructure, bringing Industrial Control Systems (ICS) to a grinding halt, abusing Low-Power Wide Area Networks (LPWAN) and device communication hijacking, system lockdown threats caused by ransomware, manipulation of sensor data to cause widespread panic (e.g., disaster detection systems) and siphoning citizen, healthcare, consumer data, and personally identifiable information (PII), among many others," explains Dimitrios Pavlakis, Industry Analyst at ABI Research. "In this increasingly connected technological landscape, every smart city service is as secure as its weakest link." (Help Net Security, 2019).

In general, smart cities use connected technology and data to :1) improve the efficiency of city service delivery; 2) enhance quality of life for all; 3) increase equity and prosperity for residents and businesses. Report by Pandey et al (2019) underlying technology infrastructure of the ecosystem comprises three layers: the edge, the core, and the communication. The edge layer comprises devices such as sensors, actuators, other IoT devices, and smartphones. The core is the technology platform that processes and makes sense of the data flowing from the edge. Whereas the communication layer, establishes a constant, two-way data exchange between the core and the edge to seamlessly integrate the various components of the ecosystem. Consequently, with the growth of smart cities that is projected to increase fourfold by 2025, cybercrime will also continue to rise within smart cities.

2. CYBER CRIME IS ON THE RISE

Across the globe, cybersecurity breaches are increasing. Recent report conducted by Symantec Internet Security Threat Report (2019) highlighted some alarming figures: an average of 4,800 websites compromised each month; Ransomware shifted targets from consumers to enterprises, where infections rose 12 percent; More than 70 million records stolen from poorly configured S3 buckets, a casualty of rapid cloud adoption; Internet of Things (IoT) was a key entry point for targeted attacks. In addition to the statistics above, Verizon Report (2019), states that the most common causes of data breach include: weak and stolen credentials; passwords; Back Doors; Application Vulnerabilities; Malware; Social Engineering; Too Many Permissions; Insider Threats; Improper Configuration; and User Error.

2.1. Smart cities and technology

Smart cities are the future, especially in urban area to boost the efficiency and effectiveness of city services. Amsterdam was one of the first European cities to launch a smart city program, with the goal of improving its economy, environment, government, living, and mobility. Some of their qualifications for becoming a smart city are: smart housing; open data; smart grids; home energy storage; connectivity; and smart mobility. Currently, only 600 urban cities contribute to 60% of global GDP. The idea of a smart city is to collect information through connected devices and make life more comfortable using these devices for logistics. It is estimated that the market for smart city technology is expected to reach $1.5 trillion by 2020. One smart city initiative is Singapore's "Smart Nation" project, that aims to apply new technologies to enhance transportation systems, health, home and business. (Sivaramakrishnan, 2017). The goal is to increase interconnectedness in all aspects of citizens' lives through digital technologies. The underlying goal for Singapore smart city is: healthier citizens mean healthier cities; a house with a heart is a home; and mobility is a shared community experience. China, is another example, where it aims to develop 103 smart cities. These smart cities seek to bring pollution, traffic congestion and widespread energy consumption under control through greater use of connected technologies. Within India, 90 cities are targeted to develop smart capabilities as part of its "Smart City Mission." Their pragmatic approach is to attack this initiative in a layered approach, solving specific issues one at a time. Since 2011, Tokyo (Japan) has focused heavily on becoming greener by integrating homes with solar panels as well as attaching those homes to a smart grid. Policies are being modified to integrate a more sustainable and an eco-friendly infrastructure. Within the United States, cities such Boston, Las Vegas, Kansas City, and Chicago are all taking advantage of smart solutions to address transportation, sanitation, connectivity, and safety issues in their communities. Other cities within the United States include, Columbus, Pittsburgh, Denver, San Francisco, and Dallas (Condliffe, 2016). They are implementing smart technology in innovative ways to incorporate transportation as part of the smart city strategy to better connect citizens to human services; encourage greater use of sustainable transportation; improve access to jobs; and provide real-time traffic information to improve commuter mobility. Investments for some cities include in millions of US dollars.

One of the consequences of building a smart city is systems, sensors and devices are not only connected to each other and to external systems around the world but also are increasingly connected with over lightning-fast networks via the public internet and a wide range of cloud computing architectures. The more things that are connected, the greater the opportunity for cyber breaches to infiltrate your systems, exfiltrate sensitive data and disrupt potentially critical systems used in law enforcement, public health and other municipal applications.

2.2. Smart cities and cyber breaches

As mentioned above, this new wave of digital transformation also brings new cyber risks that could fundamentally impact the existence of smart cities. Cyber threats have been on the rise for years, but the last few years have seen an explosion in cyberattacks that target both data and physical assets. No doubt citizens benefit from living in smart cities. However, at the same time technology brings risks and vulnerabilities to cybersecurity. There are many examples of cyber breaches in smart cities that have warrant us to think about the consequences and solutions. For example, Atlanta, capital of Georgia State in the United States, faced SamSam, a ruthless "ransomware" bug in March 2019. This lasted approximately two weeks and a cost $55,000 worth of bitcoin in payment was demanded. The aftermath of denying the demand left Atlanta City processing reports and legal documents, which cost of this attack in millions. Baltimore, another smart city in the United States faced cyber breach. The attack involved a ransomware attack that led to accessibility issues to their Computer Aided Dispatch (CAD) system of Emergency services for 17 hours. The city's Emergency services relied on this system to automatically divert calls to emergency responders who are closest in location so that emergency assistance is directed as efficiently as possible. While the system was down responders operated by taking phone calls manually, a far slower process that could have had a more sinister outcome if the cyber-attack had been prolonged. There are many other examples of cyber breaches in smart cities. Another example, San Francisco Municipal Transportation Agency example when hackers used ransomware to shut down its ticketing systems and demand payment (Condliffe, 2016).

It has been pointed out that by 2050, about 66% of the world's population is expected to live in cities. Methods to minimize, manage, and mitigate cyber breaches should be one of the first priority while using advanced technologies within smart cities. While developing solutions, it is important to keep in mind some of the realities of cybersecurity in smart cities, for example: 1) The introduction of new web and mobile apps, IoT, connected homes, connected cars and even connected logistics. The increase use of such new technologies will make more data and gadget accessible to criminals; 2) Cybercriminal motivations increase to get access to IoT. The sophistication of technology also means increases in skills and tactics of cyber criminals; 3) Lack of skill of cybersecurity experts. Statistics indicate there is more demand than supply for cybersecurity experts. This reality can become a problem as cities become more dependent on technology for everyday activities. In addition to the realities of cybersecurity and smart cities, ethical implications also become a concern. The U.S. Department of Transportation issued a "Smart City Challenge," to U.S. cities, encouraging them with funding to develop more smart technologies around transportation. The hope is that, as cities compete, they will develop ways to use digital technology to solve transportation problems and improve efficiencies. As ideas are developed, other cities can then adapt them to their needs.

3. THREE LAYERS OF ECOSYSTEM AND CYBER CRIME

To help address some of the challenges mentioned above, researchers have argued that cities should embed cybersecurity principles to minimize, manage, and mitigate cyber breaches. This paper uses the framework proposed in the article “Making smart cities cybersecure Ways to address distinct risks in an increasingly connected urban future”, by Pandey et al (2019). They outline three layers of ecosystem in context of smart cities: the edge, the core, and the communication. This paper also reviews cybersecurity challenges in the ecosystem of smart cities. According to Pandey et al, there are three factors influence the potential cyber risk in a smart city ecosystem: Convergence of the cyber and physical worlds; Interoperability between legacy and new systems; and Integration of disparate city services and enabling infrastructure.

This section reviews the challenges and reflects on the three layers of the ecosystem and cyber risks in context of smart cities. This is significant since it is estimated that the world’s urban population will rise by 72 per cent between 2011 and 2050. To combat this growing demand, it is important to keep a check on use and misuse of technology, especially within smart cities. Furthermore, it becomes critical that smart cities service providers such as networking Internet of Things (IoT) technology with existing infrastructure are balanced and reshape supply chains and manage assets and resources more efficiently.

3.1. The core layer

This layer deals with technology platform, such as cloud, Internet of Things (IoT), that process data. This layer also enables to generate business logic to make sense of data following from the edge layer. As mentioned above, more than half of globe is spending money to connect cities with technologies to provide resilient energy infrastructure, data-driven public safety or intelligent transportation. Examples above describe cities from San Francisco’s smart power grid to Barcelona’s digitized waste management systems. With the smart cites projects come unintended consequences to all those infrastructures connected with technologies. It is true to state that the more things that are connected, the greater the opportunity for cyber attackers to infiltrate such systems, exfiltrate sensitive data and disrupt potentially critical systems used in law enforcement, public health and other municipal applications.

Moving towards smart city strategy requires careful consideration of reviewing polices. Examples of cyber breach cases in smart cities of Atlanta, Baltimore and Sint Martens, recent report by DiliTrust (2019) criticized by development experts as having been “procured and developed with little coordinated consideration of privacy and security harms”.

Accurate recovery from a cyber-attack depends on fast and perfect damage assessment. According to Anderson (2019), cyberattacks, regulators that are beginning to introduce expansive rules for tech usage and data protection should be in close proximity as part of response of relentless data breaches. In addition, at the core layer, cities should define a detailed cybersecurity strategy that is in line with their broader smart city strategy to mitigate challenges arising from the ongoing convergence, interoperability, and interconnectedness of city systems and processes. Risk management assessments that identify assets, vulnerabilities and threats will help organizations manage, minimize, and mitigate breaches associated with technology adoption. The integrated view of the risks and knowledge of interdependencies of the critical assets can enable cities to develop a comprehensive cybersecurity strategy. In the Catapult Future Cities report (2017), various strategies were discussed with examples of different countries. The report suggested smart city strategies are made through collaborative stakeholder engagement with city stakeholders and citizens. It is critical that they follow an approach that is well-understood with the existing breaches and have a buy-in with the stakeholders that need to deliver it. There is no doubt that such an approach will takes longer to complete and implement smart cities core infrastructure. Some of the recommendation proposed in the article can be applied to any smart city context. For example, establish strong leadership to develop skills and capacity within local government to deliver at-scale smart city projects (Catapult Future Cities, 2017).

A great example, is Tel Aviv smart city. They have included champions and training their council staff as part of their smart city core strategy. Core city plans that include smart city strategy within existing statutory frameworks can also prove to be a beneficial and smooth transition process. Sydney and São Paulo ae two examples that have embedded their city plans when creating their smart city strategy. Modification or revising polices linked with existing cyber security laws and acts is part of the core when creating policies. For instance, Singapore launched its National Cyber Security Master plan in 2013 and followed it with a new cyber security bill in 2016. Both initiatives were an integral part of Singapore's smart nation strategy. This is because with use of technology, the misuse of technology also increases.

3.2. The communication layer

Communication layer is related to the technology gadgets like Bluetooth and wireless. Different cities have utilized different mechanism to build a smart infrastructure. The Songdo International Business District, as it's formally known, was built from scratch, on reclaimed land from the Yellow Sea. It has the state of art communication layer to support the 1,500-acre development of smart city. New Songdo City was envisioned as "a giant test bed for new technologies' that would demonstrate the country's technological prowess to help attract foreign investment" (Poon, 2018). Programs like these aim at using technology gadgets to improve life in cities, whilst attracting foreign investment to increase economic growth.

Innovations in the communication layer of smart cities are also increasing the risks of cyber breaches and data comprise. As highlighted above, statistics highlight the increases of cyber breaches. Hence it is true to say that such cities thriving to become smart will require to explore new techniques and policies to solve cyber-related issues. Managing, minimizing and mitigation cyber breaches within smart cities requires attention from strategy and design to implementation and operations. Elmaghraby et al (2014), in their article "Cyber Security Challenges in Smart Cities: Safety, security and privacy", suggest a framework linked with privacy solutions and smart cities.

3.3. The edge layer

The edge layer comprises of sensors, actuators, and smart phones. This is the front end of the smart cities. A classic example of comprise of the edge layer is the 2018 case when Emotet malware virus struck the city of Allentown, Pennsylvania. The virus quickly multiplied in a week and rendered the city's finance department system unusable by not allowing it to make external bank transactions. Also, the police department could not access databases controlled by the Pennsylvania state police. Containing the virus and getting back to operational status is estimated to have cost the city US$1 million. Some of the vulnerabilities of the smart systems can be the traffic control systems. Cases have highlighted how smart traffic control systems are vulnerable to takeover. The 2006 case of a disgruntled employees who attacked Los Angeles traffic control systems is another example of communication layer being compromised. Although an increasing number of newer smart systems have the necessary state of art technologies, however older systems currently installed lack it and would be very hard to replace without major street reconstruction. Attacks on traffic and surveillance cameras, too, could render a city blind. The 2016, ransomware attack on the San Francisco municipal rail system demonstrates that municipal transit systems are being targeted. Additionally, introducing false information on the systems could comprise privacy of citizens. Other cases of cyber breaches include, 2015 BlackEnergy attack on the Ukrainian power grid, where more than 80,000 consumers were left without power. Another case is the 2016 hack of an unidentified water treatment plant, where mass casualties were averted only because the hacktivists who attacked it did not immediately realize what toxic chemicals they were in a position to unleash on the plant's consumers (See Verizon report 2019 for details for the cases).

When it comes to smart cities, privacy concern is a topic that comes to the forefront. According to AlDairi and Tawalbeh (2017), smart cities influencers must pay more attention so security and privacy concerns in smart cities. In their paper, "Cyber Security Attacks on Smart Cities and Associated Mobile Technologies', they discuss major security problems and recommendations linked with current smart cites. Further, they present several influencing factors that can affect data and information security in smart cities. They outline five main components that are essentially required to be in a smart city: modern information and communication technologies, buildings, utilities and infrastructure, transportation and traffic management and the city itself.

4. CONCLUSION

In addition to advantages of smart cities, the tremendous data exchange and integration of technologies within the three layers of the ecosystem can create higher cyber risks and threats. The dynamically changing of innovative technologies can also result in complexities that requires attention before we play catch up. In context of cyber security, smart cities should into account the three layers of the ecosystem to manage, minimize or mitigate various challenges that could be posed due to the interconnectivity of technology that makes a city smart.

REFERENCES

Anderson, B. (2018), How to Improve Cybersecurity in a Smart City. *ReadWrite.* Retrieved from https://readwrite.com/2018/11/12/how-to-improve-cybersecurity-in-a-smart-city/

AlDairi, Anwaar, Tawalbeh, Lo'ai. "Cyber Security Attacks on Smart Cities and Associated Mobile Technologies", *8th International Conference on Ambient Systems, Networks and Technologies, ANT-2017 and the 7th International Conference on Sustainable Energy Information Technology*, SEIT 2017, 16-19.

Condliffe, Jamie (2016). Ransomware Took San Francisco's Public Transit for a Ride. *MIT Review*. Retrieved from https://www.technologyreview.com/2016/11/28/69496/ransomware-took-san-franciscos-public-transit-for-a-ride/

DiliTrust (2019), Cyber Attacks on Smart Cities: Why we Need to be Prepared. Retrieved from https://www.dilitrust.com/en/blog/cyber-attacks-smart-cities/

Elmaghraby, Adel & Losavio, Michael (2014). Cyber Security Challenges in Smart Cities: Safety, security and privacy. *Journal of Advanced Research.* 5(10).1016/j.jare.2014.02.006.

Catapult Future Cities (2017). Smart City Strategies: A Global review. Retrieved from https://futurecities.catapult.org.uk/press-release/first-global-review-smart-cities-published/

Help Net Security (2019). Cybersecurity challenges for smart cities: Key issues and top threats. Retrieved from https://www.helpnetsecurity.com/2019/08/21/cybersecurity-smart-cities/

Pandey, Piyush., Golden, Deborah., Peasley, Sean., & Kelkar, Mahesh (2019). Making smart cities cybersecure: Ways to address distinct risks in an increasingly connected urban future . *Deloitte*

Insights. Retrieved from https://www2.deloitte.com/us/en/insights/focus/smart-city/making-smart-cities-cyber-secure.html.

Poon, Linda (2018). Sleepy in Songdo, Korea's Smartest City. *City Lab*. Retrieved from https://www.citylab.com/life/2018/06/sleepy-in-songdo-koreas-smartest-city/561374/

Sivaramakrishnan, Sharmishta (2019). 3 reasons why Singapore is the smartest city in the world Retrieved from https://www.weforum.org/agenda/2019/11/singapore-smart-city/

Symantec Security (2019). Internet Security Report. Retrieved from https://www.frontiersin.org/articles/438810

Verizon (2019). 2019 Data Breach Investigations. https://enterprise.verizon.com/en-gb/resources/reports/dbir/

HOW TO BE ON TIME WITH SECURITY PROTOCOL?

Sabina Szymoniak

Czestochowa University of Technology (Poland)

sabina.szymoniak@icis.pcz.pl

ABSTRACT

The paper discusses a very important problem which is the verification of security protocols. Security protocols are used to secure users' communication living in the Smart Cities. Users in the cyber world could be exposed to dishonest users' actions. These users are called Intruders. Also, they could fall victim to cybercrime. Due to continuous technological development, the security of protocols should be regularly verified to confirm their correctness. Also, in the case of security protocols, time plays a significant role. It may turn out that a few seconds will allow an Intruder to acquire the appropriate knowledge to execute an attack and stole confidential data. Therefore, it is right to verify security protocols also in terms of the influence of time on their security. For this purpose, we propose a new method for security protocols verification including timed parameters and their influence on security. Our method includes analysis of encryption and decryption times, composing the message time, delays in the network and lifetime, using the specially implemented tool. Thanks to this, we can calculate the correct time protocol execution, indicate time dependencies and check the possibility of Intruder's attack. Our experimental results we present on well-known Needham Schroeder protocol example.

KEYWORDS: timed analysis, security protocols, cybersecurity, verification.

1. INTRODUCTION

The concept of smart cities primarily uses information and communication technologies to increase the interactivity of city infrastructure. The development of technology entails the improvement of everyone's quality of life. In turn, the use of modern IT technologies is associated with the problem of security of users and the entire urban society.

Ensuring security at the appropriate level is based on the security protocol (SP). The security protocol is a sequence of several steps during which authentication information is exchanged between computer network users. Unfortunately, SP's are exposed to attacks and activities of dishonest users, so-called Intruders. An intruder can eavesdrop on the communications of honest users, intercept their messages, and also use the knowledge thus acquired to conduct attacks. The activities of the Intruder entail the need to regularly check the operation and security of protocols.

Over the years, several methods have been developed for verifying security protocols and tools implementing these methods ((Dolev et al., 1983), (Burrows et al., 1989), (Lowe, 1996), (Paulson, 1999), (Armando A., et. al., 2005), (Nigam et al., 2016), (Blanchet B., 2016), (Steingartner et al., 2017), (Chadha et al., 2017), (Basin et al., 2018), (Siedlecka-Lamch O., et al., 2019)). These methods and tools did not take into account time parameters and their impact on the security of security protocols.

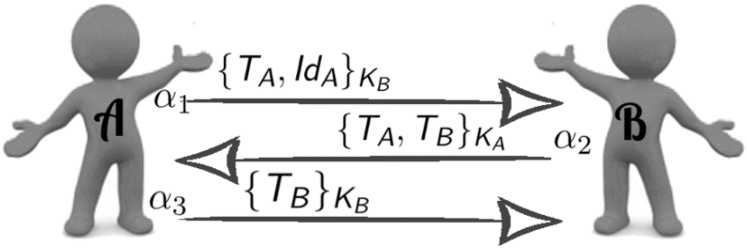
Attempts to demonstrate the influence of time on safety appeared in the works of Jakubowska and Penczka (Jakubowska et al., 2006), (Jakubowska et al., 2007). Unfortunately, this work was not continued. In turn, the model of implementation of security protocols presented in (Kurkowski, 2013) enables the generation of various versions of security protocols. We have extended this model with the mentioned time parameters. Thanks to this, we can check the duration of the session and check how time parameters have an impact on performance security. Also, we check whether the Intruder can attack the protocol for the set time parameters. In our considerations, we took into account constant and random values of time parameters.

The rest of the article is organized as follows. At the second Section we present an example of security protocol, which is Needham Schroeder Public Key protocol. We used this protocol to show the results of our research. The next Section presents our methods and materials. At the fourth Section we present our experimental results of our research. The last Section includes conclusions and plans for the future.

2. NEEDHAM SCHROEDER PUBLIC KEY PROTOCOL

As an example we will use the well-known Needham Schroeder protocol, NSPK for short (Needham et al., 1978). NSPK consists of three steps, during which two honest users (signed as *A* and *B*) try to authenticate with each other. For this purpose they exchange messages with timestamps ($T_{A,}$ T_B), IDs (I_A) encrypted by theirs public keys ($K_{A,}$ K_B).

Figure 1. Scheme of NSPK protocol.



Source: self-elaboration based on (Needham et al., 1978)

The timed version of NSPK protocol in Alice-Bob notation is presented in Figure 1. In the first step (α_1) *Alice* generate she's timestamp T_A and send it with she's ID to *Bob*. The message is encrypted by *Bob*'s public key K_B. In the next step (α_1) *Bob* generate he's timestamp T_B and send it with *Alice*'s timestamp to *Alice*. This message is encrypted by *Alice's* public key K_A. In the last step (α_3) *Alice* send to *Bob* his timestamp encrypted by *Bob's* public key.

Figure 2. Scheme of attack on NSPK protocol.

$$\begin{array}{lll}
\alpha_1 & A \rightarrow T & : \{T_A, I_A\}_{K_I} \\
\beta_1 & T(A) \rightarrow B & : \{T_A, I_A\}_{K_B} \\
\beta_2 & B \rightarrow T(A) & : \{T_A, T_B\}_{K_A} \\
\alpha_2 & T \rightarrow A & : \{T_A, T_B\}_{K_A} \\
\alpha_3 & A \rightarrow T & : \{T_B\}_{K_I} \\
\beta_3 & T(A) \rightarrow B & : \{T_B\}_{K_B}
\end{array}$$

Source: self-elaboration based on (Lowe G., 1996)

The Figure 2 presents a scheme of the attacker's intrusion on NSPK protocol. This attack was described by Gavin Lowe in (Lowe G., 1996).

The Intruder (*Trudy*, *T*) must execute additional steps (signed by *β*), to acquire adequate knowledge to complete the main execution (signed by *α*). Communication is as follows. First, *Alice* begin a communication with *Trudy* (step α_1). Therefore, she prepare message encrypted by *Trudy's* public key according to NSPK's scheme. *Trudy* decrypt this message and encrypt it again with *Bob's* public key. This message is sent to *Bob* (step β_1). Therefore, *Trudy* knows Alice's timestamp and *Bob* thinks that *Alice* try to communicate with him. Next, *Bob* generate his timestamp and send it to *Trudy*, who impersonates *Alice*, with *Alice's* timestamp. Message from step β_2 is encrypted by *Alice's* public key. *Trudy* cannot decrypt this message because she does not know *Alice's* private key. So she send whole message to *Alice* in step α_2. *Alice* decrypt it and send to *Trudy Bob's* timestamp in step α_3. Because this message was encrypted by key $K_{I,}$, *Trudy* can decrypt it and send to *Bob* his timestamp.

The effect of such protocol execution is following. *Alice* and *Bob* think that they communicated with each other, but in fact all their messages were read by *Trudy*. *Trudy* know *Alice's* and *Bob's* timestamp.

3. METHODS AND MATERIALS

To perform the full specification of the security protocol, we have extended all model definitions from (Kurkowski, 2013) by time parameters. The new model includes definitions of a set of time conditions, a protocol step, as well as the entire protocol in a timed version. In turn, the computational structure defines the current execution of the protocol and its interpretation. In the executions of the protocol, we took into account the Intruder in four models: Dolev-Yao (Dolev et al., 1983), restricted Dolev-Yao, lazy Intruder and restricted lazy Intruder. Interpretation of the protocol makes it possible to generate a set of different protocol executions. Also, the model takes into account changes in participants' knowledge during the protocol.

Also, the computational structure defines a set of time dependencies that allow to calculate the duration of a session and prepare appropriate time conditions. These dependencies are related to message composition, step and session times, and lifetime. Dependencies include delays in the network.

For the needs of our approach, we introduced the following delays in the network values for the protocol's step: minimum (D_{min}), current (D_s) and maximum (D_{max}). Such distinction determines the range of tested values delays in the network. It is necessary to enable correct protocol execution to honest users regardless of network conditions. Also, for step time and the session time we consider similar distinction. The step time (minimal T_s^{min}, current T_s, maximal T_s^{max}) is the sum of message composition's time (T_c), encryption time by the sender (T_e), delay in the network and decryption of the message's time by the recipient (T_d). The session time (minimal T_{ses}^{min}, current T_{ses}, maximal T_{ses}^{max}) consists of all steps' times. The values of step and session times depend on used delays in the network values.

To check the influence of time parameter values on the protocol's users and its security, lifetime was established. Lifetime cannot be exceeded in any of the executed steps. If this value will be exceeded, users should know that they are communicating with the Intruder. Therefore, communication should be immediately terminated. We calculate lifetime value in one step as a sum of maximal step times of this step and next steps.

For research, we created a tool which allows verifying the timed security protocols. In the beginning, the tool loads the protocol's specification from the file. Then all potential executions of the tested protocol are combinatorically generated. In the next step, using the SAT-solver, we checked whether the generated executions are possible in reality. It is possible that during one execution the Intruder will not be able to acquire the appropriate knowledge to complete it.

Then we conducted two types of research on the loaded protocol. The first of these is the so-called time analysis. This analysis enables the determination of limits for delays in the network and lifetime for which the protocol remains secure. The second type of research is simulations. In this case, we can simulate delays in the network values and encryption and decryption times to provide a real representation of the computer network.

4. EXPERIMENTAL RESULTS

Our tests were carried out using a computer unit with the Linux Ubuntu operating system with Intel Core i7 processor, and 16 GB RAM. Also, we used an abstract time unit ([tu]) to determine the time.

The experimental results will be presented on the example of Needham Schroeder Public Key protocol. According to NSPK protocol structure, we assumed that encryption and decryption times were equal 5 time units ([tu]), time of composing the message for the first and second step were equal 2 [tu], time of composing the message for the third step were equal 1 [tu]. Also, we choose the range of delays in the network values from 1 to 10 [tu] and set constant value for the current delay in the network D=1 [tu].

Next, we calculated lifetimes for steps:

- L_1=59 [tu],
- L_2=39 [tu],
- L_3=19 [tu].

Also, we calculated minimal and maximal session time:

- T_s^{min}=32 [tu],
- T_s^{max}=65 [tu].

These values were necessary to enable and set time conditions.

Table 1. Summary of NSPK protocol's executions.

No.	Send. - Rec.	Parameters	No.	Send. - Rec.	Parameters
1	*A→B*		10	*B→I(A)*	T_a, K_a
2	*B→A*		11	*I→A*	T_i, K_i
3	*I→B*	T_i, K_i	12	*I→A*	T_b, K_i
4	*I→B*	T_a, K_i	13	*I(B)→A*	T_i, K_b
5	*I(A)→B*	T_i, K_a	14	*I(B)→A*	T_b, K_b
6	*I(A)→B*	T_a, K_a	15	*A→I*	T_i, K_i
7	*B→I*	T_i, K_i	16	*A→I*	T_b, K_i
8	*B→I*	T_a, K_i	17	*A→I(B)*	T_i, K_b
9	*B→I(A)*	T_a, K_a	18	*A→I(B)*	T_b, K_b

Source: self-elaboration

For the NSPK protocol, eighteen executions have been generated. A list of these executions was presented in Table 1. Column *Send. - Rec.* relate to protocol's participants: *A*, *B* is the honest users, *I*, *I(A)*, *I(B)* is the Intruder. *I* means Intruder who occur as a regular user, *I(A)* means intruder who impersonates user *A*, and *I(B)* means Intruder who impersonates user *B.* Column *Parameters* includes cryptographic objects, which are used by Intruder during execution. Column *No.* contains an ordinal number which was assigned in order to simplify the reference to execution. For example, execution no. 9 take place between honest user *B* and Intruder who impersonates user *A*. In this case, the Intruder uses the timestamp of user *A* and also his public key.

Table 2. Timed analysis of attacking execution in [tu].

α step	β step	T_e	T_c	D	T_d	T_s	T_{ses}	Result
α_1		4	2	1	4	11	11	ok
	β_1	4	2	1	4	11	22	ok
	β_2	4	2	1	0	7	29	ok
α_2		0	2	1	4	(7) 25	36	ok
α_3		4	1	1	4	10	46	ok
	β_3	4	1	1	4	(10) 27	56	$T_3>L_3$

Source: self-elaboration

Let's analyze the attacking execution. We analysed an interlacing of 7 (α-execution) and 11 (*β*-execution) executions, which reflects attack included in Figure 2. The timed analysis was presented in Table 2. This Table consist of nine columns. First of them is α-step which is related to the steps of basic execution. The second column is called *β*-step. Here will be assigned steps from additional execution. Next six columns are connected to time parameters. These are encryption time ($T_{e)}$), composition time (T_c), delay in the network value (D), decryption time (T_d), current step time (T_s) and current session time (T_{ses}). The last column is called Result. Here we assign our comment about current step.

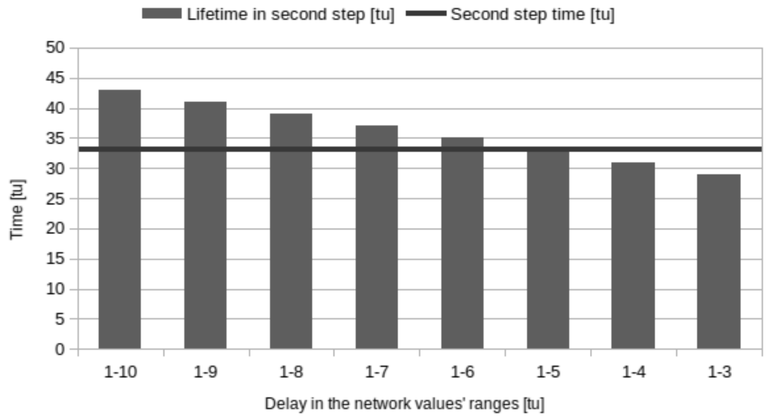
Therefore, each step consists of encryption time, composition time, delay in the network value and decryption time. Because Intruder did not have enough knowledge to execute α_2-step, he must establish *β*-execution and execute additional step from it. Steps times from additional steps were added to basic step time and also to the session time. Therefore, α_2-step includes β_1 and β_2 times. Also, β_3-step includes α_2 and α_3 times.

In the step α_2 and β_2 there were no included encryption time. The Intruder did not perform such operations, because he has not appropriate knowledge to decrypt the message from β_2, therefore he sends whole this message to *Bob* in the step α_2.

Please note that if Intruder will not end *β*-execution, the α-execution will be ended in the correct time, including additional steps' times and Intruder will know *A's* and *B's* timestamps. Therefore, the attack on this protocol is possible for assumed values of the time parameters.

Next, we calculated how changes in the delay in the network range would affect protocol security. We increased the maximum delay value in the network by 1 [tu] and checked how long the duration of the second step would be. Thanks to this, we will be able to determine what limit should be set for this step.

Figure 3. Changes in the delay in the network range.



Source: self-elaboration

Our results were shown in Figure 3. The setting the upper limit of delay in the network values to 4 [tu] protects the protocol. In such a situation, a lifetime set in the second step will end the communication and the attack on NSPK protocol is not possible.

In the next step, we performed simulations of Needham Schroeder Public Key protocol's executions. We used the randomly generated the current delay in the network values. We used normal, uniform, Cauchy's, Poisson's and exponential probability distributions to generate these values. The tool also allows random selection of values out of the accepted range to model the real work of a computer network.

We made the following assumptions:

- encryption and decryption times were equal 2 [tu],
- time of composing the message for all steps was equal 1 [tu],
- the range of delays in the network values from 1 to 10 [tu].

Next, we calculated new lifetimes for steps, minimal and maximal session times:

- L_1=44 [tu],
- L_2=29 [tu],
- L_3=14 [tu],
- T_s^{min}=17 [tu],
- T_s^{max}=44 [tu].

Table 3. NSPK executions ended in the correct session time.

No.	Session time [tu]			**Average delay in the network [tu]**
	min	**avg**	**max**	
1	18.4	30.98	43.3	5.67
2	19.1	30.83	43	5.61
3	17.4	29.24	41.8	5.41
4	40.7	42.62	43.9	3.04
7	18	29.73	41.3	5.58
8	39.6	42.31	43.6	3.1
11	17	29.72	42.2	5.57
12	39.8	42.14	43.8	3.1
15	17.4	29.77	41.7	5.71
16	38.3	41.92	43.8	2.89

Source: self-elaboration

We carried out 18,000 test series for each probability distribution. In Table 3, we presented minimal, average and maximal values of session time for several executions of Needham Schroeder Public Key protocol including delay in the network values generated according to the uniform probability distribution. Also, we presented the average delay in the network values. These sessions ended in the correct session time.

Also, we assumed two specific situations. First of them was when the current session time was lower then minimal session time (T_s^{min}). This means that Intruder may send cryptograms which were in his knowledge set. The Intruder did not encrypt or decrypt messages and also he did not generate any object, so these times were not added to session time. Session time (T_s) was lower then minimal session time (T_s^{min}).

Table 4. NSPK executions ended below the minimal session time.

No.	Session time [tu]			**Average delay in the network [tu]**
	min	**avg**	**max**	
3	16.01	16.31	16.98	1.1
7	16.02	16.3	16.97	1.2
10	16.03	16.49	16.49	1.09
11	16.04	16.3	16.99	1.1
15	16.02	16.31	16.98	1.1
18	16.03	16.51	16.99	1.09

Source: self-elaboration

In Table 4, we presented minimal, average and maximal values of session time for several executions of Needham Schroeder Public Key protocol including delay in the network values generated according to the exponential probability distribution. Also, we presented the average delay in the network values in each execution. These sessions ended in below the minimal session time. In these exeutions Intruder used his cryptographic objects and also resent whole ciphertext received from honest users. Please note that, the average delay in the network values were between 1.09 [tu] and 1.2 [tu].

The second specific situation was when the current session time was upper then maximal session time. This means that Intruder must execute additional steps to get knowledge. Additional steps' times affect the current step and session time. In this case, the execution ended incorrectly (exceeding the maximum session time), while the time conditions imposed on each step have been preserved.

Table 5. NSPK executions ended upper then the maximal session time.

No.	Session time [tu]			**Average delay in the network [tu]**
	min	**avg**	**max**	
1	44.02	50.40	76.81	12.98
2	44.01	51.81	77.61	12.74
3	44.02	51.75	70.47	17.84
4	48.08	75.26	105.99	8.81
7	44.05	52.15	71.12	21.66
8	46.05	68.58	90.61	7.19
11	44.03	52.19	75.57	13.06
12	47.81	73.72	110.78	8.37
15	44.16	52.38	72.02	13.13
16	46.48	66.0	95.99	7.1

Source: self-elaboration

In Table 5, we presented minimal, average and maximal values of session time for several executions of Needham Schroeder Public Key protocol including delay in the network values generated according to the normal probability distribution. Also, we presented the average delay in the network values in each execution. These sessions ended in below the maximal session time. Please note that, the average delay in the network values were greater then in case of exponential probability distributions. Also, we observed values out of adopted range.

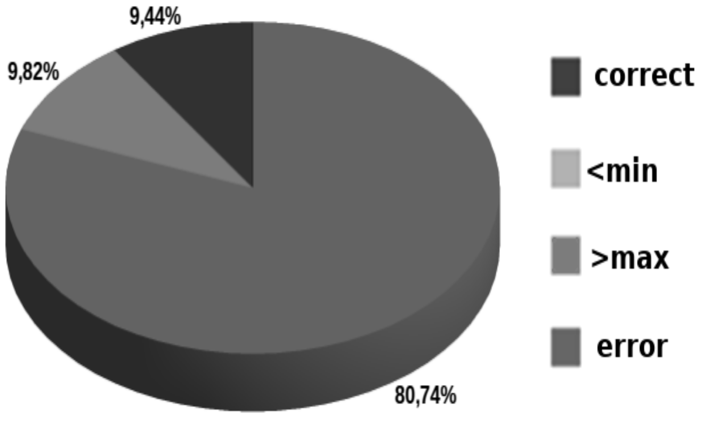
To present summary of our research we included following designations on the charts:

- *correct*, which is designated to the session ended between minimal session time (T_s^{min}) and maximal session time (T_s^{max}),
- *<min*, which is designated the session that were ended lower then minimal session time (T_s^{min}),
- *>max*, which is designated to the session that were ended upper then maximal session time *(T_s^{max}),*
- *error*, which is designated the session that were ended because one of the time conditions was not met.

We presented the percentage summary of the number of NSPK protocol executions ended with assumed statuses. Please note that there was no situation in which session ended upper then maximal session time (T_s^{max}).

On Figure 4, we presented a summary of the results for the Needham Schroeder Public Key protocol using a normal probability distribution.

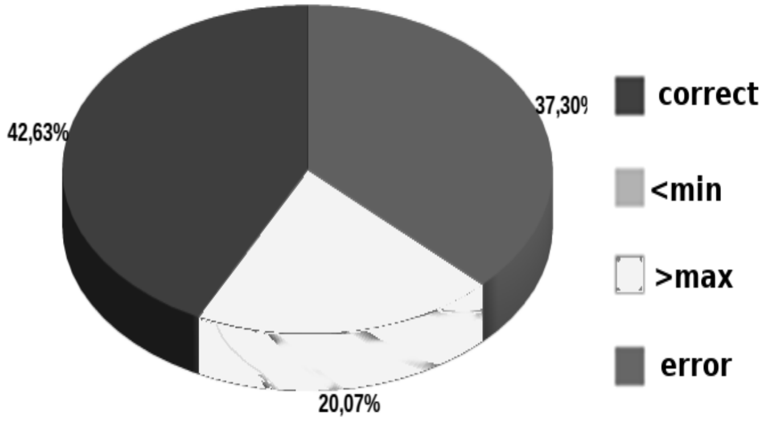
Figure 4. Summary of the results for the NSPK protocol using normal probability distribution.



Source: self-elaboration

On Figure 5, we presented a summary of the results for the Needham Schroeder Public Key protocol using a uniform probability distribution. Please note that there was no situation in which session ended upper then minimal session time *(T_s^{min}).*

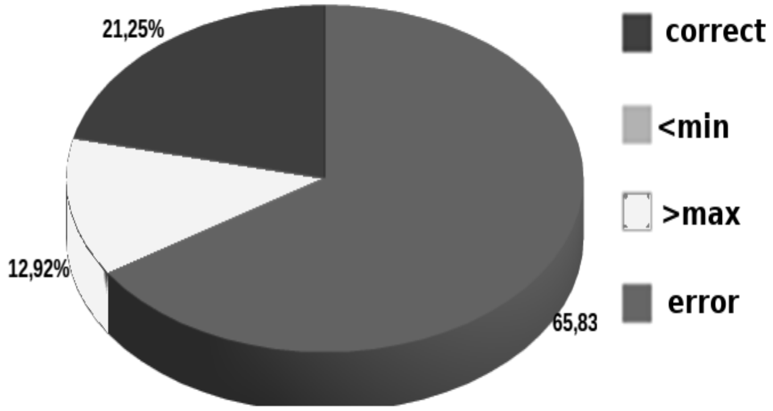
Figure 5. Summary of the results for the NSPK protocol using uniform probability distribution.



Source: self-elaboration

On the Figure 6, we presented summary of the results for the Needham Schroeder Public Key protocol using Poisson's probability distribution. Please note that there were no situation in which session ended lower then minimal session time *(T_s^{min}).* and there were a lot of error situations.

Figure 6. Summary of the results for the NSPK using Poisson's probability distribution.

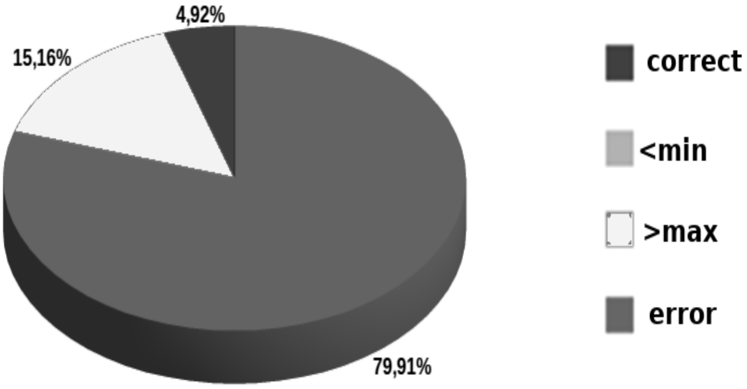


Source: self-elaboration

On Figure 7, we presented a summary of the results for the Needham Schroeder Public Key protocol using Poisson's probability distribution. Please note that there was no situation in which session ended lower then minimal session time *(T_s^{min})* and there were a lot of error situations.

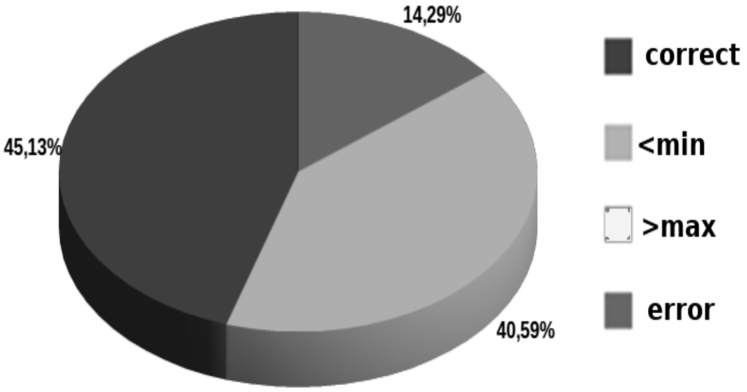
On Figure 8, we presented a summary of the results for the Needham Schroeder Public Key protocol using an exponential probability distribution. Please note that there was no situation in which session ended upper then maximal session time.

Figure 7. Summary of the results for the NSPK using Cauchy's probability distribution.



Source: self-elaboration

Figure 8. Summary of the results for the NSPK using exponential probability distribution.



Source: self-elaboration

The obtained results showed various aspects of computer network operation. We observed that usage of uniform probability distribution shows the natural operation of the network. The usage of normal probability distribution reflects the real operation of the network. The usage of Cauchy and Poisson probability distributions suggest a busy network that very often has problems. The analyze of the results obtained for the exponential probability distribution it should be stated that this distribution illustrates the fast network.

Also, we try to check what is the influence of encryption algorithms' speed and computing power. For this purpose, we perform simulations Needham Schroeder Public Key protocol's executions using randomly generated values of encryption and decryption time, according to a uniform probability distribution.

Again, we carried out 18,000 test series. We observed that for encryption and decryption times close to 1 [tu], the attacking execution ended correctly. This means that if Intruder has great computing power, his encryption and decryption times could be short and he can successfully perform an attack on the protocol.

5. CONCLUSION

This paper discussed the problem of security protocols' verification. SPs are widely used in smart cities to secure users communication. For this reason, security protocols verification is important to check if they provide an appropriate level of security.

We presented a new approach to this issue. In our research, we take into account the following time parameters: encryption and decryption times, composing the message time, delays in the network and lifetime. These parameters were used to calculate the correct protocol's execution time and designate time dependencies. The imposed dependencies should protect prevent loss of confidential information. We researched by timed analysis and simulation of delays in the network and simulations of encryption and decryption times.

We observed that time has a huge impact on protocols' security. Badly selected time dependencies could allow Intruder to perform additional actions to steal the data and threaten the security of the smart city. During our research, we analyzed how delays in the network range affect Intruder's capabilities. We took into account constant and random values of time parameters. Observed results showed that if delays in the network range will be to extensive, it will not be secure for honest users because Intruder could have enough time to compromise the protocol.

Also, we observed how the selected probability distributions illustrated the operation of a computer network. Our next research will be focused on further parameters, which may have an impact on communication security in smart cities.

ACKNOWLEDGEMENTS

The project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019 - 2022 project number 020/RID/2018/19, the amount of financing 12,000,000.00 PLN.

REFERENCES

Armando A., et. al. (2005). The AVISPA tool for the automated validation of internet security protocols and applications, In: Proc. of 17th Int. Conf. on Computer Aided Verification (CAV'05), vol. 3576 of LNCS, pp. 281–285, Springer

Basin D., Cremers C., Meadows C. (2018). Model Checking Security Protocols, in Handbook of Model Checking, Springer International Publishing

Blanchet B. (2016). Modeling and Verifying Security Protocols with the Applied Pi Calculus and ProVerif, Foundations and Trends in Privacy and Security, vol. 1(1-2) pp.1–135

Burrows M., Abadi M., Needham R. (1989). A Logic of Authentication, In: Proceedings of the Royal Society of London A, vol. 426

Chadha R., Sistla P, Viswanathan M. (2017). Verification of randomized security protocols, Logic in Computer Science

Dolev D., Yao A. (1983). On the security of public key protocols. In: IEEE Transactions on Information Theory, 29(2)

Jakubowska G., Penczek W. (2006). Modeling and Checking Timed Authentication Security Protocols, Proc. of the Int. Workshop on Concurrency, Specification and Programming (CS&P'06), Informatik-Berichte 206(2)

Jakubowska G., Penczek W. (2007). Is your security protocol on time?, In Proc. Of FSEN'07, volume 4767 of LNCS, Springer-Verlag

Kurkowski M. (2013). Formalne metody weryfikacji wlasności protokolow zabezpieczajacych w sieciach komputerowych, in polish, Exit, Warsaw

Lowe G. (1996). Breaking and fixing the needham-schroeder public-key protocol using fdr. In Proceedings of the Second International Workshop on Tools and Algorithms for Construction and Analysis of Systems, TACAS '96, pages 147–166, London, UK, 1996. Springer-Verlag.

Needham R. M., Schroeder M. D. (1978). Using encryption for authentication in large networks of computers. Commun. ACM, 21(12)

Nigam V., et. al (2016). Towards the Automated Verification of Cyber-Physical Security Protocols: Bounding the Number of Timed Intruders, Computer Security – ESORICS 2016", Springer International Publishing

Paulson L. (1999). Inductive Analysis of the Internet Protocol TLS, ACM Transactions on Information and System Security (TISSEC), vol 2 (3)

Siedlecka-Lamch O., et. al (2019) A fast method for security protocols verification, Computer Information Systems and Industrial Management, Springer

Steingartner W., Novitzka V. (2017). Coalgebras for modelling observable behaviour of programs, In: Journal of applied mathematics and computational mechanics. 16(2)

6. Meeting Societal Challenges in Intelligent Communities Through Value Sensitive Design

A HOLISTIC APPLICATION OF VALUE SENSITIVE DESIGN IN BIG DATA APPLICATIONS: A CASE STUDY OF TELECOM NAMIBIA

Emilia Shikeenga, Rosa Gil, Roberto García

Universitat de Lleida (Spain)

es16@alumnes.udl.cat; rgil@diei.udl.cat; roberto.garcia@udl.cat

ABSTRACT

In order to encourage ethical considerations and integrity in Big Data applications that incorporate Machine Learning techniques, this paper introduces a case as to how we intend to apply Value Sensitive Design (VSD) methodology in the design of a Telecom Customer Churn Prediction model. The VSD approach identifies stakeholders throughout the design process and this assists in steering clear of any biases in the design choices that might compromise any of the stakeholders' values. In this paper, we realize a VSD conceptual investigation of a churn prediction model, including stakeholder identification and the selection of human values to be included in the design.

KEYWORDS: big data, machine learning, telecommunications, human values, value-sensitive design.

1. INTRODUCTION

In recent years, big data technologies have been putting some pressure with regard to what is deemed acceptable or not acceptable from an ethical point of view. A great deal of the literature that focuses on ethical issues related to big data mostly concentrates on the following values: privacy, human dignity, justice or autonomy (La Fors et al., 2019).

Telecom Namibia is facing ever-increasing competition from new entrants such as MTN, Paratus Telecom, and Capricorn Mobile. With these new entrants, all chasing the same pool of customers and declining customer spend, Telecom Namibia needs to be able to retain its customer base in order to protect its revenues and ensure growth.

According to Harvard Business Review (Gallo, 2014), it costs between 5 times and 25 times as much to find a new customer than to retain an existing one. Thus, preventing customer churn is quickly becoming an important business function.

Telecom Namibia currently has a very basic churn model in place, which simply looks at churn on the basis of how many customers have discontinued the use of telecom services but that is too late to win back the customer. Consequently, this churn model is no longer practical nor efficient.

Telecom needs to have a more robust churn model built to predict customer churn with machine learning algorithms. Ideally, telecom can nip the problem of unsatisfied customers in the bud to keep the revenue flowing and ring-fencing its customer base.

During the development of the churn model, the team will also take the opportunity to explore the customer data by determining the different personality types of each customer through accessing their personal social media profiles. This will allow the team to provide proof that it is in fact possible for companies to "use" their customers' personal data in various ways that might be unethical.

This will therefore require that the value aspect be taken into account because the model is going to utilize data that is sensitive which might have value implications. It is for this reason why the churn model design process will employ the Value Sensitive Design approach.

Value Sensitive Design is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process (Friedman et al., 2013).

Our approach will be to apply Value Sensitive Design to design a churn prediction model for Telecom Namibia. To allow us to proactively make use of the values, we will be engaging the stakeholders throughout the design process including the prototype development. The study will consider the VSD values starting from those listed by Friedman et al. (2013) and focusing on the values for big data technologies listed by La Fors et al. (2019), as shown in Table 1.

Table 1. Integrated view of human values from different domains. Source: La Fors et al. (2019).

Technomoral values (Vallor 2016)	**Values from value-sensitive design (VSD)** (Friedman et al. 2006)	**Values from Anticipatory emerging technology ethics** (Brey 2012)	**Values in biomedical ethics** (Beauchamp & Childress 2012)	**Integration: values for big data technologies**
Care	Human Welfare	Well-being and the common good	Beneficence	Human welfare
Autonomy	Autonomy	Autonomy	Autonomy	Autonomy
Humility, self-control	Calmness	Health, (no) bodily and psychological harm	Non-maleficence	Non-maleficence
Justice	Freedom from bias; Universal usability	Justice (distributive)	Justice	Justice (incl. equality, nondiscrimination, digital inclusion)
Perspective	Accountability	N/A	N/A	Accountability (incl. transparency)
Honesty, self-control	Trust	N/A	Veracity	Trustworthiness (including honesty and underpinning also security)
N/A	Privacy; informed consent; ownership and property	Rights and freedoms, including Property	N/A	Privacy
Empathy	Identity	Human dignity	Respect for dignity	Dignity
Empathy, flexibility, courage, civility	Courtesy	N/A	N/A	Solidarity
Courage, empathy Environmental	Sustainability	(No) environmental harm, Animal welfare	N/A	Environmental welfare

Through the implementation of this case, this paper will provide support on incorporating ethics and human values in Big Data applications. The findings will outline and demonstrate how viable the VSD approach is for providing a more comprehensive view and balancing of human values and ethics in Big Data applications.

2. APPROACH

Our approach draws on the Value Sensitive Design theory and involves three types of investigations: conceptual, empirical, and technical (Friedman et al., 2013, La Fors et al., 2019). As the goal of our research is to design a churn prediction model for Telecom Namibia using VSD, our research consists of the following investigations:

1. **Conduct conceptual investigations** to find the indirect and direct stakeholders, plus the values that are implicated. To achieve this, we have to identify the different stakeholders including discovering how they are affected and the values that are implicated with regard to the implementation of the application. Applying stakeholder analysis (Friedman & Hendry, 2019) we identify:

    - **Policy Makers:** This includes the government Republic of Namibia, as well the regulators- Communications Regulatory Authority of Namibia (CRAN) which is mandated to regulate the telecommunication services and networks in Namibia. It also includes the Ministry of ICT**.**
    - **Contractors:** Any person or firm that undertakes a contract to provide materials or labor to perform a service or do a job for Telecom Namibia**.**
    - **Competitors:** Other companies in Namibia that offer the same products and services offered by Telecom Namibia. The competitors include MTC, MTN, Paratus Telecom, and Capricorn Mobile.
    - **Shareholders:** The organizations and individuals that have a stake in Telecom Namibia.
    - **Customers:** The people, organizations, businesses, etc. who buy and apply for the products and services that Telecom Namibia offers and makes use of those services.
    - **Marketing and Sales representatives**: Responsible for monitoring customer churn and coming up with relevant solutions to retain customers.
    - **Lead Data Scientist:** In charge of developing the churn model.

2. **Choosing the ethical values to consider:** the values are selected according to the Telecom Namibia company values and the value considerations for techno-social change in Big Data contexts presented by La Fors et al. (2019). Telecom Namibia's company values are (Telecom Namibia, 2017): Integrity, Care, Commitment, Accountability, Empowerment, Teamwork and Mutual Respect.

    The following VSD values are the values we will consider for the particular case of the design of the churn prediction model using Big Data techniques: human welfare, ownership, and property, autonomy, calmness, universal usability, accountability, trust, privacy, identity, courtesy and sustainability (Friedman et al., 2013).

3. ANALYZE THE TECHNOLOGY AGAINST VALUES

Following the previous approach, in Table 2 we carry out an analysis of the technology against the values. The analysis includes hashtags such as #Risk and #Need4Action to indicate any ethical risk or as an indicator where actions need to be taken to address specific challenges. The technology can be seen as INPUT, MODEL and OUTPUT, as detailed in the following table.

Table 2. Analyzing the technology against values. Source: Open Roboethics Institute. (2019).

	Value Questions	Telecom Namibia
	INPUT	
Transparency	Do the relevant stakeholders know how/when the information is collected/changed/used?	No
	Are the data provided by the stakeholders used to collect any secondary sources of information (e.g., connected to social media profiles, external online platforms)? If so, are the stakeholders informed of this?	No, so far telecom Namibia does not use data to access users' social media. A note is that we will be using the data merely to also prove that a company can access user data on their social platforms to for example find out their personalities.
Autonomy & Consent	Is there an informed consent process in place for the data collection that outlines the fact that the data can be used for this use case?	No and it is not required.
	Can the stakeholders decide not to have their data used for the algorithmic system?	Not yet #Need4Action
	Are there any elements in the data collection process (e.g., user interface used for inputting data) that could result in unintended outcomes?	No
Fairness	If people are involved in directly collecting data from someone/something, how diverse are these people in terms of race, gender, age, class, and other socioeconomic factors? Teams of people who are similar to one another can lead to similarly biased observations and data entries.	No data needs to be collected, all data to be used is already in the company's databases.
	Are certain groups of stakeholders' information collected disproportionally more than others? If so, does this fact support or conflict with the societal and stakeholder values?	No
Human Rights	Does the input data include sensitive/identifying information (e.g., gender, race/ethnicity, religion, location of work/residence, education, social and professional associations/groups)?	Yes #Risk
	Can the stakeholders opt not to enter the sensitive/identifying information?	Yes

MODEL		
Transparency	Will the model and its performance be understandable to and monitored by those training it?	We will include this premise in our design.
	If there is a questionable/erroneous outcome or an incident in the future, is it possible to explain to a third party what aspects of the model led to the outcome/incident?	Yes
Accountability	How often is the model updated/re-trained and is the frequency adequate for the use case?	Firstly we will train it every evening and ensure that the frequency is adequate for the use case.
	Who oversees the model training/updating process and are they the right people who can detect new problems and act upon them?	The Lead data scientist will be overseeing the process and will be able to detect any new problems and act upon them.
Fairness	Are there sources of bias that could lead to unfairly discriminating against individuals/groups, especially against specific gender, race/ethnicity, religion, social class or otherwise marginalised groups?	No
	Are there any parameters or technical aspects of the system that can contribute to biases in the output against specific gender, race/ethnicity, religion, social class or otherwise marginalised groups?	No
Human Rights	Is the model designed to reveal or predict an individual's identity (e.g., sexual orientation), potential (e.g., a child's probability of success in life), such that it contradicts with stakeholder and societal values, including human rights?	No
OUTPUT		
Transparency	Will the output from the algorithm presented in such a way that is understandable to its audience?	The team will be working to ensure it is understandable to its audience
	Is the output presented to the stakeholders in a way that allows them to understand how/why the system has produced the specific output? Is it important for them to understand this?	It will be important for Telecom Namibia to know how the system produces a particular output when the output is surprising.
Trust	Will the output from the algorithm translated from a probability score to a categorization (e.g., 90% probability of being X is presented as being X)? Is the translation of the probability to categorization appropriate for the use case and trustworthy?	It will be important to have the translation of the probability to categorization appropriate for the use case and trustworthy
	Will the technology and its output have the potential to lead to a destructive cycle of behaviours or operations (e.g., reinforcing gender bias of those who are the primary source of input data)?	There will be periodic supervision and monitoring of the design process.

	If someone were to take the outputs from the system and generalise it to other use cases, is it reasonable to foresee problematic interpretations or increase in distrust among stakeholders?	There is a possibility that could happen. Any bias that will be perpetuated by the algorithm may be wrongly interpreted as facts.
Accountability	Who will be responsible for acting on the output, and does this stakeholder group have ways to remedy or override erroneous or questionable output?	Telecom Namibia will be responsible for using the output appropriately. Will have to design a way for Telecom Namibia to handle erroneous recommendations
	Is there a communicated and unobstructed means for different stakeholder groups to raise an alarm on possibly dangerous usage of the technology?	No #Need4Action
	For cases where sensitive findings arise from the outcome, is there a clear means for different stakeholder groups to deal with the potentially uncomfortable truths (burden of knowledge)?	No #Need4Action
	What are the implications of false positives? What are the implications of false negatives? Are the appropriate decision makers aware of the balancing of risks between the two?	Yes, false positives can have severe impacts on any initiatives that will be undertaken to address churn since that would increase the cost of retaining a customer. For e.g if the model assigns as someone to be more likely to churn but is not the case then the organization would essentially be spending money trying to retain customers that were never really at risk of leaving the company because of the false positive predictions of high risk. On the other hand, false negatives can cost the company more than false positives. In this case, the model predicts customers as not churning, while customers actually will churn. The company will therefore lose profits by making them leave without doing any action for them not to churn.
Autonomy/ Consent	Is the output connected to another process or technology without human intervention being necessary? If so, are the risks from worst case scenarios minimal and acceptable?	No, it will not
	Will the technology be designed to replace or assist human decisions? If it is meant to replace them, is it meant to support the overall function of the stakeholders whose decisions are being replaced?	Yes, it will be designed to assist human decisions and techniques from HCI will be used.
Fairness	Will the primary users of the technology be aware of the potential biases that may have contributed to the output?	We will work towards that #Need4Action
	Will the stakeholders who are subjected to the technology be given a means of remedy?	We will work towards that #Need4Action
	Will the output produce the same result for all users? Does it lead to unfairness or discrimination?	There will be some supervision and reviewing to ensure that the output yields the same results for all users.

	Will the output lead to fair distribution of wealth, opportunity, or other positive outcomes?	There will be some supervision and reviewing to ensure that the output leads to positive outcomes.
Human Rights	Will the technology suppress or protect fundamental human rights, such as right to life, liberty, security, freedom of movement and of expression, among others?	We will follow international standards focused on technology to ensure that they are currently being developed as an IEEE Ethically Aligned Design. (IEEE, 2019)

4. CUSTOMER SERVICE SURVEY

To test the hypothesis that is based on user experience, which is: One of the reasons why customers churn is because of poor customer service (Retention Science, 2019), we carried out a customer service survey. The survey was conducted online using google forms and shared among Namibians, about 25 participants took part. We present our findings below:

- About 80% of the participants agree that they value staff at the call center/customer service are friendly and helpful
- About 68% of the participants agree that they value staff who provide them with good feedback and solutions to any issues/problems they might encounter with the products/services.
- About 68% of the participants agree that they value more products/services that fully meet their needs.
- About 64% of the participants agree that they value more products/services that are innovative.
- About 52% of the participants agree that they value more products/services are better compared to competitors' products/services (quality)
- About 44% of the participants agree that they value more the pricing of Products/services is reasonable.

Overall to sum up everything, the survey indeed revealed that customers value good customer service and should they be at the receiving end of poor service, they would most likely churn.

5. CONCLUSIONS AND FUTURE WORK

The increase in providing Ethical considerations in Big Data has become a concern and the values are also indicated in the ACM Principles for Algorithmic Transparency and Accountability (ACM, 2017). This paper introduced the application of VSD in telecom customer churn models construction. We have identified the direct and indirect stakeholders of Telecom Namibia and identified the associated human values. VSD has proven to be a promising approach in promoting ethical considerations in Big Data applications.

Our future work for this study is to clearly outline in detail how we applied VSD through the design process of the Telecom Namibia Churn Model. We will be researching and analyzing any laws or norms around the chosen values and we will define the design requirements. Another step will be to consider how we will verify/evaluate whether the designed model embodies the chosen human values.

REFERENCES

ACM (2017). Statement on Algorithmic Transparency and Accountability. Available from: https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

Customer Survey. Available from: https://forms.gle/KfhVhiBBzxYfUAC49

Friedman, B. & Hendry, D. G. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: MIT Press.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In: Early engagement and new technologies: Opening up the laboratory (pp. 55-95). Springer, Dordrecht.

Gallo, A. (2014). The Value of Keeping the Right Customers. Harvard Business Review. Available from: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

IEEE. (2019). Ethically Aligned Design. Available from: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

La Fors, K., Custers, B., & Keymolen, E. (2019). Reassessing values for emerging big data technologies: integrating design-based and application-based approaches. Ethics and Information Technology, 1-18.

Open Roboethics Institute. (2019). Foresight into AI Ethics (FAIE): A toolkit for creating an ethics roadmap for your AI project. Available from: https://openroboethics.org//ai-toolkit/

Retention Science. (2019) The Data-Driven Marketer’s Guide to Predicting Customer Churn. Available from: https://go.retentionscience.com/retention-marketing

Telecom Namibia. (2017). 2016/2017 Annual Report. Available from: https://www.telecom.na/downloads/reports/2016-17/Annual%20Report%202016-2017.pdf

EXPLORING VALUE SENSITIVE DESIGN
FOR BLOCKCHAIN DEVELOPMENT

Roberto García, Rosa Gil

Universitat de Lleida (Spain)

roberto.garcia@udl.cat; rgil@diei.udl.cat

ABSTRACT

The potential impact that blockchain technologies might have in our society makes it paramount to consider human values during their design and development. Though the blockchain community has been moved from the beginning by a set of values that are favored by the underlying technologies, it is necessary to explore how these values play among the diverse set of stakeholders and the potential conflicts that might arise. The final aim is to motivate the establishment of a set of guidelines that make blockchains better support human values, despite the initial bias these technologies might impose.

KEYWORDS: blockchain, smart contract, human values, value sensitive design.

1. INTRODUCTION

Blockchains have their roots in Bitcoin. After many attempts to create digital money, Nakamoto (2008) made a revolutionary proposal that resulted in the first cryptocurrency. The main breakthrough was that Bitcoin was completely decentralized, not requiring a central control responsible for keeping track of who owned every Bitcoin and, thus, putting too much power on it.

This is attained by implementing a distributed ledger, where all nodes participating in running the blockchain hold a copy of the ledger with all the Bitcoin transactions to date. This way, all blockchain nodes are responsible for controlling that no-one cheats, which is discouraged with an incentives system for those behaving properly, called mining rewards.

Second generation blockchains, like Ethereum (Buterin, 2014), move things one step further to create distributed ledgers that are not just capable of keeping track of currency payments, but also the transactions and current state of a shared computer. This shared computer is in fact replicated and run in every blockchain node to guarantee that it produces the same computations for everyone.

In this case, there are also application developers that can program this shared computer contributing pieces of code called smart contracts. They are contracts in the sense that it can be trusted that their code will execute as programmed. For instance, it is possible to develop an escrow payment application that does not require a trusted third party. There is guaranteed that the corresponding smart contract will make the payment if the escrow conditions are met.

Overall, the blockchain has the potential to change the ways that people and organizations trust each other, establishing a shared and tamper-proof registry of events that aims to be decentralized and neutral. This means a potential shift in money, law and government that those traditionally intermediating might perceive as a menace. Thinking even longer term, developing your own blockchain-based application you are not just making another application, it might evolve into a new form of society where humans and even machines can autonomously interact. For instance, self-driving cars that get paid and use the income to pay their energy consumption or repairs.

2. OBJECTIVES

To date, the core values that inspired blockchains design have been decentralization, transparency and neutrality. However, these values and intentions cannot be guaranteed just by the technical infrastructure alone and must be considered for each application built on top of existing blockchains. A decentralized computer network does not guarantee decentralized power, transparency does not guarantee legibility and finally, code and cryptography do not guarantee neutrality. Finally, it is important to assure that the use of blockchain technologies does not go against other values that, though no favored by the technology, should not be limited by them. For instance, transparency versus privacy.

Consequently, considering the big bias towards some specific human values, and against others conflicting, plus the enormous impact that blockchain technologies might have on our society (Tapscott & Tapscott, 2018), it is paramount to consider human values throughout the design process of blockchains and blockchain applications.

The objective of this work is to start exploring the application of Value Sensitive Design (VSD) as a way to ensure that human values are taken into account in these cases (Friedman & Hendry, 2019; Spiekermann, 2015). VSD builds on an iterative methodology that integrates conceptual, empirical, and technical investigations, which can be aligned with the development processes of information systems.

3. BLOCKCHAIN STAKEHOLDERS

Following VSD, conceptual investigations first identify the direct and indirect stakeholders affected by the considered technology. In the case of blockchain, we have made a literature review (e.g. the report by GetSmarter (2018) or the study by Nanayakkara et al. (2019) and compiled a list of stakeholders. Most of them are targeted and direct (as indicated next):

- **Miners (direct)**: run nodes looking for rewards for those that do not try to cheat. The way of proving their commitment might involve a costly task (proof of work) or require the deposit of an economic amount as a guarantee (proof of stake), among other approaches.
- **Core Developers (direct)**: create and define the evolution of the blockchains they are involved in by contributing to its codebase. For instance, they can change the rewards that miners receive or the costs of transactions that users should satisfy.
- **Entrepreneurs (direct)**: create applications on top of blockchains that benefit from its features, especially the trust mechanisms. Trust makes it possible to develop smart contracts, pieces of code that, once deployed, guarantee their execution. These

applications usually employ incentives like cryptocurrencies or tokens, which might also have economic value.

- **Investors (direct)**: buy cryptocurrencies and other tokens as an investment. They try to forecast the success of the associated blockchain or application, which might increase their demand and consequently their value.
- **Users (direct)**: employ blockchains to make cryptocurrency transactions or to use applications developed on top of blockchains, which might also include direct or indirect economic transactions but also other kinds of uses as registering agreements or voting.
- **Exchanges (direct)**: provide mechanisms to convert fiat currencies to the cryptocurrencies they have listed. Most of them are centralized and require that users move their holdings to accounts in the exchange. More recently and thanks to smart contracts, decentralized exchanges have also become available.
- **Key personalities and celebrities (indirect)**: are people that have influence in a particular blockchain community, or its associate cryptocurrency. This includes outstanding developers like the creators of some blockchains or celebrities from media that advocate in favor of particular cryptocurrencies or blockchain applications (Business Insider, 2019).
- **Regulators (indirect)**: are different kinds of organizations, public and private, that survey or regulate different kinds of economic and social systems which might be impacted by blockchain technologies. Examples of such organizations are those regulating financial systems, energy or taxes at different levels of granularity, from local to international level.

Finally, it is also possible to identify non-targeted stakeholders. Technologies are not always used in ways that the designers intended. Non-targeted stakeholders include those who might use the system for unplanned or malicious purposes. In this case, the most relevant ones are malicious hackers trying to steal assets managed using blockchain technologies, specially cryptocurrencies. Another kind of non-targeted stakeholder also very relevant and with high impact in the evolution of blockchain technologies are those using them for money laundering, including not just individuals but also organizations, for instance countries trying to circumvent trade restrictions.

4. BENEFITS, HARMS AND VALUES

Continuing with the VSD approach, we analyze the benefits and harms for the targeted stakeholders and then map them to the corresponding values using a deductive approach: human welfare, ownership and property, privacy, universal usability, trust, autonomy, informed consent, accountability or environmental sustainability (Friedman et al., 2013). The output of stakeholders analysis plus the identified benefits, harms and values are listed in Table 1.

Table 1 Mapping Blockchain Stakeholders' Benefits and Harms to Values.

Stakeholder	Benefits	Harms	Values
Miners (direct)	- Economic rewards. - Participating in the decentralization movement. - Enjoying additional privacy by interacting with the blockchain through an own node.	- Changes in rewards or costs (like electricity) might make mining not profitable. - Entry barriers, and risk of losing opportunities to earn rewards, due to the increasing investments required in computational resources or staked value because the chance of earning rewards is proportional to the commitment. - Environmental impact of mining when it is based on the intensive use of computational resources	- Human Welfare - Ownership and property - Privacy - Trust - Autonomy - Accountability - Environmental Sustainability
Core Developers (direct)	- Participating in the decentralization movement. - Public acknowledgement from the developer community, usually blockchains are open source projects to facilitate accountability and trust - Influencing the evolution of the blockchain or cryptocurrency ecosystem.	- Risk of losing the interest of miners or users that might abandon a blockchain and make it useless - Pressures from other stakeholders (including exchanges or key personalities and celebrities)	- Human Welfare - Ownership and property - Trust - Autonomy - Accountability
Entrepreneurs (direct)	- Participating in the decentralization movement. - Economic rewards from investors, including token offerings, or from users through utility tokens.	- High costs and risks of developing projects on top of a nascent technology with a lot of uncertainties - Complex technology imposes high entry barriers to potential users	- Human Welfare - Ownership and property - Universal usability - Trust - Autonomy - Accountability
Investors (direct)	- Participating in the decentralization movement, operating outside traditional and more restricted investment ecosystems - Investment returns are usually higher than other more mature markets.	- Higher risks than other more mature markets, including legal voids and potential scams	- Human Welfare - Ownership and property - Autonomy
Users (direct)	- Participating in the decentralization movement. - Economic incentives derived from cryptocurrencies and tokens earned as a reward for contributing to the application being used.	- Additional complexities introduced by an immature technology might produce economic harms - Risk of losing collected rewards if the economic volatility associated with the blockchain ecosystem makes them less valuable	- Human Welfare - Ownership and property - Privacy - Trust - Autonomy - Accountability - Informed Consent

Exchanges (direct)	- Economic profit from transaction fees. - Influencing the evolution of the cryptocurrency ecosystem, for instance choosing the currencies to be listed in the exchange. - Potential to reach a more diverse user base and reduced costs of operation	- Higher risks than other more mature markets, including legal voids and potential scams - Accumulation of value makes them very attractive to malicious hackers	- Ownership and property - Trust - Autonomy - Accountability
Key personalities and celebrities (indirect)	- Participating in the decentralization movement. - Participation in investments and other economic rewards related to cryptocurrencies and tokens.	- Higher risks than other more mature markets - Potential popularity harms due to legal or other kinds of issues associated to the blockchain or application being supported	- Human Welfare - Ownership and property
Regulators, financial systems, energy, etc. (indirect)	- Alternative mechanisms to regulate through incentives - Costs reductions - Facilitate the availability of banking and financial services	- Lack of control and enforcement measures over blockchain actors, pseudo-anonymous or outside jurisdiction - Higher risks than other more mature markets, volatility - Legal uncertainties	- Human Welfare - Ownership and property - Accountability

5. CASE STUDY

Following the previous analysis of Benefits, Harms and Values, we have studied a particular blockchain application based on smart contracts. The application is conveniently called EthicHub and geared towards becoming an ethical bridge of inclusion, as described in Figure 1.

Users can make investments that go directly, without intermediaries, straight to the involved farmers in developing countries, where access to credit to finance their farming activities is unavailable or at unaffordable rates. In many cases, these communities are not even banked.

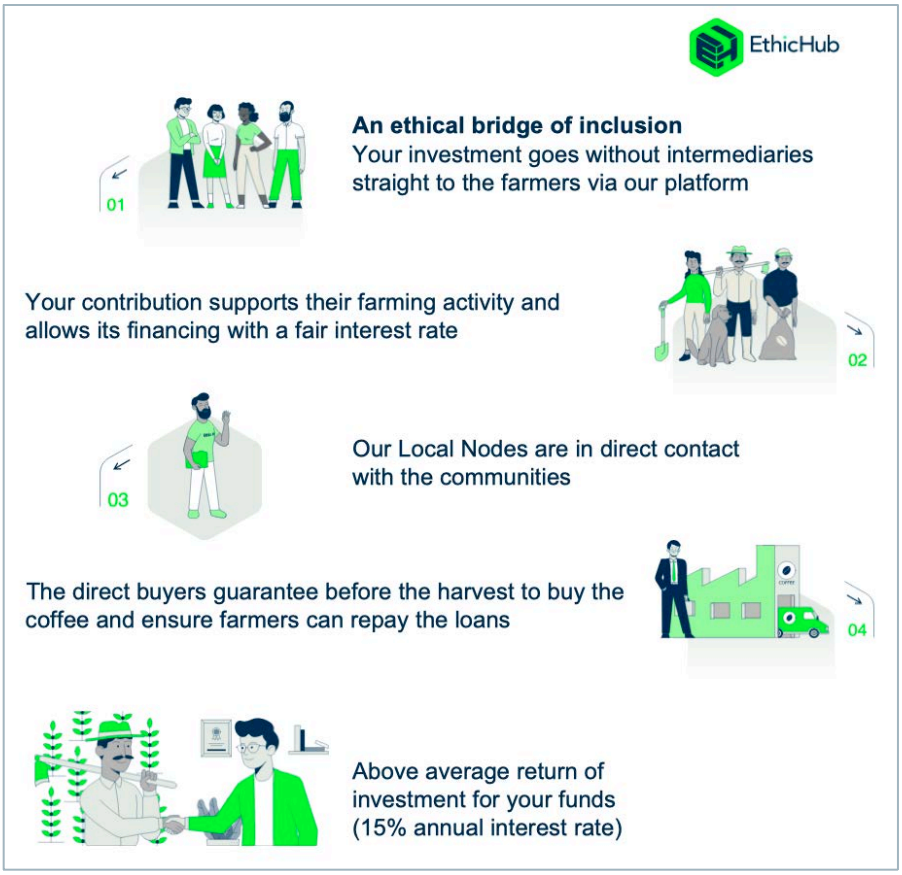
The contributions support their farming activities, as detailed in the platform, and allow their funding with a fair interest rate. There are EthicHub local nodes, persons that are in direct contact with the farming communities. They help communities define projects looking for funding, converting contributions made using blockchain assets into local currency, contacting direct buyers to guarantee purchase before the harvest to ensure farmers can repay the loans and, finally, returning the invested quantity including a 15% annual interest rate plus a 8% that goes to the platform and the local node.

The stakeholders in this particular case are:

- **Entrepreneurs**: they include all the EthicHub staff running the platform, which gets 4% of the investments, plus the local nodes, which also get another 4%. The values into play in this case are mainly Human Welfare, especially regarding wealth redistribution, plus Ownership and property.
- **Investors**: these are the users willing to invest in farming projects and looking for a 15% annual interest on the invested quantity, starting from just 20€. For investors the most relevant values are also Human Welfare plus Ownership and property, though they can also appreciate the Privacy that blockchain technologies provide them.

- **Users**: the farmers looking for funding for they farming projects at a very competitive interest rate compared to local options. In most cases, farmers are unbanked and cannot apply to commercial banks loans. Consequently, compared to the local alternatives that farmers have that can be of a 20% interest rate but monthly, it is a very convenient option that allow them to look for bigger and much longer-term project. For farmers, the target values are Human Welfare and Ownership and property. However, in this case, Autonomy is also very important.
- **Regulators**: this category of stakeholders includes the entities participating in the markets where the farmers operate, especially buyers of their harvest. Additionally, other entities like the local financial system should be also considered as EthicHub can be perceived as a competitor on the longer term. Finally, there are also the entities responsible for collecting taxes for the investors using EthicHub. As for other kinds of investments, it is likely that they will be willing to collect the corresponding taxes on the returns. This might impose the biggest value tension between the Privacy that investors through blockchain technologies and the Accountability requested by the regulators involved in this case. Currently, there is little regulation to this can of investments that occur outside traditional channels. On the longer term, is seems evident that EthicHub will need to provide mechanism to regulators to collect investments data while maintaining investors Privacy to the maximum extent possible, which in any case should not enable taxes evasion.

Figure 1. How does EthicHub Work?



An ethical bridge of inclusion
Your investment goes without intermediaries straight to the farmers via our platform

Your contribution supports their farming activity and allows its financing with a fair interest rate

Our Local Nodes are in direct contact with the communities

The direct buyers guarantee before the harvest to buy the coffee and ensure farmers can repay the loans

Above average return of investment for your funds (15% annual interest rate)

Source: EthicHub, https://ethichub.com (2020)

6. CONCLUSIONS AND FUTURE WORK

The previous study of stakeholders and values following the VSD approach highlights potential conflicts like accountability vs. privacy or trust vs. environmental sustainability. These are trade-offs among competing values in the design, implementation, and use of blockchain-based systems. For instance, blockchain technologies due to their immutability imply serious risks for privacy. From a VSD perspective, this issue is addressed during the whole blockchain application development process so it implements measures than ensure user privacy. For instance, store personal data on chain once encrypted or just a hash of it.

Remains future work to conduct further empirical investigations that help clarify the outcomes of different blockchains and applications regarding the identified values by exploring the corresponding white papers. The final target is to be able to characterize the properties and underlying mechanisms of blockchain technologies to generate a set of recommendations that make them and applications build on top of them better support human values, despite the initial bias the technology might impose.

REFERENCES

Business Insider. (2019). 13 celebrities who back cryptocurrency and may own millions in bitcoin. Available from: https://www.businessinsider.com/13-celebrities-who-back-cryptocurrency-and-may-own-millions-in-bitcoin-2019-1

Buterin, V. (2014). A next-generation smart contract and decentralized application platform. *White Paper*, 3, 37.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In: *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer, Dordrecht.

Friedman, B. & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: MIT Press.

GetSmarter. (2018). What stakeholders are involved in the blockchain strategy system? Available from: https://www.getsmarter.com/blog/career-advice/what-stakeholders-are-involved-in-the-blockchain-strategy-system/

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Nanayakkara, Samudaya & Perera, Srinath & Senaratne, Sepani. (2019). Stakeholders' Perspective on Blockchain and Smart Contracts Solutions for Construction Supply Chains. 10.6084/m9.figshare.8868386.

Spiekermann, S. (2015). *Ethical IT innovation: A value-based system design approach*. Auerbach Publications.

Tapscott, D., & Tapscott, A. (2018). Blockchain revolution: how the technology behind bitcoin and other cryptocurrencies is changing the world. *Portfolio*.

ONTOLOGIES AND KNOWLEDGE BASES: A NEW WAY TO REPRESENT AND COMMUNICATE VALUES IN TECHNOLOGY DESIGN

Kathrin Bednar, Till Winkler

Institute for Information Systems and Society,
Vienna University of Economics and Business (Austria)

kbednar@wu.ac.at; till.winkler@wu.ac.at

ABSTRACT

A promising pathway towards an ethically aligned design of technology is to consider human values such as "privacy" and "autonomy" in the design process. Value-oriented approaches have inspired the development of several unique methods for identifying stakeholders, values, and design requirements. However, none of these methods focus solely on the representation of values and the communication of value knowledge. In this paper, we outline a methodology that is inspired by techniques from the semantic web community. Based on a case study, we demonstrate how building an ontology can be used to represent and visualize value knowledge. The underlying empirical data comes from a sample of students who applied Value-based Engineering to elicit and analyse values for a telemedicine communication system. In spite of the specifics of this case study, the techniques for representing and communicating the resulting value data are compatible with any value-oriented approach. Furthermore, they support quality criteria that are essential when dealing with values both in research and design. The formal representation of value knowledge in form of semantic data ensures a high level of detail while respecting context-specificity. The underlying ontology helps to represent key concepts and their relations and supports the transparency of data analysis, including the initial coding of the value-related data. The resulting knowledge base can be shared with stakeholders and researchers, supporting the joint evolution of value-oriented approaches and technology.

KEYWORDS: values, design, engineering, ontology, knowledge base, semantic web.

1. INTRODUCTION

Since the advent of the internet, the variety of devices and applications has kept increasing. Information technology helps us to structure and organize our everyday life, but also shapes our work and social lives. Traditionally designed technological products have been optimized mainly for functionality, ignoring that functionality also depends on non-functional characteristics (Chung & do Prado Leite, 2009). Ignoring high-level non-functional characteristics during the design process can lead to harmful effects with ethical implications, such as information distortion in the form of search engine manipulations (Epstein & Robertson, 2015), filter bubbles (Pariser, 2011), and algorithm biases (O'Neil, 2016). Consequences often play out at the societal level, as in the example of social media's impact on democracy (Cadwalladr, 2017), but there are also physiological effects for individuals, which can be observed in the form of symptoms of stress and depression (Barley, Meyerson, & Grodal, 2011). Technology is a mediator for biases and human values, it moulds its use context and changes the perceptions and actions of people to the point where it creates new practices and forms of living (Verbeek, 2008). In this view, negative or positive effects do not emerge as a result of technology use, but are triggered by the affordances *inherent* in technological artefacts, systems, and infrastructures (van den Hoven, 2017).

For this reason, designers, researchers and engineers need to address the potential effects and consequences of a technology throughout its design and development process. This is especially important, as many engineers are willing to go beyond traditional functional requirements but do not have the necessary time or autonomy to implement them (Bednar, Spiekermann, & Langheinrich, 2019; Spiekermann, Korunovska, & Langheinrich, 2018). This situation might change if the consideration of values is incorporated into the design and development process of technologies as for instance Value-based Engineering requires developers to take the necessary time to think about values (Spiekermann & Winkler, 2020). A promising pathway towards an ethically aligned design of technology is to consider human values such as “wellbeing”, “privacy”, “security”, and “autonomy”.

Considering human values during system development minimizes biases by making the system more accessible to a greater diversity of users, leads to more desirable software, and increases the likeliness of new technology to be adopted (Friedman & Nissenbaum, 1996; Isomursu, Ervasti, Kinnula, & Isomursu, 2011; Spiekermann, 2016). While there is a diverse landscape of value-oriented approaches and methods (e.g. Friedman & Hendry, 2019; Spiekermann, 2016), a common framework that focuses on their commonalities is still missing. However, a common way to represent value knowledge (e.g. in form of lists or networks) could lead to a better understanding of value knowledge gained in a value-orient project and make it easier to share this knowledge *among* team members as well as *across* different value paradigms. Ideally, such a framework would support quality criteria such as transparency and preserve context-specificity, e.g. by including information on the technology under investigation or the affected stakeholders. Considering challenges of value-oriented design processes such as keeping a high level of detail when representing original data throughout the design process, such a framework could benefit the further development of value-oriented approaches. Improving the capability to maintain quality criteria could even increase the recognition of value-oriented design outside academia (Detweiler & Harbers, 2014; Miller, Friedman, Jancke, & Gill, 2007).

To the best of our knowledge, there is no method that focuses solely on the representation and communication of values independent from the underlying theoretical background (e.g. the definition of values) or the specific method used (e.g. for eliciting values). In this paper, we propose that ontology-building and the development of a knowledge base, techniques from the semantic web community, can step in here. In the following, we take a closer look at the challenges of value-oriented design processes. Then, we explore the different phases that a value-oriented project would need to run through to develop an ontology and build a knowledge base that can be shared with project members and other researchers. We present and discuss a case study on a telemedicine communication system to illustrate these steps and discuss benefits and future potentials of this method. While the case study dataset resulted from the elicitation and analysis of values in accordance with Value-based Engineering, the proposed methodology for value representation and communication is compatible with any value-oriented approach.

2. CHALLENGES OF VALUE-ORIENTED APPROACHES TO TECHNOLOGY DESIGN

The most prominent approach for considering human values during technology development is called *Value Sensitive Design* (VSD; Friedman & Nissenbaum 1996, Friedman et al. 2006, Friedman & Hendry, 2019). VSD was first conceptualized twenty-four years ago and has advanced ever since. In their recent book, Friedman and Hendry (2019) present unique methods that have been applied as part of the VSD's iterative tripartite methodology, i.e. for 1) conceptual, 2) empirical, and 3) technical investigations. The 17 methods cover value elicitation, value analysis, value source or stakeholder identification as well as various other purposes. *Value-based Engineering* (Spiekermann, 2016; Spiekermann & Winkler, 2020) has developed from the same motivations as VSD, but proposes a different methodology. In this approach, three different ethical theories (utilitarianism, virtue ethics and deontology) are applied in the value elicitation phase to identify values that are ethically salient for a specific product and context. It then conceptualizes these values and concretizes them in a technical analysis that is either iterative or risk-assessment based. Both value-oriented approaches share core ideas, such as the integration of direct and indirect stakeholders into the design and development process, the appreciation of values to go beyond functionality, the envisioning of long-term effects, and the consideration of context. At the same time, they also differ in some respects, for example, in how they understand values.

The concept of values is generally difficult to define. In psychology, values are considered to represent desirable behaviours, end states or transitional goals, and the source of a person's self-esteem (Pereira & Baranauskas, 2015; Schwartz, 1994; Verplanken & Holland, 2002). In this view, the relative importance of values depends on the person's culture, socioeconomic status, and practical context (Verplanken & Holland, 2002). The VSD community commonly refers to values as "what a person or group of people consider important in life" (Friedman, Kahn Jr., et al., 2006) with a focus on morality and ethics (Friedman & Hendry, 2019). Value-based Engineering, on the other hand, builds on the philosophical understanding of values developed by Material Ethics of Value (Hartmann, 1932; Scheler, 1913-1916/1973), which understands values as *ought-to-be* principles that should generally guide behaviour. Several scholars in different theoretical contexts have come up with lists of values to support an exemplary understanding of the concept of values (e.g. Friedman, Kahn, Borning, & Huldtgren, 2013; Winkler & Spiekermann, 2019). While value lists can provide a helpful resource for incorporating ethical considerations into technology design, but the cultural and subjective variety of values challenge the completeness of any such list. Also, a list does not provide a solution for the selection of the most relevant values that need to be taken into account for a certain technology and its specific context. Therefore, the acknowledgement of the context-specificity of values has formed a key characteristic of value-oriented projects, which usually start with the identification of values for a specific technology and its context.

While the different methodological and theoretical value frameworks might lead to the identification of similar values for a specific technology, they still influence how relevant values are selected. Thus, the underlying understanding of the concept of values as well as criteria for the elicitation and selection of values need to be made transparent. A transparent process also helps to avoid known challenges in qualitative research, including confirmation biases, culture biases and other cognitive biases (Kahneman, Slovic, & Tversky, 1982; Plous, 1993), which can endanger the validity, reliability and value consciousness of value-oriented projects. Especially since the roles of the designer, value prioritizer, interpreter, reporter and conflict solver are often subsumed in one person, keeping track of the value analysis process is important to avoid a power discrepancy among the actual affected stakeholders and those working on the data (Borning & Muller, 2012). This power discrepancy in combination with a lack of transparency could lead to resistance from managers, engineers, and designers involved in the development of the technological product and thus to a failure of the whole value-oriented project. This, in turn, requires an extremely careful treatment of value data and detailed documentation to ensure transparency.

Human values need to be discussed in detail to explore different contexts, interpretations and nuances (Steen & van de Poel, 2012). Building on rich and versatile value knowledge also helps in communicating values (Pommeranz, Detweiler, Wiggers, & Jonker, 2012) and most importantly in understanding their true meaning. For example, considering “freedom from harms” as a general definition of “security” does not provide enough information on what specifications or requirements a product needs to fulfil. Security could refer to the protection of private data using encryption as well as to the protection of a private estate using video surveillance. This necessary level of detail for (context) information poses another challenge for the representation and communication of values during the development of a technical product.

Rich coding manuals have been developed to preserve context-specific information throughout the analyses (e.g. Friedman, Kahn, Hagman, Severson, & Gill, 2006; Hendry, Abokhodair, Kinsley, & Woelfer, 2017). Such a diligent data analysis procedure is essential, but a value-oriented design project that focuses on the integration of stakeholder perspectives ideally supports transparency throughout the whole design process. This is already time-intensive work for small scale projects and becomes more and more difficult to sustain in larger value-oriented projects. In large-scale projects, the amount of generated value content coming from numerous stakeholders and potentially numerous methods can be enormous. Additionally, the content needs to be updated and extended constantly due to the iterative nature of value-oriented design processes, through which new insights are continuously analysed and validated with stakeholders. These dynamics make it difficult to apply value-oriented design in large industry projects, where high-quality technology development can only be achieved through a transparent and traceable product design and development process. In summary, value-oriented approaches face several challenges. First, there are various ways to define values and related concepts, leading to potentially different selection criteria across individuals and teams. Second, the aim to represent different stakeholder perspectives puts a lot of responsibility onto those involved in the selection and analysis of original value data, which can lead to biases and problematic power discrepancies. Third, values are always bound to specific contexts, and this context-specificity needs to be preserved in every step of analysis. Fourth, the consideration of multiple stakeholder perspectives and the iterative nature of value-oriented approaches requires constant extension of value knowledge, leading to enormous and volatile datasets, which form the basis for representing and communicating values and associated value knowledge. In a recent case study (Spiekermann-Hoff, Winkler, & Bednar, 2019), we have encountered the challenges enlisted above. This inspired us to look for a solution, which we believe can be found in building ontologies and knowledge bases. After introducing the case study, we present a methodology that allows to define value concepts, fill them with rich datasets, and track changes throughout the design process.

3. A NEW WAY OF REPRESENTING AND COMMUNICATING VALUES

For a representation of gathered value knowledge that is independent from the underlying theoretical understanding of values and specific methods of e.g. value elicitation, we borrow techniques from the semantic web community. The semantic web is an extension of the current web that aims at converting unstructured and semi-structured information into information of which the underlying semantics are expressed in a formal machine-understandable way (W3C, 2014).

Within the vision of the semantic web, ontologies play a key role in providing formally defined terms for describing resources in an unambiguous manner. The term *ontology* describes a form of semantic knowledge representation (Ehrlinger & Wöß, 2016). An ontology is an explicit description of concepts (or *classes*) and their properties within an area of interest. In our case, the area of interest comprises human values for a specific technology context and their relations among each other as well as with the stakeholders. Ontologies can be used flexibly to define concepts and relations, but also allow the definition of constraints (e.g. a value being relevant only for the affected stakeholders is a constraint). Once an ontology is "filled" with individual instances (i.e. specific values for a specific technology context), a *knowledge base* is formed (Noy & McGuinness, 2001). The information contained in the knowledge base can be visually represented with graphs, which connect single concepts (or *nodes*).

Expressing the semantics of value knowledge is necessary in order to preserve the connections and avoid the shortcomings of descriptive data analysis. The *Resource Description Framework* (RDF) is a domain-independent data model that expresses information with a specific vocabulary, for instance as a RDF Schema (RDFS; Antoniou et al. 2012). The smallest entity in the RDF is an *RDF statement*, also referred to as *semantic triple* (W3C, 2014) as it consists of three elements: the subject, the object and the predicate. An RDF statement expresses a relation between the subject and the object and the predicate represents the nature of their relationship. The *property* denotes the relationship between the subject and the object (W3C, 2014). Representing knowledge with RDF statements supports the reuse and expansion of knowledge (Antoniou et al., 2012).

3.1. Case study: A telemedicine communication system

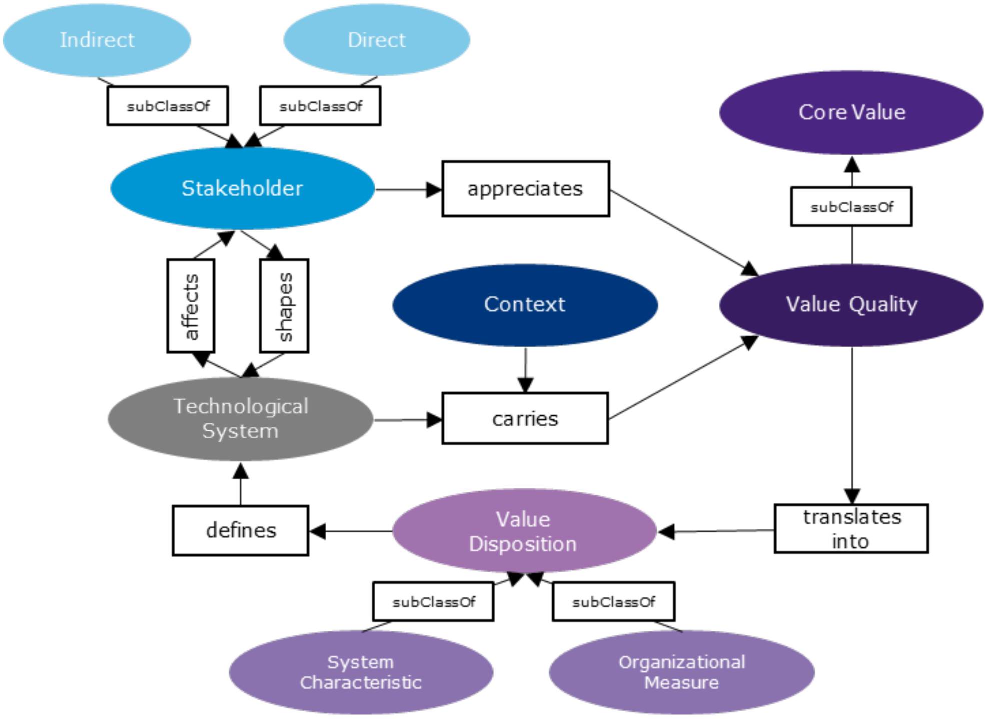
Our case study is based on an idea for a telemedicine communication system, which was analysed in accordance with Value-based Engineering (for a detailed description of the case study, see Spiekermann-Hoff, Winkler, & Bednar, 2019). This online communication system connects patients to a general practitioner (GP) who first records patients' medical history and symptoms and then refers them to a specialized doctor who was recommended by other doctors. Several values can immediately be identified with the envisioned beneficial effects of this IT product, e.g. "health". However, the underlying recommendation system and the telecommunication system also raise ethical issues, especially as they are used in a medical context. Consider, for example, the underlying motivation of recommendations among doctors (who might know each other) and the fact that a digital platform does not allow physical interaction, which could influence the GP's decision and undermine values such as "fairness" and "accuracy". Because of these important ethical implications, the envisioned system fits perfectly as a case study for a value-oriented project.

The value elicitation was conducted by 35 students (age: *M* = 24.56, *SD* = 2.61, 38.2% female, 14 different nationalities) that formed teams of two. 13 participants were female (38.2%) and 21 male (61.8%; 1 missing value). All participants were students at the Vienna University of Economics and Business. Value-related data was gathered following the Value-based Engineering approach (Spiekermann, 2016; Spiekermann & Winkler, 2020), which deploys three ethical theories to elicit values: consequentialism, virtue ethics, and deontology. Once participants had identified relevant values, they were asked to come up with design ideas to further foster beneficial value effects and to avoid negative effects.

3.2. Building an ontology for value representation

For building an ontology for Value-based Engineering, we followed steps and guidance provided by Noy and McGuinness (2001). Figure 1 shows a visualisation of the resulting ontology with the most important terms and their relations.

Figure 1. Exemplified ontology for Value-based Engineering.



As a first step, we determined that the main domain of this ontology is the representation of value knowledge for Value-based Engineering. This scope definition has several consequences for deciding on appropriate terms (step 2) and defining classes and class hierarchies (step 3).

In step 2, we accumulated appropriate terminology from the Value-based Engineering literature (Spiekermann-Hoff et al., 2019; Spiekermann, 2016; Spiekermann & Winkler, 2020), literature on material value-ethics (Hartmann, 1932; Scheler, 1913-1916/1973) and value lists (Winkler & Spiekermann, 2019). This resulted in numerous important terms, including *core value, value quality, indirect stakeholder, direct stakeholder, affects, appreciates,* or *system characteristics*.

For the third step, we mainly used a top-down approach, by first defining the most general concepts (or *classes*) and then further defining them with sub-concepts (or *sub-classes*). These classes need to be able to describe subjects and objects in an unambiguous manner (Ye et al., 2015), which is a challenge as many terms from step 2 are inherently ambiguous. For instance, according to material value-ethics, a single value can be at the same time a *core value* and a *value quality*. The value "security", for example, can be a value on its own, but also a *value quality* of the value "privacy". We solved this by allowing the attribution of both classes for one value, i.e. "security" can be defined as a value quality and as a value. In our ontology we used *core value*, *stakeholder*, and *value disposition* as main classes and *value quality*, *indirect stakeholder, direct stakeholder, system characteristic,* and *organizational measure* as subclasses.

The fourth step in ontology development is the definition of class relations and class properties. Class relations are described in the Value-based Engineering literature with terms such as *affects, shapes, defines, carries, appreciates* and *translates into*. Class properties in a design case could be, for instance, the *degree of availability*, *social background*, *mean age*, *gender*, or *degree of importance* for a certain stakeholder. Such information is not included in the illustrated example, but can easily be added. The same goes for the type definition of a property, which forms the fifth step. For example, *mean age* would be defined as a *number* here.

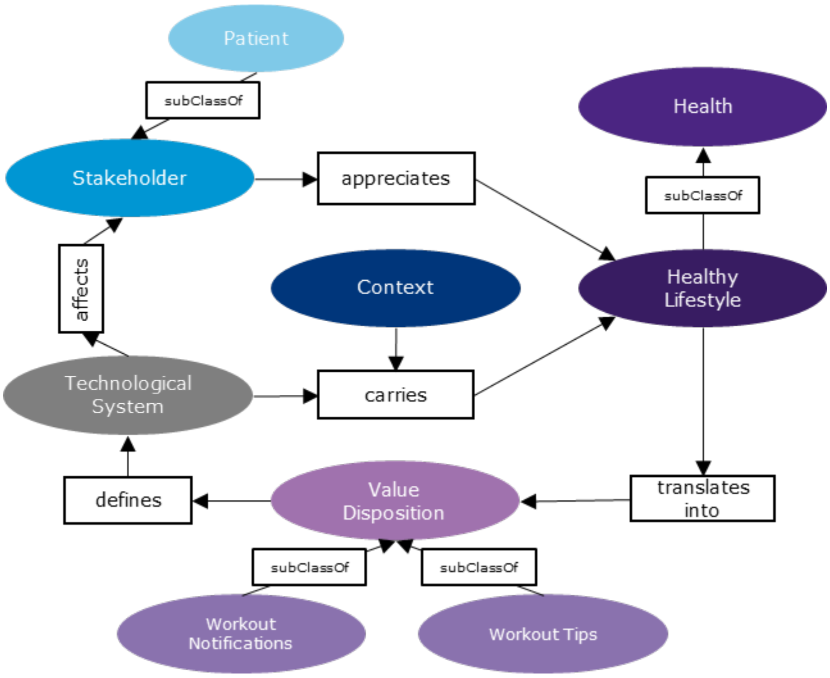
The last step, creating instances by a) choosing a class, b) creating an individual instance of that class and b) filling in the type definition, was achieved during coding. Building the underlying ontology helps to make assumptions about the data explicit and can guide the qualitative analysis of the original value data. Ontology development is necessarily an iterative process, which means that a more detailed differentiation of classes, their relation and properties can be achieved when the ontology is filled with specific instances.

3.3. Formulating instances and RDF statements

We prepared all value-related data for the formulation of machine-readable RDF statements or triples, which fill the structure of the ontology, i.e. the defined classes, relations, and properties, with specific instances. Figure 2 visualises the value "health" and related instances for the predefined ontology.

For formulating RDF statements, the data resulting from the value elicitation phase had to be divided into the defined classes (e.g. *core values* or *stakeholders*). This process equals the coding of qualitative data in any research project, as we summarized all value-related ideas, checked their logical structure and named them adequately. To produce RDF statements, we put the coded data into a *subject* – *predicate* (relation) – *object* structure. First, we developed coding rules and formulated examples, as it is standard when developing a coding manual. As a second step, a subset of the original dataset was coded and transformed into RDF statements by two independent coders. We found an inter-coder agreement of 80%, including triples that were completely identical and triples that represented the same entities with a slightly differing wording (example: coder 1: "Leaked doctor information", coder 2: "Information of doctor is leaked"). Afterwards, the coding manual was improved and the whole dataset was coded by coder 1. Coder 2 analysed the resulting semantic triples and checked their logical structure.

Figure 2. Exemplified instances related to “health” for the telemedicine case study.



The resulting RDF statements can be combined into a value knowledge base, which represents the original value-related data following the structure of the predefined ontology. This knowledge base maintains the relations suggested in the original data, but can be formulated in a machine-readable way, providing the basis for further design steps. Triples can also be visualized as connected *graphs*, consisting of nodes (representing subjects and objects) and arcs (representing the predicate; W3C 2014). Additionally, node (or class) *properties* can be displayed visually. The open-source software developed for network exploration and manipulation “Gephi” (Bastian, Heymann, & Jacomy, 2009) allows an initial visualization and exploration of these graphs and includes and an adequate spatial visualization through its included “ForceAtlas2” layout algorithm (Jacomy, Venturini, Heymann, & Bastian, 2014). The visual representation is especially powerful, as it makes the most frequent concepts and relations immediately apparent through the relative size of the nodes and the thickness of the arches. An interactive platform that facilitates such visualisations would be an especially powerful contribution to an effective communication among teams and stakeholders.

4. DISCUSSION

The benefits of ontology-building and the development of knowledge bases come from the representation of data in machine-readable form, which supports a structured representation of concepts and their relation. Even for large-scale value-oriented projects, this methodology can offer a common framework for creating high-quality value knowledge. It also facilitates the communication of such knowledge among stakeholders, teams, and across different approaches. The methodology inspired by techniques from the semantic web community offers ways to deal with several challenges in value-oriented design projects.

First, different ways of defining key concepts such as values and stakeholders can lead to different selection criteria across individuals, teams, and methods. Building ontologies helps to share information, analyse and reuse knowledge, and to make assumptions explicit (Noy & McGuinness, 2001). A common knowledge base supports the sharing of gained value knowledge in specific projects (as in our telemedicine case study) while making it possible for every approach to maintain and express its own theoretical foundation (e.g. defining value qualities related to core values). Thus, it allows the representation and communication of values independent from the underlying theoretical background and specific methods used. This supports that value-oriented projects build upon and learn from previous projects more easily. Ontologies can also help to make relations between key concepts explicit (e.g. value qualities are carried by the technological system) and thus support the coding of original value-related data (e.g. by defining value qualities and core values). From an engineering perspective, formally defined terms and classes can make the fuzzy concept of values more tangible, which might encourage a wider adoption of value-oriented approaches.

Second, the aim to represent different stakeholder perspectives puts a lot of responsibility onto those involved in the selection and analysis of original value data. The presented methodology supports transparency of this process by providing detailed information (e.g. by indicating the stakeholders for whom a value quality is important). Representing value knowledge in a formal way also makes it easier to track changes, e.g. when introducing a category in the coding process, and does so in a machine-readable form, which can be queried, visually represented, and explored by stakeholders. Enabling transparency and the exploration of existing value knowledge by any stakeholder can help to decrease power discrepancies (Borning & Muller, 2012) and potential biases (Kahneman et al., 1982; Plous, 1993). Still, researcher values should always be made explicit and considered during the design process (Steen & van de Poel, 2012).

Third, human values are bound to specific contexts, and need to be discussed in detail to explore different contexts, interpretations and nuances (Steen & van de Poel, 2012). This context-specificity needs to be maintained in every step of analysis. The methodology we propose produces coded data that is semantically coherent, that is, represents the underlying logical structure in the form of semantic triples (e.g. when expressing that *patients – appreciate – high quality medical service*). An initially specified set of rules for the formulation of semantic triples supports the completeness of value data and secures important information such as the context of meaning.

Fourth, the consideration of multiple stakeholder perspectives, the iterative nature of value-oriented approaches, and the context-specificity of values leads to enormous datasets that need to be updated constantly. This is especially challenging for large-scale projects. Drawing definitions from a sound theoretical background and following a pre-defined set of rules already form the basics of good research practices in social sciences. But with increasing complexity of a dataset, it becomes more and more difficult to keep track of changes. The machine-readable form of RDF statements supports the digital handling of data and could thus provide a solution here. Building ontologies and knowledge bases could be especially beneficial to the handling of date in large-scale value-oriented projects and increase the recognition of value-oriented approaches outside of academia.

This paper wants to offer an inspirational starting point for making value knowledge explicit, transparent, and accessible to all stakeholders. As this paper presents only first experiences in utilizing semantic web techniques for value-oriented data analysis, the results we present for the case study are only rudimentary. We also acknowledge that methods that produce less structured data might be more difficult to translate into semantic triples. Still, we hope to inspire future value-oriented projects to build more elaborate ontologies and thus to improve the sharing of knowledge among stakeholders, teams, and different approaches. Future work in this area could also explore new ways of value knowledge discovery, e.g. by including query functions, filter mechanisms, interactive visualisation or developing value knowledge patterns, counting triples, nodes, and relations (Presutti et al., 2011).

5. CONCLUSION

In this paper, we show that ontology-building and the development of a knowledge base, methods from the semantic web community, facilitate the representation and communication of values independent from the underlying theoretical framework and the specific method used to acquire value data. At the same time, the proposed methodology supports essential quality criteria for value-oriented projects. The formal representation of value knowledge in form of semantic triples ensures a high level of detail while respecting the context-specificity of any information. The underlying ontology helps to represent key concepts and their relations and supports the transparency of data analysis, including the initial coding of the value-related data. Furthermore, the resulting knowledge base can be shared with stakeholders and researchers, supporting the joint evolution of value-oriented approaches and technology. The case study of a telemedicine communication system shows the steps that a value-oriented project would need to run through to develop an ontology, which can be extended it into a knowledge base to be shared among stakeholders, teams, and across different approaches. Future value-oriented projects could apply this methodology to jointly build an extensive value knowledge base and experiment with more advanced applications such as data query and interactive visualisations.

ACKNOWLEDGEMENTS

We want to thank our PhD supervisor Sarah Spiekermann for supporting us in pursuing this project. Also, Oliver Burmeister, Batya Friedman, David Hendry and the international VSD community have provided valuable feedback and comments, which we highly appreciate.

REFERENCES

Antoniou, G., Groth, P., Harmelen, F. van, & Hoekstra, R. (2012). *A semantic web primer* (3rd ed.). Cambridge, MA: MIT Press.

Barley, S. R., Meyerson, D. E., & Grodal, S. (2011). E-mail as a source and symbol of stress. *Organization Science*, *22*(4), 887–906. https://doi.org/10.1287/orsc.1100.0573

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Bednar, K., Spiekermann, S., & Langheinrich, M. (2019). Engineering Privacy by Design: Are engineers ready to live up to the challenge? *The Information Society: An International Journal*, *35*(3), 122–142. https://doi.org/10.1080/01972243.2019.1583296

Borning, A., & Muller, M. (2012). Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)* (pp. 1125–1134). ACM. https://doi.org/10.1145/2207676.2208560

Cadwalladr, C. (2017, May). The great British Brexit robbery: how our democracy was hijacked. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

Chung, L., & do Prado Leite, J. C. S. (2009). On non-functional requirements in software engineering. In *Conceptual modeling: Foundations and applications* (pp. 363–379). Berlin, Heidelberg: Springer.

Detweiler, C., & Harbers, M. (2014). Value stories: Putting human values into requirements engineering. *CEUR Workshop Proceedings*, *1138*, 2–11.

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. In *SEMANTiCS 2016*. https://doi.org/10.1007/978-1-349-19066-9_2

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences (PNAS)*, *112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping technology with moral imagination*. Cambridge, MA: MIT Press.

Friedman, B., Kahn Jr., P. H., & Borning, A. (2006). Value sensitive design and information systems. In P. Zhang & D. Galletta (Eds.), *Human-computer interaction and management information systems: Foundations* (pp. 348–372). Armonk, NY: M.E.Sharpe.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Springer.

Friedman, B., Kahn, P. H., Hagman, J., Severson, R. L., & Gill, B. (2006). The watcher and the watched: Social judgments about privacy in a public place. *Human-Computer Interaction*, *21*(2), 235–272. https://doi.org/10.1207/s15327051hci2102_3

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Hartmann, N. (1932). *Ethics*. London: George Allen & Unwin.

Hendry, D. G., Abokhodair, N., Kinsley, R. P., & Woelfer, J. P. (2017). Homeless young people, jobs, and a future vision: Community members' perceptions of the Job Co-op. *ACM International Conference Proceeding Series*, *Part F1285*, 22–31. https://doi.org/10.1145/3083671.3083680

Isomursu, M., Ervasti, M., Kinnula, M., & Isomursu, P. (2011). Understanding human values in adopting new technology: A case study and methodological discussion. *International Journal of Human-Computer Studies*, *69*, 183–200. https://doi.org/10.1016/j.ijhcs.2010.12.001

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, *9*(6), e98679.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Miller, J., Friedman, B., Jancke, G., & Gill, B. (2007). Value tensions in design: The value sensitive design, development, and appropriation of a corporation's groupware system. *Proceedings* *of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*, 281–290. https://doi.org/10.1145/1316624.1316668

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Retrieved from http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Pariser, E. (2011). *Filter bubble: What the Internet is hiding from you*. New York: The Penguin Press.

Pereira, R., & Baranauskas, M. C. C. (2015). A value-oriented and culturally informed approach to the design of interactive systems. *International Journal of Human Computer Studies*, *80*, 66–82. https://doi.org/10.1016/j.ijhcs.2015.04.001

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.

Pommeranz, A., Detweiler, C., Wiggers, P., & Jonker, C. (2012). Elicitation of situated values: Need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, *14*, 285–303. https://doi.org/10.1007/s10676-011-9282-6

Presutti, V., Aroyo, L., Adamou, A., Schopman, B., Gangemi, A., & Schreiber, G. (2011). Extracting core knowledge from Linked Data. In *Proceedings of the Second International Conference on Consuming Linked Data* (Vol. 782, pp. 37–48). CEUR-WS.

Scheler, M. (1973). *Formalism in ethics and non-formal ethics of values: A new attempt toward the foundation of an ethical personalism [1913-1916]*. (M. S. Frings & R. L. Funk, Eds.). Evanston, Ill: Northwestern University Press. https://doi.org/10.2307/2707101

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, *50*(4), 19–45. https://doi.org/10.1111/j.1540-4560.1994.tb01196.x

Spiekermann-Hoff, S., Winkler, T., & Bednar, K. (2019). *A telemedicine case study for the early phases of value based engineering* (Working Paper Series/Institute for IS & Society No. 001_vs 1). *Working Paper Series/Institute for IS & Society*. Vienna: WU Vienna University of Economics and Business. Retrieved from https://epub.wu.ac.at/7119/

Spiekermann, S. (2016). *Ethical IT innovation: A value-based system design approach*. Boca Raton: CRC Press.

Spiekermann, S., Korunovska, J., & Langheinrich, M. (2018). Inside the organization: Why privacy and security engineering is a challenge for engineers. *Proceedings of the IEEE*, *107*(3), 600–615. https://doi.org/10.1109/JPROC.2018.2866769

Spiekermann, S., & Winkler, T. (2020). Value-based Engineering for Ethics by Design. *ArXiv*, *2004.13676*. Retrieved from https://arxiv.org/abs/2004.13676

Steen, M., & van de Poel, I. (2012). Making values explicit during the design process. *IEEE Technology and Society Magazine*, *31*(4), 63–72. https://doi.org/10.1109/MTS.2012.2225671

van den Hoven, J. (2017). Ethics for the digital age: Where are the moral specs? In *Informatics in the future: Proceedings of the 11th European Computer Science Summit (ECSS 2015), Vienna,* *October 2015* (pp. 65–67). Cham: Springer Nature. https://doi.org/10.1007/978-3-319-55735-9

Verbeek, P.-P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In S. A. M. Pieter E. Vermaas, Peter Kroes, Andrew Light (Ed.), *Philosophy and design: From engineering to architecture* (pp. 91–103). Springer Science + Business Media. https://doi.org/10.1007/978-1-4020-6591-0_7

Verplanken, B., & Holland, R. W. (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology*, *82*(3), 434–447. https://doi.org/10.1037/0022-3514.82.3.434

W3C. (2014). RDF 1.1 Primer. Retrieved from https://www.w3.org/TR/rdf11-primer/#section-triple

Winkler, T., & Spiekermann, S. (2019). Human values as the basis for sustainable information system design. *IEEE Technology and Society Magazine*, *38*(3), 34–43. https://doi.org/10.1109/MTS.2019.2930268

Ye, J., Dasiopoulou, S., Stevenson, G., Meditskos, G., Kontopoulos, E., Kompatsiaris, I., & Dobson, S. (2015). Semantic web technologies in pervasive computing: A survey and research roadmap. *Pervasive and Mobile Computing*, *23*, 1–25. https://doi.org/10.1016/j.pmcj.2014.12.009

UNIVERSALITY OF HOPE IN PATIENT CARE: THE CASE OF MOBILE APP FOR DIABETES

Majid Dadgar, K.D. Joshi

University of San Francisco (USA), University of Nevada (Reno, USA)

mdadgar@usfca.edu; kjoshi@unr.edu

ABSTRACT

In this paper we investigate the human value of hope in the self-management systems used by the patients with diabetes to manage their chronic health conditions. We use value sensitive design (VSD) framework to uncover the value instances revealed in our interviews with patients with diabetes. The value instances identified in the interview transcript map to components of hope theory: goal, agency, and pathways. We recommend technology features that allow patients with diabetes to achieve their goals in life while managing their chronic conditions.

KEYWORDS: Value Sensitive Design, Hope, Self-management, Healthcare, Diabetes, Agency, Pathways.

“Hope” is the thing with feathers -
That perches in the soul -
And sings the tune without the words -
And never stops - at all -

I’ve heard it in the chillest land -
And on the strangest Sea -
Yet - never - in Extremity,
It asked a crumb - of me.

- Emily Dickinson

1. INTRODUCTION

As information and communication technologies (ICTs) advance, their uses and applications become more diverse and complex. These complex technologies are designed and used by humans and therefore, need a human-centric approach. The human-centric ICTs in the healthcare context play a major role in improving patients' lives (Bardhan, Chen, & Karahanna, 2017). These ICTs should be sensitive to the values of the patients (Dadgar & Joshi, 2018).

The value sensitive design (VSD) framework has proven to be an effective tool in identifying and explaining the human values of technology users and their development and change over time (Friedman, Howe, & Felten, 2002). In this paper we investigate the value of hope in the patients with diabetes. Specifically, we identify the instances of the value of hope for the patients with diabetes and recommend technology features that could support them.

2. HOPE AND SELF-MANAGEMENT

Hope in the theory of hope is defined as the perceived capability to derive pathways to desired goals, and motivate oneself via agency thinking to use those pathways (Snyder, 2000). Setting and attainment of goals are central in how the construct of hope is conceptualized by Snyder. People have higher hope when they believe their goals are attainable. Pathways to desired goals are necessary for hopeful thoughts. People who can realize and pursue pathways toward their desired goals stay hopeful over time. The sense of agency in achieving their goals through purposeful pathways motivates and empowers patients. The agency and pathway components of hope are distinct but entangled. At difficult times when people face barriers towards their goals, the strong sense of agency enables them to tackle the barriers. Positive and negative emotions are the result of the perceived success in achieving goals. Perceived success in achieving goals creates positive emotions in people and perceived failure triggers negative emotions.

We investigate how these components of hope theory can be supported using ICTs. We use VSD to identify value instances of hope for the patients with diabetes who use mobile app to self-manage their chronic conditions. The value instances identified in the interviews bridge the support needed from hope interventions implicated in technology features (see Table 1).

3. METHOD

We have used VSD to develop interview strategies and criteria that will reveal the values of the patients with diabetes (Friedman & Hendry, 2019). We interviewed 20 patients with diabetes. In the first meeting, patients were introduced to a mobile app that they could use to manage their diabetes, its symptoms, and life style changes. After the first meeting, patients used the mobile app to manage their diabetes for one week. In the second meeting, patients were interviewed about their experience with the diabetes mobile app and their needs and concerns. Interviews were transcribed and analyzed based on VSD to identify value instances that map to the hope components. Next we make recommendations that how technology can support these values instances and hope components necessary to create and maintain hope in the patients with chronic diseases and conditions.

4. RESULTS AND DISCUSSION

Hope instances identified and extracted from interviews with patients with diabetes illustrate how this value manifests in different variations in patients' lives. These value instances could be supported effectively by technology features. The hope components with one example of value instance and technology features are provided in Table 1.

Table 1. An instance of the value of hope extracted from interview data, hope components mapped to the value instance, and technology features that can support this value instance and hope components.

Value Instances	Hope components	Technology features
"I felt upset [when I knew I was diagnosed with prediabetes] because all my life I had done the right things. I had exercised, I had eaten right and even when I had to stop exercising I still ate right and so I was very disappointed. I was angry at my muscle disease and I was upset, I almost started crying because I was just … One more thing that has gone wrong with my health because of my other disease, so yeah, I was upset. I at first told the doctor that I didn't want to take anything. I was mad. I didn't want to do this because it was admitting that I had diabetes or pre-diabetes or whatever."	Goal: healthy life style	Digital coaches are intelligent technology-based services that simulate human coaches and reinforce patients on their pathways toward goals and enhance patients' agency by providing motivational messages, techniques, and resources in real time.
	Agency: lack of agency reflected in negative emotions – "I was very disappointed", "I almost started crying", "I was upset", "I was mad".	
	Pathway: exercise and eating right – "I had exercised, I had eaten right and even when I had to stop exercising I still ate right"	

In Table 1 an example of a value instance mapped to hope components with support of technology features is provided to illustrate how value-sensitive technologies support goal-oriented agency and pathways in patients with diabetes. A patient diagnosed with diabetes expressing and describing negatives emotions indicates an underlying issues with patient's agency and available pathways. The available pathways towards a healthy life style for this patient have not been effective in achieving her goals to live a healthy life style. The ineffective pathways undermine patients' feeling of agency. The patient questions her abilities in achieving goals and develops negative emotions of being upset and mad. Digital coaches enhance patients' agency by motivating patients along the way in pursuit of their goals. An empowered patient with higher agency can tackle barriers and negative emotions in achieving goals. Digital coaches designed in the diabetes mobile app provide guidance, resources, and emotional support. The real time and on-demand access to digital coaches increases patients' motivations in achieving their goals by reinforcing and reaffirming patients' thoughts towards their goals.

5. CONCLUSION

In this work in progress study we being to explore the role of ICTs in supporting and enhancing patients' hope to self-manage their diabetic chronic conditions. We use VSD to design interview strategies and questions for patients with diabetes to identify their needs and desires to use a diabetes mobile app and self-manage their chronic conditions. We use hope theory components of agency, pathways, and goal to translate value instances of hope into supportive technology features.

This study provides guidance and recommendations for the healthcare providers and system developers to assist patients with diabetes self-manage their chronic conditions. The paper instantiates value of hope in the context of ICT-enabled self-management of diabetes and illustrates how system developers can design and develop technology features and healthcare providers to use those technology features to enhance the feeling of agency in the patients and provide effective pathways towards their goals.

REFERENCES

Bardhan, I., Chen, H., & Karahanna, E. (2017). The Role of Information Systems and Analytics in Chronic Disease Prevention and Management. *MIS Quarterly*, (Call for Papers MISQ Special Issue).

Dadgar, M., & Joshi, K. D. (2018). The Role of Information and Communication Technology in Self-Management of Chronic Diseases: An Empirical Investigation through Value sensitive design. *Journal of the Association for Information Systems (JAIS)*, *19*(2), 86–112.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: The MIT Press.

Friedman, B., Howe, D. C., & Felten, E. (2002). Informed consent in the Mozilla browser: Implementing value-sensitive design. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. Presented at the Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Big Island, HI.

Snyder, C. R. (2000). *Handbook of Hope: Theory, Measures, and Applications*. San Diego, Calif: Academic Press.

VALUE SENSING ROBOTS: THE OLDER LGBTIQ+ COMMUNITY

Adam Poulsen, Ivan Skaines, Suzanne McLaren, Oliver K. Burmeister

Charles Sturt University (Australia), Newcastle Pride (Australia), Charles Sturt University (Australia), Charles Sturt University (Australia)

apoulsen@csu.edu.au; iskaines@ozemail.com.au; smclaren@csu.edu.au; oburmeister@csu.edu.au

ABSTRACT

LGBTIQ+ older adults (Lesbian-Gay-Bisexual-Transgender-Intersex-Queer+others) are an under-researched community experiencing high rates of loneliness. The value sensitive design and use of social care robots provides an innovative advance toward equity for older LGBTIQ+ adults at risk of loneliness. Focusing on the LGBTIQ+ older adult and social care robot case study, values in motion design and value sensing robots are presented as solutions to the missing account of good care in value sensitive design. This constructivist study found that *LGBTIQ+ connectivity and community*, *social connectedness*, and *no special attention in care* are identified as key instrumental values for the older LGBTIQ+ community regarding social care robots.

KEYWORDS: Healthcare robotics, community, LGBTIQ+ ageing, value sensitive design.

1. INTRODUCTION

Value sensitive design (VSD) is a popular method for investigating stakeholder values and designing systems to account for those values (Friedman & Hendry, 2019). Recent VSD works (e.g., Jacobs & Huldtgren, 2018; Manders-Huits, 2011), attempt to move the methodology towards normative ethics, aiming to establish a standardised design decision framework to create technologies. In contrast, VSD pioneers were careful not to suggest that *values are either entirely normative nor descriptive*. Each value is conceptualised within its respective field and no list of values is comprehensive (Friedman & Hendry, 2019).

Similarly, good care practice is neither entirely normative nor descriptive. What each person and community needs and values in care matters (Abma, Molewijk, & Widdershoven, 2009). Descriptive principles of care hold instrumental value for individuals, and they should be considered in VSD. At the same time, there are normative principles in care expressed through applied ethics that are intrinsically good and valuable, including safety and wellbeing, as identified by duty of care, professional ethics, and law (Teipel et al., 2016).

Social care robots (SCRs) play a role in social support or care by enabling, assisting in, or replacing social interactions. For good robot-delivered care, SCRs need to ensure both normative intrinsic values and descriptive instrumental values found in real care practices. Moreover, just as good care is determinative in practice (Beauchamp, 2004), SCRs must account for changing and emerging values in care. *Value sensing robots* (i.e., robots which attempt to learn user values and adapt behaviour to suit those values) may work towards this using the VSD-adapted design approach values in motion design (VMD) (Poulsen & Burmeister, 2019; Poulsen, Burmeister, & Kreps, 2018).

The purpose of the study presented here is to put the concepts of value sensing robots, as well as VMD, into practice with a particular community with its own set of values and value interpretations. The older LGBTIQ+ community was selected for this purpose given that they are under-researched (Fredriksen-Goldsen, Kim, Barkan, Muraco, & Hoy-Ellis, 2013) and experiencing high rates of loneliness (Fredriksen-Goldsen, 2016; Hughes, 2016) which might be alleviated with SCRs. Additionally, highlighting under-surveyed LGBTIQ+ older population values alerts potential discriminatory implications of robots, which do not consider vulnerable, marginalized, silent aging populations (Poulsen, Fosch-Villaronga & Søraa, 2020). LGBTIQ+ older adults were interviewed to create knowledge about the older LGBTIQ+ community's values. At the same time, the literature was used to conceptualize the normative goods in this care context. With this information, exemplary LGBTIQ-friendly SCRs were designed. The pilot data of this study are presented here. Care with robots has been discussed in the engineering, philosophical, and design literature, but little of that discussion has so far addressed good care. It is here that this article makes its contribution.

The following sections review the literature, beginning with value sensitive design. Then the broader context of this pilot study is described, after which comes the methodology employed in this study. Next, the findings are presented and discussed, followed by a description of potential further studies.

2. LITERATURE REVIEW

2.1. Value sensitive design

Technology is not value neutral (Friedman & Hendry, 2019; Legassick & Harding, 2017); technologies and systems have an impact on stakeholder values. While 'value' typically refers to the economic worth of an object, in recent VSD theory, values are described as "what is important to people in their lives, with a focus on ethics and morality" (Friedman & Hendry, 2019, p. 24). One popular method to account for stakeholder values in technology design is VSD which aims to promote positive value impacts by design (Friedman & Hendry, 2019). In the literature, VSD has been widely applied to the design of information systems (IS) (Friedman & Hendry, 2019; Manders-Huits, 2011; van Wynsberghe, 2016).

Umbrello and De Bellis (2018) explain that VSD is a unique design approach because it is proactive in such a way that it encourages predicting emerging values and realizing solutions in designs. Another advantage of VSD is that it invites a multidisciplinary approach to better address the diverse complexities of design with the involvement of philosophers, ethicists, social scientists, behavioural scientists, computer scientists, and designers (Friedman & Hendry, 2019). VSD realises values and incorporates them into design via its tripartite methodology consisting of conceptual, empirical, and technical investigations.

Friedman and Hendry (2019) elaborate on the three VSD investigations as follows. Conceptual investigations define the IS users and other stakeholders, identify the values of all stakeholders who interact with the IS, and conceptually examine how those values are positively and negatively impacted by the IS design. An empirical investigation aims to create further knowledge about those values concerning the IS, through empirical means. Finally, the technical investigation involves designing a new IS to support the values of users as they have been understood empirically, or it involves analysing how users interact with an existing IS.

Manders-Huits (2011) suggests that VSD is too descriptive in its conceptualisation of values and that value trade-off decisions in VSD need to be grounded in normative ethical theory. Similarly, Jacobs and Huldtgren (2018) argue that for VSD practitioners to be able to legitimize value trade-offs during the design process, their approach should to be grounded in ethical theory. This trend contrasts with the traditional approach to VSD which holds the plurality of values, i.e. values are neither entirely normative nor descriptive.

2.2. Good care

Given the continued use of VSD in the healthcare IS space (Maathuis, Niezen, Buitenweg, Bongers, & van Nieuwenhuizen, 2019; Schoenhofer, van Wynsberghe, & Boykin, 2019), concern for the provision of good care emerges. As VSD studies continue to move in the direction of normative ethics and values, the importance of descriptive ethics and values in good care is not being accounted for in the realisation of healthcare technologies created using VSD.

Like values, good care practice is neither entirely normative nor descriptive. What each person and community wants in care matters, that is, care is also person-centred (Lloyd, 2005; Tronto, 1993), culturally competent (Farber, 2019; Purnell & Fenkl, 2019), and determinative in practice (Beauchamp, 2004) or concrete situations (Abma et al., 2009). Descriptive goods expressed by individuals and groups hold instrumental value in good care practice, and they should be considered in VSD. At the same time, there are normative principles in care expressed through applied ethics that are intrinsically good and valuable, including safety and wellbeing, as identified by duty of care, professional ethics and codes, and healthcare law. Not only are some normative principles required by professional standards and law, but they are reasoned to be valuable in applied ethics in healthcare. This dichotomy of care values (i.e., the need to ensure normative intrinsic values and descriptive instrumental values at the same time) is not represented in the recent VSD literature which attempts to move the methodology to a normative grounding, thus missing the essence of VSD, as well as missing what matters in good care.

One of the greatest influences on an individual’s values is each person’s cultural background (Burmeister, 2013; Huang, Teo, Sánchez-Prieto, García-Peñalvo, & Olmos-Migueláñez, 2019; Sunny, Patrick, & Rob, 2019). Thus, values should be examined through a culturally sensitive lens. In good human-delivered care, understanding what individuals instrumentally value in care requires an emphasis on cultural competence (Farber, 2019; Purnell & Fenkl, 2019), person-centred care (Kamrul, Malin, & Ramsden, 2014; Santana et al., 2018), and context (Abma et al., 2009; Beauchamp, 2004). If healthcare technologies, such as care robots, are to provide good care they also need to, in part, demonstrate these key competencies by design and in-situ.

2.3. Care robots

Globally, there is a need for healthcare IS intervention in aged care due to the growing number of older adults and lack of caregivers in this sector (Burmeister, 2016; Burmeister & Kreps, 2018; Draper & Sorell, 2017; Garner, Powell, & Carr, 2016). In the 2019 Revision of the World Population Prospects, the United Nations (2019) predict that the global population will continue to grow older throughout the century. The data suggests that in 2050 the worldwide percentage of persons aged 65+ will reach 15.9%, up from 8.2% in 2015 and 9.3% in 2020 (United Nations, 2019). Furthermore, the United Nations (2019) shows that life expectancy at birth is continuing to rise. In 2015 the life expectancy at birth was 70.9 (years old), rising to 72.3 in 2020, and continuing upward to 76.8 by 2050.

Compounding the problem is the decreasing amount of caregiver support internationally. Poor government funding, high job requirements, and low pay has created a lack of uptake in aged caregiver jobs internationally, including in the United Kingdom (The Lancet, 2014), the United States (Flaherty & Bartels, 2019), and Australia (Cope, Jones, & Hendricks, 2016). Health Workforce Australia (2012) predict that there will be a shortage of 100,000 nurses across all Australian healthcare by 2025. The scarcity of aged caregivers impacts older adults residing in remote and rural areas of Australia especially (Ervin, Reid, Moran, Opie, & Haines, 2019).

Care robots present an opportunity to supplement the shortage of caregivers and assist the growing older population (Miyachi, Iga, & Furuhata, 2017; van Wynsberghe, 2013; Wright, 2018). The International Organization for Standardization, in ISO 13482:2014, define a *personal care robot*, as a service robot (one which is programmable, autonomous, and performs useful tasks for humans or equipment excluding industrial automation applications) that performs actions contributing directly towards improvement in the quality of life of humans, excluding medical applications (ISO, 2014). In aged care, robots are taking on functional roles as physical assistants (Niemelä & Melkas, 2019), personal service assistants (Martinez-Martin & del Pobil, 2018), physical rehabilitators (Fosch-Villaronga & Özcan, 2019), and health monitors (Michaud et al., 2007).

Social care robots are being made useful in valuable roles such as social support or companionship (Birks, Bodak, Barlas, Harwood, & Pether, 2016), emotional support with affective communication (Khosla, Chu, Kachouie, Yamada, & Yamaguchi, 2012), and social connection with telepresence systems (Moyle, Jones, & Sung, 2020). Informed by existing definitions of robots (ISO, 2012), personal care robots (ISO, 2014), and assistive social robots (Kachouie, Sedighadeli, Khosla, & Chu, 2014), the novel definition of a SCR is as follows:

> A robot which operates in a caring role to assist in care, enable self-care, or replace a caregiver; interacts with care recipients on some sociable dimension, intentional or not; performs actions contributing directly towards improvement in the quality of social life of care recipients and fostering human-human connection; is programmable and has a degree of autonomy for moving within (or reacting too) its environment when performing useful tasks for humans (both caregivers and care recipients) without human operation.

On the role of SCRs in alleviating loneliness in aged care, several studies show the effectiveness of care robots in a social role. In a recent review, three studies show that SCRs, as companions for older adults, reduced experiences of loneliness (Abdi, Al-Hindawi, Ng, & Vizcaychipi, 2018). Another study in New Zealand reported that Paro, the companion robot, significantly decreased loneliness among older adults in a nursing home (Robinson, Macdonald, Kerse, & Broadbent, 2013). A different study used telepresence as a long-term tool to alleviate the sense of loneliness experienced by older adults (Cesta, Cortellessa, Orlandini, & Tiberio, 2016). The authors concluded that the psychosocial impact on the quality of life and loneliness was positive. Video-conferencing technologies have shown to alleviate loneliness experienced by older adults in other studies (Tsai & Tsai, 2011; Tsai, Tsai, Wang, Chang, & Chu, 2010).

Broadly, across aged care, studies show that care robots are helping to improve emotional state, reduce challenging behaviours, and improve social interactions (Birks et al., 2016); engage elderly in social activities and break down intergeneration technology barriers, (Khosla et al., 2012); improve quality of life (Broadbent, Jayawardena, Kerse, Stafford, & Macdonald, 2011); improve wellbeing (Kachouie et al., 2014); reduce caregiver workload and promote self-care, positive emotions, engagement, relationships, and meaning achievement (Kachouie et al., 2014); and successfully mediate conversation (Birks et al., 2016).

No other study has addressed the need to balance and concurrently respect intrinsic and instrumental values in the design and operation of care robots; this study does so with a particular case study – LGBTIQ+ older adults.

2.4. The older LGBTIQ+ community case study

LGBTIQ+ older adults experience higher rates of loneliness compared to the general older adult population within Australia (Hughes, 2016) and internationally (Fredriksen-Goldsen, 2016). In a study consisting of 312 LGBTIQ+ older adults, Hughes (2016) found that the social isolation of this population directly contributes to the high rates of loneliness reported.

As an alternative to current social supports for alleviating loneliness in the LGBTIQ+ aged care space, such as outreach services[1, 2] and LGBTIQ-friendly aged care facilities[3, 4], this study explores the use of SCRs. However, before SCRs can be realised a VSD investigation is required. The healthcare needs and values of the older LGBTIQ+ community are under-surveyed (Fredriksen-Goldsen et al., 2013) and their values concerning technology are unexplored entirely. Thus, a VSD investigation is needed to discover this community's values otherwise the descriptive, instrumentally valuable side of good care would be overlooked in SCR design.

Each community has a value framework, consisting of particular value priorities, orientations, and interpretations (Burchum, 2002; Crawley, Marshall, Lo, & Koenig, 2002); the older LGBTIQ+ community is no different (Waling & Roffee, 2017). Tenenbaum (2011) describes the older LGBTIQ+ community as one with unique values, concerns, needs, and critical and experiential interests in aged care. In the search for cultural sensitivity, many LGBTIQ+ older adults seek out services which are LGBTIQ-friendly and healthcare professionals who are sensitive to their needs and values (Jann, Edmiston, & Ehrenfeld, 2015). The difficulty of finding a doctor who is competent in, and sensitive to, LGBTIQ+ needs and values leads to this group being "significantly more likely to delay or avoid necessary medical care compared with heterosexuals" (29% versus 17%, respectively) (Khalili, Leung, & Diamant, 2015). On the values of LGBTIQ+ older adults, the value of family is often interpreted as a *chosen family* consisting of close friends, rather than relatives (Cannon, Shukla, & Vanderbilt, 2017). Furthermore, intersex older adults define the value of *non-judgemental care* concerning their intersex status as it impacts their physical,

[1] See http://www.switchboard.org.au/out-about/

[2] See http://www.umbrellacommunitycare.com.au/services/at-home-care/community-visitor-scheme/

[3] See https://arcare.com.au/qld-aged-care/parkwood-aged-care/

[4] See https://www.lintonestate.com.au/vision-linton-estate/

hormonal, or genetic differences (Latham & Barrett, 2015). This knowledge should translate into SCR design and behaviour in-situ.

To perform the VSD investigation, an innovative VSD approach - *values in motion design* - was developed to account for good care with *value sensing robots*.

2.5. Values in motion design

VMD was realised to address the limitations with value sensing robots (Poulsen & Burmeister, 2019). It accounts for the pluralistic and evolving nature of values through the design of value sensing robots which make explicit value-driven decisions to govern actions. These decisions are shaped to the values of the user in-situ, when it is safe to do so, and only within a framework of intrinsic values implicitly embedded into the design. As a starting point in VMD, designers aim to develop a basic care robot framework based on intrinsic values found in applied ethics. Thereafter, designers attempt to capture instrumental, community-based values and develop a set of initial robotic behaviours to respect those instrumental values. A value-driven decision-making process should be implemented to allow the care robot to adapt to the values of the user and shape these initial behaviours to the instrumental values of the user during run-time to provide person-centred care.

To be capable of good care, care robots must uphold both applied ethics and descriptive ethics. VMD was developed to aid HRI practitioners in realising care robots capable of good care. The principles of VMD are as follows:

> A distinction should be made between intrinsic and instrumental care values. This distinction is grounded in applied ethics (e.g., values emerging from professional ethics and codes, healthcare law, robot design standards, and duty of care) and descriptive ethics (e.g., values emerging from determinative in practice, person-centred, culturally competent care), respectively. Care robots must be ethically designed and ethically minded; designers should only make intrinsic value decisions, and value sensing robots are to make instrumental value decisions in relationship with the user.

Following VMD, one performs the VSD investigations, but additionally distinguishes between intrinsic and instrumental care values by examining applied and descriptive ethics. Thereafter, the intrinsic values should to be embedded in the care robot design and the instrumental values should to be realised as dynamic robot actions and programmed into the robot for it to decide what actions are right for each user in-situ using the principle of value sensing.

3. METHODOLOGY

To test VMD, an interpretivist, constructivist pilot study was conducted with five LGBTIQ+ older adults (three gay men, one gay gender-fluid person, and one lesbian non-binary person). Through semi-structured interviews, participants were questioned about the LGBTIQ+ experience of ageing, aged care, social isolation, and loneliness, as well as the older LGBTIQ+ community's values. These interviews were transcribed and analysed using content analysis. Ethics approval from the university and from participating LGBTIQ+ organisations, from which participants were recruited for this and the larger study, was obtained. This pilot study is a part of a larger project.

4. FINDINGS

Using content analysis, the values of LGBTIQ+ older adults interviewed were derived (see Figure 1). Figure 1 shows how LGBTIQ+ older adults prioritise values, illustrating the number of persons who cited a value and the number of times a value was referenced during all the interviews. *LGBTIQ+ connectivity and community*, *social connectedness*, and *no special attention in care* were frequently mentioned by all participants, suggesting that these are important values. Whereas *safety*, *diverse friendships and community*, and *love and attention* were cited less frequently and by fewer participants.

Table 1 shows examples of how LGBTIQ+ older adults interpret values compared to the literature. Several participants noted the value of *appreciating difference*, suggesting a key interpretation of *inclusivity* which appreciates difference rather than simply includes different people and views. *Freedom of expression* was conceptualised by all participants as *LGBTIQ+ openness*, indicating the importance of free expression of LGBTIQ+ pride (e.g., slang, symbols, and events) in the wider community.

Figure 1. The older LGBTIQ+ community's values found in five pilot study interviews. Only those values which were referenced by three or more persons have been included.

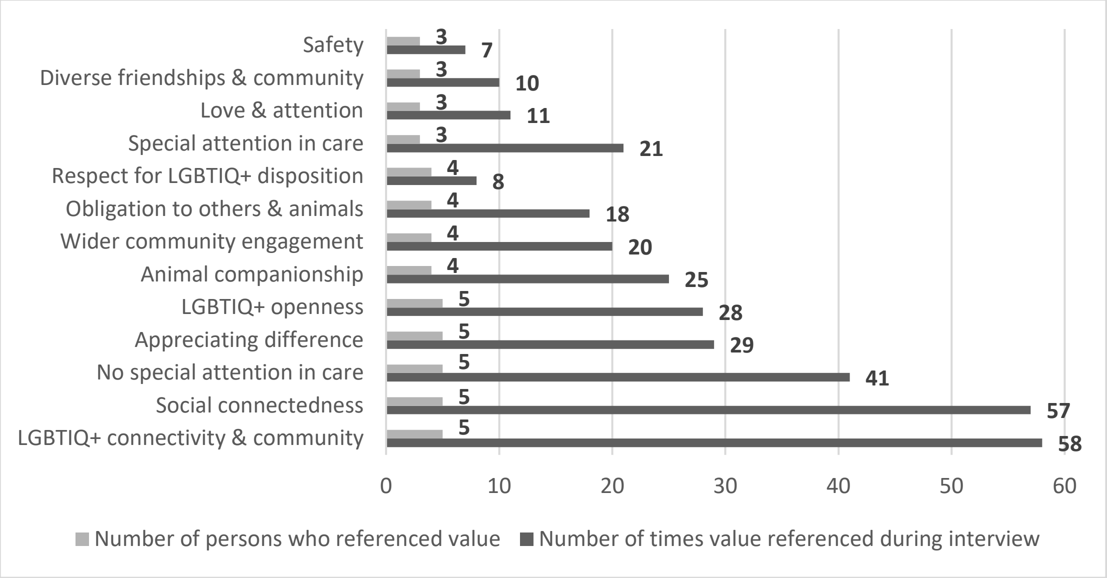


Table 1. Exemplary pilot study values compared to values found in the literature.

LGBTIQ+ older adult value interpretations	Equivalent values found in the literature
Appreciating difference	Inclusivity
LGBTIQ+ connectivity & community	Community
Diverse friendships & community	Cultural diversity
LGBTIQ+ openness	Freedom of expression
No special attention in care	Equality
Obligation to others & animals	Being needed
Special attention in care	Equity
Respect for LGBTIQ+ disposition	Respect

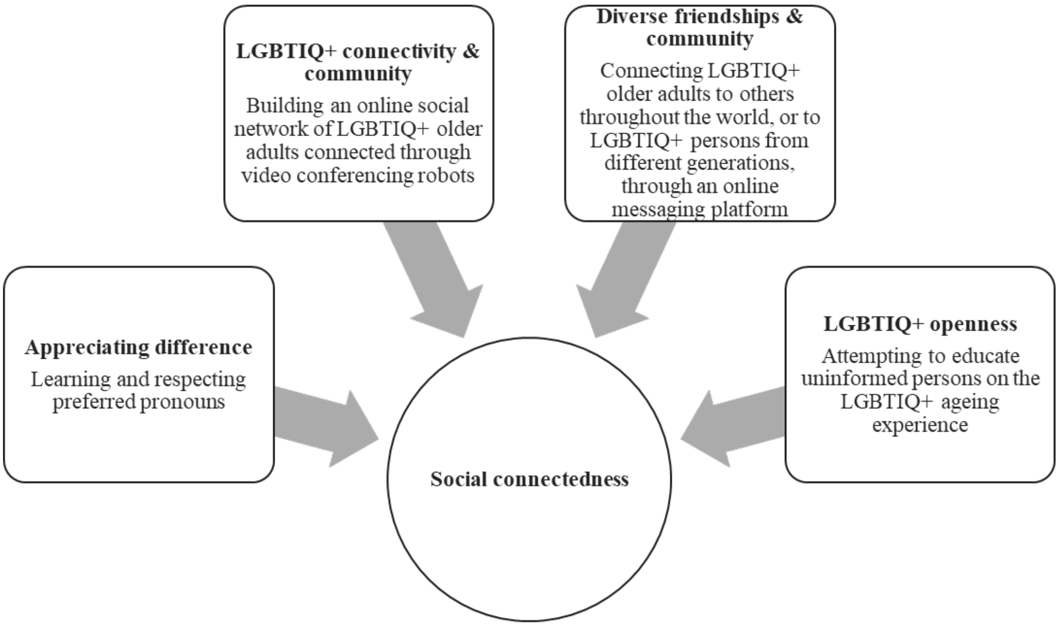
5. DISCUSSION

The values identified in this study can be used to configure an initial framework of instrumental, community-based values for value sensing SCRs intended for the older LGBTIQ+ community. Thereafter, using adaptive functions, the SCR can reprioritise those values and learn new ones in-situ with the user to provide person-centred care. To explain value sensing adaptive functions, by analogy, consider the current care robot Elli-Q[5] which examines an image, recognises the objects in an image, and provides a verbal translation of what is featured in the image. Value sensing could examine, recognise, and translate user values in a similar way.

For example, an LGBTIQ+ older adult who uses an SCR is sitting with another person, but they are no longer conversing. The SCR should to be able to understand what user values are being impacted. Is the user desiring social connectedness, but they have exhausted conversation topics? Are the user and the other person struggling to socially connect due to cultural differences? Does the user enjoy the silence and feel adequately socially connected?

Knowing what user values are being impacted, and why, helps the SCR hone its delivery of care. If the SCR understands that running out of conversation negatively impacts the value of *social connectedness*, then it will be able to support the user to better exercise this value in the future. For instance, the SCR could suggest conversation topics. However, even with this social support, perhaps the LGBTIQ+ older adult is still not feeling socially connected (e.g., giving short answers or often looking away) because the SCR is suggesting conversation topics which are not relatable for the LGBTIQ+ older adult (e.g., family or children). Arising from the results shown in Table 1, observing that LGBTIQ+ older adults interpret respect as *respect for LGBTIQ+ disposition*, the SCR is negatively impacting this value. A value sensing robot should understand the values and value interpretations of different communities and individuals to provide person-centred care.

Figure 2. Designing SCR components with the older LGBTIQ+ community's value interpretations in mind, each working to ensure the normative intrinsic value social connectedness.


LGBTIQ+ connectivity & community
Building an online social network of LGBTIQ+ older adults connected through video conferencing robots
Diverse friendships & community
Connecting LGBTIQ+ older adults to others throughout the world, or to LGBTIQ+ persons from different generations, through an online messaging platform
Appreciating difference
Learning and respecting preferred pronouns
LGBTIQ+ openness
Attempting to educate uninformed persons on the LGBTIQ+ ageing experience
Social connectedness

[5] *See* https://elliq.com/

Figure 2 further demonstrates how social connectedness might be achieved with the value interpretations of LGBTIQ+ older adults in mind. In run-time, value sensing robots could shift these values to better suit the values of individual LGBTIQ+ older adults. For instance, consider a video conferencing robot which plays a role in social care by hosting video calls across an online LGBTIQ+ social network. If the user does not utilise the existing functions designed to ensure *LGBTIQ+ connectivity and community*, then it might instead shift this value (and subsequent behaviours) to schedule cafe meetups with other local LGBTIQ+ older adults connected to the online social network.

6. CONCLUSION

The older LGBTIQ+ community hold key instrumental values regarding SCRs, including *LGBTIQ+ connectivity and community*, *social connectedness*, and *no special attention in care*. Value sensing SCRs need to adapt to the values of LGBTIQ+ older adults in a *person-centred care mode* to help overcome the loneliness that is presently widespread in this community. With adaptive functionality, SCRs can be designed to make dynamic, value-driven decisions in-situ to customise the level of care down to the person-centred level within duty of care limits.

REFERENCES

Abdi, J., Al-Hindawi, A., Ng, T., & Vizcaychipi, M. P. (2018). Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open, 8*(2). doi:10.1136/bmjopen-2017-018815.

Abma, T., Molewijk, B., & Widdershoven, G. (2009). Good care in ongoing dialogue. Improving the quality of care through moral deliberation and responsive evaluation. *Health Care Analysis, 17*, 217-235. doi:10.1007/s10728-008-0102-z.

Beauchamp, T. L. (2004). Does ethical theory have a future in bioethics? *The Journal of Law, Medicine & Ethics, 32*(2), 209-217. doi:10.1111/j.1748-720X.2004.tb00467.x.

Birks, M., Bodak, M., Barlas, J., Harwood, J., & Pether, M. (2016). Robotic Seals as Therapeutic Tools in an Aged Care Facility: A Qualitative Study. *Journal of Ageing Research, 2016*, 1-7. doi:10.1155/2016/8569602.

Broadbent, E., Jayawardena, C., Kerse, N., Stafford, R., & Macdonald, B. (2011). Human-Robot Interaction Research to Improve Quality of Life in Elder Care - An Approach and Issues. In *Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care* (pp. 13–19): AAAI Press.

Burchum, J. L. R. (2002). Cultural Competence: An Evolutionary Perspective. *Nursing Forum, 37*(4), 5-15. doi:10.1111/j.1744-6198.2002.tb01287.x.

Burmeister, O. K. (2013). Achieving the goal of a global computing code of ethics through an international-localisation hybrid. *Ethical Space: The International Journal of Communication Ethics, 10*(4), 25-32.

Burmeister, O. K. (2016). The development of assistive dementia technology that accounts for the values of those affected by its use. *Ethics and Information Technology, 18*(3), 185-198. doi:10.1007/s10676-016-9404-2.

Burmeister, O. K., & Kreps, D. (2018). Power influences upon technology design for age-related cognitive decline using the VSD framework. *Ethics and Information Technology*, 1-4. doi:10.1007/s10676-018-9460-x.

Cannon, S. M., Shukla, V., & Vanderbilt, A. A. (2017). Addressing the healthcare needs of older Lesbian, Gay, Bisexual, and Transgender patients in medical school curricula: a call to action. *Medical education online, 22*(1), 1-4. doi:10.1080/10872981.2017.1320933.

Cesta, A., Cortellessa, G., Orlandini, A., & Tiberio, L. (2016). Long-Term Evaluation of a Telepresence Robot for the Elderly: Methodology and Ecological Case Study. *International Journal of Social Robotics, 8*(3), 421-441. doi:10.1007/s12369-016-0337-z.

Cope, V. C., Jones, B., & Hendricks, J. (2016). Residential aged care nurses: portraits of resilience. *Contemporary Nurse, 52*(6), 736-752. doi:10.1080/10376178.2016.1246950.

Crawley, L. M., Marshall, P. A., Lo, B., & Koenig, B. A. (2002). Strategies for Culturally Effective End-of-Life Care. *Annals of Internal Medicine, 136*(9), 673-679. doi:10.7326/0003-4819-136-9-200205070-00010.

Draper, H., & Sorell, T. (2017). Ethical values and social care robots for older people: an international qualitative study. *Ethics and Information Technology, 19*(1), 49-68. doi:10.1007/s10676-016-9413-1.

Ervin, K., Reid, C., Moran, A., Opie, C., & Haines, H. (2019). Implementation of an older person's nurse practitioner in rural aged care in Victoria, Australia: a qualitative study. *Human Resources for Health, 17*(1), 80. doi:10.1186/s12960-019-0415-z.

Farber, J. E. (2019). Cultural Competence of Baccalaureate Nurse Faculty: Relationship to Cultural Experiences. *Journal of Professional Nursing, 35*(2), 81-88. doi:10.1016/j.profnurs.2018.09.005.

Flaherty, E., & Bartels, S. J. (2019). Addressing the Community-Based Geriatric Healthcare Workforce Shortage by Leveraging the Potential of Interprofessional Teams. *Journal of the American Geriatrics Society, 67*(S2), 400-408. doi:10.1111/jgs.15924.

Fosch-Villaronga, E., & Özcan, B. (2019). The progressive intertwinement between design, human needs and the regulation of care technology: The case of lower-limb exoskeletons. *International Journal of Social Robotics*. doi:10.1007/s12369-019-00537-8.

Fredriksen-Goldsen, K. I. (2016). The Future of LGBT+ Aging: A Blueprint for Action in Services, Policies, and Research. *Generations, 40*(2), 6-15. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28366980.

Fredriksen-Goldsen, K. I., Kim, H.-J., Barkan, S. E., Muraco, A., & Hoy-Ellis, C. P. (2013). Health disparities among lesbian, gay, and bisexual older adults: results from a population-based study. *American Journal of Public Health, 103*(10), 1802-1809. doi:10.2105/AJPH.2012.301110.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*: MIT Press.

Garner, T. A., Powell, W. A., & Carr, V. (2016). Virtual carers for the elderly: A case study review of ethical responsibilities. *Digital Health, 2*, 1-14. doi:10.1177/2055207616681173.

Health Workforce Australia. (2012). *Health Workforce 2025 – Doctors, Nurses and Midwives – Volume 1*. Retrieved from https://apo.org.au/node/154456

Huang, F., Teo, T., Sánchez-Prieto, J. C., García-Peñalvo, F. J., & Olmos-Migueláñez, S. (2019). Cultural values and technology adoption: A model comparison with university teachers from China and Spain. *Computers & Education, 133*, 69-81. doi:10.1016/j.compedu.2019.01.012.

Hughes, M. (2016). Loneliness and social support among lesbian, gay, bisexual, transgender and intersex people aged 50 and over. *Ageing and Society, 36*(9), 1961-1981. doi:10.1017/S0144686X1500080X.

ISO. (2012). ISO 8373:2012: Robots and robotic devices - Vocabulary. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-2:v1:en

ISO. (2014). ISO 13482:2014: Robots and robotic devices - Safety requirements for personal care robots. Retrieved from https://www.iso.org/standard/53820.html

Jacobs, N., & Huldtgren, A. (2018). Why value sensitive design needs ethical commitments. *Ethics and Information Technology*. doi:10.1007/s10676-018-9467-3.

Jann, J. T., Edmiston, E. K., & Ehrenfeld, J. M. (2015). Important Considerations for Addressing LGBT Health Care Competency. *American Journal of Public Health, 105*(11), e8-e8. doi:10.2105/AJPH.2015.302864.

Kachouie, R., Sedighadeli, S., Khosla, R., & Chu, M.-T. (2014). Socially Assistive Robots in Elderly Care: A Mixed-Method Systematic Literature Review. *International Journal of Human–Computer Interaction, 30*(5), 369-393. doi:10.1080/10447318.2013.873278.

Kamrul, R., Malin, G., & Ramsden, V. R. (2014). Beauty of patient-centred care within a cultural context. *Canadian Family Physician, 60*(4), 313-315. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4046555/.

Khalili, J., Leung, L. B., & Diamant, A. L. (2015). Finding the perfect doctor: identifying lesbian, gay, bisexual, and transgender-competent physicians. *American Journal of Public Health, 105*(6), 1114-1119. doi:10.2105/ajph.2014.302448.

Khosla, R., Chu, M.-T., Kachouie, R., Yamada, K., & Yamaguchi, T. (2012). *Embodying care in Matilda: An affective communication robot for the elderly in Australia*. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 295–304): ACM

Latham, J., & Barrett, C. (2015). Appropriate bodies and other damn lies: Intersex ageing and aged care. *Australasian Journal on Ageing, 34*(S2), 19-20. doi:10.1111/ajag.12275.

Legassick, S., & Harding, V. (2017). Why we launched DeepMind Ethics & Society. Retrieved from https://deepmind.com/blog/announcements/why-we-launched-deepmind-ethics-society

Lloyd, L. (2005). A Caring Profession? The Ethics of Care and Social Work with Older People. *The British Journal of Social Work, 36*(7), 1171-1185. doi:10.1093/bjsw/bch400.

Maathuis, I., Niezen, M., Buitenweg, D., Bongers, I. L., & van Nieuwenhuizen, C. (2019). Exploring Human Values in the Design of a Web-Based QoL-Instrument for People with Mental Health Problems: A Value Sensitive Design Approach. *Science and Engineering Ethics, 26*(2), 871–898. doi:10.1007/s11948-019-00142-y.

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics, 17*(2), 271-287. doi:10.1007/s11948-010-9198-2.

Martinez-Martin, E., & del Pobil, A. P. (2018). Personal Robot Assistants for Elderly Care: An Overview. In A. Costa, V. Julian, & P. Novais (Eds.), *Personal Assistants: Emerging Computational Technologies* (pp. 77-91). Cham: Springer International Publishing.

Michaud, F., Boissy, P., Labonte, D., Corriveau, H., Grant, A., Lauria, M., . . . Royer, M.-P. (2007). Telepresence Robot for Home Care Assistance. In *Proceedings of AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics* (pp. 50-55): AAAI Press.

Miyachi, T., Iga, S., & Furuhata, T. (2017). Human Robot Communication with Facilitators for Care Robot Innovation. *Procedia Computer Science, 112*, 1254-1262. doi:10.1016/j.procs.2017.08.078.

Moyle, W., Jones, C., & Sung, B. (2020). Telepresence robots: Encouraging interactive communication between family carers and people with dementia. *Australasian Journal on Ageing, 39*(1), 127-133. doi:10.1111/ajag.12713.

Niemelä, M., & Melkas, H. (2019). Robots as Social and Physical Assistants in Elderly Care. In M. Toivonen & E. Saari (Eds.), *Human-Centered Digitalization and Services* (pp. 177-197). Singapore: Springer Singapore.

Poulsen, A., & Burmeister, O. K. (2019). Overcoming carer shortages with care robots: Dynamic value trade-offs in run-time. *Australasian Journal of Information Systems, 23*. doi:10.3127/ajis.v23i0.1688.

Poulsen, A., Burmeister, O. K., & Kreps, D. (2018). The ethics of inherent trust in care robots for the elderly. In D. Kreps, C. Ess, L. Leenen, & K. Kimppa (Eds.), *This Changes Everything – ICT and Climate Change: What Can We Do?* (pp. 314-328). doi:10.1007/978-3-319-99605-9_24

Poulsen, A., Fosch-Villaronga, E., & Søraa, R. A. (2020). Queering machines. Nature Machine Intelligence, 2, 152. doi:10.1038/s42256-020-0157-6

Purnell, L. D., & Fenkl, E. A. (2019). The Purnell Model for Cultural Competence. In *Handbook for Culturally Competent Care* (pp. 7-18). Cham: Springer International Publishing.

Robinson, H., Macdonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: a randomized controlled trial. *J Am Med Dir Assoc, 14*(9), 661-667. doi:10.1016/j.jamda.2013.02.007.

Santana, M. J., Manalili, K., Jolley, R. J., Zelinsky, S., Quan, H., & Lu, M. (2018). How to practice person-centred care: A conceptual framework. *Health Expectations, 21*(2), 429-440. doi:10.1111/hex.12640.

Schoenhofer, S. O., van Wynsberghe, A., & Boykin, A. (2019). Engaging robots as nursing partners in caring: Nursing as caring meets care-centered value-sensitive design. *International Journal for Human Caring*(2), 157-167. doi:10.20467/1091-5710.23.2.157.

Sunny, S., Patrick, L., & Rob, L. (2019). Impact of cultural values on technology acceptance and technology readiness. *International Journal of Hospitality Management, 77*, 89-96. doi:10.1016/j.ijhm.2018.06.017.

Teipel, S., Babiloni, C., Hoey, J., Kaye, J., Kirste, T., & Burmeister, O. K. (2016). Information and communication technology solutions for outdoor navigation in dementia. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 12*(6), 695-707. doi:10.1016/j.jalz.2015.11.003.

Tenenbaum, E. M. (2011). Sexual expression and intimacy between nursing home residents with dementia: Balancing the current interests and prior values of heterosexual and LGBT residents. *Temple Political and Civil Rights Law Review, 21*, 459.

The Lancet. (2014). Global elderly care in crisis. *Lancet, 383*(9921), 927. doi:10.1016/s0140-6736(14)60463-3.

Tronto, J. C. (1993). *Moral Boundaries: A Political Argument for an Ethic of Care*: Routledge.

Tsai, H.-H., & Tsai, Y.-F. (2011). Changes in Depressive Symptoms, Social Support, and Loneliness Over 1 Year After a Minimum 3-Month Videoconference Program for Older Nursing Home Residents. *Journal of Medical Internet Research, 13*(4), e93. doi:10.2196/jmir.1678.

Tsai, H.-H., Tsai, Y.-F., Wang, H.-H., Chang, Y.-C., & Chu, H. H. (2010). Videoconference program enhances social support, loneliness, and depressive status of elderly nursing home residents. *Aging & Mental Health, 14*(8), 947-954. doi:10.1080/13607863.2010.501057.

Umbrello, S., & De Bellis, A. F. (2018). A Value-Sensitive Design Approach to Intelligent Agents. In R. Yampolskiy (Ed.), *Artificial Intelligence Safety and Security* (pp. 395-410): CRC Press.

United Nations. (2019). *World Population Prospects 2019: Volume II: Demographic Profiles*.

van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics, 19*(2), 407-433. doi:10.1007/s11948-011-9343-6.

van Wynsberghe, A. (2016). Service robots, care ethics, and design. *Ethics and Information Technology, 18*(4), 311-321. doi:10.1007/s10676-016-9409-x.

Waling, A., & Roffee, J. A. (2017). Knowing, performing and holding queerness: LGBTIQ+ student experiences in Australian tertiary education. *Sex Education, 17*(3), 302-318. doi:10.1080/14681811.2017.1294535.

Wright, J. (2018). Tactile care, mechanical Hugs: Japanese caregivers and robotic lifting devices. *Asian Anthropology, 17*(1), 24-39. doi:10.1080/1683478X.2017.1406576.

VALUE SENSITIVE DESIGN AND AGILE DEVELOPMENT: POTENTIAL METHODS FOR VALUE PRIORITIZATION

Till Winkler

Institute for Information Systems and Society,
Vienna University of Economics and Business (Austria)

till.winkler@wu.ac.at

ABSTRACT

A promising pathway towards increasing the recognition of value-oriented approaches outside of academia is to facilitate their integration into state-of-the-art agile development processes. This paper provides an overview on the similarities between value-oriented approaches and agile development processes. To facilitate an integration into agile development, light-weight value prioritization methods are needed. Several approaches to value prioritization are presented and insights from using a Likert-scale, as a scalable and light-weight method, in the context of a mobile navigation application are reported. It is the aim of this paper to inspire more empirical work in this area and to foster discussions about an integration into agile development processes and about the challenges of prioritizing values.

KEYWORDS: value sensitive design, value-based engineering, agile development, value prioritization, mobile navigation application.

1. INTRODUCTION

Software is a crucial element of digital technology and has become an integral part of our society (Andreessen, 2011). However, software is not neutral, but acts as a mediator for human values and biases, molding its own operational context, which in return shapes human perception and actions, creates new practices and subsequently ways of living (Verbeek, 2008). Mediation through software can have negative effects such as biases introduced by algorithms (O'Neill, 2016; Obermeyer & Mullainathan, 2019). This suggests that software developers have the responsibility to consider human values, potential for biases and ethical concerns during the development process. The idea to consider instrumental values during technology development, such as *efficiency* and *reliability*, is as old as technology itself (van de Poel, 2015). However, such instrumental values only represent a fraction of relevant human values. A value list aggregated by Winkler and Spiekermann (2019) mentions 355 values potentially important for sustainable technology development. Due to the context-sensitive nature of values (Steen & van de Poel, 2012) even such extensive lists can never be complete. This raises the question on how to choose and prioritize values during a value-oriented project.

Value sensitive Design (VSD) provides a framework for the integration of values with ethical importance into technology design (Friedman, Kahn, Borning, & Huldtgren, 2013; van de Poel, 2015). To achieve successful integration of human values into the design process, VSD employs an integrative and iterative tripartite methodology, consisting of conceptual, empirical and technical investigation (Friedman & Hendry, 2019). All three investigations are interdependent and inform each other mutually (Manders-Huits, 2011). Since its initial conceptualization, VSD has inspired several value-oriented approaches such as *Values at Play* (Flanagan et al. 2005), *Value-oriented and Culturally Informed Approach* (Pereira & Baranauskas, 2015), *Value-based Engineering* (Spiekermann & Winkler, 2020) and several others. All these subsequent approaches contribute to a growing body of knowledge on designing technology in accordance with human values.

Despite these methodological advancements, value-oriented approaches have not found widespread deployment in industry (Miller, Friedman, Jancke, Gill, 2007). Some scholars attribute this to a lack of light-weight methods (e.g. compatible with agile development), a shortage of methods for important tasks (e.g. for prioritizing values) and a lack of consistent methodological description (Miller et al. 2007; van de Poel, 2015; Burmeister, 2016). Additionally, the reliance on exhaustive stakeholder identification and participation, one key feature of value-oriented approaches, is also considered as a major obstacle (Manders-Huits, 2011).

In this paper, I argue for an integration of value-oriented approaches into agile software development processes to increase value consciousness within industry. As a first step, I point out similarities between value-oriented approaches and agile development processes and emphasise the need for value prioritization methods to facilitate an integration. Afterwards, I will present several potential methods for value prioritization. Finally, I report on results and insights from applying a light-weight method to value prioritization in the context of mobile navigation application development.

2. CONSIDERING VALUES DURING AGILE SOFTWARE DEVELOPMENT

The once popular sequential waterfall software development process is based on the assumption that software requirements can be fully specified upfront and that optimal solutions are predictable and plannable in advance; design and sub-sequential coding in this case does not start before a clear set of requirements is defined (Royce 1970). In practice, 4 out of 10 factors for project failure are related to problems with software requirements (Clancy 2014). The volatility requirements, the inability of users to formulate them upfront and the impossibility to know all software details in the beginning, are major obstacles for upfront planning (Mellis, Loebbecke, & Baskerville, 2010; Schmidt, 2016). Changing requirements emphasize the need for less formal and more flexible processes (Sommerville, 2010).

The iterative (or spiral) model to software development employs several adapted waterfall phases and adds risk assessment processes and prototyping activities (MacCormack, Verganti, & Iansiti, 2001). While iterative software processes are more flexible, they are still based on the assumption that software development can be a predictable and upfront plannable process (Schwaber, 1997). As the source of uncertainties is often outside of the development team's control, flexibility and adaptability to new situations is vital for successful software development (Schmidt 2016). Furthermore, upfront planning becomes nearly impossible as the fast evolution of technology often leads to the emergence of new practices and required skills during project realization (Schmidt 2016).

Agile development can be seen as a counter-reaction to the habit of adding more processes (e.g. risk management) to improve development, as it tries to avoid heavy-weight processes and complex documentation (Fowler & Highsmith, 2001; De Lucia & Qusef, 2010). While traditional approaches advocate extensive planning and codified processes, agile approaches rely on the development team's creativity to deal with unpredictable challenges (Dybå & Dingsøyr, 2008; Nerur, Mahapatra, & Mangalaraj, 2005). An initial goal of agile development, as stated in the "Agile Manifest" (Fowler & Highsmith, 2001), was to give developers the autonomy and flexibility needed to deal with challenges (e.g. changing requirements, uncontrollable uncertainties, emerging new practices) and to enable high quality software development. Furthermore, the goal was to strengthen the role of individuals (Fowler & Highsmith, 2001) by ascending them above being the entourage of a development process. Instead, agile development focuses on human collaboration during a development project. These goals produced desired results by increasing software and code quality, improving job satisfaction, raising productivity and customer satisfaction (Dybå & Dingsøyr, 2008).

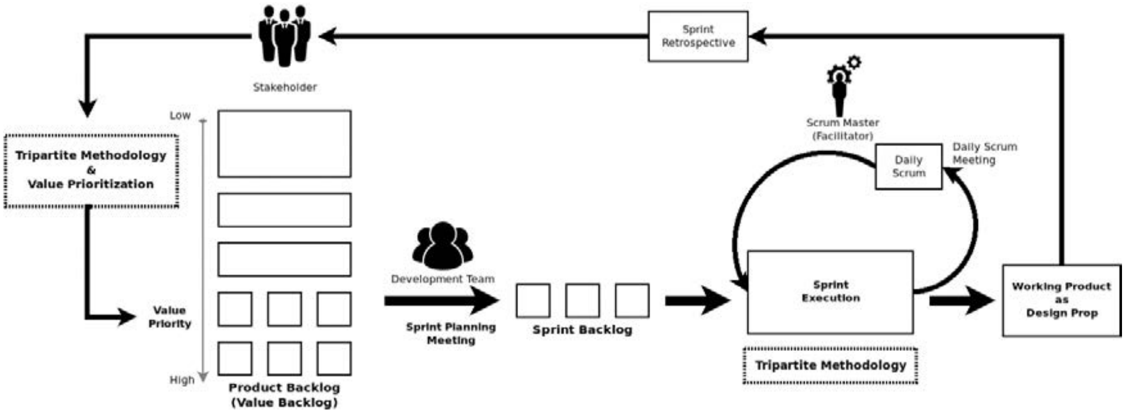
2.1. Value-oriented approaches and agile development processes

In agile development, essential requirement engineering activities are mingled together and performed iteratively throughout the whole development cycle (Paetsch, Eberlein, & Maurer, 2003). Such an iterative and integrative nature is also characteristic for the tripartite methodology of VSD (Friedman & Hendry, 2019). The general goal of any requirement engineering activity is to identify stakeholders and elicit, analyse, specify, validate, document and manage their requirements (Abran, Moore, Bourque, Dupuis, & Tripp, 2004; Sommerville, 2010). Achieving similar goals, the VSD tripartite methodology (conceptual, empirical and technical investigation) can be considered as a requirement engineering activity for human values. In VSD, direct and indirect stakeholders are identified and relevant values elicited and analysed during the conceptual investigation. The empirical investigation further analyses stakeholder perception, the use context of a system and validates relevant values. The technical investigation specifies how technology can support certain values, or what values are implicated by already existing features (Manders-Huits, 2011; Friedman & Hendry, 2019).

Agile development starts with a rough approximation of the final requirements and adds details during the whole development process (Rees, 2002). In a similar fashion, VSD's tripartite method adds details on important values during the whole development project. In general, the tripartite methodology seems to be a prime candidate for an integration into agile development processes. Figure 1 exemplifies a potential integration of the tripartite methodology into Scrum, which is one of the most common agile development processes (Sharma, Hasteer, 2016). In Scrum, the tripartite methodology can be used for developing and maintaining the product backlog - or "value backlog" in this case. Additionally, during each sprint the tripartite method can be used to further specify and understand values and related concepts.

In agile development, requirement-related activities are performed during face-to-face meetings in collaboration with stakeholders (Ramesh et al. 2010). In contrast to traditional practices (waterfall or iterative model), the development team and all stakeholders are involved throughout the whole development process in all relevant requirement engineering activities (Paetsch et al. 2003; Sillitti & Succi, 2005). Similarly, value-oriented approaches also heavily rely on stakeholder participation (Manders-Huits, 2011), as human values can best emerge during active stakeholder interaction (Borning & Muller, 2012). Especially after larger development cycles (or “Sprint Execution” in Figure 1), additional requirements are identified by presenting the working product to stakeholders. Presenting a working product as design prop can facilitate discussions about human values (Koch, Proynova, Paech, & Wetter, 2013), which could increase the value consciousness of a project and mitigate drawbacks of using prototypes (Reilly Dearman, Welsman-Dinelle, & Inkpen, 2005).

Figure 1. Integration of tripartite methodology and value prioritization into scrum.



In agile development, requirements are considered as being decoupled from each other, allowing a flexible order of implementation according to their priority (Silliti & Succi, 2005). At the beginning of each new development cycle, requirements (or values in Figure 1) are sorted by their priority (Product Backlog), discussed, and chosen for implementation (Sprint Planning Meeting in Figure 1; Sharma, Hasteer, 2016). It is questionable whether decoupling is suitable for value-oriented approaches, as treating values independently from each other prevents resolving value tensions and threatens the delicate balance between values (Friedman & Hendry, 2019). A solution can be to implement values one after the other according to their priority and resolving value tensions during the implementation. Another solution can be to use the conception of values by *Value-based Engineering* based on material value ethics (Hartmann, 1932; Scheler, 1913-1916/1973). According to Value-based Engineering, values are coming as value clusters consisting of core values and related value qualities. Value qualities are instrumental to their specific core value (Spiekermann & Winkler, 2020). For instance, the core value *trust* is supported by value qualities such as *openness*, *accountability*, *privacy* and others (Spiekermann-Hoff, Winkler, & Bednar, 2019). Implementing one value cluster after the other has the advantage that value tensions and the balance between values is encapsualed within a cluster. This tension is expressed in the relation between core values (e.g. *trust*) and related value qualities (e.g. *openness, accountability, privacy*). In any case, prioritizing values (or value cluster) at the beginning of each development cycle (or sprint) using an appropriate light-weight method is essential for a successful integration of value-oriented approaches into agile development processes.

2.2. The fall from grace of agile development

While in the beginning agile development lived up to the goal of giving developers the necessary autonomy and flexibility and increased software and code quality (Dybå & Dingsøyr, 2008), this is not the case anymore in today's industry practice. The promise of shorter development time and rapid results resonated highly in industry (Howard, 2002; Savolainen, Kuusela, & Vilavaara, 2010) and on the management floor. Especially the management saw agility as a tool to increase productivity and time pressure, which lead to an emphasis of the "lean" mentality of agile development adapted from "lean manufacturing" (Poppendieck and Poppendieck, 2007). Studies show that developers are willing to go beyond traditional functional requirements but do not have the necessary time or autonomy to do so (Bednar, Spiekermann, & Langheinrich, 2019; Spiekermann, Korunovska, & Langheinrich, 2018). The aim of agile development to strengthen the role of individuals and the foster collaboration (Fowler & Highsmith, 2001) has not become part of today's industry practice. While stakeholders are supposed to be included in all development steps, in practice, they are substituted by on-site representatives (Sillitti & Succi, 2005). In previous software development processes, developers were bound to satisfy a process, in agile practice they are bound to meet the deadline in time. *Value-based Engineering* explicitly acknowledges this industry practice by calling for "a more careful use of agile forms of system development" (Spiekermann & Winkler, 2020, p. 16) and requiring value-oriented projects to take the necessary time for value considerations.

In conclusion, due to the iterative and integrative nature of the tripartite method there is the potential for integrating value-oriented approaches into agile development processes. Such an integration can be of mutual benefit, as it raises the recognition of value-oriented approaches outside of academia and increases the success of software products by considering human values (Spiekermann 2016; van den Hoven, 2017). Furthermore, agile development could gain back its own original focus on stakeholder collaboration. However, caution seems advisable for value-oriented approaches as key aspects, such as strong stakeholder participation and value consciousness, should not be sacrificed.

3. METHODS FOR VALUE PRIORITIZATION

Methods for prioritizing values do not only enable an integration into agile development processes, but are important for any value-oriented project. For instance, considering all 355 values from a value list (Winker & Spiekermann, 2019) can hardly be achieved by a development team; the attempt to fulfil this many value obligations could lead to moral overload (van den Hoven, 2013). None of the unique 17 VSD methods (Friedman & Hendry, 2019) is explicitly recommended for the task of value prioritization. Potentially, the Value Dams and Flows methods (Miller et. al. 2007) could achieve a value prioritization by *excluding* objected design options (value dams) or *including* appealing design options (value flows) and thus implicitly giving priority to related values. The openness of VSD to use the entire range of quantitative and qualitative methods (Friedman & Hendry, 2019) creates the opportunity to employ standard social science methods such as laddering interviews, card sorting tasks or any type of ranking or scaling techniques including Likert-scales. Using a simple Likert-scale sounds especially promising, as it is a light-weight and scalable method, allowing an afford-less inclusion of many stakeholders.

While not explicitly mentioned as a unique VSD method, Burmeister (2016) uses his own approach to value prioritization. The "Burmeister method" focuses on the frequency and emotional intensity with which values are expressed. To my knowledge, this is the only approach that directly includes emotional aspects, therefore recognising that emotions are an important source of moral knowledge, understanding and awareness (Desmet & Roeser, 2015).

Value-based Engineering uses three complementary ethical investigations to prioritize values during its ethical exploration phase (Spiekermann & Winkler, 2020). During these investigations, the development team, the management and ideally all stakeholders are involved in value prioritization. The first investigation assesses the overlap between core values and the business model. During the second duty ethical investigation, personal maxims and values that each prioritizer wants to have as universal law are chosen. Furthermore, values that are not instrumental have a higher priority and the hierarchy of values from material values ethics (Hartmann, 1932; Scheler, 1913-1916/1973) is considered. The third investigation requires to check the values against existing corporate principles, legal frameworks, international human rights agreements or ethical principles. While recognising practical considerations (e.g. involving the management, considering business model) is promising, due to the amount of involved people, the necessary knowledge (e.g. material value ethics, legal frameworks, human rights) and amount of steps, this cannot be considered as a light-weight method.

In their paper, van de Kaa, Rezaei, Taebi, van de Poel, and Kizhakenath (2020) use a two-step approach to value prioritization based on expert opinions. First, experts qualitatively validate values from value lists and create a set of values for each stakeholder group. Secondly, the Best Worst Method (BWM) suitable for highly complex systems (van de Kaa et al. 2020) was used to assign weights (best and worst), determining the most important and least important values of a value set. The most important and subsequently the least important value is used to rate other values of a set using numbers between 1 (equally important) and 9 (extremely more important). Finally, the optimal weight (importance) for each value is derived. While such a type of analysis has been criticized for practically reducing all values into a single value of *utility* (van de Kaa et al. 2020), employing pairwise comparison to reduce the difficulty of the prioritization task seems reasonable. While the reduction of several values to *utility* must be seen critically, this step might make the benefits of value-conscious development graspable for the management and justify additional development time. Other potential methods for value prioritization that also focus on the utility are *discrete choice analysis* and *conjoint analysis* (Breidert, Hahsler, & Reutterer, 2006).

In summary, there are several methods available to prioritize values during a value-oriented development project. As the notion of agile development calls for light-weight methods, I decided to try out a simple Likert-rating scale.

3.1. Using a Likert-rating scale to prioritize values

Based on a value list (Winkler & Spiekermann, 2019), 48 values were identified by the author as relevant for the development of a mobile navigation application. As part of an online survey about mobility behaviour in Vienna, these values were presented in random order to participants. Overall, 264 participants rated values on a Likert-scale ranging from 1 (not important) to 7 (extremely important). The following question was presented to introduce this task: “How important do you consider the following attributes for the development of a mobile navigation app?” Participants were recruited via a university mailing list of the Vienna University of Economics and Business. Participants mean age is 26.34 years (± 6.77) of which 55.3% identified themselves as female and 44.7% as male. Of the initial 48 values, eighteen values had a mean below 4 (neither nor) and therefore were considered as unimportant. Most notably, the values *health*, *human well-being*, *dignity* and *justice*, initially considered as important by the author and considered as high values according to material values ethics, were considered as unimportant by the participants. Of the remaining 30 values, five technical values (reliability, usability, dependability, simplicity and efficiency) were considered as highly important achieving means higher than 6 (very important). Fifteen values, including privacy, trust, security, environmental protection and freedom from bias, received a mid-range importance between 6 (very important) and 5 (slightly important). The remaining 10 values achieved a rating between 5 (slightly important) and 4 (neither nor). Comments stated by the participants indicated high-level frustration and confusion during this rating task. Many participants were not able to see the connection between a mobile navigation application and for instance values such as health or human well-being.

These results are devastating for the idea to use the most light-weight and scalable method for value prioritization. In this case, especially technical values were considered as highly important, which seriously endangers the human and social value consciousness of a project. Simply letting stakeholders rate values, can therefore not be an appropriate method for value prioritization.

This bias towards instrumental values might be a sampling effect, caused by the participants' educational background as economy students. Such a bias could indicate an availability heuristic, leading people to consider recently discussed information or values (e.g. *efficiency*) as being more important than others (Wänke, Schwarz, & Bless, 1995). Implicit association tests assessing subconscious associations between concepts might be a more appropriate approach for further research in this direction (Greenwald, McGhee, & Schwartz, 1998). Rating values in an online setting facilitates fast responses but limits reflection upon the true importance of values. Serious play methods (Garde & van der Voort, 2016) to prioritize values (e.g. using bricks) could be another pathway to increase stakeholder reflection. As values are context-sensitive, setting a generic context (as was done here with the navigation application) could lead to the prioritization of a particular set of values. These preliminary results show the need for further empirical research to assess effects of stakeholder groups, used prioritization methods and established context on stated value priority.

4. CONCLUSION

An integration of value-oriented approaches into agile development processes has the potential to increase the recognition of value-conscious development outside of academia. Such an integration must be pursued with caution as stakeholder collaboration, a key feature of value-oriented approaches, is currently a practice not lived in industry. Furthermore, properly considering values during software development needs time, which is a scarce resource in the area of software development. Value prioritization methods can be a pathway towards successful integration into agile development processes. While light-weight methods would facilitate an integration, results presented here show that using a Likert-scale is not a suitable approach to prioritize values. However, there are several other methods, of which some might be capable of proofing the utility for value consciousness to managers.

REFERENCES

Abran, A., Moore, J. W., Bourque, P., Dupuis, R., & Tripp, L. (2004). Software engineering body of knowledge. IEEE Computer Society, Angela Burgess.

Andreessen, M. (2011). Why software is eating the world. *Wall Street Journal*, 20(2011), C2.

Bednar, K., Spiekermann, S., & Langheinrich, M. (2019). Engineering Privacy by Design: Are engineers ready to live up to the challenge? *The Information Society: An International Journal*, *35*(3), 122–142. https://doi.org/10.1080/01972243.2019.1583296

Burmeister, O. K. (2016). The development of assistive dementia technology that accounts for the values of those affected by its use. *Ethics and Information Technology*, 18(3), 185-198.

Breidert, C., Hahsler, M., & Reutterer, T. (2006). A review of methods for measuring willingness-to-pay. Innovative Marketing, 2(4), 8-32.

Clancy, T. 2014. "The Standish Group Report," Retrieved from: https://www.projectsmart.co.uk/white-papers/chaos-report.pdf, Jan 15, 2018

De Lucia, A., & Qusef, A. (2010). Requirements engineering in agile software development. Journal of emerging technologies in web intelligence, 2(3), 212-220.

Desmet, P. M., & Roeser, S. (2015). Emotions in design for values. van den Hoven et al, 203-219.

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. Information and software technology, 50(9-10), 833-859.

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2005, April). Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 751-760).

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *In Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer, Dordrecht.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination.* Mit Press.

Fowler, M., & Highsmith, J. (2001). The agile manifesto. Software Development, 9(8), 28-35.

Garde, J. A., & van der Voort, M. C. (2016). Could LEGO® Serious Play® be a useful technique for product co-design? *Design Research Society*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6), 1464.

Hartmann, N. (1932). *Ethics*. London: George Allen & Unwin.

Howard, A. (2002). Rapid application development: Rough and dirty or value-for-money engineering? *Communications of the ACM*, 45(10), 27-29.

Koch, S. H., Proynova, R., Paech, B., & Wetter, T. (2013). How to approximate users' values while preserving privacy: experiences with using attitudes towards work tasks as proxies for personal value elicitation. Ethics and information technology, 15(1), 45-61.

MacCormack, A., Verganti, R., & Iansiti, M. (2001). Developing products on "Internet time": The anatomy of a flexible development process. Management science, 47(1), 133-150.

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. Science and engineering ethics, 17(2), 271-287.

Mellis, W., Loebbecke, C., & Baskerville, R. (2010). Moderating effects of requirements uncertainty on flexible software development techniques. In International Research Workshop on IT Project Management.

Miller, J. K., Friedman, B., Jancke, G., & Gill, B. (2007, November). Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. *In Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 281-290). ACM.

Nerur, S., Mahapatra, R., & Mangalaraj, G. (2005). Challenges of migrating to agile methodologies. Communications of the ACM, 48(5), 72-78.

Obermeyer, Z., & Mullainathan, S. (2019, January). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 89-89). ACM.

O'Neill, C. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group.

Reilly, D., Dearman, D., Welsman-Dinelle, M., & Inkpen, K. (2005). Evaluating early prototypes in context: trade-offs, challenges, and successes. IEEE Pervasive Computing, 4(4), 42-50.

Royce, W. W. 1970. "Managing the Development of Large Software Systems," Retrieved from: http://www. cs.umd.edu/class/spring2003/cmsc838p. Aug 08, 2017

Paetsch, F., Eberlein, A., & Maurer, F. (2003, June). Requirements engineering and agile software development. In WET ICE 2003. Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003. (pp. 308-313). IEEE.

Pereira, R., & Baranauskas, M. C. C. (2015). "A value-oriented and culturally informed approach to the design of interactive systems," *International Journal of Human-Computer Studies*, pp. 66-82.

Poppendieck, M., & Poppendieck, T. (2007). Implementing lean software development: From concept to cash. Pearson Education.

Ramesh, B., Cao, L., & Baskerville, R. (2010). Agile requirements engineering practices and challenges: an empirical study. *Information Systems Journal*, 20(5), 449-480.

Rees, M. J. (2002, December). A feasible user story tool for agile software development? In Ninth Asia-Pacific Software Engineering Conference, 2002. (pp. 22-30). IEEE.

Savolainen, J., Kuusela, J., & Vilavaara, A. (2010, September). Transition to agile development-rediscovery of important requirements engineering practices. In 2010 18th IEEE International Requirements Engineering Conference (pp. 289-294). IEEE.

Schwaber, K. (1997). Scrum development process. In Business object design and implementation (pp. 117-134). Springer, London.

Steen, M., & van de Poel, I. (2012). Making values explicit during the design process. *IEEE Technology and Society Magazine*, *31*(4), 63–72. https://doi.org/10.1109/MTS.2012.2225671

Sharma, S., & Hasteer, N. (2016, April). A comprehensive study on state of Scrum development. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 867-872). IEEE.

Sommerville, I. 2010. *Software engineering,* New York: Addison-Wesley.

Spiekermann, S. (2016). *Ethical IT innovation: A value-based system design approach*. Boca Raton: CRC Press.

Spiekermann-Hoff, S., Winkler, T., & Bednar, K. (2019). *A telemedicine case study for the early phases of value based engineering* (Working Paper Series/Institute for IS & Society No. 001_vs 1). *Working Paper Series/Institute for IS & Society*. Vienna: WU Vienna University of Economics and Business. Retrieved from https://epub.wu.ac.at/7119/

Spiekermann-Hoff, S., Winkler, T., & Bednar, K. (2019). *A telemedicine case study for the early phases of value based engineering* (Working Paper Series/Institute for IS & Society No. 001_vs 1). *Working Paper Series/Institute for IS & Society*. Vienna: WU Vienna University of Economics and Business. Retrieved from https://epub.wu.ac.at/7119/

Spiekermann, S., & Winkler, T. (2020). Value-based Engineering for Ethics by Design. *ArXiv, 2004.13676*. Retrieved from https://arxiv.org/abs/2004.13676

Scheler, M. (1973). *Formalism in ethics and non-formal ethics of values: A new attempt toward the foundation of an ethical personalism [1913-1916]*. (M. S. Frings & R. L. Funk, Eds.). Evanston, Ill: Northwestern University Press. https://doi.org/10.2307/2707101

Schmidt, C. (2016). Agile software development teams. Springer International Publishing.

Sillitti, A., & Succi, G. (2005). Requirements engineering for agile methods. In *Engineering and Managing Software Requirements* (pp. 309-326). Springer, Berlin, Heidelberg.

van de Poel, I. (2015). Design for values in engineering. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 667-690.

van de Kaa, G., Rezaei, J., Taebi, B., van de Poel, I., & Kizhakenath, A. (2020). How to weigh values in value sensitive design: A best worst method approach for the case of smart metering. Science and Engineering Ethics, 26(1), 475-494.

van den Hoven, J. (2017). Ethics for the digital age: Where are the moral specs? In Informatics in the Future (pp. 65-76). Springer, Cham.

Van den Hoven, J. (2013). Value sensitive design and responsible innovation. Responsible innovation: Managing the responsible emergence of science and innovation in society, 47, 75-83.

Verbeek, P. P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In *Philosophy and design* (pp. 91-103). Springer, Dordrecht.

Wänke, M., Schwarz, N., & Bless, H. (1995). The availability heuristic revisited: Experienced ease of retrieval in mundane frequency estimates. Acta Psychologica, 89(1), 83-90.

Winkler, T., & Spiekermann, S. (2019). Human Values as the Basis for Sustainable Information System Design. *IEEE Technology and Society Magazine*, 38(3), 34-43.

Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*, 1-5.

VALUES AND POLITICS OF A BEHAVIOR CHANGE SUPPORT SYSTEM

Janet Davis, Buyaki Nyatichi

Whitman College (USA)

davisj@whitman.edu; nyaticmb@whitman.edu

ABSTRACT

We report on an analysis of 27 media opinions concerning *Just Not Sorry*, a tool intended to influence word choice when composing email. Designed for and by women in business, *Just Not Sorry* persuades users to avoid apologies such as “sorry,” hedge words such as “just,” and intensifiers such as “very.” Our media analysis explicates differing positions on whether *Just Not Sorry* supports gender equality, while confirming the relevance of values such as achievement, autonomy, privacy, and politeness, and implicating further values of mindfulness and sincerity. We propose media analysis as a “discount” method for empirical discovery of values and value tensions.

KEYWORDS: Persuasive technology, values, email, language, gender, feminism.

1. INTRODUCTION

Just Not Sorry is a Gmail plug-in that highlights when the user writes apologies such as “sorry,” hedge words such as “just,” and intensifiers such as “very” (Def Method, 2019b). Red underlines appear as if the words had been misspelled. For each underline, a motivational quote appears as a tooltip. One such quote follows as an example: “Using ‘sorry' frequently undermines your gravitas and makes you appear unfit for leadership - Sylvia Ann Hewlett.”

In this case study, we consider *Just Not Sorry* as an example of a persuasive technology—that is, a technology designed to change attitudes and behaviors (Fogg, 2002). *Just Not Sorry* came to our attention through a survey of technologies designed to influence speech and writing (Twersky & Davis, 2017). When we describe *Just Not Sorry* to others, it elicits strong and opposing reactions. Some say, “I need that!” while others say, “I would never use a tool like that.” Our research question for this case study: What might explain such strong, opposing reactions?

To address this question, we adopted value sensitive design (Friedman et al., 2006) as our guiding theory and methodology, intertwining conceptual, technical, and empirical investigations. In our initial conceptual investigations, we considered direct and indirect stakeholders and the values implicated by those relationships. In parallel technical investigations, the authors each installed *Just Not Sorry* and used it for at least one month. We examined the source code on GitHub (Def Method, 2019a), as well as the system image presented in the Chrome Web Store (Def Method, 2019b). To understand the intentions behind

Just Not Sorry, we read Tami Reiss's (2015) story of the tool's conception and design. Finally, to address our research question, we conducted a content analysis of media opinions about *Just Not Sorry*. The main goal of this paper is to report on the findings of that analysis.

In the next section, we present as background our initial conceptual investigations. After refining our research question, we describe the method and results of the media analysis. We conclude with a brief discussion of the findings and implications for future value sensitive design studies.

2. PRELIMINARY CONCEPTUAL INVESTIGATIONS

2.1. Stakeholders and values

Just Not Sorry was inspired by a conversation among women in leadership positions about recent satire exaggerating women's stereotypical overuse of words such as "just" and "sorry" (Reiss, 2015). According to Reiss (2015), the group agreed these stereotypes were true: "The women in these rooms were all softening their speech in situations that called for directness and leadership. We had all inadvertently fallen prey to a cultural communication pattern that undermined our ideas." Reiss proposed a tool to help women change their behavior. In this way, *Just Not Sorry* was designed by and for women, to promote the status of women. However, the description of *Just Not Sorry* on the Chrome Web Store (Def Method, 2019b) does not mention gender, and we found masculine as well as feminine names among the authors of public reviews. Therefore, while gender and thus *identity* is salient to the design of *Just Not Sorry*, it is likely that users (or direct stakeholders) include both women and men.

Just Not Sorry is intended to enhance users' *achievement* and *social power* through its support for behavior change. From the Chrome Web Store description and our inspection of the tool, we see that *Just Not Sorry* respects users' *autonomy* in that it suggests changes but does not compel them. Finally, *Just Not Sorry* protects users' *privacy* in that it leaves no trace of its use when emails are sent.

Indirect stakeholders include email recipients, colleagues, and feminists. Email recipients may perceive an email composed in accordance with *Just Not Sorry* as rude or abrupt, implicating the value of *politeness* or *courtesy*. If *Just Not Sorry* does indeed promote workplace success, colleagues who do not use the tool may find their own achievement and social power lessened. Changes in communication style across a workplace or industry may enhance either *collaboration* or *competition*.

While still more indirect, feminists constitute a large stakeholder group with a substantial interest in *Just Not Sorry*. Bucholtz (2014 [1991]) offers the following definition of feminism:

> A diverse and sometimes conflicting set of theoretical, methodological, and political perspectives that have in common a commitment to understanding and challenging social inequalities related to gender and sexuality.

Hence, considering feminists as a stakeholder group implicates the value of *equality* with respect to gender. We refined our research question as follows: Could conflicting perspectives on gender equality explain strong, opposing reactions to *Just Not Sorry?* In the next section, we consider the diversity of feminist perspectives, drawing on Bucholtz (2014 [1991]) as our primary source.

2.2. Feminist perspectives on language

We find that it is one particular kind of feminism that motivates *Just Not Sorry*. Reiss (2015) seeks to enhance women's positions within existing structures by helping them address a perceived deficiency in their attitudes and behaviors. This is a textbook example of a liberal feminist approach: "Given its concern to bring women into men's spheres, liberal feminism has generally aimed to eradicate gender inequality by eradicating or at least reducing gender difference," efforts that "have sometimes resulted in societal expectations that women must adapt to male norms" (Bucholtz, 2014 [1991]). Bucholtz goes on to write that Robin Lakoff is the linguist most associated with liberal feminism. Lakoff's groundbreaking work, *Language and Woman's Place* (2004 [1975]), describes how a culture that expects women to use "women's language," including hedge words, vacuous modifiers, and superpolite forms, is one means by which "women are systematically denied access to power." Lakoff also applies the concept of the double bind to women's language: if women adopt the language of power, they are still denied access to power due to their unfeminine behavior.

Contemporary feminist scholars Gill and Orgad (2017) associate *Just Not Sorry* with the movement they call "confidence culture," in which the persistence of gender inequality is attributed to individual shortcomings that can be addressed through projects of self-improvement. Where Bucholtz (2014 [1991]) observes that liberal feminism "is often no longer recognized as feminism at all," Gill and Orgad describe confidence culture as a popular, postfeminist remaking of feminism.

While liberal feminism seeks equality through the reduction of gender differences, cultural feminism views women's communication practices as distinctive and of value equal to or greater than men's (Bucholtz, 2014 [1991]). Bucholtz cites Deborah Tannen's work as emblematic of a cultural feminist approach. Tannen's work, including bestsellers *You Just Don't Understand* (1990) and *Talking from 9 to 5* (1994), attributes miscommunication between men and women to gendered communication practices, including men's cultural preference for "report talk" and women's for "rapport talk."

Beyond liberal and cultural feminism, Bucholtz (2014 [1991]) explains a third and less common branch of difference feminism: radical feminism, in which all inequality is viewed has having gender inequality at its root. And beyond theories that view gender as an essential difference, we find

- material feminism, which "holds that women's subordination is a consequence of class oppression" (Bucholtz, 2014 [1991]);
- critical race feminism, which "challenges the field's tendency to marginalize the distinctive experiences of women of color" (Bucholtz, 2014 [1991]);
- queer feminisms, which challenge heteronormativity and the gender binary.

3. METHODS

3.1. Forming the corpus

To form the corpus, we conducted a Google Search for the keywords "Gmail 'Just Not Sorry'" on December 13, 2019. We did not use a news database, such as NewsBank, ProQuest, or NexisUni, because we found these databases do not include relevant women's publications such as *Vogue* and *Marie Claire*, and do not consistently index popular online news sources such as *Slate*.

Amongst over 100 search results, we identified 26 articles that meet the following criteria:

1. The article is published in a blog, magazine, or newspaper that is recognized as notable by its inclusion in Wikipedia. This excludes not only personal blogs and web sites, but also small business blogs and minor news sources such as high school newspapers. We wanted to ensure that all the articles included in our analysis were clearly intended for a public audience.
2. The article's author expresses a substantive opinion regarding *Just Not Sorry*, including some rationale. This excludes purely factual reporting, including reporting on others' opinions, as well as articles that follow a factual report with a brief, unsupported evaluation (e.g., "Genius!")

Most articles are the work of a single author, but Day and Ellen (2016) wrote in conversation with each other. Their segments are coded as two separate cases, leading to the 27 cases listed in Table 1.

Table 1. All cases, classified as supportive, critical, or equivocal with respect to *Just Not Sorry.*

(a) Supportive tone

Author	Gender	Genre
Brandon (2015)	M	Business & Tech
Curtis (2018)	F	Business & Tech
Day (in Day & Ellen, 2016)	F	News (UK)
Erikson (2016)	F	Women
Fessler (2018)	F	Business & Tech
Gillespie (2016)	F	Women
Ginn (2016)	F	News (UK)
Hines (2016)	F	Women
Guest (2016)	F	News (UK)
Lastoe (2016)	F	Business & Tech
Lord (2015)	F	Women
Paul (2016)	M	Business & Tech
Scott (2016)	F	News (UK)
Stevens (2016)	F	News (US)
Wills (2016)	F	News (UK)
Ye (2016)	F	Business & Tech

(b) Critical tone

Author	Gender	Genre
Grose (2016)	F	News (US)
Horobin (2016)	M	News (UK)
Minter (2016)	F	News (UK)
Sawyer (2016)	F	Business & Tech
Tejada (2016)	F	Women

(c) Equivocal tone

Author	Gender	Genre
Cauterucci (2015)	F	News (US)
Dishman (2016)	F	Business & Tech
Ellen (in Day & Ellen, 2016)	F	News (UK)
Garcia (2016)	F	Women
Levine (2016)	F	News (US)
Turk (2019)	F	Business & Tech

3.2. Coding and analysis

Coding and analysis was performed by the first author, using a hybrid approach guided by the recommendations of Lazar et al. (2017). After reading the articles several times, I classified the overall tone of each author as supportive, critical, or equivocal with respect to *Just Not Sorry*, as shown in Table 1. I inferred each author's gender from their name and classified the publications by genre. I then proceeded to open coding, with attention to authors' treatment of gender as well as their arguments supporting, opposing, or critiquing *Just Not Sorry*. From prior experience, I was sensitized to statements alluding to values as framed by Friedman et al. (2006) and by Schwartz (1994). From our preliminary conceptual investigation, I was also sensitized to feminist concepts as presented by Bucholtz (2014). After the initial coding was complete, I identified open codes with feminist perspectives and classified each author's perspective according to the predominance of the evidence. In parallel, I identified open codes as arguments supporting or opposing *Just Not Sorry* and grouped similar arguments through axial coding, an iterative process in which some open codes were revisited.

4. RESULTS

4.1. Does the corpus include feminist viewpoints?

The composition of the corpus confirms the relevance of gender: the vast majority of the authors (24/27) have women's names, and six articles appear in publications for women. But few opinions come from an explicitly feminist position. Only Ellen (2016), Turk (2019), and Wills (2016) refer directly to "feminism" or a "feminist issue," and just six (Cauterucci, 2015; Curtis, 2016; Ellen, 2016; Ginn, 2016; Horobin, 2016; Lastoe, 2016) mention gender equality or inequality, the central concern of feminism.

Yet the content analysis reveals many statements consistent with an underlying perspective of liberal or cultural feminism. Statements typical of liberal feminism include problematizing women's language, taking masculine behavior as the norm or ideal, arguing that women need to change to achieve success in traditionally male workplaces, and articulating a double bind.

> Acting like a spellcheck for sorries, the app was designed to help women who are prone to using "soft" language at work and thereby sending emails which are about as effective as putting on a baby voice when asking for a pay rise. (Wills, 2016)
>
> Way back when, my high school English teacher, Mrs. Skoog, told my all-women's class to stop qualifying our speech and our writing with the word "just." … Mrs. Skoog encouraged us to…say and write what we meant with confidence and authority—the way men did. (Lastoe, 2016)
>
> [L]anguage is the key element in helping women progress to the top and succeed there. The question is then, how can women change their language to convey confidence? (Ginn, 2016)
>
> If emailing "like a woman" is to perpetuate stereotypes about how women should act, but emailing "like a man" is to reinforce the idea that professionalism should aspire to male corporate culture—and if either approach can be held against you anyway—then what's a female emailer to do? (Turk, 2019)

By contrast, cultural feminists value women's language and behavior as much as or more than men's.

> [W]omen can also be superb, hyper-intuitive people-readers and managers… So, while the female approach isn't always softer, when it is, let's own it. (Ellen, in Day & Ellen, 2016)

> [I]t is also my hope that "Just Not Sorry" isn't indicative of yet another way women need to act more like men in order to succeed. Sometimes, it's okay to soften language if that feels more authentic. In many cases, it's paramount to consider people's reactions and sensitivities before speaking. (Levine, 2016)
>
> Rather than holding our hands up and apologising for our choice of words, let's stand up for them. Let's stand up for taking people's feelings into consideration when we speak, for not seeing arrogance as a virtue, for thanking people for their contributions and for being sorry for putting our work onto other people. Let's stop apologising for being women and instead demand that men behave differently. (Minter, 2016)

Other feminist stances beyond liberal and cultural feminism do not seem to be represented in the corpus. Fessler (2018) and Wills (2016) mention "the patriarchy," but colloquially and not in the context of radical feminism. No one writes from the standpoint of critical race feminism; only Turk (2019) briefly considers biases beyond gender bias. And while many disagree with drawing a stark distinction between women's and men's behavior, all assume gender is binary. For example:

> I'm disinclined to think this is a clear male v female divide - we've all experienced colleagues, both men and women, who fall on both ends of the spectrum. And although the idea that men and women speak a different language is age old, I've seen no solid evidence that women write 'sorry' and 'just' in emails more than men. (Hines, 2016)

Classifying all cases according to their predominant sentiments (Table 2) reveals many more liberal feminists (17) than cultural feminists (3), as well as two more who advocate for gender equality without clearly adopting a stance of liberal or cultural feminism. Three discuss gender only to downplay its relevance, while two writing for business publications avoid discussing gender altogether.

Table 2. Cases by position on gender equality.

Liberal feminists	Cauterucci, Curtis, Day, Dishman, Fessler, Garcia, Gillespie, Ginn, Hines, Lastoe, Lord, Sawyer, Scott, Stevens, Tejada, Turk, Wills
Cultural feminists	Ellen, Grose, Minter
Other feminists	Horobin (makes neither kind of statement), Levine (makes both equally)
Downplay gender	Erikson, Guest, Paul
Do not discuss gender	Brandon, Ye

4.2. Do feminist positions predict opinions about *Just Not Sorry?*

Reviewing Table 1 shows that we cannot predict authors' opinions about *Just Not Sorry* based on either their gender or their audience. But can we predict their opinions from their positions on gender equality? Table 3 shows that, to some extent, we can. In this corpus, authors who downplay or ignore gender all recommend *Just Not Sorry*, cultural and other feminists all critique or oppose *Just Not Sorry*, and liberal feminists are divided. The remainder of this section explicates these differences in opinion.

Table 3. Position on *Just Not Sorry* vs. position on gender equality.

	Supportive	Equivocal or Critical
Liberal feminists	Curtis, Day, Fessler, Gillespie, Ginn, Hines, Lastoe, Lord, Scott, Stevens, Wills	Cauterucci, Dishman, Garcia, Sawyer, Tejada, Turk
Cultural feminists	-	Ellen, Grose, Minter
Other feminists	-	Horobin, Levine
Downplay gender	Erikson, Guest, Paul	-
Do not discuss gender	Brandon, Ye	-

4.3. Disagreements about the value of the target behavior change

Brandon (2015) and Ye (2016) argue for *Just Not Sorry* without reference to gender. Like the majority of the authors (21/27)—including all 16 who ultimately recommend *Just Not Sorry*—Brandon and Ye begin by problematizing some of the words and phrases flagged by the app, characterizing these words as “wishy-washy,” “weak,” “timid,” even “submissive,” in contrast with the “confident,” “strong,” “authoritative” language they see as key to effective business writing. Along with Erikson (2016) and Paul (2016), who mention gender only to say that the app can be used by men as well as women, Brandon and Ye recommend *Just Not Sorry* to support individual achievement. Hines (2016), here classified as a liberal feminist, most explicitly addresses achievement:

> [I]f it helps me be part of the high-achieving crew, I’m happy to use every tool out there to suppress my natural tendency to put ‘not offending anyone’ ahead of ‘getting the job done’. (Hines, 2016)

On the other hand, as we have already seen, those taking a cultural feminist perspective disagree that the behavior targeted by *Just Not Sorry* is problematic. They argue instead that a “softer” or more considerate approach has value. Grose (2016) adds that *Just Not Sorry* ignores social context:

> What is appropriate, effective language when writing to a boss might not work with a subordinate; … you might not even use the same style of writing when communicating with a woman colleague as with a man.

Grose (2016) goes on to argue that feminine communication styles “can be incredibly useful in the realpolitik of the workplace,” including the use of “sorry” as a “conversational smoother.” Ellen (2016) also links apologies to “good manners,” which she values in “both sexes”:

> Like me, I’m sure you’ve come across presumptuous braggarts, shameless buck-passers and dreary blame-dodgers of both sexes; the types that think basic good manners are for other people. … Is it really so bad to say “Sorry for bothering you…” at the start of an email? Where some see feeble and self-defeating, I see human and perceptive – and how about a round of applause for just plain nice?

Horobin (2016) argues that apologies are not just valuable but often necessary, independent of gender:

> The problem with an app designed to discourage the use of apologetic language is the assumption that saying sorry is always an act of contrition – one that undermines one's case, or assumes an inferior position. Since saying sorry can be a valuable means of refusing a request: "Sorry, but I'm just too busy right now"; or enlisting someone's support – "I'm sorry to bother you when I know you're busy." Removing this useful word runs the risk of making you appear plain rude. There are many business contexts in which an email to a customer or a boss requires apologetic language; to avoid such politeness strategies when explaining why a report is late, or asking for a pay rise, would be a risky policy indeed.

To sum up, Brandon, Erikson, Paul, and Ye recommend *Just Not Sorry* because they value "strong" language in support of achievement. Ellen, Grose, Horobin, Levine, and Minter oppose *Just Not Sorry* because they value feminine language, or politeness irrespective of gender.

In the bottom four rows of Table 3, just Guest remains. Similar to Ye, Guest (2016) writes about apologizing as "a hard habit to shake." But where Ye (2016) recommends *Just Not Sorry* for supporting "repetition and practice" towards "improving your writing," Guest (2016) sees other values at stake:

> I do use "sorry" liberally when I notice I've caused damage or am in the wrong, but I no longer apologise for myself, my opinions, or other people's mistakes. … When "sorry" is used, it is more meaningful for being meant. … The app should make us stop and think – whether we are British, or women, or in the workplace – about when a thing is really worth apologising for.

Guest values the role of *Just Not Sorry* in promoting mindfulness and eliminating habitual apologies, leading to greater sincerity. Some liberal feminists make similar claims about mindfulness and sincerity:

> Anything that makes us more aware of the language we use and the effect it has can only be a good thing "in my opinion" (whoops). (Wills, 2016)
>
> Of course, every now and then you really do screw up and ought to send an apology, and every now and then you're not entirely sure about something — which is why we have the words "I think" in the first place. Fortunately, the email doesn't reflect any of the suggested changes or stay underlined once it's sent. It merely encourages you to take a second glance at an email to make sure that the words you're using are actually words that you mean. (Lord, 2015)

Note that Lord links sincerity and mindfulness to *Just Not Sorry's* support for autonomy and privacy.

Returning to Table 3, we see that we cannot predict whether a liberal feminist will recommend *Just Not Sorry*. Like Brandon, Erikson, Paul, and Ye, liberal feminists problematize words or phrases targeted by *Just Not Sorry*. Most go on to implicitly or explicitly tie the problematic language to gender. Curtis (2018) and Fessler (2018) directly characterize such writing as "lady language," while some like Stevens (2016) cite "evidence" that women write or speak differently from men. Others foreground their identities as women in connection with such language. For example, Day (2016) writes about how she removed "filler words" from her emails after writing a novel with a "bombastic male protagonist," while Garcia (2016) and Lord (2015) address their presumed female audience as "you" or "we."

Some liberal feminists, including Hines, Wills, and Lord, recommend *Just Not Sorry* for reasons we have already seen. Others critique the completeness of *Just Not Sorry* with respect to the intended behavior change. Dishman (2016) points out that the app does not distinguish between the many uses of the word "just," not all of which are diminishing, while Garcia (2016) finds that *Just Not Sorry* does not flag all of *Vogue*'s "six things every working woman should avoid when writing an email." Sawyer (2016) agrees that women should change, but disagrees that *Just Not Sorry*'s focus on email is effective:

> An email that omits words like "sorry" will do little to improve how you're viewed in the workplace, especially if you still use such words during in-person meetings and presentations.

4.4. Disagreements about the effects of the intervention on the user

Reiss (2015) intends *Just Not Sorry* to help women in leadership write "with the confidence of their positions." Some liberal feminists also hope the app will help make confident women:

> Our favourite part is that when you hover your cursor over the highlighted word, up pops a quote from a successful woman to remind you that you, too, are a strong woman who doesn't need to be apologising for anything. (Gillespie, 2016)
>
> With less than 10% of executive directors at FTSE 100 companies being women there are areas that need improvement to reach gender equality; one of these areas is women's ability to assert themselves and confidence. Changes to how we speak and how we hold ourselves to confront these subtle, yet important behaviors will begin to make a big difference. (Ginn, 2016)
>
> While I think some softening language is sometimes needed, I'd prefer to sound in control and definitive and have my words matter—rather than sound soft and sweet, careful of not coming across as demanding. (Lastoe, 2016)

But some liberal feminists believe *Just Not Sorry* will have the opposite effect, undermining rather than bolstering women's confidence. This is the crux of Cauterucci's (2015) critique:

> [P]art of me always cringes when people tell women that the way they speak or write is wrong. … Making fun of the way women speak, when they've been socialized for a lifetime to take up as little physical, temporal, and aural space as possible, is not productive and can further erode their self-confidence.

Cauterucci (2015) seems to recommend *Just Not Sorry* in the end, but her recommendation is equivocal:

> This app relieves women of a bit of the sizeable burden of realigning their subconscious word choices though the hover-over explanations could be tweaked to read as more encouraging than blame-y.

Tejada (2016) goes further and rejects *Just Not Sorry* for engendering shame around women's language:

> When someone types these kinds of words, they get underlined in red. (Get it? It's as if the words are misspelled! For shame, ladies!) … Instead of creating apps that shame women…we should be writing our own stories about intelligent women who don't care whether they're liked.

Horobin (2016) names this concern directly, citing cultural feminist Deborah Cameron's objections to "efforts to police women's language." Cultural feminists Ellen and Grose also argue this point:

> It's arguable that the female "sorry" communication tic has been overplayed – overexamined, exaggerated and distorted in a way that traditional "masculine" mannerisms would never be. It errs on the patronising – "This can help you with that pathetic girly thing you do that renders you a disgrace to modern feminism." (Ellen, in Day & Ellen, 2016)
>
> My fervent hope for 2016 is that there are fewer articles and tech hacks preaching at women — particularly young women — about how they should be speaking, writing and presenting themselves to the world. Maybe if their communications weren't constantly picked apart, even by well-meaning observers, they'd have more of the deeply felt confidence they need to succeed. (Grose, 2016)

Hence, this is an objection that unites some liberal and cultural feminists.

4.5. Disagreements about implications for gender equality

As we saw earlier, Ginn (2016) implies that *Just Not Sorry* will help boost women into positions of authority and thereby promote gender equality. Her opinion is by far the most hopeful. Four other liberal feminists recommend *Just Not Sorry* as "a step in the right direction" while acknowledging there are bigger problems for gender equality that it does not address:

> Gender inequality has had its influence on both women's and men's speech for centuries. A Gmail plug-in won't completely dismantle linguistic gender stereotypes, but pointing out simple words that unintentionally undermine certain voices is a step in the right direction. (Curtis, 2018)
>
> [T]his plug-in is not the most important stride that's ever been made in the ongoing battle for women's rights, but it's still a nice bit of wood thrown on to the bonfire. (Day, 2016)
>
> No Gmail plugin will topple [the patriarchy], but calling out the ways in which we unintentionally diminish our voices is a meaningful step forward. (Fessler, 2018)

> [W]hile we agree…it's rubbish that 'female' patterns of speech are still seen as weak, and it'd be great if we could change how people read our speech rather than having to change the speech itself – the app does make a good way to tackle the current struggle until we sort out the bigger issues. (Scott, 2016)

Turk, too, reports that she uses *Just Not Sorry* begrudgingly in the absence of "real empowerment":

> I use [the] plugin, but I don't always change the words it suggests. Instead, it helps me decide whether I'm saying what I actually want to say. It's the difference between writing a certain way because you want to, or because you feel you have to. Ultimately, real empowerment would be not having to think about how we come across at all. Needless to say, there are conspicuously few think pieces out there about how men should email. (Turk, 2019)

But some oppose *Just Not Sorry* because it "misses" or "ignores" a larger problem. For Sawyer (2016), the larger problem is that women "lack confidence and have few mentors to help them navigate the professional landscape." Horobin (2016) claims that *Just Not Sorry* is beside the point altogether:

> To advocate that women imitate male speech in order to gain equality in the workplace is to ignore the real problems about gender inequality which have nothing to do with the way women dress or speak. … I'm sorry, but I just don't think that an app will help.

And Ellen (2016) fears that *Just Not Sorry* could even "mask" bigger problems of gender inequality:

> My concern is that email drives such as these, while good for awareness, aren't going to tackle the far more entrenched issues, and might even serve as another mask. As in, "Oh look, there's a 'Just Not Sorry' email plug-in – gender disparity in the workplace is solved!"

5. DISCUSSION

As we have just seen, differing perspectives on gender equality go some distance towards explaining disagreements about *Just Not Sorry*. While some agree that apologies and hedges are signs of weakness, others value this softening and attribute it to gender. While some see the app boosting women's confidence, others predict the opposite effect. While some view the app as a small step towards gender equality, others see it as a distraction from more fundamental issues.

But perspectives on gender do not explain all critiques of *Just Not Sorry*. Some value politeness separately from femininity, and others agree with the premise of *Just Not Sorry* but think it doesn't go far enough. Moreover, the media analysis confirms the relevance of a range of values: not only gender *identity* and *equality*, but also *achievement*, *autonomy*, and *privacy*. Guest

(2016) points towards further values of *mindfulness* (also evident in Reiss's 2015 essay) and *sincerity*.

We have shown that media analysis can contribute to value sensitive design as a form of empirical investigation. Our understanding of the role of gender in opinions about *Just Not Sorry* is more nuanced than before, and we discovered two further values to consider. But the approach has both advantages and disadvantages, for example, in comparison to semi-structured interviews. Media opinions constitute a naturally occurring, public dataset for which human subjects review is not required; data collection takes hours rather than weeks or months. Media opinions inform and are informed by public opinion. However, an article is not a conversation. We can't ask questions; we can only interpret what is written. Moreover, media opinions do not necessarily represent all opinions: we see only what writers and editors deem appropriate for publication. Finally, media opinions are not necessarily independent of each other. Where it might be surprising and meaningful to see the same words (e.g., "lady language") across several interviews, in media opinions it may only mean that one author read another's work.

The next step for this study of *Just Not Sorry* is to plan empirical investigations focused on the adoption of the app and its effectiveness in promoting behavior change. This media analysis will inform the design of a survey or interview protocol through the identification of additional values and viewpoints to address. Media analysis could take a similar role in other value sensitive design studies. While this is a retrospective study of an existing technology, which let us search for the tool's name, we can also imagine a role for media analysis in formative value sensitive design studies, particularly in the design of behavior change support systems. If the target behavior is one that people already try to change in the absence of supporting technology, a media analysis could focus on opinions about that behavior change. Media analyses could also consider opinions about related or competing tools.

6. CONCLUSION

We have contributed a value sensitive design case study of a persuasive technology concerned with language and gender. Through a media analysis focused on opinions about the tool, we confirmed the relevance of several values identified in a preliminary conceptual investigation, discovered two further values, and characterized disagreements about the tool grounded in disagreements about gender equality. In the context of value sensitive design, media analysis may work as a "discount" empirical method informing (or if necessary, replacing) surveys or interviews. While this study is a retrospective analysis of an existing tool, media analysis might also contribute to formative value sensitive design.

ACKNOWLEDGEMENTS

Thanks to Whitman College, and specifically the Louis B. Perry Summer Research Endowment and the Microsoft Chair in Computer Science, for supporting this work.

REFERENCES

Brandon, J. (2015, December 29). Use This Free Chrome Tool to Delete Weak Words From Your Emails. *Inc.Com*. https://www.inc.com/john-brandon/use-this-free-google-chrome-add-on-to-remove-weak-words-from-your-emails.html

Bucholtz, M. (2014). The Feminist Foundations of Language, Gender, and Sexuality Research. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The Handbook of Language, Gender, and Sexuality* (2nd ed., pp. 23–47). John Wiley & Sons.

Cauterucci, C. (2015, December 29). New Chrome App Helps Women Stop Saying “Just” and “Sorry” in Emails. *Slate Magazine*. https://slate.com/human-interest/2015/12/new-chrome-app-helps-women-stop-saying-just-and-sorry-in-emails.html

Curtis, C. (2018, November 4). Sorry to bother you, but I apologize too much over email. *The Next Web*. https://thenextweb.com/distract/2018/11/04/sorry-to-bother-you-but-i-apologize-too-much-over-email/

Day, E., & Ellen, B. (2016, January 9). Is new email app Just Not Sorry good for women? *The Guardian*. https://www.theguardian.com/commentisfree/2016/jan/09/is-new-email-tool-just-not-sorry-good-for-women-plug-in

Def Method. (2019a). *Defmethodinc/just-not-sorry* [JavaScript]. Def Method, Inc. https://github.com/defmethodinc/just-not-sorry

Def Method. (2019b). *Just Not Sorry—The Gmail plug-in*. Chrome Web Store. https://chrome.google.com/webstore/detail/just-not-sorry-the-gmail/fmegmibednnlgojepmidhlhpjbppmlci?hl=en-US

Dishman, L. (2016, January 5). New Gmail Plug-In Highlights Words And Phrases That Undermine Your Message. *Fast Company*. https://www.fastcompany.com/3055071/new-gmail-plug-in-highlights-words-and-phrases-that-undermine-your-messag

Erikson, J. (2016, January 5). Sorry, but Gmail’s New Plug-in Isn’t Just for Women. *CafeMom*. https://thestir.cafemom.com/good_news/194598/sorry_but_gmails_new_plugin

Fessler, L. (2018, June 27). There’s a new Gmail plugin to make you stop apologizing so much. *Quartz*. https://qz.com/work/1314825/theres-a-gmail-plug-in-to-make-you-stop-apologizing-so-much/

Fogg, B. J. (2002). *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann.

Friedman, B., Kahn, P. H., & Borning, A. (2006). Value sensitive design and information systems. In P. Zhang & D. Galletta (Eds.), *Human-computer interaction in management information systems: Foundations,* (pp. 348–372). M.E. Sharpe.

Garcia, P. (2016, January 4). Can This App Help Women Become More Empowered Over Email? *Vogue*. https://www.vogue.com/article/just-not-sorry-plugin

Gill, R., & Orgad, S. (2017). Confidence culture and the remaking of feminism. *New Formations*, *91*(91), 16–34. https://doi.org/10.3898/NEWF:91.01.2017

Gillespie, C. (2016, January 5). “Just Not Sorry” could be the best plugin ever invented, because we apologise far too much. *SheKnows*. https://www.sheknows.com/living/articles/1108059/email-tool-just-not-sorry-stops-women-apologising/

Ginn, A. (2016, March 1). The Art of Not Saying Sorry. *HuffPost UK*. https://www.huffingtonpost.co.uk/adeline-ginn/the-art-of-not-saying-sor_b_9353222.html

Grose, J. (2016, January 4). Telling women to apologize less isn't about empowerment. It's about shame. *Washington Post*. https://www.washingtonpost.com/posteverything/wp/2016/01/04/sorry-language-shamers-but-women-just-dont-need-your-new-email-policing-app/

Guest, K. (2016, January 9). *Sorry, but I'm going to give up apologising—And here's why*. The Independent. http://www.independent.co.uk/voices/sorry-but-im-going-to-give-up-apologising-and-heres-why-a6804371.html

Hines, S. (2016, January 6). An app that strips out "sorry" from my emails? I'm on board. *Good Housekeeping*. http://www.goodhousekeeping.co.uk/news/a558687/an-app-that-strips-out-sorry-from-my-emails-im-on-board/

Horobin, S. (2016, January 12). Why people say "sorry" so much. *The Independent*. http://www.independent.co.uk/life-style/why-do-people-say-sorry-so-much-a6807831.html

Lastoe, S. (2016, January 4). You're Actually Not Sorry—And This App Will Help You Stop Saying it So Much in Email. *The Muse*. https://www.themuse.com/advice/youre-actually-not-sorryand-this-app-will-help-you-stop-saying-it-so-much-in-email

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Analyzing Qualitative Data. In *Research Methods in Human-Computer Interaction* (2nd ed.). Morgan Kaufmann.

Levine, M. (2016, January 4). Does "Just Not Sorry" Hurt Women? *The Forward*. https://forward.com/sisterhood/328587/does-just-not-sorry-hurt-women/

Lord, E. (2015, December 30). This Plugin Prevents You From Undermining Yourself. *The Bustle*. https://www.bustle.com/articles/132730-just-not-sorry-gmail-plugin-points-out-when-you-diminish-yourself-in-emails

Minter, H. (2016, January 14). The Just Not Sorry app is keeping women trapped in a man's world. *The Guardian*. https://www.theguardian.com/women-in-leadership/2016/jan/14/the-just-not-sorry-app-is-keeping-women-trapped-in-a-mans-world

Paul, I. (2016, January 8). Strengthen your email writing with this Chrome extension. *PCWorld*. https://www.pcworld.com/article/3020059/improve-your-email-writing-with-this-chrome-extension.html

Reiss, T. (2015, December 22). Just Not Sorry! (The backstory). *The Medium*. https://medium.com/@tamireiss/just-not-sorry-the-backstory-33f54b30fe48

Sawyer, E. (2016, January 10). Gmail's Just Not Sorry Campaign Misses Some Painful Truths. *Fortune*. https://fortune.com/2016/01/10/gmail-just-not-sorry-campaign/

Schwartz, S. H. (1994). Are There Universal Aspects in the Structure and Contents of Human Values? *Journal of Social Issues*, *50*(4), 19–45.

Scott, E. (2016, January 5). This app will stop women writing "sorry" and "just" in their emails. *Metro*. https://metro.co.uk/2016/01/05/this-app-will-stop-women-writing-sorry-and-just-in-their-emails-5601821/

Stevens, S. (2016, November 28). Gmail plug-in means never having to say you're sorry. *Mother Nature Network*. https://www.mnn.com/green-tech/computers/stories/new-gmail-plug-means-never-having-say-youre-sorry

Tannen, D. (1990). *You Just Don't Understand*. Ballantine Books.

Tannen, D. (1994). *Talking from 9 to 5*. Harper Collins.

Tejada, C. (2016, January 13). Why being polite fails women. *FASHION Magazine*. https://fashionmagazine.com/lifestyle/health/why-being-polite-fails-women/

Turk, V. (2019, March 11). The Problem With Telling Women to Email Like Men. *Vice*. https://www.vice.com/en_us/article/8xyb5v/how-to-write-professional-work-email-women

Twersky, E., & Davis, J. (2017). "Don't Say That!" *Persuasive Technology*, 215–226.

Wills, K. (2016, January 7). Just Not Sorry email app aims to help women be more assertive. *The Independent*. http://www.independent.co.uk/life-style/gadgets-and-tech/just-not-sorry-gmails-new-app-aims-to-help-women-be-more-assertive-a6801506.html

Ye, L. (2016, January 13). The "Just Not Sorry" App Will Improve Your Sales Emails in 5 Seconds Flat. *HubSpot*. https://blog.hubspot.com/sales/just-not-sorry

VALUES IN PUBLIC SERVICE MEDIA RECOMMENDERS

Maaike Harbers, Lotte Willemsen, Paul Rutten

Rotterdam University of Applied Sciences (The Netherlands)

m.harbers@hr.nl; l.m.willemsen@hr.nl; p.w.m.rutten@hr.nl

ABSTRACT

Public Service Media (PSM) organizations show an increased interest in using recommendation systems to bring their audience in contact with new content, such as television shows, series and films. Recommendations of these PSM recommenders should not only match their users' interests and preferences, but coming from publicly funded organizations, they should also serve goals like informing the public and exposing them to a balanced mix of different views and perspectives. This yields the question what metrics (e.g., diversity, serendipity, accuracy) PSM recommenders should optimize for. This abstract follows a Value Sensitive Design (VSD) approach to map the most important values at stake in the design of PSM recommenders. This value analysis shows that a multifaceted technology, such as a PSM recommender, would benefit from an VSD approach in which 'the system' is not treated as a single unit, and 'the designer' of the system is not viewed as a single role.

KEYWORDS: Public service media, recommendation system, recommender, values, value sensitive design, pluriformity.

1. INTRODUCTION

Recommendation systems, recommenders in short, selecting and filtering content are widely used by companies in order to provide suggestions for items to users (Ricci et al., 2011). These 'items' range from songs (e.g., Spotify), series (e.g., Netflix), and movies (e.g., YouTube) to messages (e.g., Facebook), job vacancies (e.g., LinkedIn) and products (e.g., Amazon). Public Service Media (PSM) organizations, publicly funded organizations that offer radio and television content to a general audience, can also benefit from recommenders by using them to bring their audience in contact with new content. However, whereas recommenders used by commercial parties often aim to maximize profit or engagement, which is often achieved by recommending items in line with the user's views and interests, PSM organizations have other goals, such as informing the public and exposing them to a balanced mix of different views and perspectives, that could conflict with these commercial recommendation practices. The European Broadcast Union (EBU) acknowledges the tension between serving the audience with recommenders and the responsibilities of PSM organizations (EBU, 2017).

Recently, increasing attention has been paid to the development of recommenders for PSM (Sørensen et al., 2017; Fields et al., 2018; Van den Bulck et al., 2018; Sørensen, 2019). However, though the need for PSM recommenders is acknowledged, research into their design and development is still in its infancy. One of the open questions is what metrics (e.g., diversity or serendipity) PSM recommenders should optimize for (Fields et al., 2018). As a first step towards answering this question, following a Value Sensitive Design (VSD) approach (Friedman et al., 2019), this extended abstract describes a value source analysis (Friedman, 2017), in which an overview of the most important values at stake in the design of PSM recommenders is provided, including a description of where these values come from. The overview is based on a literature study and empirical investigations performed at NPO, the Dutch national public broadcasting organization (NPO, 2019a). Furthermore, some observations regarding the (value-sensitive) design of information systems in general are made.

2. VALUES AT STAKE - LITERATURE

The first set of values relevant to PSM recommenders can be found in literature on PSM. One of the most prominent lists of values for PSM is provided by UNESCO, consisting of universality, diversity, independence and distinctiveness (UNESCO, 2001). *Universality* refers to the accessibility of media content to all citizens in the country, *diversity* involves diversity in content, audience targeted, and subjects discussed, *independence* involves the freedom to express ideas and circulate information, and *distinctiveness* refers to the distinction of one PSM organization from other media organizations.

In addition to PSM values, there are values related to the use of the technology underlying recommenders. As public organizations such as PSM generally have the goal to 'serve the public', they often take public values into account (Jørgensen et al., 2007). Multiple values have been identified as relevant to the responsible design of information systems (Winkler et al., 2019). In relation to recommenders, most notably, this could mean a responsible use of personal data to protect *privacy* (Hoepman, 2014), and responsible use of machine learning, a technology often used in recommenders, supporting the values of values of *fairness*, *accountability* and *transparency* (ACM FAT).

3. VALUES AT STAKE - IN PRACTICE

We studied values at stake in PSM recommenders in a real-world setting at NPO, the organization that oversees public broadcasting services in the Netherlands. One of the ways in which NPO brings content produced by public broadcasters to the Dutch audience is via its website NPO Start (www.npostart.nl), which makes limited use of a recommendation algorithm. Most of the recommendations on NPO Start are manually curated, but for website visitors with an account (the minority of the visitors), a small part of the recommendations is personalized and generated by an algorithm. NPO is currently working on improving and expanding their recommender. For our study, we attended several meetings at NPO in which the design of the new recommender was discussed, conducted interviews with stakeholders within and out of NPO, and studied project documentation and reports produced by NPO.

NPO's mission is to connect and enrich the Dutch audience with content that informs, inspires and entertains (NPO, 2019a, 2019b), which is broadly in line with the PSM values described in the previous section. Project documentation showed that the most prominent value in the recommender design project was *pluriformity* (the 'explicitly supported value' in VSD terminology). In meetings, a lot of time was spent on discussing what, exactly, pluriformity means with respect to the recommender to be designed. Other values that came up during the meetings were accuracy, privacy and transparency. *Accuracy* of recommendations was deemed important, as users receiving too many recommendations that are not interesting to them would disengage. With respect to *privacy*, it was agreed upon that the recommender should not collect explicit personal information such as age, gender or ethnicity, but only use watching behavior. *Transparency* to users about the origin of recommendations was also deemed important.

In an interview with the head of the development team, responsible for the implementation of the recommendation algorithm (and also part of the project team), we learned that the current algorithm weights five factors: novelty, clickthrough rate, personalization, fraction watched and public values.

The last factor, public values is composed of users' ratings of content based on eight values, out of which one is pluriformity (p.62, NPO, 2019c). There is thus a discrepancy between the focus on pluriformity in the redesign project and the (minor) role of pluriformity in the current recommender. With respect to the planned increased importance of pluriformity, the development team neither knew how to translate pluriformity into an implementation, nor did they see it as their responsibility.

Interviews with users, people who watch content produced by public broadcasters on NPO Start, revealed that the majority of users is interested in personalized recommendations, but that most of them were not or only vaguely familiar with the term pluriformity.

4. DISCUSSION

Several insights can be drawn from the results so far. There are several values that the organization wants to embed in the new recommender, most notably pluriformity. Yet, problems are encountered in translating these values into a concrete implementation. Whereas the development team refers to others to operationalize pluriformity so that they can implement the algorithm, other members of the project team have trouble providing such an operationalization, partly because they have limited programming knowledge and have troubles imagining what developers need. At the same time, the term pluriformity does not appeal to users of the recommender, which may be problematic in providing transparency (another value at stake) about the system. These differences seem to indicate a mismatch between knowledge, culture and languages spoken by different groups of people: (most members of) the project team, developers and users.

A mismatch in understanding of the design challenge and its implicated values between teams in the organization is possibly reinforced by an organizational structure in which employees with different expertise and backgrounds are organized different teams. This is problematic when the goal is to embed values in technology. For example, embedding the value of transparency in a recommender has implications for both the recommender's algorithm and its user interface. On the technical, backend side, the algorithm should be explainable, which may imply avoiding certain deep learning algorithms (Samek et al., 2017). On the user-facing, front-end side, there should be a way to communicate explanations to users in the interface, e.g. a textbox or a button for requesting an explanation for why an item was recommended (Tintarev et al., 2011). If the system does not meet requirements on both of these sides, it will not support transparency. In order to align different components of a system, teams responsible for the creation of these different components need to be aligned as well.

The insights above lead to a more general observation. In VSD analyses, when describing value implications, 'technology' is often treated as a single system and 'the designer' is often treated as a single role (Friedman et al., 2019). However, this is a simplification of (the creation of) a lot of technologies, as systems often consist of different components, which are developed by different teams, consisting of a variety of individuals, with different backgrounds and cultures. We believe that a VSD process could benefit from a more nuanced view on the 'technology' and 'designer', doing justice to their complexities. This may be particularly relevant for complex and intelligent systems, which have a heavy technical component, as well as a user interface.

5. FUTURE WORK

This paper forms a first step towards designing a PSM recommender. Next steps involve analyzing value tensions (Miller et al., 2007); selecting metrics based on these values (Fields et al., 2018);

operationalizing these metrics, including weighing them against each other; and designing and evaluating prototypes. This process will be performed iteratively, involving multiple cycles of prototyping and collecting user feedback. In this process, attention will be paid to the multifaceted nature of recommenders as well as their designers.

ACKNOWLEDGEMENTS

The authors thank NPO for their collaboration and support in this research project. This extended abstract reflects the authors' interpretations of information and events, and the authors are solely responsible for the contents of this extended abstract.

REFERENCES

ACM FAT. Conference on Fairness, Accountability, and Transparency. Retrieved from: https://fatconference.org/.

EBU (2017). *Big data initiative report: time to invest.* Technical report. European Broadcast Unit (EBU).

Fields, B., Jones, R., & Cowlishaw, T. (2018). The case for public service recommender algorithms. Proceedings of *FATREC Workshop on Responsible Recommendation*.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2), 63-125.

Hoepman, J. H. (2014). Privacy design strategies. In *IFIP International Information Security Conference* (pp. 446-459). Berlin, Heidelberg: Springer.

Jørgensen, T. B., & Bozeman, B. (2007). Public values: An inventory. Administration & Society, 39(3), 354-381.

Miller, J. K., Friedman, B., Jancke, G., & Gill, B. (2007). Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 281-290). ACM.

NPO (2019a). Nederlandse Publieke Omroep. Retrieved from: https://over.npo.nl/.

NPO (2019b). Jaarverslag 2018. Annual report. Retrieved from: https://over.npo.nl/organisatie/onze-waarde-voor-nederland/jaarverslag.

NPO (2019c). Terugblik 2018. Technical report. Retrieved from: https://over.npo.nl/organisatie/onze-waarde-voor-nederland/terugblik.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Boston, MA: Springer.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Sørensen, J. K. (2019). Public Service Media, Diversity and Algorithmic Recommendation: A Europe-wide Implementation Study. In *RecSys 2019: 13th ACM Conference on Recommender Systems*.

Sørensen, J. K., & Hutchinson, J. (2017). *Algorithms and public service media*. Public Service Media in the Networked Society RIPE, 91-106.

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In Recommender systems handbook (pp. 479-510). Springer, Boston, MA.

UNESCO (2001). *Public broadcasting: Why? How?* Technical Report (pp. 1-28). Paris: UNESCO.

Van den Bulck, H., & Moe, H. (2018). Public service media, universality and personalisation through algorithms: mapping strategies and exploring dilemmas. *Media, Culture & Society*, 40(6), 875-892.

Winkler, T., & Spiekermann, S. (2019). Human Values as the Basis for Sustainable Information System Design. *IEEE Technology and Society Magazine*, 38(3), 34-43.

7. Monitoring and Control of AI Artifacts

AN EMPIRICAL STUDY FOR THE ACCEPTANCE OF ORIGINAL NUDGES AND HYPERNUDGES

Yukari Yamazaki

Seikei University (Japan)

yyamazak@econ.seikei.ac.jp

ABSTRACT

While nudges have been paid attention to, they have also been criticized by several studies especially for their ethicality. Sunstein (2016, 2018) considered several aspects of nudges and claimed that if nudges guarantee decision-makers autonomy and dignity as well as transparency of choice architecture, they are extremely beneficial for individuals and society and are therefore not unethical.

In the recent past, the neologism 'hypernudge' has drawn attention, arousing much controversy. It is said that a hypernudge is a 'kind of' nudge utilizing artificial intelligence (AI) or machine learning. While the ethicality of nudges is likely to be certified, keeping autonomy, dignity, and transparency in mind, it is certainly arduous to warrant these three conditions in hypernudges driven by AI artefacts. That is, the acceptance of the original nudges, which have become popular and been adhered to, are likely to be distinct from that of hypernudges.

To address the issue, this study tested the acceptance of both interventions and revealed that the acceptance of the original nudges differed from that of hypernudges, and that the latter was less acceptable than the former. Notably, in hypernudges, while individuals tended to accept the less flexible and forbidden intervention, they rejected the ones that utilized their children's personal data. However, neither typical nor common features were confirmed that could identify the acceptance level of hypernudges, such as categories of interventions, individual sociodemographic factors, political attitudes, and mobile phone usage histories. The findings from this study suggest a kind of alert for spreading hypernudges that utilize AI-driven artefacts in the future.

KEYWORDS: AI-driven artefacts, original nudge, hypernudge, acceptance.

1. INTRODUCTION

The paradigm word, nudge (Thaler and Sunstein, 2008) and its strategy has been attracting people in various fields. One breakthrough attempt is changing the law around organ donation (Max and Keira's law) from the year 2020 in the UK. All adults in England will be considered organ donors when they die unless they have recorded a decision not to donate or are in one of the excluded groups (BBC News, 2019). Nowadays, other countries such as Denmark and the Netherlands will also change to or consider adopting an opt-out organ donation system, a nudging technique. This paternalistic strategic intervention tries either presenting people in a more salient or impressive light, or making them the easier or default option, rather than enforcing restrictions or drawing out people's rational behaviour. Such a selection system is called the choice architecture. Notably, a nudge promotes people's choice and behaviour and is assumed to benefit target individuals and society as a whole; therefore, it should never be applied to marketing or particular profit pursuing activities.

While the fact that a nudge steering peoples' behaviour in desirable directions through milder choice interventions has drawn attention, it has also received blistering critiques of ethicality (e.g., Goodwin, 2012; O'Neil, 2011), diminishing human wisdom (Furedi, 2011), troubles and pitfalls (Bovens, 2009), and manipulations (Wilkinson, 2012). In response to these misunderstood critiques, C. R. Sunstein, one of the advocates of the nudge, has discussed the validity and considered the benefit and ethicality of nudges (Sunstein, 2015, 2016). According to his consideration, there are neither neutral ways to present options nor can choices be made in a vacuum, and one cannot avoid the choice architectures that influence choice in many ways. It might be easy to promote purchasing by altering the presentation order of alternatives and attributes, ease of picking them up, selection of defaults, and naming just a few of the design options available. Therefore, it is essential to choose alternatives while paying attention to the structures and effects of the choice architectures.

He also said, 'When nudges are fully transparent and subject to public scrutiny, a convincing ethical objection is less likely to be available'. In addition, he also stated that, 'if people have not consented to them; such nudges can undermine autonomy and dignity' (p.1). Furthermore, Sunstein (2018) insisted that 'Nudges always respect, and often promote human agency; because nudges insist on preserving freedom of choice, they do not put excessive trust in government; nudges are generally transparent rather than covert or forms of manipulation' (p.1). Indeed, it has been already examined that nudges utilized the defaults setting to be transparent and yet effective (Bruns et al., 2018). According to the above considerations, two of the prominent elements in ethical nudges should be transparency and autonomy. In other words, maintaining transparency and decision-makers' autonomy in nudging must be recognized as ethical and beneficial.

Currently, the neologism 'hypernudge' has highlighted and aroused much controversy. It is thought of as a 'kind of' nudge utilizing big data, personal data, AI algorithms, deep learning, and so on. While the ethicality of nudges would be certified, keeping autonomy, dignity, and transparency in mind, it is certainly arduous to warrant these three conditions in hypernudges driven by AI artefacts. That is, the acceptance of the original nudges, which have become popular and been applied to, would be distinct from that of hypernudges. While empirical evidence regarding the acceptance of the original nudges has appeared, the ones of hypernudges have remained unproven.

In this study, the comparison with acceptance of the original nudge and hypernudge is examined. It is found that while hypernudges are less acceptable than the original nudges, the representative features showing which hypernudges are more acceptable than others are still veiled. The findings are discussed, and conclusions are drawn at the end.

2. HYPERNUDGE

Recently, artificial intelligence (AI)/machine learning (ML) has drawn attention among mass media and academic fields not only because of their attractive, tremendous, and hyper functions as well as efficiency and effectiveness, but also because of their ethicality and riskiness. The IEEE, for example, has taken the ethicality of AI designing, utilizing, and prevalence as a serious problem and given an alert for AI systems as nudging tools. (IEEE, 2018). In the section Affective Computing of the 2nd draft version of *Ethically Aligned Design*, the following six recommendations have been pointed out: 1) systematic analysis for ethical design of AI systems before deployment, 2) showing the types, effects, and purposes of nudges towards users, 3) analysing the possibility of infantilization of those who were nudged by AI, 4) making default settings opt-in, 5) giving additional protection for vulnerable users who do not pay attention to informed consent; and 6) keeping transparency and accountability. These are consistent with the certified requirements of nudges mentioned above, guaranteeing transparency and autonomy.

Nowadays, nudges through AI/ML-driven new technologies are coined as ‘hypernudges’ (Yeung, 2017) or ‘digital nudges’ (Weinmann et al., 2016). While it is being gradually known that AI/ML systems have beneficial traits, there are several specific features as hypernudges. Some of them are, for example, self-tracking of past behaviour (in some cases) without getting the agreement for utilizing it, presenting immediate feedback based on self-tracking and big data as recommendations for each user, making some judgment on behalf of human autonomy, and steering people to use AI artefacts repeatedly (e.g., Google navigation system). Based on prior studies that had paid attention to hypernudge, this study considers several differences between the original nudges and hypernudges (Table 1).These typical features with AI/ML-driven hypernudges might make users blind and depend too much on them. Therefore, the manipulative aspects of data-driven personalized communication, big data utilization, and behavioural targeting in the online realm have been regarded as problems (e.g., Lanzing, 2018).

Table 1. Comparison of the original nudges and hypernudges.

	Original nudge	**Hypernudge**
Purpose	Steering people towards a better direction to nudge softly and mildly utilizing human judgmental tendencies.	Designing and programming the choice architecture utilizing bigdata, personalized information, and computer algorithms.
Methods	• Presenting visible, available, noticeable information or options. • Using the default setting to do something automatically, omit doing something, or avoid forgetting. • Emphasizing (the possibility of) suffering losses, gaining incentives, and so on. • Utilizing social influence such as normative message, showing others behaviour, or giving approval.	• Presenting desirable options such as recommendations, alerts, advertisements, or notices that are suited to each individual based on the person’s behaviour history.
Examples	• Labelling healthy food packages, setting fruits and vegetables nearer to a person in the buffet counter, providing information about the amount of sodium or sugar, serving food with a small plate, and so on. • Web shopping sites recommend storing of customers’ credit card numbers to avoid entering these details again at the next shopping visit.	• Recommending purchasing healthier, lower sodium and sugar food based on the user’s purchase history and health condition in the following shopping visit. • Web shopping sites recommend which credit cards the customer should use in this shopping.

Specific features	Giving a general message, not using specific customized data.	• Giving customized recommendation or predictive message based on personal information or bigdata. • Vague request for consent for using personal information.
	One to a few nudges. (A few people are nudged.)	One to many nudges. (It is possible to nudge many people at once. The presented information is different for individuals.)
	Transient and does not influence the next or other choices.	Continuous and repetitive and influences the next or other choices, as well as the possibility to change the choice architecture.
	No particular feedback, generally.	Immediate and continuous feedback based on personal data and bigdata.

In another argument, however, whereas various services by AI/ML, such as vehicle navigation systems, positional information on the digital map, recommendations based on purchase history, personalized chatting with bots, various apps (e.g., health care, saving and investing money with FinTech, and smart home with IoT), and so on have spread among people recently, it is hard to ascertain these new nudges as the original ethical nudges, whose validity has been discussed in Sunstein (2015, 2016) because of several serious reasons (Yamazaki, 2019). On the one hand, the following three factors have been pointed out as the hampering factors of autonomy. The first is decreasing users' motivation, responsibility, and morality. The second is indulging in of bad habits because of attractive recommendations from AI/ML artefacts. The third is a semiconscious repetition of habitual behaviour against their real will (meta preference). On the other hand, transparency is infringed by the complicated algorithm, dynamic feedback that confuses users' preferences, unpredictability, and complete unavailability of users' informed consent. Because of the lack of autonomy and transparency, it is difficult to certify hypernudges as original nudges at the moment.

3. PRIOR STUDIES

In the past years, several studies have surveyed the acceptance, trustworthiness, and consensus for various types of interventions in various countries under different contexts. These surveys have shown that, on the one hand, citizens in various countries perceived that nudges were being inconsistent with their interests or values of most choosers; on the other hand, they generally tended to approve of almost all nudges.

As for the difference of countries (nationality), three prior studies (Reisch and Sustein, 2016; Sunstein et al., 2018) investigated the approval of the same 15 interventions for 15 nations. According to their survey, the approval rates of many western democratic countries (Australia, Canada, French, Germany, Italy, Russia, the UK, and the US) were similar (around 68-75%). In contrast, Brazil, South Africa, China, and South Korea showed overwhelmingly high approval rates for all nudges (around 80%), while rates for Denmark, Hungary, and Japan were extremely low (around 60%). They could find neither the reasons nor countries' specific features, therefore, requiring further examination.

The fields of health and safety nudges would be approved for people and the levels of acceptance of nudging techniques depended on the countries of the participants as well as the depth, types, contexts, and prosociality of nudges. As in other empirical studies on the acceptance of nudges, in the fields of medicine and health, choice and public policies (e.g., Diepeveen et al., 2013; Felsen et al., 2013; Junghans et al., 2015, 2016; Jung and Mellers, 2016) showed similar results; whereas almost all subjects generally approved of nudges, they did not accept the nudges inconsistent with their preferences, improper nudges such as political and religious favouritism and perceived manipulation.

Other causes of different acceptances are the types and contexts of nudging. Felsen et al. (2013) tried a decisional enhancement nudging program that contained two types of nudging questions on five scenarios. One is overt and conscious nudging which the decision maker is aware of and can consciously process, while another is covert and unconscious nudging such as subconsciously decreasing hunger, which in some situations, can be related to people's autonomy. The five scenarios promote healthier eating, prudent online purchasing, encouraging exercising, investing in retirement, and improving productivity. The results showed that conscious nudges are more acceptable than subconscious nudge processes except for improving productivity scenarios. The subconscious nudges might infringe upon individual autonomy, while covertly trying to protect their own autonomous decision making.

Jung and Meller (2016) examined the effects of types of nudges (automatic and unconscious vs. effortful and conscious), individual dispositions (e.g., level of empathy, conservative, desire for control, reactance, and individualist vs. communitarian) and benefit from nudges (i.e., nudges for societal vs. personal). They divided several nudges into System 1 (automatic and unconscious), and 2 (effortful and conscious). Whereas the automatic enrolment for something such as retirement savings or medical coverage plans and default settings were classified as System 1, providing of reminders, alerts, or messages were classified as System 2. The results showed that the effect of the nudge type on support was significant. The effortful nudges were significantly more accepted than the automatic ones. In addition, the level of support for nudges was not significantly affected by benefits from nudges but was affected by participants' individual dispositions.

Although the acceptance of the original nudge as mentioned above has been gradually positive, empirical studies that have paid attention to the hypernudge utilized AI/ML artefacts have been scarce.

4. RESEARCH QUESTIONS AND METHODS

The aim of this research is to examine whether the acceptance of interventions differs from the original nudge to hypernudges. In particular, because hypernudges driven by AI artefacts would keep neither transparency nor autonomy for decision-makers, the high level of acceptance can come under doubt. Namely, while the original nudges have revealed almost acceptable results among almost all people, hypernudges that are new, neither transparent nor autonomous might have lower acceptance than the original nudges. In addition, according to several prior studies, the acceptance levels differed in the categories of intervention as well as several demographic factors. In sum, hypotheses concerning people's acceptance of interventions are as follows:

H1: The acceptance rate might differ in the original nudge and the hypernudge.

H2: The original nudge might be more acceptable than a hypernudge.

H3: The categories of interventions might be related to the peoples' acceptance rate of both the original nudge and hypernudge. Deeper and subconscious interventions might be less approved.

> H4: Sociodemographic variables, political attitudes, and years of mobile use might be related to the acceptance rate of both the original nudge and the hypernudge. Whereas age might positively relate to acceptance, the years of mobile use might have a negative relationship with it.

To examine these hypotheses, this study compares and enlarges the results of prior studies. The same 15 survey interventions in the background literature (e.g., Sunstein et al., 2018) were selected, with one new intervention on security protection added for both the original nudges and hypernudges. The summary of each content is shown in Table 2. An example of a question on hypernudge is 'Online food shopping sites are required to show good or bad effects of each food on each user's health condition (such as decreasing body fat, body pressure, no effects, and so on) based on AI recommendation'. Because participants might not be familiar with and imagine for each different AI-driven service, general and uniform expression about AI services, 'artificial intelligence will be expected to present various recommendations using your personalized information such as your age, health and medical history, sleeping time, career, status of income and assets, purchase histories, and mobile phone use history in the future' was included at the beginning of the nudging questions. Participants were asked to answer 'agree' or 'disagree' to 16 questions on the social systems that promoted health and enriched society as well as to comment on nudges by AI.

Table 2. Categories of 16 interventions.

No.	Summary of contents	Depth	Type	Context	Prosociality	Percentage of agreement※
1	Showing calorie labels in restaurants' menu.	Mandatory information disclosure	Conscious	Health	Personal	79.40
2	Showing food bad effects for health.	Mandatory information disclosure	Conscious	Health	Personal	80.87
3	Enrolling green energy suppliers automatically, possible to opt out.	Mandatory default	Subconscious	Ecology	Social	73.47
4	Asking to be organ donors in obtaining driver's licence.	Mandatory default	Subconscious	Charity	Social	62.53
5	Placing healthy foods at prominent visible places in grocery stores.	Mandatory default	Subconscious	Health	Personal	73.00
6	An education campaign to reduce distracted driving.	Government Campaign	Conscious	Traffic safety	Social	87.67
7	An education campaign for promoting healthier choice for parents to reduce childhood obesity.	Government Campaign	Conscious	Health	Social	89.73
8	Providing prohibited subliminal advertisements in theatres to discourage smoking and overeating.	Non-nudge (Forced)	Subconscious	Health	Personal	52.00
9	Charging a specific amount with offset opt out option for carbon emission.	Mandatory default	Subconscious	Ecology	Social	44.00
10	Labelling unhealthy food making notice it is harmful.	Mandatory information disclosure	Conscious	Health	Personal	83.67
11	Asking to donate the Red Cross refund automatically, possible to opt out.	Mandatory default	Subconscious	Charity	Social	40.33
12	Requiring movie theatres to provide public education messages to discourage smoking and overeating.	Mandatory information disclosure	Subconscious	Health	Personal	67.13

13	Requiring large electricity providers to make people enrol in green energy suppliers automatically, possible to opt out.	Mandatory default	Subconscious	Ecology	Social	72.27
14	Keeping cashier areas in supermarkets chains free of unhealthy foods to halt obesity.	Mandatory choice architecture	Subconscious	Health	Personal	62.86
15	Requiring public institutions to have meat-free day per week.	Mandatory choice architecture	Conscious	Health	Personal	55.14
16	Installing security software automatically to avoid viruses and hackers, possible to opt out.	Mandatory default	Subconscious	Information security	Personal	

※The average rates of agreement that prior studies examined.

In addition, this study refers to the categories of interventions as laid out in prior studies (Jung and Meller, 2016; Felsen, 2013; Sunstein et al., 2018) (Table 2). It is categorized into five levels of depth: the shallowest is campaign Nos. 6 and 7); second, mandatory information disclosure (Nos. 1, 2, 10, and 12); third, mandatory default (Nos. 3, 4, 5, 9, 11, 13, and 16), fourth, mandatory choice architecture (Nos. 14 and 15), and the deepest is forced (No. 8). There are also two types (conscious vs. unconscious), five contexts (health, ecology, charity, traffic safety, and security), and two prosocialities (for society or personal) of interventions.

Participants in this study were recruited from university students in Japan and the Japanese consulting company *Kiccoe Survey* and students attending university in Tokyo. A total of 1,192 participants were asked 16 questions, half of which were on original nudges and the other half on hypernudges. The original nudge had *n*=596 and the hypernudge had *n*=596. The percentage of males was 54.6%. The mean age was 36.64: *SD*=18.2, range=13–87 years, under 20=332 (27.85%), the 20s=209 (17.53%), the 30s=140 (11.74), the 40s=175 (14.68%), the 50s=162 (13.59%), the 60s=118 (9.90%), the 70s=49 (4.11%), and the 80s=7 (0.59%). The participants also gave responses on their years of education (*M*=14.98 years), years of mobile phone use (never=134 or 11.24%, under a year=44 or 3.78%, under 5 years=374 or 31.38%, under 10 years=515 or 43.21%, and over 10 years=124 or 10.40%), and political attitude (ruling=308 or 25.84%, opposition=173 or 14.51%, and non-partisan=711 or 59.65%).

The first was designed to compare the acceptance levels of each original nudge and the hypernudges. The second was designed to test the differences among the categories of interventions, followed by examining the effects of participants' individual dispositions as the third.

5. RESULTS

5.1 Overall acceptance rates

The results indicated that overall, the original nudges were more accepted than hypernudges (Table 3), and the acceptance level of the original nudge and hypernudge differed in the categories of interventions (Table 4). In line with prior studies on the original nudges, more than half of the original nudges (10 out of 16) were significantly more acceptable. On the other hand, as new notions, less than half of the hypernudges (6 out of 16) were significantly acceptable. In addition, eight interventions (Nos. 1, 2, 3, 4, 6, 7, 9, and 12) of the original nudges, half of the total, found significantly higher acceptance than hypernudges. These results indicate that people would not be more receptive to AI-driven interventions and suggest H1 and H2 are supported. One possible reason for these results is that using their personal information or behavioural history might have been considered creepy and untrustworthy; although hypernudges are customised for the individuals, their acceptance level was lower than the original nudge. The results would suggest that AI artefacts designers, service providers, choice architectures as well as users, should be careful of hypernudges, not a little.

Table 3 Significant difference of acceptances within each original nudge and hypernudge.

Original nudges (*n*=596)

No.	1	2	3	4	5	6	7	8	
Agree Disagree % of agree	514** 82 86.24	395** 201 66.28	444** 152 74.50	349* 247 58.56	272 324*※ 45.64	407** 188 68.29	469** 127 78.69	249 347** 41.78	***Means of original nudges*** Agree 347.68 Disagree 248.31 % of agree 58.34
No.	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
Agree Disagree % of agree	327* 269 54.87	341** 255 57.21	281 315 47.15	394** 202 66.11	294 302 49.33	215 381** 36.07	183 413** 30.70	429** 167 71.98	

Hypernudges (*n*=596)

No.	1	2	3	4	5	6	7	8	
Agree Disagree % of agree	306 290 51.34	327* 269 54.87	352** 244 59.06	315 281 52.85	269 327* 45.13	281 315 47.15	199 397** 33.39	328* 268 55.03	***Means of hypernudges*** Agree 300.19 Disagree 295.75 % of agree 50.38
No.	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
Agree Disagree % of agree	285 311 47.82	379** 217 63.59	291 305 48.83	295 301 49.50	329* 267 55.20	227 369** 38.09	162 434** 27.18	459** 137 77.01	

Chi-square value significant at alpha * $p < 0.05$

** $p < 0.01$

※The numbers with underlines suggest significantly more unacceptable results.

Table 4. Cross Tabulation with each nudge pairs.

	O1	H1	O2	H2	O3	H3	O4	H4	O5	H5	O6	H6	O7	H7	O8	H8
Agree	**62.7**※	37.3	**54.7**	45.3	**55.8**	44.2	**52.6**	47.4	50.3	49.7	**59.2**	40.8	**70.2**	29.8	43.2	**56.8**
Disagree	22.0	78.0	42.8	57.2	38.4	61.6	46.8	53.2	49.8	50.2	37.5	62.5	24.2	75.8	56.4	43.6
$\chi^2(1)$	169.062**		16.243**		32.007**		3.930*		.030		54.576**		248.254**		20.964**	
	O9	H9	O10	H10	O11	H11	O12	H12	O13	H13	O14	H14	O15	H15	O16	H16
Agree	**53.6**	46.4	47.4	**52.6**	49.1	50.9	**57.2**	42.8	47.2	**52.8**	48.6	51.4	53.0	47.0	48.3	**51.7**
Disagree	46.2	53.8	54.0	46.0	50.8	49.2	40.2	59.8	53.1	46.9	50.8	49.2	48.8	51.2	54.9	45.1
$\chi^2(1)$	6.500*		5.065*		.336		33.710**		4.119*		.518		1.799		3.974*	

Chi-square value significant at alpha * $p < 0.05$

** $p < 0.01$

※The significant higher acceptances are shown with bold letters

It might be easy to imagine that because No. 8 (subliminal advertisement) were, as Sunstein et al. (2018) also mentioned, not a nudge, there were significantly higher disagreements in the original nudge. In addition, No. 5 (placing visible healthy food), 14 (avoiding unhealthy food), and 15 (requiring meat-free day per a week) were significantly lower acceptance in both the original nudges and hypernudges because they were strongly obtrusive for people.

Unlike prior studies, the highest accepted intervention in the original nudges was the 'calorie label' (No. 1, 86.24%), with the next being 'education campaign for childhood obesity' which was in highest agreement in prior studies. The lowest accepted interventions in the original nudges, on the other hand, were 'requiring a meat-free day per week for health' (No. 15, 30.7%), which was likewise the lowest in hypernudges (27.18%) and relatively lower acceptance in prior studies of hypernudges. This might have been considered as excessive interference in diet. One of the higher accepted interventions in both nudges was 'information security' (No. 16) that might have been the most familiar matter for the largest number of youths in this study. Overall, the acceptance tendency of the original nudges shown in this study is similar to that of Sunstein et al. (2018), where the Japanese consensus was remarkably lower than in other countries and in averages.

As for comparison with the original nudges and hypernudges for acceptance percentages (Table 4), the number of acceptance of the eight original nudges (Nos. 1, 2, 3, 4, 6, 7, 9, and 12) were exceed that of hypernudges, as mentioned, while the one of the four hypernudges (Nos. 8, 10, 13, and 16) exceed that of original nudges, exceptionally. The other four interventions (Nos. 5, 11, 14, and 15) were not significantly different between the original and the hypernudge. Surprisingly, whereas No. 8 (subliminal advertisement) intervention is significantly higher in the original nudges, it is significantly in higher agreement in hypernudges. It might be thought that, on the one hand, individuals tend to take care of interventions by hypernudges; on the other hand, they might agree with this intervention even prohibited one because it is easy for them to stop watching customized advertisements. In contrast, the exact opposite result is shown in No. 7 (education campaign for childhood obesity), which shows a significantly higher disagreement in the hypernudge, even though higher agreement in the original nudge. It seems that individuals would hesitate to present and utilize their children's personal data. No. 16 (security protection) is easier to accept in hypernudges because it is easy to imagine that AI-driven systems are always exposed through viruses and hackers. While several typical reasons can be suggested for these results, it is difficult to give interpretations of why No. 10 (labelling unhealthy food) and 13 (green energy consumption) that are similar to Nos. 1, 2, and 3, have obtained more agreement in hypernudges.

5.2. Effects of intervention categories

The next step has been designed to compare the differences of intervention categories with a one-way analysis of variance. The rates of agreement for each intervention are shown in Table 5.

Table 5. The rates of acceptance for each intervention (%).

Depth	original	hyper
Campaign	73.49	40.27
Man. info.	40.27	68.96
Default	68.96	54.82
Choice arch.	54.82	57.43
Forced	57.43	55.13

Type	original	hyper
Cons.	64.57	46.25
Subcon.	46.25	54.60

Context	original	hyper
Health	56.52	46.46
Ecology	46.46	59.56
Charity	59.56	54.03
Traffic safety	54.03	52.85
Info. security	52.85	50.84

Prosoci.	original	hyper
Social	55.78	51.30
Personal	51.30	61.63

Comparison of five levels of intervention depth only showed marginally significant tendencies. Except for No. 8 (forced), the deeper nudge (Nos. 14 and 15, mandatory choice architecture) is more disagreeable, and the shallowest interventions (Nos. 6 and 7 campaigns) are the most acceptable among them (*F*=4.991, *p*=.015). However, this tendency was confirmed in the original nudge, but not in the hypernudge (*F*=3.117, *p*=.61). This means that it would be more difficult to consider whether and which interventions by AI-driven hypernudges would be accepted by people because the depth of interventions is not related to the acceptance. Other categories are inconsistent with prior studies and do not show significant differences. Type – original nudge: *F*=1.534, *p*=.236; hypernudge: *F*=1.208, *p*=.290; context – original nudge: *F*=.307, *p*=.867; hypernudge: *F*=2.133, *p*=.145; prosociality – original nudge: *F*=.517, *p*=.484; hypernudge: *F*=.121, *p*=.733. These results show that H3 is partly supported, and only the depth of interventions has significantly different effects in the original nudges.

5.3 The effects of individual difference

Further, we estimated the logistic regression for the five levels of depth of interventions with significant approval rates of the 16 interventions being dependent variables in both the original and the hypernudge. Age, gender (number of male), educational years (of schooling), and political attitude (support for ruling party, opposition, or non-partisan) were used as independent variables.

Notably, Table 6 shows that political attitude has a unique influence on participants' approval of nudges: non-partisan people significantly disapprove three out of five types of original nudges (mandatory information, default, and choice architecture), but significantly approve (negative influence) four types of hypernudges (campaign, mandatory information, default, and forced). Political independents tend to, on the one hand, be more doubtful on general interventions by the original nudges, while on the other hand, being acceptable for new technology and customized data. However, because the rate of non-partisan people was highest among the participants (59.65%), more than half of them had applied for this tendency.

In addition, the age of participants has a negative impact: older people tend to disapprove both one original nudge (default) and three hypernudges (mandatory information, default, and choice architecture). This is in line with Sunstein et al. (2018). Several other factors have significant influence. For example, males tend to less favour original nudges on mandatory information, people who have used a mobile phone for approximately 5 to 10 years tend to disapprove the original nudges on mandatory information, people who have used a mobile phone under one year tend to support the original nudge on choice architecture; also years of education had a significantly negative correlation on the choice architecture. These results suggest that H4 is marginally supported.

	Campaigns (Nos. 6 and 7)				Mandatory Information (Nos. 1, 2, 10, and 12)			
Intervention type	Original		Hypernudge		Original		Hypernudge	
	β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)
Constant	2.389*	10.904	-1.209	.299	.978*	2.659	.392	1.479
Age (in years)	.015	1.015 (1.006-1.024)	-.002	.998 (.991-1.006)	.002	1.002 (.969-1.008)	-.015**	.986 (.980-.991)
Gender (male)	-.209	.812 (.615-1.070)	-.028	972 (.762-1.240)	-.208*	.812 (.676-.976)	.021	1.201 (.860-1.213)
Political attitude								
Ruling	-.168	.846 (.609-1.174)	—	—	—	—	—	—
Opposition	.179	1.196 (.799-1.792)	-.167	.846 (.583-1.228)	.209	1.232 (.905-1.677)	.022	1.023 (.775-1.348)
Non-partisan	—	—	.427*	1.532 (1.168-2.010)	-.248*	.781 (.633-.963)	.457**	1.579 (1.295-1.926)
Education (in years)	-.021	.964 (.901-1.032)	-.031	.969 (.905-1.038)	.013	1.013 (.967-1.062)	-.001	.999 (.956-1.034)
Mobile use years								
No use	-.103	.928 (.517-1.665)	.461	1.586 (.968-2.597)	-.235	2.427 (1.053-1.170)	.206	1.229 (.861-1.754)
Under one year	-1.034	.463 (.187-1.149)	.642	1.901 (.911-3.965)	.068	1.070 (.610-1. 778)	.043	1.044 (.640-1.704)
Less than 5 years	-.661	.699 (.434-1.127)	.169	1.184 (.785-1.787)	-.162*	.851 (.611-1.184)	-.051	.950 (.707-1.277)
Less than 10 years	-.340	.695 (.437-1.105)	.287	1.332 (.898-1.978)	-.325*	2.659 (.525-.994)	-.067	.935 (.702-1.246)
Over 10 years	—	—	—	—	—	—	—	—
Obs.	1,192		1,192		2,384		2,384	
Omnibus test	$\chi^2(9)$ =20.211*		$\chi^2(9)$ =21.084*		$\chi^2(9)$ =26.038**		$\chi^2(9)$ =67.401**	

Default (Nos. 3, 4, 5, 9, 11, 13, and 16)				Choice Architecture (Nos.14 and 15)				Forced (No.8)			
Original		Hypernudge		Original		Hypernudge		Original		Hypernudge	
β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)	β	OR (95% CI)
.545	1.724	.332	1.394	-2.190*	.112	-.544	.580	-.637	.529	-1.104	.331
-.004*	.966* (.992-1.000)	-.009**	.991 (.987-.995]	.008	1.008 (1.000-1.0162)	.009*	1.009 (1.001-1.017)	.004	1.004 (.993-1.015)	.000	1.000 (.989-1.011)
-.052	.950 (.835-1.081)	.060	1.062 (.933-1.209)	.020	1.020 (.791-1.314)	.060	1.062 (.817-1.382)	.027	1.027 (.723-1.457)	-.152	.859 (.611-1.207)
—	—	.218*	1.243 (1.008-1.534)	—	—	.073	1.076 (.792-1.462)	.117	1.124 (.748-1.688)	—	—
.088	1.092 (.888-1.343)	—	—	-.049	.952 (.651-1.393)	-.061	.941 (.646-1.370)	.264	1.303 (.775-2.189)	-.367	.693 (.398-1.206)
-.183*	.833 (.719-.965)	.473**	1.605 (1.324-1.944)	.380*	1.463 (1.103-1.939)	—	—	—		.474*	1.607 (.398-1.206)
.007	1.007 (.974-1.041)	-.015	.985 (.953-1.018)	-.010	.990 (.923-1.062)	-.068*	.934 (.878-.993)	-.046	.955 (.878-1.038)	.005	1.005 (.912-1.106)
-.324*	.724 (.554-.945)	.258	1.294 (.988-1.695)	.116	1.123 (.684-1.845)	.467	1.596 (.925-2.753)	.535	1.707 (.805-3.621)	.503	1.653 (.823-3.321)
.037	1.037 (.714-1.506)	-.100	.905 (.625-1.309)	.851*	2.341 (1.051-5.213)	-.378	.685 (.343-1.370)	-.432	.649 (.241-1.748)	.699	2.012 (.735-5.507)
-.021	.979 (.783-1.225)	-.035	.966 (.772-1.208)	.263	1.301 (.852-1.989)	.292	1.339 (.860-2.085)	.387	1.473 (.800-2.711)	.197	1.217 (.673-2.200)
-.043	.958 (.771-1.190)	-.056	.761 (.484-1.176)	.201	1.223 (.816-1.833)	.409	1.506 (.976-2.323)	.339	1.404 (.775-2.545)	.280	1.323 (.751-2.331)
—	—	—	—	—	—	—	—	—	—	—	—
4,172		4,172		1,192		1,192		596		596	
$\chi^2(9)$ =28.488**		$\chi^2(9)$ =62.235**		$\chi^2(9)$ =23.179*		$\chi^2(9)$ =29.934**		$\chi^2(9)$ =.9470		$\chi^2(9)$ =17.298*	

* p<.05
** p<.001

Table 6. Estimates of selected individual demographic, political attitude, and mobile usage of selected approval of interventions per five levels of depth: Results of a logistic regression analysis.

6. DISCUSSION AND CONCLUSION

Overall, the results of this study indicated that the approval and disapproval for hypernudges were dramatically different from the original nudges. In this study, it was revealed that hypernudges were not more acceptable than the original nudges. Notably, in hypernudges, while individuals tended to accept the less flexible and forbidden intervention, they rejected the ones that utilized their children's personal data. However, neither typical nor common features were confirmed that could identify the acceptance level of hypernudges, such as categories of interventions, individual sociodemographic factors, political attitudes, and mobile phone usage histories.

As predicted, the deeper the intervention (too much meddling), the less acceptable for people. However, this tendency was seen only in the original nudge. Compared with the original nudge, though only four interventions on 'education campaign for childhood obesity', 'labelling unhealthy food', 'requiring large energy provider to enrol green energy', and 'installing security software' could get more approval in hypernudge, there could be found neither typical nor common features among them. In addition, while several prior studies had investigated that consciousness, contexts, and prosociality of interventions had different effects among people, this study does not recognize the same effects as well. Insignificant effects might stem from a nationality such as Japanese or individual differences that this study did not consider. Sunstein et al. (2018) surveyed the acceptance of nudges in several countries, and it was observed that Japanese had one of the lowest acceptance rates among all. In contrast, Americans, British, and Chinese would favour various types of nudges. Japanese, Hungarians, and Danish tended to hesitate to accept nudges. Although it is esoteric to assert the reasons, we, as Japanese, should take care of the numerous types of interventions, especially by AI-driven artefacts.

This suggests a kind of an alert or dark cloud for the introduction, utilization, and spreading of hypernudges because of neither lower acceptance than the original nudges nor no common and specific traits among the accepted interventions in hypernudges. We should continuously consider the effects of the various categories of hypernudges on various types of people. This study might serve as an onset of prevalence for appropriate AI-driven artefacts.

ACKNOWLEDGEMENTS

This work was supported by the Japan Society for Management Information Grant-in-Aid for SIG 'Monitoring and Control of AI Artefacts' and the Seikei University Grant-in-Aid 2020-21, for the research on 'Monitoring and Control of AI Artefacts: Consideration from Economics, Social, and Legal Perspectives'. I also thank the two anonymous reviewers in ETHICOMP 2020.

REFERENCES

BBS News (2019). Organ donor law change named after Max and Keira, https://www.bbc.com/news/health-47359682

Bovens, L. (2009). The Ethics of Nudge, In T. Yanoff-Grüne and S. O. Hansson. (Eds.) *Preference Change* (pp. 207–209). Dordrecht: Springer.

Bruns, H., Kantorowicz-Reznichenko, E., Klement, K., Jonsson, M. and Rahali, B. (2018). Can Nudges Be Transparent and Yet Effective? *Journal of Economic Psychology*, 65, 41-59.

Diepeveen, S., Ling, T., Suhrcke, M., Roland, M. and Marteau, T.M. (2013). Public acceptability of government intervention to change health-related behaviours: A systematic review and narrative analysis, *BMC Public Health*, 13, 756.

Felsen G, CasteloN, Reiner PB (2013). Decisional Enhancement and Autonomy: Public Attitudes Towards Overt and Covert Nudges. *Judgment and Decision Making*, 8, 202-213.

Furedi, F. (2011). Defending moral autonomy against an army of nudgers, *Spiked.* Retrieved from http://tinyurl.com/6kfafka

Goodwin, T. (2012). Why we should reject 'nudge', *Policy and Politics*, 41(2), 159-182.

Jung, J. Y. and Mellers, B. A. (2016). American Attitudes toward Nudges. *Judgment and Decision Making*, 11, 62-74.

Junghans, A. F., Cheung, T. T. L. and De Ridder, D.T.D. (2015). 'Under consumers' scrutiny: An investigation into consumers' attitudes and concerns about nudging in the realm of health behavior', *BMC Public Health* 15, 336.

Junghans, A. F. and Marchiori, D., and De Ridder, D. T. D. (2016). The Who and How of Nudging: Cross-national Perspectives on Consumer Approval in Eating Behaviour. Unpublished Manuscript, Utrecht: Utrecht University.

IEEE standard association (2018). Affective Computing, *Ethically Aligned Design*, ver. 2nd., pp. 162-181. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Lanzing, M. (2018). Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies, *Philosophical Technology*, 32(3), 549-568.

O'Neil, B. (2011). 'Nick Clegg's sinister nannies are 'nudging' us towards an Orwellian nightmare", *The Telegraph*. http://tgr.ph/g7gLsp.

Reisch, L. A. and Sunstein, C. (2016). Do Europeans Like Nudges? *Judgment and Decision Making*, 1(4), 310-325.

Schnellenbach, J. (2012). Nudges and norms: On the political economy of soft paternalism, *European Journal of Political Economy*, 28, 266-277.

Sunstein, C. R. (2015). *Nudging and Choice Architecture: Ethical Considerations*, (Harvard John M. Olin Discussion Paper Series Discussion Paper No. 809, Jan. 2015, Yale Journal of Regulation.

Sunstein, C. R. (2016). *The Ethics of Influence: Government in the Age of Behavioral Science.* CUP, New York.

Sunstein, C. R. (2018). Misconceptions about nudges, *Journal of Behavioral Economics for Policy*, 2(1), 61-67.

Sunstein, C. R., Reich, L. A., Rauber, J. (2018). A worldwide consensus on nudging? Not quite, but almost, *Regulation & Governance,* 12, 3-22.

Thaler, R. H. and Sunstein C. R. (2008). Nudge: *Improving Decisions about Health, Wealth, and Happiness.* Yale University Press, New Haven, CT.

Yamazaki, Y. (2019). Certified Requirements of Nudging by AI Artefacts: Hypernudges Driven by AI/ML Artifacts will not be Recognized as Nudges, XXVIII AEDEM International Meeting Tokyo (Japan).

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design, *Information Communication and Society*, 20(1), 118-136.

Weinmann, M., Schneider, C. and Brocke. J. V. (2016). Digital Nudging, *Business and Information Systems Engineering*, 58(6), 433-436.

Wilkinson, T. M. (2012). Nudging and Manipulation, *Political Studies*, 61(2), 341-355.

APPROACH TO LEGISLATION FOR ETHICAL USES OF AI ARTEFACTS IN PRACTICE

Yasuki Sekiguchi

Hokkaido University (Japan)

seki@econ.hokudai.ac.jp

ABSTRACT

AI artefacts, how ethical they are by design and development, must be ethically used in practice in order to keep the coming AI society ethical. This study insists on the necessity of an approach to find practical use cases of AI artefacts requiring legislation against unethical use. The concept of the Mixed World is introduced in order to make it easier to find the corresponding cases in the real world that are related to cases requiring legislation. In order to find AI artefacts used in practice, definitions of AI based on functions but not on technology and related parties are proposed. Four factors that are useful to describe and estimate use cases of AI artefacts are also proposed, and they are applied to some popular systems on the Internet. Finally, all these proposals are applied for making sample regulations to protect and foster sound minors from harm caused by AI artefacts.

KEYWORDS: use-ethics, mixed world, legislation, AI artefact, minor protection.

1. INTRODUCTION

Artificial intelligence (AI) technology is one of critical technologies that drastically transform the modern society. Its penetration to the society and the economy could change every facet of our lives into the bad as well as good directions. AI technology can be applied for diverse regions. In this sense, it is one of the general purpose technologies (GPTs) of Bresnahan & Trajtenberg (1995). There are a wide variety of regulations related to every GPT (e.g., farming, steam engine, internal-combustion engine and electric power) so that the society and the economy can be kept ethical. AI technology must be added to such GPTs. The purpose of this study is to propose an approach for efficiently and systematically finding unethical use cases of AI artefacts and composing regulations effective in preventing them.

Ethics principles or guidelines of AI and AI artefacts were discussed by public and private initiatives and quite a few reports have been published. See for example, European Committee (AI HLEG, 2019), Japan (the Conference toward AI Network Society, 2017) and IEEE (the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017). Refer to survey papers such as Floridi and Cowles (2019) and Leikas et al. (2019) for more information. R&D and production of AI artefacts are expected to comply with these ethical requirements and supply only ethical products and services on the market. This is labelled as ‘supply-ethics’. Supply-ethics can only be complied with in the context that is supposed during the process of R&D. Therefore, regardless of how firmly supply-ethics were complied with, any technological artefacts can be used both ethically and unethically. This is because the context where a technological artefact is used in practice can be different from the supposed one. In order to keep the coming AI society ethical, unethical uses of AI artefacts in practice must be surely prohibited. This is labelled as 'use-ethics'. This is why regulations on practical uses of AI artefacts are necessary. Thus, we need to find unethical use cases of AI artefacts and compose some regulations to prevent them.

In the next section, a concept of the 'mixed world' is introduced. It will help to find situations where some regulations might be required. In section 3, we briefly look into the regulations that seem to have been effective for keeping the motorization society ethical. The inquiry discovers four factors important to describe situations or contexts where some regulations are needed. Section 4 is devoted to define AI artefacts and parties related to practical uses of AI artefacts, and some cases are listed to show effectiveness of the definitions.

As an example application of the approach proposed in the sections 2 to 4, section 5 investigates protection of sound minors from unethical uses of technological artefacts. Firstly, a set of regulations for protection of minors in the real world is investigated and then some ideas for protection of minors from application of AI artefacts targeting at their mental vulnerability.

2. FROM THE RW TO THE MW

Our living space is sometimes labelled as the real world (RW). It is the physical world we live and recognize. Information systems (ISs) composed of computers and relevant technologies also exist therein. The Internet and various information network systems are included in Iss. Diverse data, information and knowledge about the RW, and also data, information and knowledge which describe diverse ideas created by human are stored and distributed on Iss in various forms such as numbers, characters, pictures or animations. They are all stored as digital data and electro-magnetic existence in the RW. We call them simply "data" hereafter.

However, once a person understands and perceives the meaning of data, it is related to some things or some creations in the RW. The things or creations related to the data are images emerged from understandings of the data by the person. Such images are existence in the person's mind. The space composed of such images in the person's mind is labelled as the cyber world (CW). Note that the CW of one person differs from those of the others.

These two worlds have been separated in the sense that the CW was at the outside of the RW, i.e. our everyday life. We only "used" objects in the CW in order to add interests to our lives. This is why ethics in the CW have not been worried about up to recently.

The Internet and mobile devices have changed the situation. It seems not rare now that an individual consumes a few hours a day to communicate with social network services, roll playing games, smart speakers, etc. The CW has become an inevitable part of our everyday lives. This situation is described by the HM Government (2019) as "the power and influence of large companies has grown, and privately-run platforms have become akin to public spaces". Newport (2019) describes this transition as follows: We added new technologies to the periphery of our experience for minor reasons, then woke one morning to discover that they had colonized the core of our daily life. Thus, our living space has become a mixture of the RW and the CW. The new space is named the mixed world (MW).

The coming AI society would be ethical, only if the MW is ethical. For example, the purpose of the Act on the Protection of Personal Information of Japan is the protection of the rights and interests of individuals in the RW who are linked to individuals (in the CW) described by personal data used practically. Thus, the protection of personal information should be understood as an extension of the protection of privacy in the RW into the MW.

Remember that the currently existing law and regulation barely keep the RW ethical. Because the MW is an expansion of the RW into the CW, it seems reasonable to expect some extension of the existing regulations would be effective for keeping the MW ethical, too. When there are some regions or groups especially requiring ethical considerations or protection against application of technological artefacts in the RW, i.e., if there are some regulations concerning such regions and groups, necessity of similar regulations against application of AI artefacts causing similar harm should be investigated. This view is one of our proposals here.

3. LESSONS FROM THE RW: USE ETHICS OF MOTOR VEHICLES

Developers of an AI artefact suppose its use context in order to specify technological requirements. A resultant AI artefact might not be used in the supposed context, but often used in different contexts. The supply-ethics by R&D is effective only in the supposed context, and the AI artefact used in practice can happen to have unethical effects on the MW.

The situation of the traditional technological artefacts such as motor vehicles and services based on them was similar. Some regulations have been necessary for motor vehicles to be ethically used even though they are ethical in the context supposed by R&D. Thus, there exist diverse regulations so as to make motor vehicles to be ethically used in practice even when the scope is restricted within road traffic. To survey them exhaustively is out of the scope of this study, so only those related to Road Traffic Law (RTL) in Japan are briefly investigated below.

RTL concerns the usage of motor vehicles mainly on roads, although it does a little R&D and production of technological artefacts related to motor vehicles. This is very contrastive to ethics principles and guidelines which deeply refer to R&D and production but little to usage.

First, we look into the regulations that set the prerequisite conditions of RTL. Vehicles are used to transport people and things. Road networks are also used for the same purpose. Individuals and businesses are users of vehicles. An individual user may be a driver as well as an owner of a vehicle. Typical businesses are the passenger or freight transportations. It is common among transportation businesses to own necessary motor vehicles and employ their drivers.

All these imply that some regulations to control roads and their networks, transportation businesses, owners and drivers as well as vehicles are necessary in order to comply with use-ethics. Because the knowledge of traffic rules and driving techniques is necessary to drive motor vehicles, some rule of driver’s licenses is also required. Moreover, some regulation about the test and maintenance of finished vehicles are necessary too, because it is also necessary to keep the quality of vehicles above a sufficient level of safety. Table 1 shows some of such and regulations with brief explanations.

Table 1. Examples of Regulations Related to the RTL.

Law and Regulation	Brief Explanation
Road Transport Vehicle Act	Determines matters related to possession and maintenance that are necessary for securing safety and preserving environment such as prevention of pollution and etc., including determination of the range of road traffic vehicles, obligation for registration, procedure of registration. Determines also maintenance of roads where road transport vehicles run and placement of traffic signs.
Type Designation Rule of Motor Vehicles	Determines implementation details of the RTL such as procedure of type designation, criterion of inspections and format of certificate of completion inspection.
Road Act	Determines matters related to promoting development of road networks such as route designation and accreditation, administration, formation, maintenance, cost division, etc.
Road Transportation Act	Determines the taxi and bus businesses as passenger transportation by motor vehicles, and the exclusive road business of motor vehicles.
Consigned Freight Forwarding Business Act	Aims at the sound development of consigned freight forwarding business and to ensure the smooth provision of freight forwarding business that meets the needs of higher and diversified demand of users in the field of freight distribution through ensuring the fair and reasonable management of the consigned freight forwarding business, thereby contributing to protection of the users' interests and their convenience.
Road Traffic Law Enforcement Decree	Determines maximum and minimum speeds, point system for traffic violations and accidents, etc.

Source: self-elaboration

Type Designation Rule of Motor Vehicles concerns R&D of motor vehicles. Because of this rule, automakers must execute a certificate of completion inspection at the time of transfer of every motor vehicle of their products after a successful inspection by a certified inspector who approves the motor vehicle satisfies the quality standard of its designated type. This makes it possible to comply with supply-ethics.

Road Act is aimed at realizing the compliance with supply-ethics of roads where motor vehicles run. Businesses that are operated with utilization of roads are regulated by Road Transportation Act and Consigned Freight Forwarding Business Act.

RTL is aimed at preventing road hazards and otherwise ensure the safety and fluidity of traffic, as well as contributing to preventing blockages arising from road traffic (Article 1). It determines the obligations of drivers, businesses employing drivers, motor vehicle owners and the functions of the Public Safety Commission and the Police concerning regulation and management of road traffic. Articles on the maximum and minimum speeds also exist, and they are complemented by Road Traffic Law Enforcement Decree. Vehicles available for road traffic are restricted within road transport vehicles, and the rule is complemented by Road Transport Vehicle Act and Type Designation Rule of Motor Vehicles. There are articles concerning proper labour management when passenger transportation businesses by motor vehicles and consigned freight forwarding businesses employ drivers. RTL determines that a driver must get a driver’s license and carry it while driving. It also determines requirements for driver’s license of each type of road traffic vehicles. In addition, it determines traffic rules of pedestrians and motor vehicles.

We have briefly observed that use-ethics of motor vehicles is complied with by a set of regulations. Act on Punishment of Acts Inflicting Death or Injury on Others by Driving a Motor Vehicle, etc. is applied to especially poor driving. Moreover, technologies necessary for enforcing these laws and regulations are developed, e.g. automatic speed control devices.

As for legislation for ethical uses of technological artefacts such as motor vehicles, note that RTL quotes a set of related laws and regulations in order to specify circumstances of practical uses, and at least the following four factors are specified for clear description of use situations: types of artefacts to be regulated (such as motor vehicles and roads), relevant parties (such as drivers and businesses), intended functions of use of the artefact (such as individual travel and business), and the purpose of use such as maximization of benefit implicit in articles related to labour management.

4. A METHOD TO DESCRIBE USE SITUATIONS OF AI ARTEFACTS AND APPLICATIONS

The result of the preceding section shows that AI artefacts and related parties must be defined visibly for effective regulation. Establishing a method to do it is the objective here.

4.1. A functional definition of AI artefacts

AI in this study implies a system which can autonomously change its outputs or the way to determine its outputs through its operation, based on “learning etc.” of data. Learning etc. here includes learning of data, inference based on data, exploitation of data obtained during its operation, etc. Data to be explored can be obtained from responses of its users or though interactions with its circumstances by sensors, actuators, etc. The way of determining outputs includes recognition, inference, judgement, decision making, etc.

This is almost the same as the definition of “AI system” in the Conference toward AI Network Society (2017). It is not much different from the definition of A/IS in the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017) and the definition of AI by AI HLEG (2019).

The importance of this definition is the fact that it is possible to estimate whether AI is used in a system only by observing if it autonomously makes intellectual decisions or judgements on the basis of accumulation of data or communication with its circumstances, and without wondering its algorithm.

4.2. Composition of AI artefacts and relevant parties

The following are the definitions of terms shown in Figure 1.

> “Learning etc. data” are those that are used for learning etc. of AI. One type of learning etc. data is teaching or training data of machine learning. Another is data to be explored as noted in the definition of AI. Metadata is also included.
>
> “Outcome data” includes data obtained as the result of learning etc. and changes outputs and the way to determine outputs of AI. Outcome data also includes the basic forms of algorithms that determine outputs of AI, i.e. the ways of recognition, inference, judgement, decision making, etc. The algorithm(s) affected by the result of learning etc. determine(s) outputs of AI. It also includes metadata.

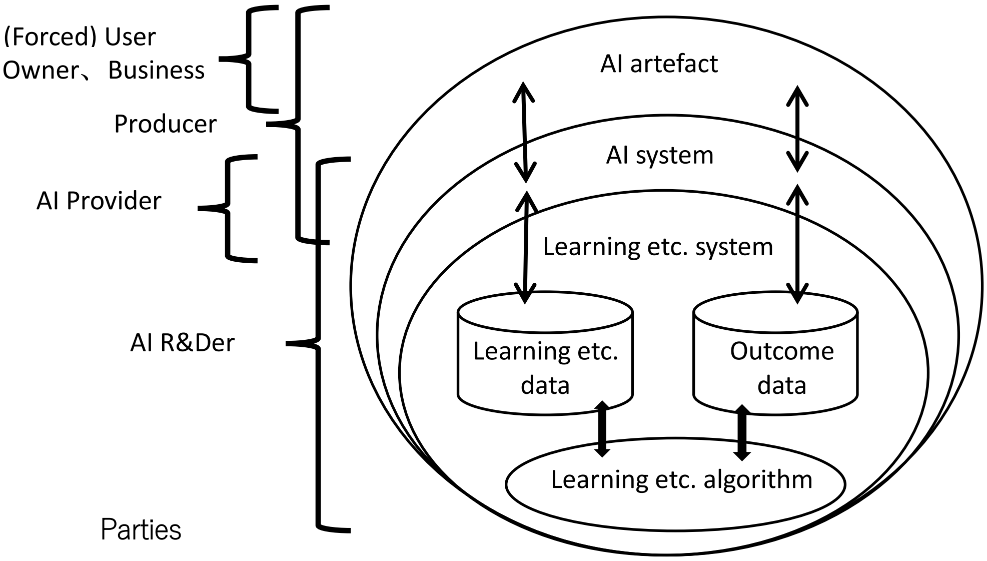
"Learning etc. algorithm" is one that performs learning etc., and originates the intellectual and adaptive function of AI. It may be a composition of plural kinds of algorithms such as machine learning, statistical analysis, optimization, etc. If there are some parameters used to adjust learning etc., they are also included.

"Learning etc. system" is one that is composed of the above three components and some interface to outside systems. When the learning etc. algorithm refers to a part of or all of the outcome data of preceding learning etc., the outcome data is supposed to be stored as parameters of the learning etc. algorithm, too.

"AI system" is one that has at least one learning etc. system and some components which add functions necessary to interact with both its learning etc. system(s) and other components of artefacts utilising it. That is to say, AI system is one made by adding application programming interfaces to its learning etc. systems. One AI system could be utilized by plural artefacts in parallel, because it is an IS. This implies that an AI system enables artefacts using it to apply the function of its learning etc. system. As a whole, AI system is identical with AI.

"AI artefact" is one that cannot function without at least one AI system. Its AI systems might work independently, or work in corporation with each other as a network.

Figure 1. Composition of AI artefact and relevant parties.



Source: self-elaboration

An AI artefact may be either an IS, or a physical system which utilizes at least one AI system. Example ISs are SNSs like Facebook and Twitter, search sites like Google and Yahoo, on-line shops like Amazon and Rakuten, etc. Example physical systems are AI speakers and intelligent robots.

Because AI systems are ISs, if a learning etc. system (or an AI system itself) is on the cloud, every artefact that utilizes it through the Internet is an AI artefact. The current IS devices such as smartphones, tablets, personal computers, etc. are all AI artefacts in this sense. Likewise, AI artefacts are ubiquitous now.

Parties relevant to AI artefacts are shown on the left side of AI artefact in Figure 1. AI R&Ders are designers and developers of AI systems and their APIs. AI providers provide their customers with functions of AI systems as services. Some producers of AI artefacts would operate AI systems by themselves and do not need services by AI providers. Owners of AI artefacts would use them for various purposes. If an owner is an individual, the owner is at the same time a user, e.g. AI speakers. If an owner is a business, the business would use an AI system in order to realize a service to others, i.e. customers. Examples are SNSs and online shops. Customers of such services are inevitably applied outputs from AI artefacts. Then, customers of such services would be called forced users.

An AI artefact might use plural AI systems, and their effect on ethics in the MW can be different from each other. Therefore, when the effect of an AI artefact on ethics is assessed, each AI system must be separately evaluated. This is another of our proposals.

4.3. A method to describe AI artefacts in practical use and example applications

Referring to the case of legislation for ethical use of motor vehicles in section 3 and the definitions in the preceding subsection, the following four properties must be estimated in order to clearly describe the situation of practical use of an AI artefact: function of AI system (i.e., results of recognition, inference, judgement, decision making, etc. by the AI system), AI artefact that uses the AI system, owner/business, and user/forced user. Based on such description, necessity of some regulations for preventing unethical use can be investigated.

AI artefacts estimated and listed in Table 2 are those that offer services or additional functions to services mainly on the Internet. A surrogate variable named “values of” in Table 2 is used for owner/business, because AI systems of such AI artefacts are used to acquire the values of owner/business. Likewise, the label “Example” is used for services offered by AI artefacts that use the corresponding function, and “Targeted to” for forced users.

Table 2. Examples of AI Systems used in services.

Function	Targeted to	Values of	Example	Learning etc. data	Issue
Use incentive	Individual user (User group)	Service provider	Like!, Share, Retweet	Use record, Individual response record	Vulnerability
Item sale	Individual user (User group)	Service provider	On-line game, Social game	Use record, Personal data	Vulnerability
Ads targeting	Individual user	Service provider/ Advertiser	SNS, EC	Use record, Personal data, Ads specs	Vulnerability
Recommend	Individual user	Service provider / Seller	EC	Use record, Personal data, Goods data	Vulnerability
Input assistance	Individual user (User group)	Service provider	Auto complete	Input record	Filter bubble
Display order	Individual user	Service provider	Search Engine	Personal query record, Personal data	Filter bubble
Medical diagnosis	Diagnostician	Appropriateness	Heart disease on ECGs	Medical data, Diagnosis record	Moral hazard
Propriety judgement	Applicant	Business	Loan, Employment	Applicant data, Judgement record	Fairness

Source: self-elaboration based on literature such as Alter (2017), Cooper (2017) and Vlahos (2019)

Table 2 has two additional columns: “Learning etc. data” is important because it shows the source of the corresponding function, “Issue” because it is useful for investigating necessity of legislation. The values in cells of Table 2 are estimated from author’s understandings of cited literature and experiences because details of AI systems have not been made public, and would be neither exact nor precise. For example, the AI system that determines display order can happen to use query record of all users as well as personal query record and personal data, or it can aim at acquiring values of business compromised with estimated user values.

Note that Table 2 does not refer to the algorithmic side of AI systems at all. Then, how these systems estimated that they use AI systems? It is because these systems make different outputs to each target objects (i.e., each forced users) or at each occasion. The fact fits to the definition of AI in 4.1. Because of this nature, this study does not conflict with its stance to promote AI technology. Use-ethics does not concern technology itself, but usage.

5. PREVENTING UNETHICAL USES OF AI ARTIFACTS TO MINORS

An example of application of the proposed approach is described below. First some regulations in the RW that prevent unethical uses of technological artefacts on minors in Japan are explained. Three types of methodology of prevention are indicated. Second, considering the protection of minors in the RW, some of the cases in Table 2 seem to require legislation against unethical uses in the MW. Finally, tentative plans of legislation are proposed by applying one of the three methodologies.

5.1. Regulations to protect minors from traditional technological artefacts

Minors are those less than 18 years old in Japan. Minor Protection Ordinance is provided by each prefectural government for the purpose of arranging a good local circumstance adequate for protecting and fostering sound minors. This ordinance regulates use of traditional technologies to minors.

There are some laws protecting minors such as Drinking Prohibition Act for People Underage, Smoking Prohibition Act for People Underage, Child Welfare Act, Child Prostitution and Child Pornography Prohibition Act, Child Abuse Prevention Act, etc. These acts also define punishment for violation of prohibitions, and they are sometimes applied to cases of violation of Minor Protection Ordinance.

Children in the above acts are the same as minors. People underage are those less than 20 years old in Japan. The purpose of these acts is typically stated in Article 1 of the Child Abuse Prevention Act that it aims to serve protection of child’s rights and interests. In other words, they aims to prevent harms to users (or forced users) caused by technological artefacts or services offered by using them, or to compensate immaturity and insufficiency of people underage.

Looking into these regulations, the methodologies used can be classified into three below:

> [Prohibition, restriction] Prohibit use of artefacts to child, minor or people underage or restrict place, occasion and/or age group of use, in accordance with nature and function of artefacts. Examples are cigarettes, drink, (adult) movies, motor vehicles, firearms and swords, etc.

> [Business permission, registration] Obligate license and/or registration of business to sell service provided by using technological artefacts (such as gambling) as well as technological products (such as firearms, swords)
>
> [Possession registration, use license] Obligate possession registration and/or user license for artefacts with high risk to harm others such as motor vehicles, craft, planes, powder, drugs, firearms and swords.

The fact that the meaning and effect of use of technological artefacts varies depending on circumstances and user properties is surely recognized in the RW, and there are a variety of regulations in order to comply with use-ethics for the purpose of making it possible both to protect minors and to foster them soundly.

5.2. Regulation against AI artefacts capitalizing on vulnerability of minors

It has been become evident that mobile devices and use of SNSs cause really harmful influence on minors from remarks by CEOs of giant IT companies like Nadella (2019) and by software engineers developed SNSs like Bowles (2018) and from some studies like Carr (2017) and Moscaritolo (2018). Such harmful influence seems especially strong where AI systems capitalizing on vulnerability of (forced) users are applied against minors. Examples of such application are shown in the first four rows in Table 2. Remember that minors are one of most vulnerable parties.

Considering that they are ubiquitous already, investigation of legislation against such uses is an urgent issue for compliance with use-ethics. There have already been some movement toward this direction in Japan as shown by the following two examples:

Act on Development of an Environment that Provides Safe and Secure Internet Use for Young People revised in 2017 promotes use of filtering software for devices that are possessed by minors and used for the Internet surfing. ISPs are obligated to supply filtering software and explain about it to minors and their curators. Minors and their curators are obligated to set up filtering software. Effectiveness of the act depends solely on the quality of filtering software and the action took by curators.

There have been increasing incidents where minors are requested to send their selfies by correspondents on SNSs and get harmed as a result. To prevent such incidents, some prefectures have revised Minor Protection Ordinance.

Another important issue is brain drain or brain hijack. It harms human intelligence, especially of minors. See for example Carr (2017), Kawashima (2018) and Sakurai (2019). It seems that brain drain is strengthened by the use of AI systems in some services such as those shown in the first four rows in Table 2. Sociality that is thought a unique human ability develops mainly up to 9 year old, and it is also harmed.

This study proposes three regulations that are formed by applying the first methodology resulted from 5.1 on use cases of AI systems observed in Table 2.

(1) Prohibit AI systems from using personal profile data for use incentive, items sale, ads targeting, and recommendation directed to minors.

(2) Prohibit applications like SNSs opened to minors from implementing AI systems for

strengthening use incentive like validation feedback (e.g. Likes!, Share, Retweet).

(3) Obligate service providers to set the lowest allowable age on each application and content, and to implement a mechanism to prevent minors below the age from using them.

The age of users must be identified to enforce these regulations if they are legislated, because they need to classify users into minors and the others. This is not possible currently. The three regulations can be legislated with a deadline for enforcement, as it was the case of the control of exhaust gas.

6. CONCLUSION

This study proposed an approach to legislation against unethical uses of AI artefacts. The approach is based on the idea to utilise regulations for preventing unethical uses of traditional technological artefacts in the RW and extend them into regulations in the MW against unethical uses of AI artefacts. This study also proposes a functional definition of AI, and applied it for defining AI artefacts and relevant parties. The functional definition was used to analyse and estimate practical use cases of AI artefacts on the Internet. As an example application of the proposed approach, some regulations against unethical uses of traditional technological artefacts to minors were investigated and three methodologies were induced. One of the methodologies was applied to use cases that capitalize vulnerability of minors and three sample regulations were formed.

The proposed approach made it possible to estimate effect of AI systems on user’s thought and mind without referring to technological properties, and to find focal points requiring some regulations in the MW.

When the proposed approach is applied to protect workers affected by practical use of AI systems in the business scene, some reliable methods to evaluate actual changes emerged from introduction of AI systems must be established. The interpretive approach (Orlikowski, 2000) seems promising. The approach proposed here is reactive and fragmental in contrast to the proactive and comprehensive one used for the study of ethics principles and R&D guidelines. An Approach that complements these two is strongly expected to appear.

ACKNOWLEDGEMENTS

This study was supported by the Japan Society for Management and Information Grant-in Aid for SIG “Monitoring and Control of AI Artefacts”, and the Seikei University Grant-in-Aid 2020 for the research on “Monitoring and Control of AI Artefacts: Consideration from economics, social, and legal perspectives”.

REFERENCES

AI HLEG (April 8, 2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/ newsroom/dae/document.cfm?doc_id=60419

Alter, A. (2017). *Irresistible: The rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Press.

Bowles, N.（February 4, 2018）“Early Facebook and Google Employees Form Coalition to Fight What They Built.” New York Times. Retrieved from https://www.nytimes.com /2018/02/04 /technology/ early-facebook-google-employees-fight-tech.html

Bresnahan, T.F., & Trajtenberg, M. (January, 1995). General purpose technologies 'Engines of growth'? *Journal of Econometrics*, 65(1), 83-108.

Carr, N. (October 6, 2017) How Smartphones Hijack Our Minds. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/how-smartphones-hijack-our-minds 1507307811? mod=djcm_fcfeml1&campaign=djmc_WSJ_CEmail_Article2

Floridi, L. & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* Issue 1,1-13. Retrieved from https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Cooper, A. (2017, April 9). What is ’Brain Hacking’? Tech insiders on why you should care. Retrieved from https://www. cbsnews.com/news/brain-hacking-tech-insiders-60-minutes/

HM Government (April 8, 2019). *Online Harms White Paper*. Retrieved from https://assets. publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793 360/Online_Harms_White_Paper.pdf

Kawashima, R. (2018). *Smartphones Destroy Scholastic Ability* (in Japanese). (Shueisha)

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(1), 1-12. Retrieved from https://doi.org/10.3390/joitmc5010018.

Moscaritolo, A. (February 22, 2018). Almost Half of Parents Say Their Kid Is Addicted to Tech. Retrieved from https://www.pcmag.com/news/359402/almost-half-of-parents-say-their-kid-is-addicted-to-tech

Nadella, S. (2016) The Partnership of the Future. Retrieved from https://slate.com/technology/ 2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

Newport, C. (2019) *Digital Minimalism: Choosing a Focused Life in a Noisy World*. Kindle version, Penguin Random House LLC.

Orlikowski, W.J. (2000). Using Technology and Constituting Structures: A practice lens for studying technology in organizations. *Organizational Science*, 11(4), 404-428.

Sakurai, T. (2018). *How does Mind Emerge?* (in Japanese). (Kodansha) pp.37-39.

The Conference toward AI Network Society (2017). *Report 2017-To Activate International Debate on AI Networking* (in Japanese). Retrieved from http://www.soumu.go.jp/main_content/000499624.pdf

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2. IEEE. Retrieved from https://standards.ieee.org/content/ieee-standards/en/ industry-connections/ec/autonomous-systems.html

Vlahos, J. (2019). *Talk to Me: Amazon, Google, Apple and the Race for Voice-Controlled AI*. An Eamon Dolan Book, Houghton Mifflin Harcourt, Boston, New York.

ARTIFICIAL INTELLIGENCE AND MASS INCARCERATION

Leah Rosenbloom

The Workshop School (USA)

leah.rosenbloom@gmail.com

ABSTRACT

Artificial intelligence (AI) is now common throughout the criminal justice system. Police use predictive algorithms to target locations and individuals for surveillance. Judges use risk assessment algorithms to determine whether defendants should be granted bail or parole. Prosecutors use the results of forensic analysis algorithms to accuse and convict defendants of crimes, including those punishable by death. These algorithms are considered intellectual property and are closed off from public scrutiny.

In this paper, we explore the impact of AI on mass incarceration. A comprehensive survey of existing practice reveals the ways in which algorithms perpetuate systemic injustice and violate defendants' legal rights. We argue the need for solutions that integrate technical and legal perspectives, including novel ways to shift the focus of the algorithms from punitive to restorative practices. With due oversight, transparency, and collaboration between experts in technology, law, and government, we can leverage existing algorithms to combat systemic injustice.

KEYWORDS: artificial intelligence, predictive policing, risk assessment, machine testimony, restorative justice, technology and the law.

1. INTRODUCTION

While existing literature concerning algorithms in criminal justice applications is extensive, research is scattered between the scientific and legal communities. In order to form a complete picture of the impact of AI on mass incarceration, which touches problems in machine learning, data science, law, and governance, it is necessary to integrate these perspectives. To that end, this paper explores sources, issues, and proposed solutions in each area of research.

Several ethical concerns emerge from the survey of existing literature. First is the issue of data that reflect existing racial and socio-economic bias in the criminal justice system. Data science experts have proposed statistical models that remove racial bias from the data, which leads us to consider the implications of "objective" black-box algorithms operating within contexts of deeply entrenched bias. We argue that the use of black-box solutions to racial discrimination encourages law enforcement and judiciaries to defer their responsibilities, preferring automatic arrests and convictions over critical consideration.

Without transparency and oversight, it is impossible to examine the underlying mechanisms that process and objectify the data. Furthermore, even if the underlying data is scrubbed clean, the "correctness" metrics and implementation of the algorithms may still reflect systemic bias. Comprehensive solutions to problems with algorithms in criminal proceedings must include technical, practical, and legal components. We introduce novel, integrated analyses for each application of artificial intelligence in the criminal justice system: predictive policing, risk assessment, and machine testimony. Our conclusion is a call to action: technologists, legal experts, and government officials alike have a responsibility to face—and hopefully fix—these issues.

2. MACHINE LEARNING

Artificial intelligence can be reduced to a machine's ability to learn (Russell and Norvig, 1995). Machine learning is defined as the ability to process input data such that the categorization of new data is correct to some degree of approximation. While a specific analysis of closed-source algorithms is regrettably impossible, all machine learning algorithms must necessarily "learn" from pre-existing data. Therefore, we can use our understanding of the data to draw conclusions about the effectiveness of these algorithms in practice.

2.1. How machine learning works

Learning algorithms break down into two stages: a learning or "training" stage, and an extrapolation stage. In the learning stage, the machine classifies human-configured data into categories, which are either human or machine-generated. We can visualize this process as points in space, where data with similar contexts and outcomes are plotted closer together. The machine then draws boundaries around close clusters of points, segmenting the space into regions. In the case of predictive policing, the algorithms classify existing data on policing. This might include the type and location of reported crimes, existing patrol routes and routines, or background information on individual suspects and arrests. If the goal of the algorithm is to locate future crimes, the algorithm might draw boundaries around groups of points with high crime rates.

Once the machine has processed the training data, it is ready to classify new data in the extrapolation stage. It plots the new point into the existing space, and the point takes on the outcome of the spatial region in which it is plotted. For example, in the case of risk assessment, the new data point is a new defendant who may share similar attributes and criminal history with existing defendants. The algorithm places the defendant among defendants with similar histories, and projects the new defendant's risk based on the previous defendants' outcomes.

2.2. Measuring correctness

The accuracy of the algorithm in the extrapolation stage reduces to the correctness of the boundaries drawn around the initial training data set. The more voluminous, diverse, and correctly classified the training data set, the better the algorithm will do. If the training data are malformed—if points are incorrectly classified or there are not enough points in a particular area—the classification of new data can be incorrect or unpredictable. For human-verifiable problems like image recognition, it is possible to run the algorithm on new data in the extrapolation stage and directly measure the accuracy of the results.

Correctness is difficult to measure for complex human systems because the input data are generally too large to evaluate on a case-by-case basis, and they are full of error and inconsistency. For example, DNA samples can be easily contaminated, intermixed, and degraded (DiFonzo, 2005). Algorithms built on millions of these inconsistent samples are similarly inconsistent. Unlike image classification, it is not trivial for a human to step in and verify the correctness of a DNA match; analysts cannot comb through millions of samples to verify their reliability, nor go back in time and preserve evidence from the crime scene.

2.3. Existing criminal justice practices are not correct

Logically we can expect that any errors, inconsistencies, and biases in the underlying training data will carry over into the algorithms. Communities that are already hyper-targeted by law enforcement will be similarly targeted by predictive policing. Risk assessment algorithms will work better for defendants who have been treated fairly in the past. Machine analysis of forensic evidence will only yield correct results if similar evidence has been impeccably collected, stored, and analysed. We know from existing problems in all of these areas that the training data is far from accurate and unbiased. Law enforcement has a serious and long-standing problem with racist policing (Langan, 1995; Alexander, 2010; Lum & Isaac, 2016). Judges are known to make racially biased decisions about bail, sentencing, and parole (Johndrow & Lum, 2017; Goel et al., 2018). Forensic evidence is often contaminated, and analysts are known to make mistakes and collude with prosecutors to guarantee convictions (DiFonzo, 2005; Shaer, 2016; Mettler, 2017). Rather than acknowledging and examining these issues, law enforcement and legal systems continue to plow forward with the integration of machine learning into criminal justice proceedings (Mohler et al., 2015; Danner et al., 2016; Saunders et al, 2016; Kaufman et al., 2017; Winston, 2018; Human Rights Watch, 2018). The result, as we will discuss in the rest of our paper, is the blind perpetuation of injustice.

3. PREDICTIVE POLICING

The U.S. National Institute of Justice (NIJ) describes predictive policing as a law enforcement approach that “leverages computer models…for law enforcement purposes, namely anticipating likely crime events and informing actions to prevent crime” (2014). These computer models are trained on existing reports of criminal and police activity. One of the most widely-used algorithms, PredPol, uses only the “three most objective data points” of time, location, and type of previously-reported crime in each precinct’s regional area (PredPol, 2020). Others, like Chicago’s “Strategic Subjects List”, focus on identifying groups and individuals (Saunders et al., 2016). The NIJ confirms that predictive algorithms can focus on “places, people, groups, or incidents” (NIJ, 2014). Each of these models has been shown to perpetuate existing racial and socio-economic bias (Saunders et al., 2016; Lum & Isaac, 2016).

3.1. In practice, predictive policing gets personal

The Chicago Police Department (CPD) uses predictive policing to curate a “heat list” of people who are likely to be involved in violent gun crime, either as victims or perpetrators. The 2013 pilot program, which was funded by the NIJ, used an algorithm on “co-arrest networks” along with human intelligence to produce a “Strategic Subjects List” (SSL) of 426 high-risk individuals.

Saunders et al. evaluated the pilot in 2016 and noted these individuals were "not necessarily under official criminal justice supervision nor were they identified through intelligence to be particularly criminally active" (p. 349).

The list was disseminated to commanders in each police district, who decided on an individual basis how and when to use the list to inform practice. In 10 of 22 districts (45.4%), officers made contact with named individuals only if they spotted one acting suspiciously. In 7 of 22 districts (31.6%), officers made regular visits to named individuals' homes. In the remaining 5 of 22 districts (22.6%), officers used a combination approach. Otherwise, there was "no practical direction about what to do with individuals on the SSL" once they were contacted (Saunders et al., 2016, p. 356).

The study found that the pilot had no statistically significant effect on the homicide rates in Chicago (Saunders et al., 2016, p. 361). Out of 405 total homicide victims between 2013 and 2014, the pilot identified only three (0.74%). During the same time period, police made contact with almost 90% of the individuals on the list with an average of 10.72 interactions each, a 39% increase over a matched control group (p. 363). While increased police contact did not affect an individual's likelihood of being arrested for a shooting, individuals on the SSL were 2.88 times more likely to be arrested for a shooting (p. 362-363). The CPD explained this phenomenon to the study's authors by admitting that the list was used to come up with suspects for unsolved shootings (p. 365).

3.2. The positive feedback loop of racially biased policing

A subsequent study by Lum and Isaac in 2016 begins with a grim anecdotal account of the Chicago pilot. A CPD commander visits the home of a 22-year-old black man on the South Side of Chicago. The commander tells him, this is a warning: you'd better not commit any more crimes. The man is confused. He is not involved in crime. He is on a heat list.

Black and Brown people have been disproportionately targeted for arrest, incarceration, and police brutality in the United States going back hundreds of years (Alexander, 2010). One of the most well-studied examples of racial discrimination in policing is the War on Drugs. A Department of Justice report published in 1995 found that while only 16% of black people reported selling drugs, they accounted for 49% of drug distribution arrests, with similar numbers of discrepancy for drug possession (Langan, 1995, p. 3). Lum and Isaac demonstrated a similar discrepancy using data from the 2011 National Survey on Drug Use and Health. They compared the demographics of arrest records in Oakland, California to the estimated demographics of drug users in Oakland according to the survey. What they found is that while drug use was roughly evenly distributed over the population, low-income and minority neighborhoods had 200 times more arrests than their middle- and upper-class white counterparts (Lum & Isaac, 2016, p. 17).

Lum and Isaac further simulated what impact the popular predictive policing algorithm PredPol would have had on the Oakland population. PredPol claims that its omission of personal information in the training data "eliminat[es]…profiling concerns" (PredPol, 2020), but the simulation illustrated how PredPol would continue and potentially worsen the disproportionate targeting of minority communities. Most notably, the authors created a positive feedback loop with PredPol's algorithm: the more the algorithm sent police to particular neighborhoods, the more crimes would be reported in those neighborhoods, the more likely the algorithm would be to send police back to those neighborhoods on a repeated basis (Lum & Isaac, 2016, p. 18-19).

Machine learning experts have begun to address PredPol's positive feedback loop. One such study by Ensign et al. proposes filtering input data to obtain a more representative sample (2018). The authors admit, however, that their model does not address the underlying problem of biased reporting and arrests. Rather, they falsely assume that crimes identified by the police are equivalent to true crime rates (Ensign et al., 2018, p.11).

3.3. Algorithms currently facilitate bad practices

There are two fundamental issues with predictive policing. First, policing data reflects institutionalized racism (Langan, 1995; Alexander, 2010; Lum & Isaac, 2016). Another study from the same time period compared Los Angeles districts using predictive algorithms against those using traditional methods. They found no statistically significant difference in the racial and ethnic breakdown of arrests between the two (Mohler et al., 2015). This is not surprising: programmed correctly, an algorithm will echo, but not enhance or diminish, existing racism.

Companies like PredPol selling their algorithms as "objective" solutions to racial bias are misinformed. This perpetuates the existence of "colorblind" racism, whereby people are convinced they are race-blind despite overwhelming evidence to the contrary (Alexander, 2010). Reliance on these algorithms will encourage law enforcement to stop thinking about the problem of biased crime reporting and arrests, even as those same biases continue to dictate their actions. If we were to ask the police commander mentioned above—a figure with authority over an entire district on the South Side of Chicago—why he knocked on that 22-year-old black man's door, would he think critically about his behavior? Would he consider it necessary to provide independent justification? Or would he respond indignantly that the man was flagged by cutting-edge technology?

The second major issue is therefore implementation—what actually happens as a result of the algorithm's predictions. Even if the algorithms really did provide an objective analysis of future criminal activity, police commanders and officers would still be the final arbiters in deciding how the results are applied. The stated purpose of the heat list in Chicago was to deter gun crime, but in reality, the CPD used the list to harass people and produce suspects for open shootings (Danner et al., 2016). A recent civilian audit of the Los Angeles Police Department found that the department's data-driven predictive policing programs "lacked oversight and that officers used inconsistent criteria to label people as 'chronic offenders'" (Puente, 2019). Some of those programs have since been decommissioned.

In order to ensure police are acting responsibly, advocates stress the need for transparency surrounding how and when police are relying on algorithms, and for stricter regulation of predictive policing technology.

3.4. From bad to worse

Perhaps the world's most comprehensive predictive policing system is China's Integrated Joint Operations Platform (IJOP), which is widely deployed in the Xinjiang Uygur Autonomous Region. IJOP is a data-driven system that receives constant, real-time input from the following "sensors":

> CCTV cameras, some of which have facial recognition or infrared capabilities…entertainment venues, supermarkets, schools…"wifi sniffers," which collect the unique identifying addresses of computers, smartphones, and other

> networked devices…license plate numbers and citizen ID card numbers from some of the region's countless security checkpoints and "visitors' management systems" in access-controlled communities. (Human Rights Watch, 2018)

In addition to the sensors, IJOP also receives data on criminal history and prior police contact, purchase history and financial records, family planning, legal records, and religious practices, including whether or not the person is an Uyghur. IJOP then issues a daily forecast to law enforcement, including the names of people to investigate further. Unnamed sources report that IJOP is also able to produce a "round-up" list of people to detain immediately. Some of the people flagged are "detained and sent to extralegal 'political education centers' where they are held indefinitely without charge or trial, and can be subject to abuse" (Human Rights Watch, 2018).

3.5. Algorithms can be used to facilitate good practices

Police contact is associated with negative mental and physical health consequences (Sewell & Jefferson, 2016). PredPol, which claims to "help protect one out of every 33 people in the United States" (PredPol, 2020), is still subject to a positive feedback loop that increases police contact in targeted areas. The "heat list" model, which explicitly targets individuals, was shown to steeply increase police contact with those individuals (Saunders et al., 2016). Communities that already experience high levels of police contact will experience even more contact in areas that employ either model. Increased policing will likely contribute to the further deterioration of community-police relations, and help to perpetuate a cycle of violence, poverty, and crime.

While predictive policing algorithms are currently employed to inform policing, it is possible to use the same algorithms for community healing and restorative justice. These algorithms reveal bias: we can use them to identify communities that are likely to have broken relationships with law enforcement. Police can wield predictive policing algorithms to confirm their bias and decide where to focus patrol and arrests, or they can harness those same algorithms to face their bias and decide where to focus outreach, mediation, and social service referrals. These algorithms are already deeply embedded in global policing systems; rather than work to patch them up or decommission them completely, the path of least resistance and greatest efficacy is to re-purpose them for methods that can repair the harm of dehumanizing police practices (Marshall, 1999). Until we see movement towards restorative justice, predictive policing, and policing in general, will continue to plague communities in need.

4. RISK ASSESSMENT

After someone is arrested, a judge determines the conditions of that person's release. Typically, the judge makes some kind of "risk assessment" to determine how likely the defendant is to commit more crimes. These assessments can influence bail, sentencing, and parole. While a judge gets the final say, risk estimates have been increasingly performed by machine learning algorithms.

Similarly to how predictive policing algorithms run on biased arrest data, risk assessment algorithms run on biased arrest data *and* biased judicial data. The data used in risk assessment, however, is uniquely biased by selective outcome representation; if the defendant in the input data set was not released on bail, there is no way to determine whether or not they would have committed an offense if they had been released. This would suggest that if a particular group was disproportionality arrested and detained, or detained for inconsistent reasons, the algorithm would be more unpredictable and less accurate for that group.

4.1. A case study of selective unpredictability

An evaluation of the Virginia Pretrial Risk Assessment Instrument (VPRAI) found evidence of the algorithm's unpredictability for People of Color (Danner et al., 2016). While they did not find race to be a statistically significant predictor of risk, they did find a statistically significant difference in the predictive ability of the algorithm based on race, "with the model performing better for Whites" (p. 8). The authors attributed this disparity in part to the inclusion of risk factors that could "over-classify the risk" for People of Color, and found that if they "weighted, summed, and collapsed [the risk factors] into risk levels, the difference…is no longer statistically significant" (p. 8). It is unclear what specific operations they performed to "collapse" the risk factors, and whether or not these mitigations are used in practice.

The study also found that weighting certain risk factors led to "overclassifying pretrial failure risk for females" (p. 9). Despite the unexplained unpredictability for marginalized groups, the authors conclude that VPRAI is race and gender neutral (p. 8-9). Similar assessments and a possible explanation for this discrepancy, which we discuss later in this section, were outlined by data scientists Johndrow and Lum (2017).

4.2. Risk assessment algorithms are unregulated and regionally inconsistent

A comprehensive review of risk assessment algorithms suggests that while algorithmic results are not free of data bias, they can be "more accurate and less biased than clinical decision making" (Goel et al., 2018, p. 2). The review cites "extremely vague" legal requirements for risk prediction testimony, which have led to traditional verdicts steeped in judicial bias. These vague requirements were upheld for risk assessment algorithms by the Wisconsin Supreme Court, which rejected the transparency concerns raised in *Wisconsin v. Loomis* (Goel et al., 2018, p. 17).

Even with added accuracy and reduced bias overall, these algorithms pose a threat to fair legal proceedings. Without transparency and regulation, we cannot know where the algorithms work and where they do not. We are left in the dark until the failures surface, at which point people have already been locked up, denied bail, and over-sentenced. The Defender Association of Philadelphia, which represents approximately 70% of the people arrested in Philadelphia, testified in Harrisburg that the Pennsylvania Sentencing Commission's algorithm correctly identifies a "risky" defendant "only 52% of the time"—hardly better than a coin toss (Defender Association of Philadelphia, 2018).

One such incorrect assessment was made for Defender Association Bail Navigator LaTonya Myers, who was incarcerated as a juvenile. Myers spoke at the hearing about being the victim of domestic violence. She described the night she stepped in to defend her mother from her mother's live-in boyfriend, who was assaulting her. Her mother's boyfriend called the police and the police took both women into custody, where they were held in separate cells. The police offered Myers a deal—they would let her and her mother go if Myers agreed to probation. Myers knew she wasn't guilty, but she took the deal anyway because she was scared and alone.

Myers then described her life in and out of the criminal justice system all the way through her twenties, when she became involved in criminal justice advocacy and defense. Even though Myers had long ceased to be of concern to parole officers and was by all human accounts a model citizen, the algorithm still classified her as a "risky" defendant.

Myers cited concerns that these algorithms encourage a judge to "overlook individual circumstances and experiences, and preclude the possibility for personal growth and rehabilitation" (Defender Association of Philadelphia, 2018).

4.3. Removing race from the input data

One proposed solution to algorithmic bias is to eliminate race as a variable. Johndrow and Lum show it is possible to do this by first identifying variables that "encode" for race, then creating a transformed set of variables that are mutually independent of the race-encoding variables (Johndrow & Lum, 2017, p. 3). For example, the new algorithm might notice that People of Color are more likely to be re-arrested for a particular crime, and would adjust re-arrest rates to correct for that bias. The model was found to somewhat equalize the predictive ability of a sample algorithm with respect to race, for an overall predictive accuracy that is close to the same as the unadjusted model (p. 16-17).

In order to justify the use of racially-independent training data, the authors argue that "the most reasonable approach is to treat all races as though they are the same with respect to recidivism" (Johndrow & Lum, 2017 p. 4). Like most risk assessment algorithms, however, the "accuracy" of the model is still defined as the algorithm's ability to predict whether a defendant will be re-arrested. The authors admit that re-arrest is a racially biased measure of criminal behavior, and that People of Color are disproportionately stopped, arrested, and incarcerated (p. 3).

A truly "accurate" risk assessment model would therefore have to predict that People of Color would be at higher risk for re-arrest—not because they are inherently prone to criminal behavior, but because they are disproportionately subjected to police attention and incarceration. An algorithm with re-arrest as its correctness metric can only measure the risk of re-arrest; it cannot measure or quantify a defendant's risk of criminal behavior. The removal of race as an input variable is a step toward colorblind data, but it cannot account for the racial bias inherent in conviction and arrest.

4.4. Rethinking risk

A better standard might be to consider the predictors of recidivism, such as income insecurity, education, mental health, and drug addiction (Makarios et al., 2010). Rather than classifying people as "high risk" for re-arrest, the algorithm might classify people as being "high risk" for unemployment, depression, or relapse. This would help judges make recommendations or choose programs for defendants that benefit them and reduce their long-term risk of recidivism.

5. MACHINE TESTIMONY

Prosecutors have become increasingly reliant on algorithms that classify forensic evidence, especially DNA, to secure convictions. Unlike more traditional methods, where an analyst might compare two forensic samples in a lab, machine learning algorithms allow analysts to compare samples against millions of other samples in a DNA database, and obtain the probabilistic estimates of various matches. Problems with traditional forensics carry over into the millions of samples in DNA databases. For instance, samples can easily become contaminated, intermixed, or decomposed, which leads to false positive matches (DiFonzo, 2005). There is also growing evidence of dishonest actors in crime labs, who have rushed testing, intentionally contaminated samples, faked results, and colluded with prosecutors to guarantee convictions (Shaer, 2016; Mettler, 2017). Rogue actors could further compromise results if they were to exploit flexibility or vulnerability in the algorithm's input parameters.

There is one problem unique to forensic algorithms that poses a grave threat to defendants' right to a fair trial. Traditionally, analysts and experts would be able to testify to each step of forensic analysis in detail. If a particular chemical reagent in a DNA experiment was called into question by the defense, an expert would be able to draw on direct knowledge of the reagent to confirm or refute concerns about its reliability. Forensic analysts that handle biological and chemical samples are understandably educated in biology and chemistry; they are not educated in machine learning, and cannot attest to the reliability of machine learning tests. Moreover, even machine learning experts cannot attest to the reliability of these tests, because the details of the algorithm are obscured behind copyrights and corporate policy. Without the source code, it is impossible for defendants to hear, understand, and question the evidence against them.

5.1. Crime labs are a mess

The scope and volume of problems with crime labs are well summarized in "The Crimes of Crime Labs" (DiFonzo, 2005). While forensic analysis works well under perfect conditions, there are a myriad of real-life conditions that hinder correct analysis. For example, the sample must be adequately sized, isolated, collected, and maintained. This is difficult to achieve in practice due to the mixing of evidence at crime scenes. Once a sufficient sample is obtained, DiFonzo describes the analysis itself as "slapdash…often performed by untrained, underpaid, overworked forensic technicians" (DiFonzo, 2005, p. 2). He cites a lack of oversight on education, certification, and lab accreditation. Mishandling and incorrect classification of historical samples undoubtedly influences the accuracy of any machine learning algorithm trained on those samples.

Even when technicians are educated and proficient, crime labs as institutions are often closely associated with police departments and prosecutors. They have a history of dishonesty and corruption, from faking the results of drug tests (Mettler, 2017) to compensating labs for DNA analysis that ends in conviction (Shaer, 2016). These bad-faith convictions would have an even worse result on training data: whereas the honest mishandling of evidence might cause the algorithm to behave erratically, the dishonest mishandling of evidence could bias the algorithm towards false positives, continuing the cycle of unjust conviction.

DiFonzo also highlights long-standing issues with the gross misrepresentation of statistical evidence in court, citing in particular an example in which the prosecution claimed the probability of a match between crime scene DNA and DNA from database subjects was one in 694,000, when in reality it was independently determined to be one in eight (DiFonzo, 2005, p. 5). This false claim and many others led to false convictions, and the crime lab responsible was subsequently shut down. There have been hundreds of similar cases, including "'perjury by expert witnesses, faked laboratory reports, and testimony based on unproven techniques'" (p.

5). The public is largely unaware of the unreliability of DNA testing and analysis, which becomes crucially important in cases where a DNA match is the only piece of accusatory evidence.

5.2. Forensic algorithms are a mess inside of a black box

The public is even less aware of the reliability of algorithms that perform forensic analysis, and the results are easier to obfuscate. In the United States, the use of closed-source software in criminal trials potentially violates several laws. One such law, the Confrontation Clause of the Sixth Amendment of the Constitution, guarantees defendants the right to "be confronted with the witnesses against him" (U.S. Const. amend. VI). In support of the defendant in *California v. Johnson*, attorneys at the American Civil Liberties Union (ACLU) argued that the "witness" to accusatory DNA evidence included the designers of DNA analysis software TrueAllele, the system's programmers, and the code itself (Kaufman et al., 2017, p. 21). Failure to produce the code—the specific procedure by which the DNA was matched—was therefore failure to produce a complete witness for the defense to confront.

Black-box witness testimony calls into question the overall "fairness" of the trial. The Due Process clause of the Fourteenth Amendment has been interpreted to include the defendant's right to perform "adversarial testing" on any evidence—that is, both sides should be able to examine the evidence and reach an independent conclusion (Kaufman et al., 2017, p. 21). In the case of an algorithm, adversarial testing might include tweaking input data and parameters to reveal bias or inconsistency, verifying the legitimacy of each function, and documenting bugs or vulnerabilities. All of these tests are impossible to perform without the code itself.

5.3. Criminal justice proceedings should not be hidden from public scrutiny

The general public has a right to "observe and evaluate the workings of the criminal justice system" (Kaufman et al., 2017, p. 21). In the United States, this is guaranteed by the First Amendment, which includes the right of the public to "petition the government for a redress of grievances" (U.S. Const. amend. I). Logically and according to Constitutional interpretation, this gives us the right to identify and confront our grievances before we suggest solutions. In a stand-alone paragraph, ACLU attorneys summarize the legitimate demand for civil rights in the digital age: "Algorithms used to produce evidence introduced to prove the guilt of a criminal defendant fit well within the broad reach of the First Amendment right of access" (Kaufman et al., 2017, p. 32). Many countries have laws that require the transparency of criminal proceedings. Until source code access is granted, people all over the world will be subject to illegal black-box convictions.

Regardless of legal basis, let us briefly consider the "fairness" of a trial that allows black-box evidence. One party, the government, has a great deal of power and resources, including this mysterious black box. The other party, the defendant, has relatively little power, few resources, and no way to understand what is in the government's secret box. The government pulls a name out of the box and declares the defendant guilty, the punishment for which can include death. Government prosecutors and Silicon Valley giants would have us believe this trial is "fair", and that the conviction is reasonable without a shadow of a doubt. In actuality, this black-box standard of evidence echoes courtrooms of the Dark Ages.

6. CONCLUSION

As we continue to integrate "intelligent" machines into society, it is worth considering what it is we actually want from these machines. Algorithms are currently employed to aggressively digest and maintain the status quo. Despite what proponents say, they do not offer any real solutions to the underlying problems of systemic bias and corruption. While the algorithms themselves are similarly unlikely to worsen bias, they create a facade of objectivity and fairness that discourages people from facing reality. Still, we believe it is possible to use these algorithms to change society for the better. We propose the following changes to the current use of machine learning algorithms in the criminal justice system.

1. **Use machine learning as a tool to understand systemic bias.** Instead of striving to remove bias in data, effectively obscuring and ignoring the root causes, we could use it to better understand the root causes. We will not be able to effectively address systemic racism until we can perform an intersectional analysis of where and how it occurs. Machine learning can help us do that.

2. **Shift the focus of implementation of machine learning results from punitive to restorative practices.** Once we determine where the bias is, we can start to address it with practices that heal, rather than practices that divide.

3. **Law enforcement, crime labs, and courtrooms should create positions for people who understand machine learning.** We need experts to help determine the accuracy and reliability of the algorithms currently used throughout the criminal justice system. These experts could also curate input data, testify clearly about the inner-workings of the algorithms in court, and act as a liaison between criminal justice offices and tech companies.

4. **Machine learning and data science researchers and developers should take responsibility for the impact of their creations.** Research and development is advancing at an unprecedented rate. It is up to the technical experts to help explain and evaluate the impact of new technology on human systems. This help could include application testing and analysis, research collaboration with legal experts, and recommendations for government policy.

5. **Government officials should regulate the use of artificial intelligence in the criminal justice setting.** We cannot count on tech companies to regulate themselves. We need transparency and government oversight in order to fully understand what is happening inside of these algorithms. This is necessary for everyone, but especially for defendants, who are being arrested, convicted, and sentenced to prison based on evidence they can neither see nor contend with.

By collectively shifting our frame of reference on machine learning in criminal justice applications, we can work toward addressing the problem of systemic injustice rather than perpetuating or ignoring it. Each stakeholder has an important role, and it will take all of us to create a better future.

ACKNOWLEDGMENTS

Thank you to Vera Eidelman, Andrea Woods, and Daniel Kahn Gillmor for encouraging this work and providing excellent ideas and expertise. You taught me a lot. Thank you to the Speech, Privacy, and Technology Project at the ACLU for the partial funding of this work in 2017. Thank you to Anna Lysyanskaya for recommending me for the ACLU program. Thank you to William Fleischman for encouraging me to write this paper, and for helping me with the citations. And finally, a big thank you to all of my family for your feedback and suggestions.

REFERENCES

Alexander, M. (2010). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York, NY: The New Press.

Danner, M., VanNostrand, M. & Spruance, L. (2016). Race and gender neutral pretrial risk assessment, release recommendations, and supervision: VPRAI and PRAXIS revised. *Luminosity, Inc.* Retrieved from https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/race-and-gender-neutral-pretrial-risk-assessment-release-recommendations-and-supervision.pdf

Defender Association of Philadelphia (2018). Press Release: Risk Assessment. Retrieved from https://www.philadefender.org/risk-assessment/

DiFonzo, J. (2005). The Crimes of Crime Labs. *Hofstra Law Review*, 34(1) 1-12. Retrieved from https://scholarlycommons.law.hofstra.edu/cgi/viewcontent.cgi?article=2372&context=hlr

Ensign, E., Friedler, S., Neville, S. Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency*, 81, 1-12. Retrieved from https://arxiv.org/pdf/1706.09847.pdf

Goel, S., Shroff, R., Skeem, J, & Slobogin, C. (2018). The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment. *Social Science Research Network* [SSRN]. Retrieved from https://ssrn.com/abstract=3306723

Human Rights Watch (2018). China: Big Data Fuels Crackdown in Minority Region. Retrieved from https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region

Johndrow, L. & Lum, K. (2017). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1), 189-220. Retrieved from https://arxiv.org/pdf/1703.04957.pdf

Kaufman, B., Buskey, B., Goodman, R., Eidelman, V., Woods, A., & Bibring, P. (2017). Brief of *Amici Curiae* In support of Defendant. *American Civil Liberties Union* [ACLU], Case No. F071640. Retrieved from https://www.aclu.org/sites/default/files/field_document/2017-09-14_billy-ray-johnson_amicus-full_accepted.pdf

Langan, P. (1995). *The Racial Disparity in U.S. Drug Arrests.* Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice. Retrieved from https://bjs.gov/content/pub/pdf/rdusda.pdf

Lum, K. & Isaac, W. (2016). To predict and serve? *Significance*, vol. 13, no. 5, October 2016. http://doi.org/10.1111/j.1740-9713.2016.00960.x

Makarios, M., Steiner, B., & Travis III, L. (2010). Examining the Predictors of Recidivism Among Men and Women Released from Prison in Ohio. *Criminal Justice and Behavior*, 37(12), 1377-1391. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0093854810382876

Marshall, T. (1999). *Restorative Justice: An Overview.* London: Home Office. Retrieved from http://www.antoniocasella.eu/restorative/Marshall_1999-b.pdf

Mettler, K. (2017). How a lab chemist went from 'superwoman' to disgraced saboteur of more than 20,000 drug cases. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2017/04/21/how-a-lab-chemist-went-from-superwoman-to-disgraced-saboteur-of-more-than-20000-drug-cases

Mohler, G., Short, M., Malinowski, S., Johnson, M., Tita, G., Bertozzi, A., & Brantingham, P. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1-12. Retrieved from https://escholarship.org/content/qt1br4975j/qt1br4975j.pdf

National Institute of Justice [NIJ] (2014). *Predictive Policing*. Washington, DC: National Institute of Justice. Retrieved from https://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/Pages/welcome.aspx

PredPol (2020). Overview. Retrieved from https://www.predpol.com/about/

PredPol (2020). Proven Crime Reduction Results. Retrieved from https://www.predpol.com/results/

Puente, M. (2019). LAPD to scrap some crime data programs after criticism. *Los Angeles Times.* Retrieved from https://www.latimes.com/local/lanow/la-me-lapd-predictive-policing-big-data-20190405-story.html

Russell, S. & Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*. Prentice-Hall. Retrieved from https://pdfs.semanticscholar.org/bef0/731f247a1d01c9e0ff52f2412007c143899d.pdf

Saunders, J., Hunt, P., & Hollywood, J. (2016). Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology*, 12(3), 347-371. Retrieved from https://link.springer.com/article/10.1007/s11292-016-9272-0

Sewell, A. & Jefferson, K. (2016). Collateral Damage: The Health Effects of Invasive Police Encounters in New York City. *Journal of Urban Health*, 93(1), 42-67. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4824697/

Shaer, M. (2016). The false promise of DNA testing. *The Atlantic*, June 2016. Retrieved from https://www.theatlantic.com/magazine/archive/2016/06/a-reasonable-doubt/

U.S. Const. amend. I, VI. Retrieved from https://www.law.cornell.edu/constitution/index.html

CAPTURING THE TRAP IN THE SEEMINGLY FREE: CINEMA AND THE DECEPTIVE MACHINATIONS OF SURVEILLANCE CAPITALISM

Fareed Ben-Youssef, Kiyoshi Murata, Andrew Adams

Texas Tech University (USA), Meiji University, Centre for Business Information Ethics (Japan), Meiji University, Centre for Business Information Ethics (Japan)

fbenyous@ttu.edu; kmurata@meiji.ac.jp; aaa@meiji.ac.jp

ABSTRACT

Shoshana Zuboff's concept of the Big Other offers a means to understand the paradigm shifts provoked by surveillance capitalism—by social networking systems collecting user data with little government oversight or end user understanding. The Big Other's inescapable annihilating power comes in part for how it evades legibility. In her analysis, Zuboff does not fully historicize the Big Other's rise, only broadly comparing its logic of total conquest to that of former imperial powers. Our paper disrupts the pervasive illegibility of the Big Other using three key examples in global cinema. In the process, we productively fill in historical blind spots in Zuboff's framework.

To underline the new subjugations of the Big Other, our interdisciplinary paper traces the line between what constitutes just and the unjust surveillance within business. Our examples feature enthused surveillance capitalists as well as confused, even terrified end users. We end by framing the historical roots of such an unquestioned form of mass surveillance by situating studies about the role of bureaucracies and Big Data in the Holocaust. Our comparison illustrates the troubling consequences of the Big Other's emergence: to be reduced to data is to accept the possibility of being deleted.

KEYWORDS: surveillance capitalism, cinema, big other, control, the holocaust, banality of evil.

1. INTRODUCTION

Shoshana Zuboff's concept of the Big Other offers a means to understand the paradigm shifts provoked by surveillance capitalism—by social networking systems collecting user data with little government oversight or end user understanding. For Zuboff, the Big Other represents "an intelligent world-spanning organism" which brings with it "new possibilities of subjugation... as this innovative institutional logic thrives on unexpected and illegible mechanisms of extraction and control that exile persons from their own behavior" (Zuboff, 2015: 85). The Big Other's inescapable annihilating power comes in part for how it evades legibility. In her analysis, Zuboff does not fully historicize the Big Other's rise, only broadly comparing its logic of total conquest to that of former imperial powers. Our paper disrupts the pervasive illegibility of the Big Other using three key examples in global cinema. In the process, we productively fill in historical blind spots in Zuboff's framework.

To underline the new subjugations of the Big Other, our interdisciplinary paper traces the line between what constitutes just and the unjust surveillance within business. Our examples feature enthused surveillance capitalists as well as confused, even terrified end users. We end by framing the historical roots of such an unquestioned form of mass surveillance by situating studies about the role of bureaucracies and Big Data in the Holocaust against Quentin Tarantino's WWII film Inglourious Basterds (2009). Our comparison illustrates the troubling consequences of the Big Other's emergence: to be reduced to data is to accept the possibility of being deleted. Cinema, we will ultimately show, is especially well-primed to visualize the trap in such seemingly free services, to make visible the often-invisible machinations of surveillance capitalism.

Our study employs a methodology which combines theories from social science with humanities-style close reading. Our framework, first developed in a study on the real-life security lessons in superhero media recently published in *Security Journal*, allows us to situate the formal construction of these media texts within an array of aesthetic and political contexts (Adams et al., 2019). Texts that may initially seem distant from debates surrounding surveillance capitalism gain vital relevance through such a prism, permitting for an expansion of the canon of surveillance-oriented cinema. Such highly textured viewing allows us to fully sus out of the contradictions and tensions in these cinematic examples. In so doing, we show how these films are not simple entertainment; rather, they frame a contradiction—the allures of unregulated surveillant power as well as the root horror of its dehumanizing potential.

2. *THE CIRCLE* – THE ATTRACTION OF SURVEILLANCE CAPITALISM

We begin our analysis by exploring the attraction of surveillance capitalism for the business world as expressed in the film adaptation of *The Circle* (2017) directed by James Ponsoldt and written by the novel's author Dave Eggers. The work focuses on the growing disillusionment of a young employee who begins a job at a Google-proxy known as The Circle. Her first encounter with the company's Steve Jobs-like executive, Eamon Bailey (played by Tom Hanks, a knowing perversion of his benign screen persona for the villainous role), centers on the introduction a line of hidden cameras that is sold with the slogan: "Knowing is good, knowing everything is better!" The way his words resonate with the (laboring) masses are made clear by how his employees leap out of their chairs to catch the cameras that he casually tosses out. This kind of vision is desired, worth fighting over.

Even as the film captures the laudatory Silicon Valley rhetoric around such practices, it also winks at data mining's costs. The executives' admission that he "stuck [the camera] near the dunes. No permit, nothing," evokes the unregulated reality that many such businesses operate. Eggers' original novel places such invasive behavior in an even more personal sphere when the exec confesses to clandestinely placing the cameras in his mother's home, joking "Forgive me! Forgive me! I had no choice. She wouldn't have let me do it otherwise. So I snuck in, and I installed cameras in every room. They're so small she'll never notice" (Eggers, 2013: 68). Even the strictures of family bend to the whims of a capitalist – even a mother's eye cannot regulate or control him.

In both novel and film, Bailey goes on to frame such pervasive sight as a resistant political tool for accountability, noting how even human rights discourses can be co-opted by Zuboff's 'world-spanning organism.' He proposes a kind of just surveillance tied to accountability, even as he runs a business that operates with impunity. The visual juxtapositions achievable in the cinematic medium permit the screen adaptation to sharply undermine the businessman's closing statement: "We will see and hear everything. If it happens, we'll know." While the executive heaps praise upon his company's total technological vision, the screen behind him shows footage of riot in a city. Fire extends across the frame, and the Silicon Valley executive seems to be at the urban inferno's center. Whistles from the recording lightly punctuate his dialogue, so that the film winks to its attempts to figuratively blow the whistle on such business practices. Moments after we see the city in flames the image fades out to a white background emblazoned by the initiative's name "Seechange." The crowd cheers – no one appears to see the true sea changes at work in conceptions of user privacy. With such a tension between the executive's valorizing words and the stark imagery of destruction that is ultimately effaced, the satirical film gestures to the unseen devastation of such unquestioned surveillance while signaling how difficult it is to see beyond the gloss of dataveillance technology.

3. *PULSE* – THE TERRIFYING DECEPTIONS ON THE END USER

Our second reading shows how cinema has represented how end users are deceived by technology corporations. Like the Spanish conquistadors before them, Zuboff argues that early surveillance capitalists, "relied on misdirection and rhetorical camouflage, with secret declarations that we could neither understand nor contest" (qtd. in Naughton, 2019). Kiyoshi Kurosawa's horror film *Pulse* (2001) features a haunted internet browser which serves as a metaphor for unscrupulous, intrusive corporations. The crude English pun inherent its name, UR@NUS or 'your anus,' implies that these entities seek to penetrate their users.

One scene portrays the user who must agree to the browser's Terms of Use Agreement. It is worth reiterating that the man believes he is downloading a standard browser and remains oblivious to its malignant supernatural intents. The film highlights his utterly naive status by showing him with an instruction manual for the Internet then searching for his modem jack before slotting in the installation CD. Before they begin their terrorizing, as with the Spanish conquistadors and the early surveillance capitalists that Zuboff references, the ghosts behind the UR@NUS browser force the unsuspecting user to agree to a contract that he can neither understand nor contest. A popup message appears that declares "Have Fun!" metaphorically suggesting how corporations begin to haunt the consumer via the promise of entertainment. Upon encountering the agreement, he asks, "What is this crap? Yes, I agree.". He then quickly clicks through it. The film adopts a point-of-view shot of the monitor from the user's perspective for the installation sequence, creating a visual monotony that evokes the boredom and confused reaction of the end user. It illustrates how corporations trick the consumer with the prospect of fun and overwhelm him with impenetrable legal language.

After consenting to the browser's terms, he faces the horror of surveillance capitalism. We see and hear the UR@NUS suddenly dial-up. There a brief shot where we see the user reacting with an unsettled expression as the computer seems to be acting on its own accord. The monitor's image suddenly fills the screen. The browser takes hold of the film's form indicative of its dominating force. The user clicks through a slide show of eerie Web-Cam videos, featuring anonymous persons sitting in spaces drained of color and filled with deep shadow. Several figures stare out at the camera. The user is now watched. Then, the screen fades to black as the question appears: "Do you want to meet a ghost?" Only after agreeing to the terms does the true nature of the accord come to life. Troubled, the character shuts off the monitor, but he will not be able to shake off its ghostly hold.

We learn the ghosts, otherwise described as the shade, will ultimately desiccate users. Resembling a morbid variant of Japan's Hikikomori, youth who isolate themselves, those subjected to the interference of the browser withdraw from real life and connect only to its shade. Notably, the shade draws on real life connections to pass on its haunting infection. What remains after the shade has sucked the life out of people is their "data shadow". The body (or corpse) "digitizes" and (sometimes) leaves a physical shadow behind, like a stain on the walls. Those being drawn in connect with the shadow of the person and are infected by the shade. The apocalyptic dimension of the horror film, its world essentially ends, functions as a kind of prescient allegory for the ways SNS systems like Facebook create "shadow profiles" where data is collected even on non-users. Director Kiyoshi Kurosawa has defined horror films as "that family of films that take as their subject matter the fear that follows one throughout one's life" (Kurosawa). For our purposes, it useful to note that he creates a horror representation of the "Big Other," one that follows user *and* non-user to the ends of the earth. No one is safe from its grasp, even those who resist the siren call to 'have fun' and who choose not to install its applications on their computers. While *Pulse* begins to gesture to the violence of the Big Other, Tarantino's *Inglourious Basterds* more fully reveals its insidious destructive potential.

4. *INGLOURIOUS BASTERDS* – UNCOVERING THE BIG OTHER IN THE HOLOCAUST

Zuboff, as mentioned, sees the historical echoes of surveillance capitalism in the violence of Spanish colonialism. Through this paper's final reading of Tarantino's *Inglourious Basterds*, we reveal how corporations' purposely occluded corrosive intention needs to be traced back to Big Data's not only colonial but genocidal beginnings during the Holocaust. At first glance, the war film seems very distant from AI or information ethics concerns. However, close analysis of the film permits us to historicize the development of the Big Other within Big Data's imbrication in the Holocaust.

IBM's complicity in the genocide has been well-documented. Their technology was used in everything from processes of identification to the management of German railway lines (Dillard, 2003: 5). Jessie F. Dillard notes that the actions of such data companies in the genocide reveals "the ability of ideology to control and destroy through the professionals who develop, implement and service technology" (Dilliard, 2003: 3). They are not morally blameless for the technology they develop. Dillard argues that IBM pursued profit without any consideration or care for the ways their technology would be deployed or "the ultimate social or human consequences." (Dillard, 2003: 5). His emphasis concerning IBM's privileging of profit above all else recalls the aforementioned scene of *The Circle* wherein images of obliteration were projected behind the smiling surveillance capitalist.

For her part, Zuboff does not situate the Big Other within such histories. Zuboff employs Karl Polanyi's idea of 'commodity fiction' where people are subordinated to the market. Polanyi noted that such fictions "disregarded the fact that leaving the fate of soil and people to the market would be tantamount to annihilating them" (qtd. in Zuboff, 2015: 83). Zuboff continues, "in the logic of surveillance capitalism there are no individuals, only the world spanning organism and all the tiniest elements within it" (Zuboff, 2015: 83). Here, while ignoring any specific examples from the past, Zuboff points to an over-riding logic of annihilation where users are reduced into the Big Other's tiniest elements. Her language acts as a starting point for our analysis of how Tarantino allegorically depicts historian Raul Hilberg's Bureaucratic Process of Destruction. By culminating upon a film about the Holocaust, we show how being a good businessman under surveillance capitalism may have the same ethical standing as a good bureaucrat in a destructive surveillance state.

Raul Hilberg has argued about the importance of bureaucracy within Nazi genocide noting, "at first sight the destruction of the Jews may have the appearance of an… impenetrable event. Upon closer observation it is revealed to be a process of sequential steps that were taken at the initiative of countless decision makers in a far-flung bureaucratic machine" (Hilberg, 1985: 53)." In the opening scene in *Inglourious Basterds*, a Nazi colonel named Hans Landa transforms a French farmhouse into an office when searching for hiding Jews, closing the windows and carefully placing his bureaucratic tools on the dinner table (Ben-Youssef, 2017. 819-820).[1] He carefully lays out his papers and fills his fountain pen. The sequence enacts the three steps of Hilberg's Bureaucratic process – Definition, Concentration, Annihilation. First comes identification where the targeted people are bureaucratically identified on paper as Jewish (Hilberg, 1985: 18). A tight cut-in on his pen as he checks off the name of each Jewish person emphasizes the sharp point of his pen, its wield the power to cut those it checks off. Second comes concentration, where the targeted group is trapped in a ghetto and is controlled "through the watchful eyes of the entire German population" (qtd. in Hilberg, 1985: 50). Landa reveals his all-seeing eye when noting that he knows exactly where the Jews are hidden. Without blinking, he asks, "They're hiding under the floorboards, aren't they?" Finally, the third step is annihilation. As Hilberg sums up, "Most bureaucrats composed memoranda, drew up blueprints, signed correspondence…they could destroy whole people by sitting at their desks" (Hilberg, 1985: 288). In the film, the Jewish populace is executed by the soldiers. To be visible is to be capable of being part of the destructive process. The film thus illustrates the stakes when surveillance, be it on the state level or that of private enterprise, goes unchecked and unmanaged. It pushes us to reflect upon what banality of evil both the distant bureaucrat and the algorithm might share.

It is worth examining, for a moment, the film's vital play with language which recalls the ignorance of the consumer in the film, *Pulse*. Landa sets up the execution of the hiding Jewish family by speaking in English so that his French victims cannot comprehend. The film metaphorically indicates that the bureaucratic system driving the murder remains similarly incomprehensible. After hearing a confession of the Jewish family's whereabouts, the colonel says, "I am going to switch back to French, and I want you to follow my masquerade. Is that clear?" He says "Adieu" while directing his soldiers to fire their guns. Those affected by the system's violence cannot understand the true meaning behind the language of bureaucracy. They cannot make sense of the possibilities of subjugation that define mass surveillance by the state or in more unfettered forms of surveillance capitalism.

5. CONCLUSION

Our analysis of these key examples of world cinema confirm that the root danger of surveillance capitalism is that most consumers do not realize the trap they fall in—they do not realize their

[1] This is an expanded reading of *Inglourious Basterds* which appears in Ben-Youssef's earlier article, "'Attendez la Crème!': Food and Cultural Trauma in Quentin Tarantino's *Inglourious Basterds* and *Django Unchained*."

status as data to be deleted. Like the workers in *The Circle* who scrabble to catch a new camera tossed out by an apathetic executive, consumers desire to be in such a web. As *Pulse* reminds us, they often just sign away their rights and agree to be haunted by corporations-to be reduced to phantoms that are mined for profit. The consequences of this kind of uncontrolled surveillance when humans are turned into data in an incomprehensible system, where the state takes on the categorizing logic of an IBM machine, has been powerfully visualized onscreen by Quentin Tarantino. Zuboff notes that "We were caught off guard by surveillance capitalism because there was no way that we could have imagined its action, any more than the early peoples of the Caribbean could have foreseen the rivers of blood that would flow from their hospitality toward the sailors who appeared out of thin air waving the banner of the Spanish monarchs. Like the Caribbean people, we faced something truly unprecedented" (qtd in Naughton, 2019). Her own violent imagery recalling a colonial past reminds us that it was not 'truly unprecedented' – business and colonialism share the same carnivorous appetites, a point only confirmed when lingering on Big Business and Big Data's complicity in the Holocaust. If people can be reduced to data, they can be both controlled and eliminated.

Our analysis renders newly legible the Big Other's illegible processes, highlighting how cinema can frame the allure and costs of such control. In so doing, these key films show how media drives home the paradigm shift of surveillance capitalism and unveil its under-explored history. We have moved from a substrate of mediated relations, a village society wherein the state had limited and exceptional access to encoded information, to a new stratum of communication with the emergence of social networks. Now we have platforms that can see everything, an unregulated entity that can access all. Cinema tracks these shifts and the ensuing danger when businesses follow mantras like: "Knowing is good, knowing everything is better!" In so doing, these films demand scholarly attention for how they offer the public a viscerally affecting and disruptive critical understanding: they permit viewers to see how our rights of privacy come to burn up in the light of seemingly free social networks. These films ultimately give scholars of surveillance, AI, and information ethics a new language to map out surveillance capitalism's trap.

REFERENCES

Adams, Andrew A., Ben-Youssef, Fareed, Schneier, Bruce and Murata, Kiyoshi (2019). "Superheroes on screen: Real life lessons for security debates." *Security Journal*. https://doi.org/10.1057/s41284-019-00193-7

Ben-Youssef. Fareed (2017) "'Attendez la Crème!': Food and Cultural Trauma in Quentin Tarantino's *Inglourious Basterds* and *Django Unchained*." *The Journal of Popular Culture*. 50(4). 814-834. https://doi.org/10.1111/jpcu.12578

The Circle. (2017). [Film] *Directed by James Ponsoldt*. US: STX Films.

Dillard, Jessie F. (2003) "Professional services, IBM, and the Holocaust." Journal *of Information Systems*, 17(2), 1-16. https://doi.org/10.2308/jis.2003.17.2.1

Eggers, David (2013). *The Circle*. NY: Vintage Books.

Hilberg, Raul (1985). *The Destruction of the European Jews: Student Edition*. NY: Holmes & Meier.

Inglourious Basterds. (2009). [Film] *Directed by Quentin Tarantino*. US: The Weinstein Co.

Kurosawa, Kiyoshi. "What is Horror Cinema?" Translated by Kendall Heitzman. *Kurosawa Kiyoshi*.

https://ceas.yale.edu/sites/default/files/files/events/past/20060323kinemaclubwksp_kurosawa.pdf. Accessed 15 April 2020.

Naughton, John (2019) "'The goal is to automate us': welcome to the age of surveillance capitalism." *The Guardian*. 20 Jan. 2019. Web. Accessed 10 July 2019.

Pulse. (2001). [Film] Directed by Kiyoshi Kurosawa. Japan: Toho.

Zuboff, Shoshanna (2015). "Big other: surveillance capitalism and the prospects of an information civilization." *Journal of Information Technology*, 30, 75-89. https://doi.org/10.1057/jit.2015.5

DIFFERENCES IN HUMAN AND AI MEMORY FOR MEMORIZATION, RECALL, AND SELECTIVE FORGETTING

Sachiko Yanagihara, Hiroshi Koga

University of Toyama (Japan), Kansai University (Japan)

sachiko@eco.u-toyama.ac.jp; koga@res.kutc.kansai-u.ac.jp

ABSTRACT

IT, including AI, has exceeded the humans ability in several areas, including memorization. Though many people aspire to be able to memorize everything recall it at will, humans have the advantage of intentionally forgetting things while learning new information. For people working in organizations, physical activity complemented by human intelligence is an advantage that humans have over AI. In this study, we examine the mechanisms of repeated memorization and selective forgetting of memory through the game of competitive karuta, which is a traditional Japanese card game. First, we outline the general differences between memories of humans and those of IT artifacts, including AI, confirming previous research of memorization and forgetting. Next, we outline competitive karuta and examine human memory in that context. Finally, we consider mechanical differences of memory between AI, which can remember everything, and human beings, who forget, in reference to the “intentional forgetting” of Golding and Macleod (1998). Moreover, we suggest that "the right to be forgotten" is a necessary perspective to consider regarding the use of AI for information, the substitution of people with androids, and the cyborgization of humans to provide the capabilities of AI.

KEYWORDS: intentional forgetting, memory, competitive karuta, artificial intelligence.

1. INTRODUCTION

Artificial intelligence (AI) mimics the learning behavior of humans; however, there are significant differences in the capabilities of AI and humans. Tasks exemplifying these differences are more appropriate for humans to perform. One of these significant differences is the capability of AI to store and use all information, in contrast to the imperfect memory of humans. Humans have struggled with the problem of forgetting, and most people want their memory to be reliable. Ricoeur (2000, p.197) commented on the consistency of this problem by stating that “In this regard memory defines itself, at least in the first instance, as a struggle against forgetting.” The pervasiveness of forgetting leads humans to fear their memory failing over time. Therefore, IT (including AI) is often relied upon for its superior capabilities of memorization. However, there is a concern that this increased reliance on AI will result in jobs currently performed by humans being lost to AI replacements.

This concerned is already reflected in board games. An AI specialized for Go (AlphaGo, produced by DeepMind) and an AI specialized for Chess (Deep Blue, produced by IBM) have both defeated human champions. Facial recognition systems using AI have become increasingly sophisticated, and can find target faces among a crowd. AI systems never forget data and information after its initial memorization. While the persistent memory of AI can be considered valuable, it can also be considered to threaten “the right to be forgotten.” The imperfect memory of humans and their ability to forget has protected humans from the ethical threat of a perfect memory, considering the capacity of humans to be judgmental. Therefore, we consider a perfect memory to be inappropriate as an equivalent for human memory in a future society. Moreover, the complexity of a human's ability to intentionally memorize and forget has significant value in particular situations which AI cannot currently fulfill.

In this study, we examine the mechanisms of repeated memorization and selective forgetting through the game of competitive karuta, which is a traditional Japanese card game. First, we outline the general differences between human memory and memory of IT artifacts including AI, confirming previous research of memorization and forgetting. For the purposes of this paper, we refer to IT artifacts and AI simply as “IT,” when used collectively. Next, we outline competitive karuta and examine human memory in that context. Finally, we consider the mechanical differences of the perfect memory of AI, and that of human beings, in reference to the “intentional forgetting” of Golding & Macleod (1998).

2. MEMORIZING AND FORGETTING

2.1. Human forgetting

As stated by Gorry (2016), “Who would not appreciate an improved memory? Misremembering or forgetting frustrates all of us to one degree or another.” Completely accurate memorization was called “total recall” by Bell and Gemmell (2009), and total recall is generally viewed positively. However, the value of total recall in reality is contested. For example, Murata and Orito (2011) asked, “Is the society composed of people who have perfect (at least semantic and episodic) memory a good society?”

In humans, symptoms such as temporary amnesia are common regardless of age. People also often have experiences that they find difficult to remember even when they try, but that are then remembered later. This is similar to the case of IT when it is temporarily unable to access an index for information. However, human memory is not as simple as IT, and does not always remember by the same mechanism. Some things can be remembered by trying, and some memories are formed without trying. On the other hand, sometimes memories are formed and recalled even when you would rather not remember them. For example, a person who finds a memory to be unbearably painful can maintain his or her mental stability by partially or completely forgetting that memory. While such events could be considered unfortunate, and the person losing that memory may appear to be unhappy, it actually feels worse for that person to remember those memories. Still, we usually have negative feelings regarding memory loss. Ricoeur (2000) said “Forgetting is experienced as an attack on the reliability of memory.” Having a good working memory is valued in our society, and jobs generally require us to know many things. Humans express concern about forgetting as a potential sign of amnesia or dementia. However, forgetting need not be a negative phenomenon entirely. In this case, "forgetting" is not consciously performed as an intentional act, but is instead performed as a natural biological function.

Such forgetting mechanisms are generally considered important and necessary. The gradual dementia of the elderly is also considered in some ways to be an important mechanism for reducing the mental burden of dying humans. This mechanism of memory and forgetting can be explained through neurological functions such as neuronal mechanisms (Yasuda et. al, 2018) and neurotransmitters (Inoue et al., 2013). However, studies on how humans subjectively perceive the mechanisms of memory have been limited to philosophical perspectives, such as those of Ricoeur and others, and there is no study that objectively demonstrates the actual state of memory and forgetting.

2.2. Differentiating the concepts of unlearning and forgetting

The concept of "unlearning" is similar to that of forgetting. Unlearning is mainly discussed from the perspective of organizational learning in the field of business organization. In business administration, it is often said that it is necessary to manage with consideration for past learning. However, in the case of "unlearning" in a new business, there are situations where past learning is rejected or ignored to allow for new learning.

Klammer and Gueldenberg (2019) reviewed 63 papers dealing with unlearning or/and forgetting, and they clearly distinguished unlearning and forgetting. They regard unlearning as "a purposefully initiated process of discarding knowledge" and forgetting as "involuntary knowledge loss." They also state that "especially in comparison to organizational learning, the concepts of unlearning and forgetting have received very little attention in scholarly research."

Though many papers focus on either unlearning or forgetting, from our research, only eight works emphasize both unlearning and forgetting. Among those that discuss both unlearning and forgetting, Wensley and Navarro (2015) provided one of the few papers focused on individuals, rather than on organizations. They defined forgetting as "intentional knowledge erasure or unintentional knowledge loss." In a review of several papers focusing on the similarities of unlearning, Klammer and Gueldenberg (2019) define the act of unlearning as "an organization actively (voluntarily) engages in giving up old knowledge". In other words, unlearning is "intentionally disregarding prior learning." Forgetting is not the same as unlearning. Forgetting is the temporary loss of multiple pieces of information from memory simultaneously, whereas unlearning is selective. Certainly, unlearning is very similar in concept to forgetting, but they differ significantly for individuals and organizations.

Rupčić (2019) focused on the relationship of the organization's learning dynamics, stated as follows: "when practitioners face challenges, they often find that certain knowledge must be abandoned or discarded, redefined and modified, or updated with new information. This process exists on the individual, team, and organizational level as every building block of a learning organization is continuously engaged in the learning-forgetting-unlearning-relearning dynamics." In other words, the "learning-forgetting-unlearning-relearning" exists not only in organizations but also individuals, and "unlearning" exists as a completely different process from "forgetting" in this mechanism. However, we think that memory and learning are different processes in human behavior, as memorization does not always lead to learning. In many cases of human behavior, long-term memory is required for learning better behavior, but memories being momentarily forgotten and then restored occurs frequently in humans. Moreover, players of some games are skilled in creating momentarily strong memories that are later forgotten. Learning may be performed based on long-term memory, but if the memory is intentionally forgotten, neither memory nor learning may be performed.

Learning is inherently different from memorization. While learning can be intentional or unintentional, learning manifests differently in behavior and conditioning. Though memorization can be achieved both actively and unconsciously, memory does not necessarily result in action. As a contrasting example, though we understand that studying consists largely of memorization, we also recognize the conceptual differences of unlearning and forgetting. We agree with the definition of "unlearning as a purposefully initiated process of discarding knowledge and forgetting as involuntary knowledge loss" (Klammer and Gueldenberg, 2019). Therefore, we consider unlearning to be sufficiently differentiated from memory and out of scope for this paper, and we will focus only on "forgetting."

2.3. Differences between humans and IT in memories and forgetting

Tulving (1974) defines human "forgetting" as "the inability to recall something now that could be recalled on an earlier occasion." One theory of forgetting is co-dependent forgetting, known in psychology as context-dependent forgetting, and defined by Tulving as "reflecting the failure of retrieval of perfectly intact trace information." This theory is famous for describing episodic memory. Episodic memory is subjective and autobiographical, and it is characteristic of the mechanisms of memory for sentient beings. Except AI, IT artifacts cannot memorize episodically, and instead memorize all information through concrete commands given by a human. Therefore, they forget memorized information only in particular situations. Moreover, they can recall all memories as needed in ordinary situations, having total recall as previously explained.

Hardware defines the only storage limitation of IT. Though forgetting is an everyday occurrence for humans, it is special situation for IT because they cannot intentionally forget. Previous studies have shown that human memory requires the ability to intentionally forget (Macleod & Golding, 1998). "Strategic forgetting" (Kearns et.al, 2010) and "organizational forgetting as strategy" (de Holan & Phillips, 2004) are concepts similar to "intentional forgetting." The former is used in the context of reclaiming value through transition by confidently forgetting stigmatized memories of cultural history. The latter, on the other hand, is used in the context of managing organizations and human resources efficiently and effectively. The latter also emphasized the context of organizational behavior, though the behaviors of an organization are created by a set of individual actions and behaviors. Hence, these particular types of intentional forgetting do not differ between individuals and organizations. These concepts and intentional forgetting can also be used for IT because they coexist with humans in "the constitutive entanglement of the social and the material in everyday organizational life" (Orlikowski, 2007).

In addition, Ricoeur (2000) says "Forgetting is bound up with memory," and "Forgetting can be considered one of the conditions for it." Though IT artifacts do not intentionally forget, they share this relationship between memory and forgetting. Kluge and Gronau (2018) have defined "intentional forgetting" as "the motivated attempt to limit the future recall of a defined memory element." Timm et al. (2018) explained further that "intentional forgetting" is a significant mechanism in an AI system. For example, Nuxoll et al. conducted a study for an algorithm of forgetting to artificially perform episodic memory (2010). The importance of this is demonstrated by technical methods for forgetting being studied considering the "Right to Be Forgotten" (Villaronga, Kieseberg, Li, 2018). To develop this idea further, we consider the processes of memory and forgetting used in competitive karuta as a way to observe the "intentional forgetting" that AI cannot do, but which humans can, through the repetition of memorization and forgetting.

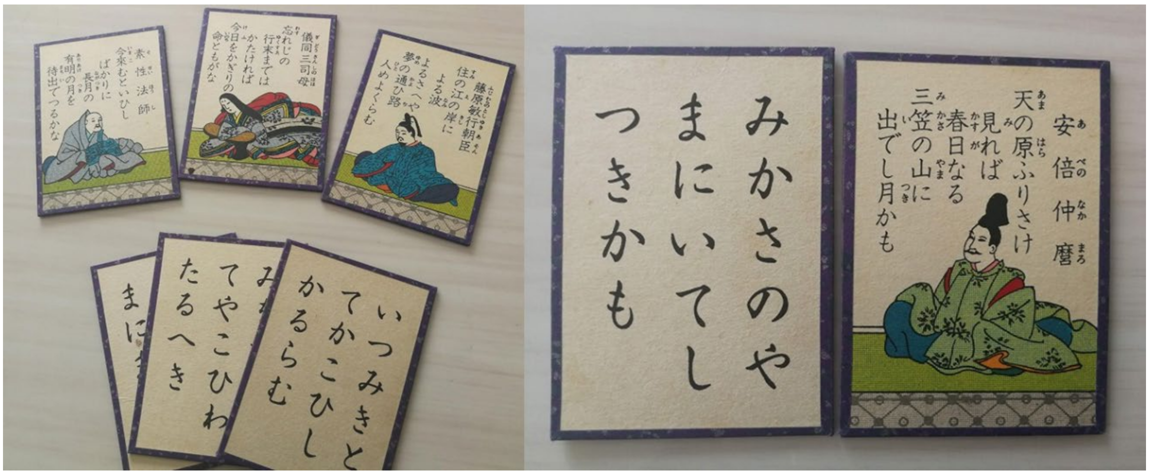
3. MEMORY AND FORGETTING MECHANISMS THAT CAN BE SEEN IN "COMPETITIVE KARUTA”

3.1. History of competitive karuta as Japanese traditional game from karuta as literature

There are many traditional indoor Japanese games. For example, the famous Japanese board game “Shogi (将棋),” which is similar to “Go (碁),” is known to have been played as early as the 16th century. Many Shogi games have been developed as software for PC, and AI has recently been used against professional Shogi players. Traditional Japanese games such as Shogi often challenge players to memorize and recall the state of a game. This is especially true of competitive karuta, which is sometimes referred to as “*Kyōgi karuta* (競技かるた)”. Competitive karuta was established based on a 13th century literary work called *Ogura Hyakunin Isshu* (小倉百人一首), or *Hyakunin Isshu* (百人一首) when abbreviated, which translates to “One Hundred Poets, One Poem Each” in English. In competitive karuta, poems from Hyakunin Isshu, each comprising 31 syllables, are written on cards called “*karuta*.” Of these, "*yomifuda* (読み札)" cards contain an entire poem, whereas "*torifuda* (取り札)" cards contain only the second half of the poem (See Figure 1). As competitive karuta is originally based on a literary work, it is not widely known as a game abroad. Moreover, the style of Japanese text used in Karuta is traditional rather than modern, so it can be very difficult for people who are not native Japanese readers. However, as both a work of Japanese literature and as the game of competitive karuta, the familiarity of Hyakunin Isshu is gradually growing worldwide.

Figure 1. Cards of Ogura Hyakunin Isshu.

Left: Three sets of yomifuda (top) and torifuda (bottom) cards. Yomifuda cards are for reciting, and include an illustration of the author and a poem written in full. Torifuda cards are to be taken by players, and include only the ending of the corresponding poem. Right: One set of corresponding yomifuda and torifuda cards for the same poem.



Source: Photos taken by the author

Despite its difficulty, competitive karuta is played in several international communities. Clubs belonging to All Japan Karuta Association exist in the USA, Thailand, Singapore, China, Brazil, France, and Germany. Furthermore, the International Festival for Competitive Karuta (“International Ogura Hyakunin-ishhu Karuta Festival 2020”) will be held preceding the 2020 Olympic and Paralympic Games in Tokyo (All Japan Karuta Association, 2019). The website for this festival is available in both Japanese and English. Moreover, there is a website available for players who speak a foreign language (Stone, 2018). This is one of the positive characteristics of competitive karuta combining both sports and culture (Bull, 1996). Though rules can be very difficult, it is played by people of many cultures, and its significance in Japanese culture differentiates it from other board games or card games.

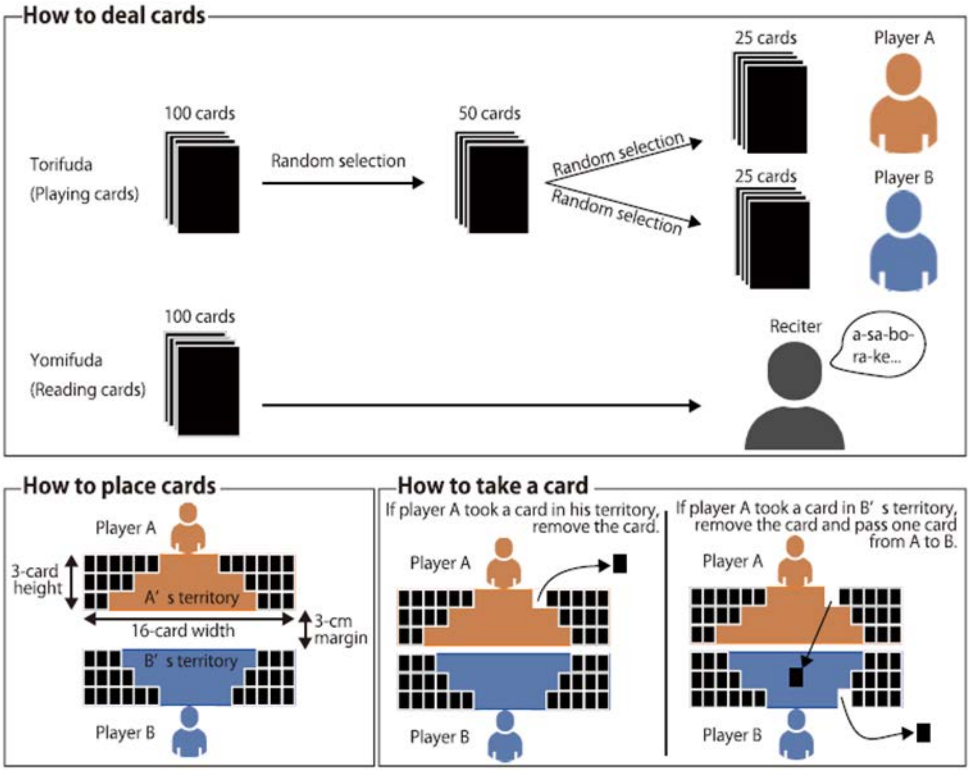
As a collection of poetry, the contents of Hyakunin Isshu are important for most Japanese readers. Currently, most Japanese high school students learn the unique sound of Japanese poetry from Hyakunin Isshu, and it is used in the curriculum of Japanese language classes for interpreting the cultural background of its time. Therefore, Hyakunin Isshu Karuta is popular among Japanese people as part of a common culture. It is also popular as a card game to be played during the New Year, and there are many schools that hold championship events in January. Karuta games played in such tournaments are simpler than competitive karuta games, but competitive karuta is also well known. Championship games are televised through programming on NHK, the Japanese state channel, and the results of championship games are broadcast on the NHK news program. In recent years, manga (Japanese graphic novels) depicting competitive karuta have become popular, one of which has been developed into a movie, further increasing its recognition. As a sport, the population of players is relatively small and the detailed rules of Competitive Karuta are generally not understood in depth, but it may be considered a minor sport in the sense that most Japanese people have at least a casual understanding, and a win or loss during a match can be understood as it is observed.

3.2. Outline of competitive karuta

The basic rules of the game are simple, and the playing area is as shown in Figure 2. After a yomifuda card is selected, a reciter reads the first half of the poem given on the card, and the players must select the corresponding torifuda card that contains the second half of the poem. The objective of the game is to reduce the number of cards in own territory from 25 to 0, and this is done by correctly remembering and identifying torifuda, which match the recited yomifuda (See Figures 2–3). When the correct torifuda is selected, the player who identified the correct torifuda either removes a card from their territory or adds a card their opponent's territory. If a player selects a torifuda incorrectly, this is known as an "*otetsuki* (お手つき)," or "foul" in English, and results in a card being taken from the opponent's territory, increasing the mistaken player's number of owned cards. Before the start of the game, the positions of the cards in the field are memorized in advance, and players attempt to memorize the cards by "*kimariji* (決まり字)," which is the first syllable of a card by which a correct torifuda can be identified. However, because the placement and kimariji of the cards change as the competition progresses, it is necessary to re-memorize the placement and kimariji of the cards quickly. Because this process is easy for IT, game apps for competitive karuta have already been launched.

Players must make decisions and move quickly to identify and select cards according to the progress of the game. This decision-making and action based on the game information has a direct impact on the outcome of the game. Though the physical aspect of the game can be performed reliably by a robot arm following the derivation of an optimal solution using an AI, the most important aspects of the game are memorizing the initial position and kimariji of the cards and then re-memorizing them as the game state changes over time. For this task, past memories are forgotten, and new information is repeatedly stored. No matter how fast the robot moves, it cannot win if its memory is weak. Even if an application is developed with total recall and accurate operation, a first-class player is difficult to surpass unless the application has the ability to hear slight differences in the voice of the reciter.

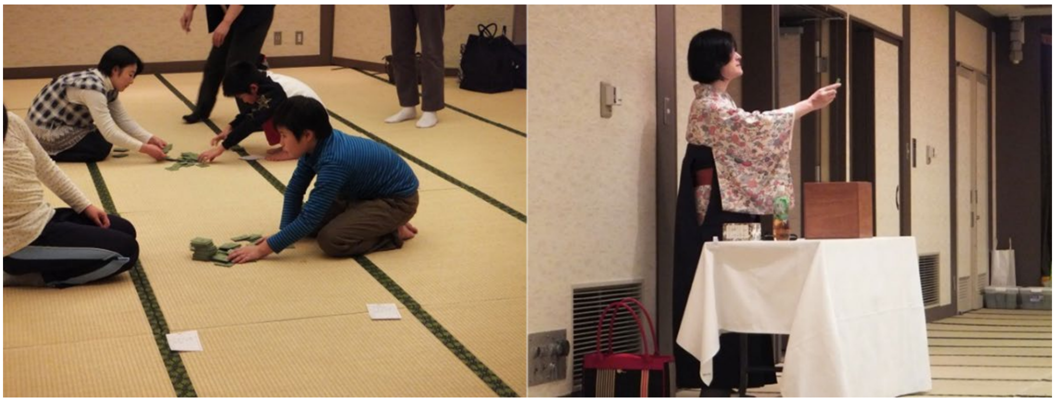
Figure 2. Rules of competitive karuta.



Source: Yamada, Murao, Terada, and Tsukamoto (2018)

Figure 3. Preparation for beginning a match.

Left: Each player takes 25 cards face-down from 100 cards available. Right: The official reciter reading a yomifuda card for competitive karuta.



Source: Photos taken by the author from a match on January 3rd 2013

3.3. Mechanism of memory and forgetting of competitive karuta

In a competitive karuta game, memorization and forgetting are repeated frequently after every reading of a yomifuda card. A player wins by being ahead by reducing own all cards, so in the case of a close match, as many as 100 yomifuda cards are read. In addition, as many as 5 or 6 the games can be played in a day leading to the final stage of tournament matches, so the winner and runner up will memorize and forget up to 500 times before the tournament is finished. Because the 50 cards used for each match are drawn from the same collection of 100 cards, cards will be seen multiple times, and if the states of previous matches are not forgotten, confusion will result from inconsistencies in a player's memory during a match.

The task of playing competitive karuta is complex and requires using information and understanding while completely forgetting past game information. Moreover, the amount of information to memorize when taking cards is high, and includes the placement of torifuda in the bases of both players, the kimariji of each torifuda while considering the situation leading to the current game state, which poems have already been read, and the current kimariji of torifuda that have not yet been read. Because these items change each time a card is read, it is necessary for players to overwrite their current memory and forget older memories in a short time before the next card is read.

For example, there are only two poems that begin with the syllable "U": "U-ka-ri-ke-ru" and "U-ra-mi-wa-bi". At the beginning of a match, players can identify the appropriate card to select by recognizing the second syllable of a poem if it begins with "U", depending on if it the poem begins with "U-ka" or "U-ra". However, after one of those poems has been recited and the corresponding card has been removed from the field, players can get identify the correct card after only hearing "U", as the first syllable, so they adjust their memory of the kimariji. Furthermore, if both the "U-ka" and "U-ra" cards are held by the opposing players, they must listen for the second syllable to avoid choosing the wrong card.

Though only two poems begin with the syllable "U", eight poems begin with "Na", and 16 poems begin with "A". Players memorize the positions of each card their fixed letter (kimariji) for 15 minutes before the game begins, but as the fixed letters change over the course of the game, it is necessary for the players to rewrite their memory as each poem is recited. In addition, players change the locations of the cards during the match as poems are recited and cards are selected. Yomifuda cards are read frequently, and players must memorize the new card positions in a very short amount of time between each reading of a yomifuda card.

Competitive karuta uses a mechanism of memory that is similar to that used for the called game called "Concentration". Players use only the essential information of past memories. However, previously memorized information is not used directly because using it becomes an obstacle to accessing current information and ensuring the accuracy of that information. This is not the case for IT artifacts, which have the capability of total memorization and recall. To accurately model human memory, though it is important to store information in an extended area that can be separated and forgotten, searched when necessary, and accessed, it is also important to block the access of some information to improve accessibility to prioritized information. In "Concentration", it is difficult for players to win against IT, which can memorize everything completely. The difference between competitive karuta and Concentration is that there are psychological tactics used between humans, and the content of those psychological tactics includes the frequency of repetitive memorization and intentional forgetting. The psychological tactics used also depend on the opponent. This phenomenon is difficult to observe against IT because of the advantage of total recall.

Of course, even in competitive karuta, if a human lets an AI learn how to play matches using a game app, it will learn various battle patterns. An AI for Shogi can play more matches in a short period of time than humans can, and its learning makes it stronger than humans. However, they cannot learn differences in the speed card selection due to players having more familiarity with certain cards. Because an AI with perfect memory does not have preferences or familiarity, it will always respond the same to all cards. By contrast, human players can deduce or intuit their opponent’s favorite cards and strategically arrange the cards for an advantage.

The reason why you can compare your opponent's small actions with the memory you have in an instant and make use of it in a match is that you can use your past knowledge instantaneously while performing intentional forgetting. If you use prior information at the same time, it will fail. It is therefore necessary to forget past information and data intentionally, using only knowledge.

This mechanism of intentional forgetting through omission can be observed in humans, and it improves memory reliability. We know that repeated learning prevents forgetting and improves long-term memory. However, it is important for humans to intentionally forget while repeatedly memorizing information, which differs from the behavior of AI. Modern IT is capable of total recall, and we recognize this capability as useful. Though humans would like the ability to memorize and recall at will, memory may be strengthened through intentionally forgetting.

Players must remember certain information in the long term while forgetting specific details of previous memories. The requirements of a player’s memory can be summarized as follows:

- Strong memory at the start of the game
- Forgetting new and old memories as the game progresses
  - ➢ Players must repeatedly rewrite a strongly established memory roughly every 30–120 seconds over the 15-minute length of each game while forgetting their memory of previous game
- Creating new memories by organizing forgotten memories
  - ➢ If players forget previous events in the course of a game, they cannot confirm the current kimariji
  - ➢ It is necessary to memorize which poems do not exist in the current game
  - ➢ The memories used for learning, organized for each moment, must be forgotten
- Forgetting previous game information when remembering the next game
  - ➢ Accounting for the uniqueness of changing conditions even though the physical appearance of the placed cards is the same
  - ➢ The use of memorization is similar to that used in the card game called “Concentration,” but the selection of cards based on fixed characters may not always be shortened, and may be longer depending on the player’s strategy

4. DIFFERENCES IN MECHANISMS OF INTENTIONAL FORGETTING AND MEMORY BETWEEN HUMANS AND AI

The case of competitive karuta suggests that we can observe the human mechanism of memory and forgetting referred to as “intentional forgetting” by Golding & Macleod (1998). “Intentional forgetting” is an old concept and is also known as “motivated forgetting” (Weiner, 1968). However, the recent development of IT can retain all information, which is seen by many as desirable. Moreover, while most people agree that human beings can never surpass IT in terms of memory, IT can never consciously forget because IT do not have consciousness. As IT cannot forget consciously and intentionally, this demonstrates a problem of “the right to be forgotten.”

Though these aspects of memory and forgetting are difficult to directly observe in the game of competitive karuta, they can be observed objectively to some extent. Memory is instantaneously renewed during the game, and previous memories are forgotten. However, these forgotten memories are not something to be permanently deleted when the scene is finished, and they are revived as material for future decision-making. Some memories should be established as knowledge later, if necessary. When a person intentionally hides important information in a strategic situation called a "match," which consists of the accumulation of momentary memorization and forgetting, it can be called "strategic oblivion."

In other words, in the game of competitive karuta, only an essential understanding is extracted from past memory and used as knowledge, but the details of information in the memory are not used in order to prevent them from becoming an obstacle in accurately recalling information of the current memory. In IT, information becoming a barrier to accessing other information generally does not occur, excepting events such as data collision or programing mistakes. By following instructions, the necessary information can be accessed. However, that information is a set of momentary fragments of memory, and even if an AI learns the flow of the game, it is difficult to learn things such as the opponent's otetsuki due to a memory error, the unique exchange of cards performed by their opponent, and card familiarity.

This is different from today's world of IT where everything is always stored in a searchable form. It is a game which cannot be solved when all things are considered retrieval candidates, even things that we would prefer to forget. A world including the "right to be forgotten" is incompatible with requiring total recall. Replicating human memory with IT is not an issue of simply having a larger memory capacity. Human memory is more similar to the use of bank switching and Expanded Memory Specification, which was used in the early 1990s to exceed the limitations of computer processors, allowing them to directly address different configurations of memory depending on the situation. Rather than constantly improving capacity, a more accurate replication of the reliability of human memory will store information in an extended area that can be separated, ignored, and made accessible when necessary so that the information can be intentionally blocked. Because the current state of AI is capable of total recall, it does not support the right to be forgotten, which presents an ethical problem when considering the introduction of AI into human society. An additional ethical problem is presented when considering the use of AI as an extension of the human brain through cyborgization. The ability of humans may be limited, but the limitations of humans are very significant in defining humanity; thus, exceeding the limitations of humanity could potentially be considered as a loss. This study suggests that because modern AI does not include "intentional forgetting", an ethical coexistence between humans and AI in society will remain limited.

5. CONCLUSION

In summary, we suggest that modern AI cannot currently replicate the memory of human beings, and does not have an equivalent relationship between memory and forgetting. However, an equivalent relationship may eventually be a possibility. When considering people working in organizations, humans are superior in activities that require collaboration to perform physical activities. This is a necessary consideration in the relationship between AI and humans. Though IT artifacts have a capacity for memory that surpasses human memory, the difference between AI and humans is significant in modeling human intelligence accurately. To accurately model human intelligence, AI must make human-like decisions about whether to forget a given piece of learned information. Until AI has the ability to intentionally forget, AI intelligence cannot be considered analogous to that of human beings.

Lastly, we consider the limitations of this study. Competitive karuta is a very particular case, and is not being widely played throughout the world. It also has a special background of Japanese culture. It is possible that this idea is only appropriate to this particular case. However, the most important point to be made in this paper is that this case shows that the ability of IT in terms of memory is currently not appropriate for replicating the human mechanism of memory and forgetting. On that point, this research is just one case of observing the unique mechanisms of human memorization and forgetting. In the future, further studies are needed to determine the relationship between memory and forgetting for communication between human beings and AI, and for examining this mechanism when considering the further limitations of AI and human potential. Thus, although there are limitations, this study suggests the importance of forgetting as a mechanism in future AI development.

ACKNOWLEDGEMENTS

We are particularly grateful to Professor Kiyoshi Murata and Associate Professor Yoko Orito for their encouragement and continuous support. Without their encouragement, this paper would not have materialized.

This study was supported by the Japan Society for Management and Information Grant-in Aid for SIG “Monitoring and Control of AI Artefacts”, the Seikei University Grant-in-Aid 2020 for the research on “Monitoring and Control of AI Artefacts: Consideration from Economics, Social, and Legal Perspectives,” and JSPS KAKENHI Grant Number JP17K03909.

REFERENCES

All Japan Karuta Association (2019) International Ogura Hyakunin-isshu Karuta Festival 2020 Official Website, Retrieved from https://karuta-fes.com/en/

Bull, D. (1996). Karuta: Sport or culture? Japan Quarterly, 43(1), 67.

Bell, G., & Gemmell, J. (2009). *Total recall: How the e-memory revolution will change everything.* New York: Dutton.

Golding, J. M., & Macleod, C. M. (1998). Intentional forgetting: interdisciplinary approaches. L. Erlbaum Associates, Mahwah, N. J.

Gorry, G.A. (2016). Memory machines and the future of knowledge management. *Knowledge Management Research & Practice*, 14(1), 55-59. http://doi.org/10.1057/kmrp.2014.19

de Holan, P. M., & Phillips, N. (2004). Organizational forgetting as strategy. *Strategic Organization*, 2(4), 423-433. https://doi.org/10.1177/1476127004047620

Inoue, A., Sawatari, E., Hisamoto, N., Kitazono, T., Teramoto, T., Fujiwara, M., Matsumoto, K., & Ishihara, T. (2013). Forgetting in C. elegans is accelerated by neuronal communication via the TIR-1/JNK-1 pathway. *Cell Reports*, 3(3), 808-819.

Kearns, R., Joseph, A. E., & Moon, G. (2010) Memorialisation and remembrance: on strategic forgetting and the metamorphosis of psychiatric asylums into sites for tertiary educational provision, *Social & Cultural Geography*, 11(8), 731-749. http://doi.org/10.1080/14649365.2010.521852

Klammer, A., & Gueldenberg, S. (2019). Unlearning and forgetting in organizations: a systematic review of literature. *Journal of Knowledge Management*, 23(5), 860-888. https://doi.org/10.1108/JKM-05-2018-0277

Kluge, A., & Gronau, N. (2018). Intentional forgetting in organizations: the importance of eliminating retrieval cues for implementing new routines. *Frontiers in Psychology,* 9, 51. http://doi.org/10.3389/fpsyg.2018.00051

Murata, K., & Orito, Y. (2011). The right to forget/be forgotten. CEPE 2011: Crossing Boundaries, 192-201.

Nuxoll, A., Tecuci, D., Ho, W. C., & Wang, N. (2010). Comparing forgetting algorithms for artificial episodic memory systems. Remembering Who We Are- Human Memory for Artificial Agents Symposium at the AISB 2010 Convention, 14-20.

Orlikowski, W. J. (2007). Sociomaterial practices: exploring technology at work. *Organization Studies,* (28)9, 1435-1448.

Rupčić, N. (2019). Learning-forgetting-unlearning-relearning – the learning organization's learning dynamics. *The Learning Organization*, 26(5), 542-548. https://doi.org/10.1108/TLO-07-2019-237

Ricoeur, P. (2000). LA MEMOIRE, L'HISTOIRE, L'OUBLI, Editions du Seuil, (Memory, History, Forgetting, Translated by Blamey, K, and Pellauer, D. The University of Chicago Press, 2004)

Stone, M. (2018) World of Kyogi Karuta, Retrieved from http://karuta.game.coocan.jp/

Timm, I. J., et al. (2018). Intentional forgetting in artificial intelligence systems: Perspectives and challenges. Joint German/Austrian Conference on Artificial Intelligence, 111117, 357-365.

Tulving, E. (1974). Cue-dependent forgetting: when we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist,* 62(1), 74-82. Retrieved from http://www.jstor.org/stable/27844717

Villaronga, E. F., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2), 304-313.

Yamada, H., Murao, K., Terada, T., & Tsukamoto, M. (2018). A method for determining the moment of touching a card using wrist-worn sensor in competitive karuta. *Journal of Information Processing*, 26, 38-47.

Yasuda, H., Kojima, N., Hanamura, K., Yamazaki, H., Sakimura, K., & Shirao, T. (2018). Drebrin isoforms critically regulate NMDAR- and mGluR-dependent LTD induction. *Frontiers in Cellular Neuroscience*, 12, 330. http://doi.org/10.3389/fncel.2018.00

Weiner, B. (1968). Motivated forgetting and the study of repression. *Journal of Personality*, 36(2), 213-234. http://doi.org/10.1111/j.1467-6494.1968.tb01470.x

Wensley, A. K. P., & Navarro, J. G. C. (2015). Overcoming knowledge loss through the utilization of an unlearning context. *Journal of Business Research*, 68(7), 1563-1569.

MONITORING AND CONTROL OF AI ARTIFACTS: A RESEARCH AGENDA

Hiroshi Koga, Sachiko Yanagihara

Kansai University (Japan), University of Toyama (Japan)

koga@res.kutc.kansai-u.ac.jp; sachiko@eco.u-toyama.ac.jp

ABSTRACT

The purpose of this paper is to find a future research agenda through examination of the concept of AI artifacts. To that end, this paper is organized as follows: First, First, what AI artifacts are discussed. Next, the characteristics of AI artifacts are clarified, that is, the following two points. (1) AI artifacts contain organizational context and human agency, (2) AI artifacts fuse boundaries with natural objects. Finally, the impact is examined and future research agendas are proposed.

KEYWORDS: IT artifacts, AI artifacts, responsibility, privacy, sociomateliarity.

1. INTRODUCTION

The purpose of this paper is to focus on the idea of "AI artifacts", which has recently attracted attention in information systems research, to examine its significance, and to present a research agenda for AI artifacts.

In recent years, the expression "IT artifacts" (*signifiant*: symbolic expression) is often used instead of "information systems" in research on information systems (especially management information systems).

One of the reasons for this is the need to expand the traditional concept of information systems, as devices such as smartphone and software (app) such as SNS have penetrated daily organizational life. In other words, behind the expression "IT artifacts," we can find an attitude of daring to focus on the technical aspect under the situation where the conventional organizational behavior and various information systems are being fused.

Let's take a look back at the conventional discussion of information system development. Once upon a time, the "introduction" or "implementation" of information systems was discussed. And what happens after the introduction and implementation has been perceived as an organizational problem. However, as mentioned above, information devices are now deeply involved in all organizational behaviors. Therefore, "transformation of the organization and information systems" through the use of information systems has come to be recognized as an important management issue. This should be easy to understand, for example, by referring to the discussion of DX (digital transformation) and IoT (internet to things).

In this paper, therefore, we would like to revisit and consider the idea of AI artifacts, and then present a future research agenda, including its relevance to information ethics and business ethics.

2. PRIOR RESEARCH ON IT/AI ARTIFACTS

Herbert A. Simon was probably the first person to use the term "artifacts" in information systems research. The winner of the Alfred Nobel Memorial Prize for Economics at the National Bank of Sweden, he was not only a pioneer in modern management theory, but also an erudite and versatile person who made outstanding achievements in artificial intelligence theory and information systems research.

2.1. Revisiting the science of artifacts by H. A. Simon

Simon's work on artifacts is the well-known "The Science of Artifacts". In the same book, the differences between the "natural sciences" and the "science of artifacts" are repeatedly emphasized. According to him, natural science is a "nomothetic science," a discipline in search of universal truths. There, natural laws are assumed to exist "objectively". Moreover, natural law can be thought of in terms of purpose, i.e., value in isolation (there is, of course, no denying that science has developed in relation to the problem of God's existence).

The science of artifacts, on the other hand, aims to "construct useful things". Therefore, the evaluation of the constructed artifacts is more important than the elucidation of objective laws or the validation of the theory. For this reason, evaluation criteria (objectives and values) are important.

Simon distinguishes between these two very different views of science, and argues that science of artifacts should aim at the "science of design".

Here, we would like to expand on Simon's argument and consider it. The natural sciences are heavily influenced by "Christianity". The object of the natural sciences is the "world of God's making". He probably thought that by analyzing the natural world, he could objectively prove God's existence if he could reveal the "laws of nature," which are the blueprints of the world-building that God used. Newton, the last alchemist, is said to have discovered the "Law of Universal Attraction" as proof of the existence of God.

On the other hand, the science of artifacts covers artifacts (things made by people); if we take artifacts broadly, we can say that humanity, such as poetry and literature, and social science, such as laws and institutions, are "artifact sciences. Here, "satisfaction" is important rather than optimization, and "description (idiographic)" and "evaluation of the artifact" are more important than "law-establishment".

Of course, as Schön (1983) critiques, it must be said that Simon's concept of design is only an effective strategy for "structured problems" and falls within the bounds of "technical rationality". Schön argued that attention should be paid to trial and error and reflection in the execution phase. In the field of information systems research, from the late 1980s onwards, a research trend emerged to focus on the execution process as pointed out by Schön.

2.2. IT artifacts as s a representation of an identity for information systems research

One of the reasons why the information systems researchers refer to "IT artifacts" is that they are aware of the difference between them and computing or computer science. If we only focus on the technical aspects, we can call it a "computer system".

However, this is because we want to emphasize the aspect of an inherent social construct that is colored by background factors such as organizational culture in the place where it is used. In other words, we understand that the feature of an artifact is that it does not function as we expect it to.

In addition, information systems have "interpretive flexibility". It depends on the context of the organization whether e-mail is seen as a tool for freely expressing opinions or as a surveillance tool that censors the content of speech to identify problematic people. In other words, "the meaning of an artifact can be interpreted very differently depending on the situation in which it is placed.

From the above, IT artifacts are characterized by the fact that their uses and effects are not known until they are used. There is a reason why objective laws are so hard to figure out.

Such an understanding is close to the position of Schön's critique of Simon. Nevertheless, the result follows Simon's position of "making something useful" and "idiographic rather than nomothetic science". That's why Simon is being re-evaluated. And so information systems research, which focuses on "practice," emerged as the science of artifacts that went beyond Simon.

Wanda J. Orlikowski a leading commentator on the flexibility of interpretation, defines IT artifacts in a paper co-authored with Iacono as follows It is, they say, "those bundles of material and cultural properties packaged in some socially recognizable form such as hardware and /or software" (Orlikowski & Iacono, 2001, p.121).

Furthermore, Orlikowski & Iacono (2000) offer the following five premises of IT artifacts; That is, (1) IT artifacts, by definition, are not natural, neutral, universal, or given. (2) IT artifacts are always embedded in some time, place, discourse, and community. (3) IT artifacts are usually made up of a multiplicity of often fragile and fragmentary components, whose interconnections are often partial and provisional and which require bridging, integration, and articulation in order for them to work together. (4) IT artifacts are neither fixed nor independent, but they emerge from ongoing social and economic practices. (5) IT artifacts are not static or unchanging, but dynamic.

Thus, IT artifacts tend to emphasize organizational aspects rather than technical characteristics. Therefore, it should be called “social artifact” or "socio-materiality". From such a perspective, Lee et al. (2015) referred to the subject in the field of information system research as "IS artifacts", which are subclasses: (1) information artifacts, (2) technology artifacts, and (3) social artifacts. It points out the need to focus on the interaction between them (Lee, Thomas & Baskerville, 2015).

As described above, a characteristic of the research approach that focuses on IT artifacts is that the significance of IT artifacts is considered to be constructed in organizational practice, based on the social constructive perspective of information systems. Therefore, the main focus of such a research approach is to focus on organizational information practices and to describe the actions of IT artifacts in the process.

Nowadays, information systems and organizations are inseparably related to each other in the actual organizational information behavior such as DX. Based on this premise, IT artifacts can be said to focus on the IT elements in an organization. This has led to the emergence of a position that emphasizes two distinct characteristics of IT artifacts: organization and technology (materiality). It is the so-called socio-material school.

For example, Orlikowski (2008) used the term "entanglement" to refer to the inseparability of organizational and technical elements, which is a term from quantum mechanics. Similarly, Leonardi (2012) used the word "entanglement" to describe such a relationship, referring to the way the tiles overlap (or the way the stones in the riverbed are rounded and aligned with the flow of the river).

In Japanese Buddhist terminology, the aspect in which two things are united and inseparable is called *Ni-Ni-Fu-NI*(而二不二). Because there is light, there is shade. We can't eliminate the light and take out only the shadows. The mind and body cannot be separated (*Shiki-Shin-Fu-Ni*:色身不二). There are ideas such as the inability to separate the environment from the organism (*E-Sho-Fu-NI*:依正不二). In this respect, the idea of social materiality can be said to be extremely oriental.

2.3. The concept of AI artifacts: embedding the Organizational Context

On the other hand, it is for AI artifacts that technical aspects are often emphasized. For example, “such as artificial neural networks, specifically focusing on deep neural networks” by Tuncali *et.al*. (2018, p.1) or “data and AI models being used in the process of AI system development” by Maksimov *et. al*. (2018, p.2).

However, Rankin's argument can also be understood as "incorporating the social aspect as a function involved in intellectual judgment. If this is the case, AI artifacts can be seen to contain the social contextual or practical aspects of IT artifacts as their own functions.

Behind such a strong technical orientation, it is thought that the organizational context is embedded in machine learning. In other words, embedding organizational context means that the intelligence activities that have been entrusted to human beings have been entrusted to artifacts.

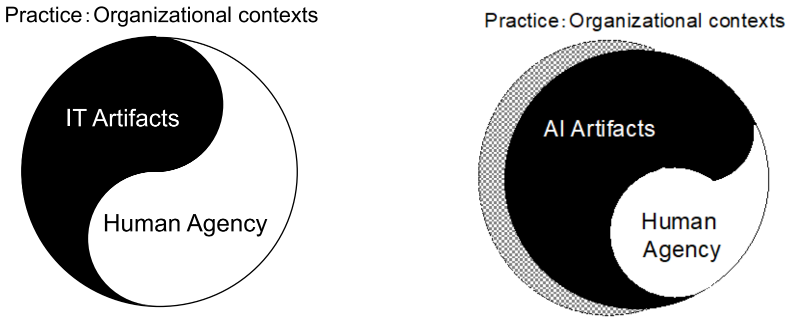
Traditionally, without fear of misunderstanding, IT artifacts have been responsible for "mechanical processing" in the field of organizational practice related to information processing. In the case of AI artifacts, however, the execution of value judgments will be automated through machine learning. In other words, the aim of artifacts has shifted from "mechanical processing of formal information" to "mechanical generation of semantic information". For example, it has become possible to predict the number of months of pregnancy based on purchase histories that are little more than formal information, or to sell the behavioral patterns of people who have declined job offers, in order to expand business opportunities through the generation of "personal information, that is, semantic information".

As a result, the ethics of the operation of AI artifacts have come to be questioned. At this time, AI ethics is indeed an organizational practice and is deeply related to organizational culture and values. So far, it can be said that the key to informatization practice using AI artifacts lies in organizational practice.

The left figure of Figure 1 illustrates the relationship between IT artifacts and organizational factors in organizational practice. This is in support of the Yin-Yang diagram. The whole (circle) indicates the organizational practice. A schematic representation of the complex intertwining of IT artifacts and tissue elements, with yin and yang occurring in the circle.

For organizational members, there is no awareness of yin and yang in the process of practice. No, you can't practice it consciously. When riding a bicycle, being aware of the movement of the pedals and chain can lead to accidents. Even in organizational practice, you are probably doing it without being aware of IT artifacts and organizational practices. Hence, The left figure of Figure 1 is shown as a complex shape of a yin-yang diagram rather than a semicircle to represent the nature of the IT artifacts and organizational factors interfering with each other rather than sharply distinguishing them.

Figure 1. Differences between AI artifacts and IT artifacts.


Practice : Organizational contexts
IT Artifacts
Human Agency
Practice : Organizational contexts
AI Artifacts
Human
Agency

Source: Drawing by authors

By the way, the Yin-Yang diagram seems to clearly distinguish between IT artifacts and organizational agency. Indeed, from a microscopic point of view, we can distinguish between devices such as smartphones and software. From a macroscopic perspective, however, they work together to form a hybrid.

Then, we want you to imagine a meal scene, for example. Like the attitude of picking up a camera thinking, "Photogenic! (or, to use a more recent expression, *instagrammable*)" while holding the smartphone (that is, camera) in his hands. There, the smart phone and the human being have become one, and the act of eating itself is being transformed. In other words, like the "pointillism" of an impressionist painting, the image of a unified whole is the basis of the idea of "IT artifacts", in the same way that the dots of various paintbrushes may appear up close, but the landscape appears from a distance.

Next, the right figure in Figure 1 is a simplified diagram of the relationship between AI artifacts and organizational factors in organizational practice. AI artifacts can be used to transcend specific contexts (as in the old Empty Systems in Expert Systems), and AI algorithms sometimes face ethical issues, as in the case of the U.S. Target's Pregnancy Index, in order to surrender the values of a particular organizational context. We consider these features to be the transcendence of the organizational context of AI artifacts. In the right figure of Figure 1, these characteristics are shown in the "shaded area" as "the expansion or transcendence of the context of organizational practice".

In addition, as pointed out in the previous section, AI artifacts may encompass organizational factors. Hence, AI artifacts in organizational contexts are not only intricately intertwined with organizational factors, but also serve as substitutes for some organizational factors. Therefore, the balance of yin and yang is different in comparison to the left diagram in Figure 1.

The implication shown schematically is that when considering AI artifacts, the scope of information ethics will be expanded because parts traditionally positioned in business and organizational ethics will be embedded in AI artifacts.

3. RETHINKING THE CONCEPT OF AI ARTIFACTS

In the previous section, we discussed the differences between IT and AI artifacts, focusing on previous research. The perspective of IT artifacts can be understood as a concept proposed for organizational practice as a hybrid of IT and organization, in order to focus on the micro-IT and organizational relationships therein.

So far, it can be said that the concept of IT artifacts is closely related to "*sociomateriality*", which is a recent keyword in the field of information systems research. Now, AI artifacts are an extension of IT artifacts. Therefore, I would like to assume that the focus is on everyday practice.

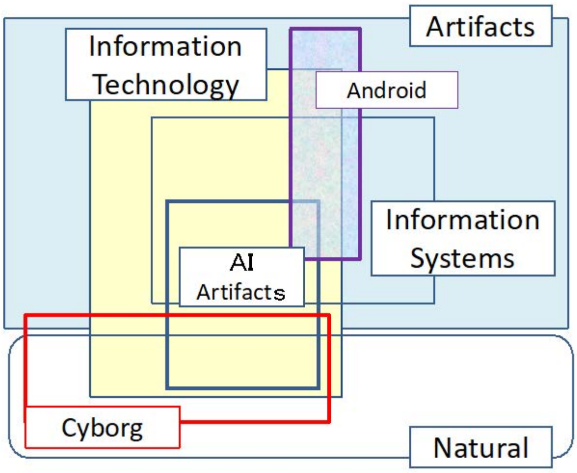
AI artifacts are more closely intertwined with organizational agencies than IT artifacts. As mentioned above, some of the organizational agency is directly incorporated into the AI artifacts, since decisions, etc., traditionally taken by humans are delegated to the system. In short, AI artifacts are encroaching on the realm of traditional organizational agency.

At this time, in the realm of computing, there is an idea of "objects" that focuses on organizational routine acts. By everyday practice, however, we mean "here and now" practices that are situated in personal, organizational, and social contexts, unlike objects as abstracted procedures. In order to do so, we would like to confirm that the assumptions are different from those of object-oriented approach.

Of course, the core algorithmic part of an AI artifact can be context-free, like the abstracted procedures of an object-oriented approach. Therefore, in the right figure of Figure 1, the shaded area shows the possibility of AI artifacts to go out of their original place of practice.

In this section, we consider the comprehensive relationship between AI and IT artifacts and natural objects and humans. Then, figure 2 was drawn by the authors to show the comprehensive relationship between AI and IT artifacts.

Figure 2. Differences between AI artifacts and IT artifacts.



Source: Drawing by authors

First, IT artifacts are concepts that encompass information systems (IS). Of course, it is also true that there are some IS that do not use IT. It should be noted that there is a non-IT based IS for providing and sharing information, for example, hallway bulletins and various documents.

Among the ISs that use IT, AI artifacts are the ones that use AI. However, some IT artifacts, including AI artifacts, may "cross the border" into natural objects. This is the major difference between IT artifacts and AI artifacts: humans with IT artifacts (e.g., artificial organs) are referred to as "cyborgs" in this paper. In particular, when AI is functioning as a substitute for the brain, the brain, which is in charge of decision-making and action, is an artifact even though it is natural in appearance, and it is considered to be a part of the fusion of the natural and the artifact, which should be the original meaning.

Conventional research on brain function has suggested that the brain is in charge of human behavior and that the human body is driven by the brain's control. In fact, it is common for brain diseases to cause a loss of functions that would have been controlled by the site.

On the other hand, the brain may be able to substitute functions in other parts of the body. In addition, it has recently become known that the brain does not control the brain top-down, but rather that the brain gives instructions for the first time when transmitters are transmitted to the brain from each organ (Valadi et.al., 2007).

Furthermore, recently, "bio-3D printers" that use cells to artificially create tissues have entered the stage of practical application, with some companies expecting to release them by the spring of 2020 (Ohshita, 2019).

However, this is the first time that live cells containing the genetic information of a patient can be grown and used like ink to create the organs themselves. Considering the fact that we are entering the clinical trial phase of artificial organs (even if they are only for experiments, there is the problem of destroying living cells), it can be said that we have completely surpassed the conventional concept of artificial organs and have entered into the field where IT can create natural organs. However, if it becomes possible to create neurons in the same way, the day will come when IT-created brain neurons will be implanted, in other words, brains as AI artifacts will be implanted in humans.

In organizational practice, IT artifacts and organizational agency formed a hybrid. However, AI artifacts may be integrated with human organs such as the brain to form hybrids (cyborgs). This in itself is likely to be an important issue for cyborg ethics. This point will be discussed in more detail in the next section.

4. THE RAINBOW CHALLENGE FACING AI ARTIFACTS: THE RESEARCH AGENDA

In this section, we present the research agenda for the management of AI artifacts. As we have repeatedly emphasized, AI artifacts have the following characteristics.

1. Two functions that organizational agencies have assumed: cultural constraint functions and intelligence functions such as judgments, which are now being delegated to AI artifacts
2. The phenomenon of hybridization (fusion) with human beings, that is, a boundary fusion with the human world, is emerging.

From these points of view, the management of AI artifacts will face the following seven challenges. Since there are seven challenges, we will refer to them as rainbow challenges in this paper. This is because, in Japan, the rainbow is considered to be composed of seven colors, and these issues are interrelated, just as the boundaries between the colors of the rainbow are ambiguous. Hence, I dared to title it Rainbow.

(A) Issues Related to System Delegation

(A-1) Privacy violations related to data use: Snooping for confidential information through lifelog analysis

(A-2) Responsibility for the accident: AI developers vs. users

(A-3) Employment Issues through System Delegation: Unemployment vs. Job Creation

(A-4) The loss of personal development opportunities: the Bernard's organizational theory perspective

(B) Issues related to human-machine hybrids

(B-1) Discussion of the view of artificial life

(B-2) The New Disparity Problem

(B-3) International Comparison of Attitudes toward Hybridization of Humans and Artifacts

The following is a brief description of each issue.

4.1. Privacy violations related to data use

First, there is a risk that data analysis generates "sensitive information". For example, the invasion of privacy will be seen as a problem, such as an US company (e.g. Target Corporation) predicting the number of gestation weeks of customers. Needless to say, in the field of business administration, customers are also considered to be organizational members (contributors) (Barnard, 1938). This is because it is difficult to continue organizational activities if customer contributions are supported.

4.2. Responsibility for the accident

As is often pointed out, there is an "*aporia*" in which the responsibility for traffic accidents in the case of automated driving lies with the system developer or user. On Japanese legal issues, for example, there are papers like the following (cf. Kawasi, 2020; Ninomiya, 2020). While there are a variety of moral, legal, and technical issues, we will only point out this point.

4.3. Employment Issues through System Delegation

The challenge of whether AI will take away jobs is also important, as has been widely debated in the past (cf. Frey & Osborne, 2017). In Japan, Arai (2017) argue that the more important issue is the decline in human basic academic skills, rather than the replacement of many jobs with AI. Others argue that data scientists will be replaced by AI in the near future. In any case, the key issue is whether technology induces deskilling (cf. Braverman, 1974).

4.4. The loss of personal development opportunities

Delegation of intelligence /judgment functions creates the following problems. For example, where is the responsibility for accidents in autonomous driving? In the academic field of business administration, "responsibility" has been considered a key factor for the simultaneous development of individuals and organizations (Barnard, 1938). Therefore, if the responsibility is ambiguous, the organization may collapse.

4.5. Discussion of the view of artificial life

The question is whether artificial objects should be treated as life forms. This is a conundrum. Here, we would like to present a Buddhist episode that is suggestive in considering whether or not cyborgs can be considered human.

A traveler was resting in a cave. Demon/鬼 A comes carrying a corpse on his back. Demon A was about to eat with the corpse. Then another demon, B, came along. He said, "It's mine," and tried to take the body from the first demon A. The first demon A said, "Well then, let's ask the traveler there who this corpse belongs to. The traveler honestly replied that this was what the first demon A had brought. After hearing this, Demon B got angry and tore off the traveler's right arm and ate it. Then the first demon A, took pity on him and tore off the corpse's right hand and put it on the traveler. However, demon B, who came later, tore off a traveler's leg and ate it. Then the first demon, A, tore the leg off the corpse again and attached it to me. This happened repeatedly, and the demons left the cave.

The traveler thought that he had been saved, but on reflection, he couldn't tell if the corpse had really become him or not.

This may be an extreme example. However, when cyborgs become more advanced, it will be a challenge to consider whether we are a living organism or a robot.

4.6. The New Disparity Problem

In the world of sports today, the development of tools has also contributed greatly to record breaking. In the old days, an innovation in the material of the bar vaulting pole was a huge boost to the record. In recent years, marathon shoes have encouraged good records.

At the Paralympics, it is difficult to compete without a dedicated prosthetic leg or wheelchair. If these are regarded as cyborgs in the broadest sense, there is a danger of creating a disparity in the possibility of introducing AI artifacts (and by extension, economic power) as a requirement for *cyborgization*.

4.7. International Comparison of Attitudes toward Hybridization of Humans and Artifacts

The hybrid of natural and artificial objects can be rephrased as a hybrid of real and virtual. Incidentally, as with cricket, baseball and "*Ya-Kyu* (Japanese/野球) ", there are regional differences in attitudes towards hybrids. In baseball, the United States, it is no exaggeration to think that privacy can be infringed if it can provide an excellent customer experience. In cricket, that is, in Europe, it is important not to infringe on individual rights such as the right to be forgotten. In any case, in these areas, it seems to understand that AI artifacts should be under human control. On the other hand, in *Ya-Kyu*, that is, in Japan, the attitude toward hybrids is affinity.

For example, in Japan, industrial robots are given names (for example, names of female idols such as Momoe/百恵, Junko/淳子 and so on). This is because, like humans, robots are considered "comrades". In Buddhism, it is considered "all things have the Buddha nature/一切衆生悉有仏性". Therefore, AI artifacts are also considered to be equal to humans and have little resistance to accepting AI artifacts as friends.

5. CONCLUSION

AI artifacts differ in nature from traditional IT artifacts. AI artifacts (1) merge with natural objects and the real world, and (2) come to include organizational context. Therefore, the danger of producing unintended results cannot be denied. This is a reason why AI artifacts need to be monitored and controlled. The following agendas can be pointed out as specific monitoring and control issues.

As it has been described above, we have definitely confirmed that more research is urgently needed to explore a variety of new phenomena of AI artifact monitoring and control.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP 17K03909 and JP20K01899, the Japan Society for Management and Information Grant-in Aid for SIG "Monitoring and Control of AI Artefacts", and the Seikei University Grant-in-Aid 2020 for the research on "Monitoring and Control of AI Artefacts: Consideration from Economic, Social and Legal Perspectives".

REFERENCES

Arai, N. (2017) Could an AI pass the entrance exam for the University of Tokyo? (Ted presentation) https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam?language=ja (2020.3.31. access)

Barnard, C.I. (1968). *The Functions of the Executive*. Harvard university press.

Braverman, Harry (1974) *Labor and monopoly capita*l. New York: Monthly Review.

Frey, C.B., & Osborne, M.A. (2017) The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254-280.

Kawai, Y.(2020). Attribution of Causality and Responsibility to a Robot. *Journal of the Robotics Society of Japan*. 38(1), 32-36 (in Japanese).

Leonardi, P.M. (2012). *Car crashes without cars: Lessons about simulation technology and organizational change from automotive design*. MIT Press.

Maksimov, Y.V., Fricker, S.A. and Tutschku, K. (2018). Artifact Compatibility for Enabling Collaboration in the Artificial Intelligence Ecosystem. *International Conference of Software Business*, pp.56-71.

Ninomiya, Y.(2020). A Direction of Automated Vehicles for Social Implementation. *Journal of the Robotics Society of Japan*. 38(1), 47-51(in Japanese).

Ohshita, J. (2019). 3D Printer, Human Tissue Using Cells, Evaluating Ricoh and JSR's New Drugs, including Blood Vessels and Dura, Nihon Keizai Shimbun Morning Edition, August 26, 2019. (in Japanese).

Orlikowski, W.J. & Iacono, C.S. (2000). The Truth Is Not Out There: An enacted view of the digital economy. Brynjolfsson, E. and Kahin, B. eds. *Understanding the Digital Economy: Data, Tools, and Research*. MIT Press, 352–380.

Orlikowski, W.J. & Iacono, C.S. (2001). Research commentary: Desperately seeking the "IT" in IT research: A call to theorizing the IT artifact. *Information Systems Research*, 12(2), 121-134.

Rankin, T. (1987). The Turing paradigm: A critical assessment. *Dialogue*, 29(2-3), 50-55.

Schön, D. A. (1983) *The Reflective Practitioner: How Professionals Think in Action*, Basic Book

Simon, H.A. (1969) *The Sciences of the Artificial*. MIT Press, Cambridge, Mass

Tuncali, C.E., Ito, H., Kapinski, J. and Deshmukh, J.V. (2018). Reasoning about safety of learning-enabled components in autonomous cyber-physical systems. *2018 55th ACM/ESDA/IEEE Design Automation Conference* (DAC), pp. 1-6.

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., & Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature cell biology*, 9(6), 654. [https://www.nature.com/articles/ncb1596 (2020.3.31. access)].

POST-TRUTH SOCIETY: THE AI-DRIVEN SOCIETY WHERE NO ONE IS RESPONSIBLE

Tatsuya Yamazaki, Kiyoshi Murata, Yohko Orito, Kazuyuki Shimizu

Meiji University (Japan), Meiji University (Japan), Ehime University (Japan), Meiji University (Japan)

tyamazaki@meiji.ac.jp; kmurata@meiji.ac.jp; orito.yohko.mm@ehime-u.ac.jp; shimizuk@meiji.ac.jp

ABSTRACT

This study deals with a post-truth society, which would advent due to the widespread use of artificial intelligence (AI)-based information systems using machine learning methods such as deep learning. In that society, the truths about individuals, groups, organisations, communities, society, nations and the world would become meaningless or worthless, and the situation surrounding the four factors that erode accountability in computing – many hands, bugs, blaming the computer or the computer as a scapegoat and ownership without liability (Nissenbaum, 1996) – would become worse due to the unpredictability and uncontrollability of the behaviour of AI-based systems, leading to the lack of responsibility and accountability in AI computing. To prevent the emergence of the post-truth society and regain responsibility and accountability in computing, everyone – not only ICT engineers but also end-users – has to acquire the sufficient knowledge and skill for good computing practices, in particular the ability to consider socially and ethically, through undergoing well-organised ICT educational programmes.

KEYWORDS: post-truth society, AI-based systems, unpredictability, uncontrollability, responsibility, accountability.

1. INTRODUCTION

This study deals with a post-truth society, which would advent due to the widespread use of artificial intelligence (AI)-based systems using machine learning methods such as deep learning. In that society, people would be encased in filter bubbles (Pariser, 2011) in various aspects of their everyday and social lives where what they know is unconsciously controlled by machine learning algorithms, and thus it would become very difficult for them to discover the real truth about the world. It’s well known that personalised political advertisements delivered by Cambridge Analytica at the US presidential election and in the UK national referendum on membership of the EU in 2016 have allegedly contributed to the advent of post-truth politics and the resultant social fragmentation, although many have cast doubt on the effectiveness of the ads used to control voting behaviour. However, the wave of ‘post-truth’ ripples across society and individual lives, as well as politics.

In fact, information people can acquire in their everyday lives tends to be controlled by AI-based information systems which analyse large-scale personal databases to provide individual users with pseudo-personalised data services. Search results, postings and ads individuals view online have already been pseudo-personalised. Such data services are intended to steer individual behaviour in a way that is convenient for organisations which operate those systems. As people become increasingly dependent on the systems in terms of their information acquisition and decision making, people's thought, speech and behaviour would strongly be affected by algorithms and data used in the AI-based systems, and ultimately the systems would determine what people can know and create people's own pseudo-personalised truth.

Additionally, it has become hard for an individual to successfully control his/her identity, because information on him/her created by AI-based systems, which might actually contain stigmatic one, remains accessible online and/or in organisational databases for long periods of time, and many of others who access it can easily believe in the contents of it as the reality of him/her regardless of whether they are true or not. The truths about individuals, groups, organisations, nations and so on would become meaningless or worthless for society resulting in the emergence of the post-truth society, and people would be forced to live their post-truth lives in despair. An actual example of an AI application which functions as a threat to personal identity is one to create a deepfake, a doctored video in which a person can be made to appear as if they are doing and saying anything (Cook, 2019a). Many people including politicians and famous figures have become victims of the AI applications to masterfully edit deepfakes, being distorted their digital identities. The utilisation of this sort of AI software which can be used to conceal the truth and replace it by fakes could threaten democracy and suppress individual freedom. When it comes to deepfake porn videos, the AI applications could lead to curtailing freedom of expression and violating human dignity – especially of women – although some take a negative attitude towards regulating such contents, ironically on the ground of the protection for freedom of expression. Eventually, deepfake AI applications have not been effectively regulated so far, whereas technological efforts to fight against deepfake videos are continued (Kemeny, 2018). Here, a serious problem is that it is very difficult to find people responsible for the victims' damage created by deepfakes (Cook, 2019b).

The difficulty in clarifying the locus of responsibility is quite characteristic of the post-truth society. In this society, AI-based systems tend to function as black-boxes because their autonomous behaviour based on machine learning is not only unintelligible but also unpredictable and uncontrollable even for engineers who engage in the development and operation of the systems. When the systems are networked and work with other AI-based systems, the unpredictability and uncontrollability can be exacerbated. In addition, free/libre and open-source software (FLOSS) is often incorporated in the systems. Consequently, it is not unusual that it's very difficult to decide who is responsible, accountable and/or liable for harm caused by operations of AI-based systems. However, we cannot overlook such a technology-driven vacuum of responsibility/accountability in society.

2. A VACUUM OF RESPONSIBILITY/ACCOUNTABILITY IN AI COMPUTING

2.1. Nissenbaum's four barriers to accountability in computing

The autonomous functioning of AI-based systems using machine learning techniques leads to the unpredictability and uncontrollability of the behaviour of the systems, and provides parties relevant to the development and use of the systems, such as software engineers and system developers, with a good excuse to evade their responsibility and/or accountability for harm the systems can bring. In fact, for example, it is not easy to decide who has to take a responsibility for a fatal traffic accident caused by an autonomous car. No one would be willing or able to be responsible for anything happen owing to the systems in the post-truth society.

More than twenty years ago when a computerised society centred on the Internet was emerging, Nissenbaum (1996) pointed out that there are four factors which erode and obscure accountability or answerability for failures, risks and harm computing brings about. This means that those who are involved in information system development and deployment work in an environment where it's hard to clarify the locus of accountability or answerability in computing, and thus it is difficult for them to appropriately take accountability for negative outcomes related to their work, no matter how conscientious they are. Consequently, developing and maintaining a professional attitude in the field of computing are extremely difficult.

According to her, the four barriers to accountability or answerability in computing are as follows:

(a) Many hands: Information systems are developed not by single programmers working in isolation but by groups or organisations. Such groups or organisations are composed of various people with a diverse range of skills and expertise such as designers, engineers, programmers, managers and salespeople. Consequently, when a system gives rise to harm, it's hard to identify who is accountable.

(b) Bugs: It is commonly recognised that bugs – a variety kinds of software errors including modelling, design and coding errors – are inevitably exist in a computer system, especially as it grows larger in scale. Therefore, harm and inconveniences caused by bugs can't be helped, and it is unreasonable to hold programmers, system engineers and designers to account for imperfections in their systems.

(c) The computer as scapegoat: When some kind of error or damage occurs, the computer systems, not human agents, associated with it are blamed. This would result in underestimating human agents' roles in and responsibility for it, and end up in the situation where none is accountable for an error or a damage.

(d) Ownership without liability: The software industry tends to demand maximal protection of the property rights to their products while denying accountability, as well as liability, for any harm their software would bring to the extent possible.

We are now in the early days of an AI-driven computerised society. However, the situation surrounding the four factors has become worse rather than better with the development and spread of information and communication technology (ICT) centred on AI.

2.2. Possible controversial scenarios

Let us consider the following possible scenarios, which would be or have been realised by introducing AI-based systems:

(a) An autonomous car killed a pedestrian. However, the growing use of driverless cars has dramatically cut down on traffic fatalities.

(b) A robot security guard accidentally killed a burglar. However, the introduction of security guard robots has highly enhanced security at office buildings and alleviated a chronic shortage of nightwatches.

(c) A robot soldier killed an opposing human soldier. Since the setting up of robot troops, the number of war dead has sharply decreased.

(d) An AI-based advertisement delivery system sent an irrelevant ad to a person based on customer profiling.

Much controversy seems to exist over Scenarios (a) – (c). Each homicide committed by the AI-controlled autonomous robot conflicts with Isaac Asimov's first/zeroth law of robotics that states a robot may not injure a human being/humanity or, through inaction, allow a human being/humanity to come to harm, whereas the homicide may be justified from a utilitarian perspective. On the other hand, Scenario (d) may seem less controversial. Receiving an irrelevant and unsolicited ad is annoying for anyone, but one just has to ignore it. However, the irrelevancy of the ad may mean incorrect personal profiling of the person was conducted, and this can lead to the distortion of his/her digital identity which would cast a negative influence over his/her life for years.

The feature common to the four scenarios is that it's very difficult to clarify who is responsible to what extent. Behind this is the unpredictability and uncontrollability of the behaviour of AI-based systems and the worsened situation surrounding Nissenbaum's four factors in the age of AI.

2.3. Obscured accountability due to the usage of FLOSS as a programme module

The recent circumstances surrounding responsible development and use of ICT have been complicated. One of the causes which have brought about the complication is the unpredictability and uncontrollability of the behaviour of AI-based systems mentioned above. Another cause is the widespread use of free/libre and open source software (FLOSS). In general, the quality of FLOSS is believed high on the ground of Linus's Law, which asserts 'given enough eyeballs, all bugs are shallow' (Raymond, 1999). However, this belief was questioned when serious bugs, the Heartbleed and Freak bugs, were discovered in the OpenSSL cryptographic software library in the mid-2010s. The discovery of those bugs revealed the fact that it was hard to ensure enough eyeballs in the processes of developing and revising this widely-used open source software (Yadron, 2014). Nevertheless, FLOSS is widely used in AI-based systems as a core programme module. Actually, for example, Hadoop and Spark have been incorporated into many of those systems for big data processing and machine learning. Additionally, it is not unusual that the source codes of programmes for AI-related data processing developed by for-profit ICT companies are disclosed so that the further development of the programmes can be conducted as a FLOSS project.

Modular design of computer programmes, which has been adopted in software development for a long time as a standard software design concept, has also contributed to the complication. This design concept assumes that a computer programme is a set of modules which are functionally independent with each other. The adoption of the concept is expected to make a software development faster, less expensive and more secure, owing to the reusability of software modules whose quality and safety have been demonstrated, despite the unfortunate accidents caused by bugs hidden in the reused software module of Therac-25 (Leveson & Turner, 1993). Therefore, using FLOSS as a software module is considered as a good way to ensure the high quality and low-cost development of AI-based systems. This means that many people who are not necessarily personally identified can contribute to the development of AI-based systems, and FLOSS in which bugs are hidden can be incorporated into those systems. In addition, FLOSS providers usually disclaim responsibility for any damage or harm brought by the use of the software. Consequently, many hands and bugs still remain as barriers to accountability in AI computing in a more serious fashion, and it is extremely difficult to fill the vacuum of accountability in AI computing.

2.4. Scapegoated end-users

Those who engage in the development of ICT-based products and services including FLOSS, AI-related technologies and social media seem to be compelled to shift the responsibility regarding the quality of them to end-users, because the responsibility is too heavy to bear. In fact, online service users are required to agree to a detailed terms and conditions imposed by service providers prior to using the services. Such an informed consent scheme enables providers of online services to bear no responsibility for any trouble their users would face while using the services and to avoid the associated litigation risks. In addition, they can attribute responsibility for negative impacts or harm caused, for example, by personal data leaks, flaming and the spread of disinformation to end-users. Computing professionals working for such service providers can free themselves from accountability in computing. However, those who can fill this vacuum of accountability are only those computing professionals.

2.5. Autonomous systems as scapegoat

Pasquale (2015) pointed out that society has been becoming a black box due to the use of cutting-edge ICT such as Internet of Things (IoT), big data, AI and robots. The confusion of responsibility and accountability in computing seems to already be intractable. In particular, the operation of autonomous systems into which AI technology is incorporated would lead responsible people to claim that ‘it’s the system’s fault’ to evade their accountability when it causes harm. The unpredictability and uncontrollability of the behaviour of AI-based systems would create an opportunity to justify this claim and promote relevant people’s attitude of dodging responsibility by shifting the blame to those systems. These tendencies would become stronger, when an AI-based system is networked and works with other AI-based systems.

Needless to say, AI-based systems cannot become responsible or accountable agents even if they behave completely autonomously. It seems to be reasonable that the owners of such systems take responsibility in system behaviour. However, it’s quite usual that users of a system are forced to accept absolving its owner from his/her responsibility and using the system on their own responsibility in advance of using the system. This means that the barrier of ownership without liability exists in AI computing.

Autonomous robots controlled by AI-based systems will increasingly be used in various places such as production sites, offices, hospitals, nursing homes and schools, working symbiotically with people. However, if no one can take any responsibility in malfunction of those robots and resultant property destruction and physical or mental harm as well as in unexpected harm, our future society in which the truth of the malfunction or harm has no meaning and no value would entail serious risks.

2.6. Changes in the meaning of a bug

Bugs hidden in AI-based systems may exert significant negative impacts on people's everyday and/or social lives, given the increasingly pervasive use of such systems. The trouble is that it's unclear who is responsible to explain the circumstances surrounding such negative impacts and who is liable to compensate for the resultant losses. It seems to be necessary to reconsider the meaning or definition of a bug in AI computing.

That is, even if there is neither logical nor coding error in the programmes of an AI-based system, we need to consider that there exist bugs when the system behaves in a manner that the developers of it have not intended and expected and the behaviour harms people, society and/or the environment. This type of bug may remain hidden, or the elimination of it may be prohibitively costly. If this is the case, the only possible way of debugging is to stop operating the system. This seems to undermine the value of the systems. However, continuous operation of such an AI-based system ignoring harm it brings would far more force down the value of it, and lead to losing the public's interest in AI. Responsible operation of AI-based systems is the only way of preserving the social value of them.

3. REGAINING RESPONSIBILITY/ACCOUNTABILITY IN COMPUTING

The risks entailed in the emergence of the post-truth society, where the truth has become less meaningful and worthy and no one is willing or able to accept his/her responsibility, demonstrate the social significance of the accountable management of AI artefacts through properly monitoring and controlling them, though this is really a tough challenge. However, if such management is failed, we would face the disruptions of social lives of individuals, the erosion in local communities, social fragmentation and the ruin of democracy, because AI-based systems are increasingly exerting significant influences over what we can know about our friends, acquaintances, communities, society and the world.

It is unrealistic and impractical to provide AI artefacts with legal personality and to question their responsibility. Instead, of course, organisations and/or individuals which engage in the development and use of AI systems should take their responsibility and accountability for the technological and social quality of them. Nowadays, a large majority of ICT-based system developments and operations are conducted by business organisations. The speed of ICT developments is very fast often being referred to as dog year or mouse year, and cutting-edge ICT is rapidly deployed by business organisations without disclosing sufficient information about the deployment because it is conducted in a competitive environment. Therefore, unless business organisations develop and use ICT based on the idea of 'ethics by design' taking their responsibility and accountability to the current and future generations, responses to the harm brought about by novel ICT can be made only afterwards. Only those working for organisations which engage in the development and operation of ICT-based systems can proactively address the risk of harm the operation of the systems would bring. People outside the organisations can just respond to ethical – not to speak of legal and technological – issues related to the development and use of ICT.

However, major players in the ICT industry who lead the development and use of cutting-edge ICT including AI technology seem not to willingly take their responsibility commensurate with the tremendous impact of their business activities on society. Actually, many ICT companies have maintained an attitude of 'innovative first, consider consequences afterwards'. But, if they fail to behave as professionals, their development and use of cutting-edge ICT may bring about serious social harm.

It is alleged that, in the current computerised society, there is a chronic shortage of qualified – well-trained and high-skilled – ICT engineers. As AI-based systems penetrate into society and economy, such engineers are expected to play a pivotal role in proactively addressing ethical and social issues which can be caused by AI-based systems. However, it is not easy for them to fill such a role, because of the troublesome features of AI-based systems – the unpredictability and uncontrollability of their behaviour – and the resultant obscured locus of responsibility and/or accountability in AI computing. In addition, the majority of them work for for-profit organisations, whose working environment often make it hard for them to develop their professional outlook (Murata, 2013). These suggest that not only software engineers, who have been required to build up an attitude of professionalism, but a wider range of people who are involved in computing, including end-users, need to accept their professional responsibility depending on where they stand in the AI-driven information society. In order to prevent the advent of the post-truth society, the attitude we need to develop is 'everyone has to take his/her respective responsibility in computing'.

For a wide range of people to cultivate such an attitude, appropriate ICT educational programmes must be developed. The contents and methods of the education have to be carefully examined and regularly revised, given that ICT engineers, let alone end-users, have not necessarily studied computer science and engineering at their schools and that their skill and knowledge have to be continuously renewed in accordance with the rapid advancement of ICT. The ability to consider socially and ethically has to be acquired by everyone through undergoing the educational programmes. These cannot be effective only for particular people, groups, organisations, communities and countries. In this respect, the educational programmes should be developed and revised by a non-profit body independent from any for-profit organisation and government agency, and setting up a system to issue various levels of ICT professional licences authorised by the body may be effective to encourage people to develop their professional outlook in computing suitable to their positions.

Even if engineers who engage in the development and operation of ICT-based information systems have sufficient technological knowledge and skill as ICT professionals, their lack of the knowledge and skill, as well as work habits, to ethically and socially consider would lead to serious social harm caused by their well-meaning development and operation of information systems. As a practical matter, however, it's not necessarily so easy for engineers to develop and maintain their professional outlook and appropriately address ethical and social issues. Even those ICT engineers who are well-trained and full-fledged to behave as responsible professionals would encounter difficulties in accurately predicting and properly dealing with the long-term social consequences, as well as even the immediate social impacts, of their development and operation of information systems. ICT engineers are required to humbly face up to their ineludible cognitive and intellectual limitations.

4. CONCLUSIONS

This study has examined ethical and social issues we would need to address in a post-truth society, which would advent due to the widespread use of AI-based systems using machine learning methods such as deep learning. In that society, the truths about individuals, groups, organisations, communities, society, nations and the world would become meaningless or worthless, and the situation surrounding the four factors that erode accountability in computing would become worse due to the unpredictability and uncontrollability of the behaviour of AI-based systems, leading to the lack of responsibility and accountability in AI computing. To prevent the emergence of the post-truth society and regain responsibility and accountability in computing, everyone – not only ICT engineers but also end-users – has to acquire the sufficient knowledge and skill for good computing practices, in particular the ability to consider socially and ethically, through undergoing well-organised ICT educational programmes.

We are now experiencing a hard time due to the coronavirus (COVID-19) disease pandemic. One of the most serious problems with the pandemic is a lack of accurate information as to the characteristics of the new virus. Some people who occupy professional and responsible positions in healthcare or infectious disease prophylaxis have provided inaccurate and/or wrong information about the disease. However, no one seem to have taken accountability for their misinformation delivery. Medical policies to prevent the spread of the disease differ from community to community as well as from country to country, causing general confusion as to how people can prevent themselves from being infectious. Many lay people freewheelingly deliver their irresponsible criticism to those policies online, and spread questionable or false information on the disease using social media.

The current messy situation surrounding the coronavirus is similar to the post-truth society depicted in this paper in terms of the meaninglessness and worthlessness of truths and the absence of responsible and accountable people. The pandemic will end when an effective therapy is established or a specific medicine is developed. However, the widespread use of AI-based systems will continue to be expanded due to the irresistible convenience those systems provide to the general public, although we cannot expect to have a specific cure for the social pathology which comes into existence in the post-truth society.

ACKNOWLEDGEMENTS

This study was supported by the Japan Society for Management and Information Grant-in-Aid for SIG ‘Monitoring and Control of AI Artefacts’, the Seikei University Grant-in-Aid 2020 for the research on ‘Monitoring and Control of AI Artefacts: Consideration from Economic, Social and Legal Perspectives’, and JSPS Grants-in-Aid for Scientific Research (C) 20K01920, 19K12528 and 17K03879.

REFERENCES

Cook, J. (2019a, June 12). Deepfake videos and the threat of not knowing what’s real. *Huffpost*. Retrieved from https://www.huffpost.com/entry/deepfake-videos-and-the-threat-of-not-knowing-whats-real_n_5cf97068e4b0b08cf7eb2278.

Cook, J. (2019b, June 23). Here’s what it’s like to see yourself in a deepfake porn video: there’s almost nothing you can do to get a fake sex tape of yourself taken offline. *Huffpost*.

Retrieved from https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0faee4b0a3941861fced.

Kemeny, R. (2018, July 10). AIs created our fake video dystopia but now they could help fix it: new software developed by artificial intelligence researchers could help in the fight against so-called deepfake videos. *Wired*. Retrieved from https://www.wired.co.uk/article/deepfake-fake-videos-artificial-intelligence.

Leveson, N. G. & Turner, C. S. (1993). An Investigation of the Therac-25 Accidents. *IEEE Computer*, 26 (7), 18-41.

Murata, K. (2013). Construction of an Appropriately Professional Working Environment for IT Professionals: A Key Element of Quality IT-Enabled Services. In Uesugi, S. (ed.), *IT Enabled Services* (pp. 61-75). Wien: Springer.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(1), 25-42.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Raymond, E. S. (1999). *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, CA: O'Reilly Media.

Yadron, D. (2014, April 11). Heartbleed Bug's 'Voluntary' Origins: Internet Security Relies on a Small Team of Coders, Most of Them Volunteers; Flaw Was a Fluke. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/programmer-says-flub-not-ill-intent-behind-heartbleed-bug-1397225513.

REDISCOVERY OF AN EXISTENTIAL-CULTURAL-ETHICAL HORIZON TO UNDERSTAND THE MEANINGS OF ROBOTS, AI AND AUTONOMOUS CARS WE ENCOUNTER IN THE LIFE IN THE INFORMATION ERA IN JAPAN, SOUTHEAST ASIA AND THE 'WEST'

Makoto Nakada

University of Tsukuba (Japan)

nakada.makoto.ga@u.tsukuba.ac.jp

ABSTRACT

In this paper I will make an attempt to find (rediscover) the potentially broader or alternative horizon to understand the meanings of robots, AI and autonomous cars which we encounter in the life in the information era in Japan, Southeast Asia and the 'West.' In this case, the broader or alternative (horizon) refers to the situations which would be beyond the narrowly interpreted views on human life under the strong influence of techno-determinism, Cartesian dualism of body and mind, the Western presupposition to put an emphasis on the limited aspects of human existence (i.e. rationality, intelligence separated from our existence in time, phronesis, bodily existence and vulnerabilities of life). I will make this attempt mainly by focusing on my researches performed in the past decade in Japan, Southeast Asia and some of the Western countries and also by examining the related discussions on this kind of encounter of technologies and our existence. The point in this paper will be that robots, AI and others are found to belong to the realm of technologies, the logical thinking and objective scientific methodologies and also to the realm of existence including the 'profound' relation between the logos and awareness of finitude of life.

KEYWORDS: existential horizon, robots, AI, autonomous car, meanings of life and death.

1. INTRODUCTION

In this paper I will make an attempt to find (rediscover) the potentially broader or alternative horizon to understand the meanings of robots, AI and autonomous cars which we encounter in the life in the information era in Japan, Southeast Asia and the 'West' beyond the limited perspectives combined with the techno-determinism, Cartesian dualism of body and mind and the Western presuppositions on rationality, intelligence and reasoning based on logical languages as the first principles for human existence.

I will make this attempt mainly by focusing on the following points. 1) First, we will see the research findings which I gained by my researches performed in the past decade in Japan, Southeast Asia and some of the Western countries. Through this analysis we will see that people in Japan, Asia and the West tend to understand or evaluate the meanings of life by depending on their existential-cultural-ethical perspectives on 'what is a good-virtuous life?' And our research data show that these ethical and existential ways of thinking and feeling about the meanings of our life in the modern, complicated and informatized environments are found to be correlated with people's views on the meanings of robots, AI and self-driving cars in our life.

This suggests us that the robots, AI and self-driving cars would enter the inside of our life through a pathway or in-between place where some sort of important mediating procedures might work. To put this in another way, the interpretation of the findings of my research, i.e. the correlation between people's views on life and their views on the technological products, would presuppose the presence of this kind of pathway to enable two different kinds of existence, one related to the humans beings and another related to the technological products (and the scientific/ logical reasoning/ rational thinking) would gather together.

2) Secondly, I will examine the cultural and existential background lying behind these research data. In spite of the general belief that Japanese people of today tend to show their strong interest in the material aspects such as the material wealth, the welfare associated with material happiness, degree of one's social reputation as measured by educational background and career advancement, the ranking of international competiveness based on GDP or consumption of cars or other commodities, what I have found in my researches is that Japanese people's minds of today are still occupied by various sorts of cultural-existential views on this life with cultural-historical backgrounds. These backgrounds seem to be associated with the history of Japanese interpreting of Confucianism into the direction of 'internalization,' with *Kokugaku* (indigenous cultural studies on the Japanese culture and language in the Tokugawa Era focusing on Japanese 'inner and fluid' cultural traditions in contrast to Chinese formalized and stable traditions), *Shinto*, Buddhism or the language improvement such as *Genbunicchi* Movement in the Meiji Era (= i.e. the movement in the Meiji 20s (1887–96) for unification of the written and spoken language needed for modernized internalized world) and so on.

3) Thirdly, I will examine 'why can this kind of cultural and ethical meanings be found in other Asian countries and even in some Western countries?' What I have found is that people in other Asian countries and some Western countries tend to show some sort of sympathy for the views on the meanings of 'requiem service for broken tools or robots' or 'beauty of our life or the world through its transience(*Mononoaware*)' and other views which seem to be related with Japanese cultural traditions. I will try to interpret this phenomenon, the potential cultural-existential meanings in the minds of Asian people and Western people, by examining the related discussions by various authors.

2. JAPANESE PEOPLE AND EAST ASIAN PEOPLE HAVE POTENTIAL CULTURAL-ETHICAL ATTITUDES TOWARD THE PROBLEMS IN THE MODERNIZED SOCIETY

2.1. Backgrounds of Japanese cultural and ethical views on life

In this section, we want to see the main points of my research findings and their potential implications. In order to do so, we need some preliminary explanation or interpretation on the characteristics of Japanese cultural and ethical views on life.

The important findings through my researches (as shown in Table 1) is that Japanese people of today are found to show a strong or fairly strong empathy with the views on life in the modernized and informatized environments, i.e. the views reflecting their cultural and existential experiences in the realm of the world which remains being a kind of 'inner world' or 'intermediate world.' In my view, this 'inner world' or 'intermediate world' is mediating two aspects of our life, i.e. the aspect of life associated with the logos, scientific understanding, technological procedures, the objective observation of the various problems in the society and the realistic approach to the those problems and so on and the aspect of life associated with the cultural and ethical evaluation on people's life.

For example, according to one of my researches in Japan (the one performed in 2020 which is in the process of analysis at this moment), people tend to show the sympathy with the view, 'The value of small money such as 10 Yen or 100 Yen depends on how you use the money and in this sense the small money symbolizes your personality reflected in its use.' And the degree of sympathy with this view is found to be correlated with the degree of interest in or sympathy with the other views including the views related with the meanings of job, the meaning of sharing struggles with colleagues in the same workplace, the sympathy with victims in the accidents or disasters.

And this means that various matters are thought to be located in some place in people's mind and they consist of a sort of horizon to make various matters be associated with each other, i.e. those matters coming from various aspects of life (business, human relationship, sympathy with other people's tragic situations, encountering with the technological products and so on). In this sense, this horizon is related to a wholeness of our life. The meaning of 'ethical' or 'cultural' is thought to be related to this kind of wholeness of life.

This doesn't necessarily mean that Japanese people of today or in the modern era are not/have not been interested in the material aspects of life. But rather, the important point about this is that the material wealth is thought to have its own ethical, virtuous or cultural dimension along with its realistic and material aspect. To put this in a more simple form, Japanese cultural-existential attitudes are characterized by its plurality. The money has its objective meanings and its 'inner' meanings. The social systems have their own formalized, standardized and rule-based meanings and their ethical aspects. The AI machine for diagnosis of latent disease has its own scientific/ practical meanings and at the same time it makes people feel some sort of uncomfortableness as a result of the transformation of their existence from 'Wer' (who) to 'Was' (what) (concerning Wer and Was in the informatized environments, see Capurro et al., 2013).

This kind of plurality might not be limited to Japanese culture(s). But it seems that Japanese cultural traditions are /have been very sensitive to this plurality and in this sense the awareness of this plurality seems to be an important part of self-identity for Japanese people.

The discussions by various authors seem to reflect or influence this point. The eyesight of Kitaro Nishida, which is thought to reflect or influence Japanese views in general, seems to point directly to the intermediary place that works between subjectivity and objectivity. And some scholars in the research area on Japanese language pay attention on the emerging place between *Shi* (noun) and *Ji* (suffix). This emerging place as somewhere in-between would be also the one enabling people to reconsider the relation between 'the logic and the imagination,' 'the subject and the object' or 'the logic of subject and the logic of predicate' in a more broad way. (The logic of predicate is a kind of logic or ways of thinking based on association of matters, words or imaginations.)

The use of expressions by adjectives can be found to be located in this intermediate place. For example: Matsuri (festivals) no (of) koro(season, time) ha (at or is) ito (very) okashi(fun). (The season of festivals is very fun.) According to Fukada, who is citing the interpretation of Japanese linguist Motoki Tokieda, 'fun' simultaneously refers to the state of the subject and the state of the object. The logic of predicate seems to be related to the work of *Shi* (noun) and *Ji* (suffix) in

Japanese language characterized by this kind of mediating work. If we follow Akira Mikami's explanation (Mikami 1960 :118), 'ha' in the sentence 'Zou ha hana ga nagai' is not a grammatical copula such as 'is' but shows a role to present a theme regarding Zou (elephant) or a theme including Zou as part of theme. This is another case of Japanese plural ways of thinking. In the case of disasters, when we say, 'the disasters are due to the punishment from heaven,' similar things happen. 'Disaster's is the subject which dominates a certain line of topic such as 'this disaster happened by the work of strain of Pacific Plate.' At the same time, 'Disaster' is the topic which people can talk about following the cultural or ethical association. The views in Table 1 seem to be similar in this sense, i.e. the combination of 'the concrete, objective, descriptive aspect (expression) of the matter' and 'the imaginative, associative and allegorical aspect (expression) of the matter.'

2.2. Japanese views on life

The following table (Table 1) shows that Japanese people tend to think about the meanings of the matters and things they face in the everyday life in the informatizing environments by following some sort of existential-cultural-ethical perspectives which seem to reflect their cultural-historical traditions. (The views used in these researches are adopted from examinations through (a) the qualitative researches on the characteristics of Japanese cultures and values based on discussions by authors such as Kitaro Nishida, Tetsuro Watsuji, Hideo Kobayashi, Daisetsu Suziki, Motoki Tokieda, Bin Kimura, Yujiro Nakamura and others, (b) the findings of my depth interviews with Japanese and Asian people, (c) a content analysis on newspaper reports and magazine reports about Japanese attitudes toward disasters and wars and (d) the discussions with some Western scholars such as Rafael Capurro.)

These data in this table show that people in Japan are still under the strong influence by the existential-cultural-ethical perspectives which seem to derive from Buddhism, Confucianism, Taoism, *Shinto* (in Japan), *Kokugaku.* In fact, as the table shows, the percentage of the respondents with the positive attitudes toward the views (measured by the degree of acceptance of or sympathy with the contents of these views) is surprisingly high in most of the cases. For example, the percentage of people showing the positive attitudes toward the view, 'People will become corrupt if they become too rich' is 75.7% (2018HG research) and 83.7% (1995G research). Similarly, the majority of Japanese people (respondents) show sympathy for various ethical and cultural views on 'destiny,' 'natural disasters as warning from heaven,' 'denial of natural science' and others.

But this doesn't necessarily mean that Japanese people live (only) in a traditional, 'feudalistic' or even 'superstitious spiritual' world. My interpretation is that Japanese mind consists of two different but/and coexisting aspects in some ways (as I suggested above), i.e. the combination of 'objective and realistic world views' and 'ethical, existential or "existential categories based" world views.' 'Existential categories' is the term used to explain Heidegger's explanation about the difference between two ways of understanding, knowledge based on metaphysical or scientific categories (or a priori concepts) and knowing or being awareness based on existential categories(these are a priori too by Heidegger). (We can get a useful hint to this point by Michael Gelven's critical reviews on Heidegger's *Being and Time* (Gelven, 1970).)

And in addition to this finding, we have found that people's attitudes toward the meaning of robots, AI and others are found to be (fairly) strongly correlated with these cultural and ethical views in the minds of Japanese people today.

2.3. Finding of cultural-existential views on life in Japan

The following table (Table 1) shows that, as I suggested above, the 'inner' part of Japanese minds today is still filled with sympathy for the existential and cultural views on the matters they face in the concrete environments, i.e. the views reflecting or determining the existential and cultural interpretation of those matters.

And this (the experience corresponding with the cultural, existential views) seems to also reflect the 'dual' structures (aspects) of the world itself and people's inner structure(s) being associated with this (experience), as suggested by Valera, Merleau-Ponty, Jyunichi Murata and others.

Our experiences in this world are supported by two different but internally corresponding aspects. When we see the blue sky, this experience of blue colour depends on the material aspects which reflect the laws of science. But the experience of 'blue' of the sky depends on another aspect, i.e. our interpretation of the matter or our position in the world including material aspect. The brown colour and yellow colour are the same in regard to the spectral component of light but they become a different colour through arrangement of light and the object or the positional relationship between illumination light and illuminated object.

In the case of social matters, this kind of dual situations might be observed too. This is what Heidegger, Husserl and Merleau-Ponty insisted on in their writings. And some Japanese authors such as Kitaro Nshida, Tetsujiro Watsuji and others also discuss similar points.

Table 1. People's sympathy for values deriving from their shared traditional cultural-existential-ethical experiences in Japan, Asian countries and the West.

	1995G (Japan)	2000G (Japan)	2016 HG (Japan)	2018 HG (Japan)	2003T (Germany) (N=300)	2018 Indonesia (N=250)	2017 Vietnam (N=300)
Distance from nature	73.6%	-	68.9	69.6	83.1%	82.4	91.0
Honest poverty	83.7	81.5	75.9	75.7	68.7	81.6	87.0
Destiny	84.4	79.0	76.0	75.5	35.7	85.6	82.3
Denial of natural science	88.5	88.3	80.2	82.9	60.6	89.2	96.4
Criticism of selfishness	85.5	88.3	75.6	77.5	89.9	80.8	89.0
Powerlessness	71.9	64.8	71.3	72.3	57.5	-	-
Superficial cheerfulness	73.3	65.6	71.3	72.8	40.1	78.8	86.4
Belief in kindness	-	68.1	65.2	68.7	91.5	86.4	94.6
Scourge of Heaven	62.7	49.5	-	-	-	-	-
Natural disasters as warning from heaven	-	-	56.4	58.8	-	93.6	92.7
Sensitivity to beauty through transience (*Mononoaware*)	-	-	37.2	48.5	-	56.8	69.3

1) Table1 shows the percentages of the respondents who said 'agree or somewhat agree' to various views or statements indicating people's inner values and also their critical attitudes toward phenomena or matters they face in their everyday life. These statements are such as: "Within our modern lifestyles,

people have become too distant from nature"(Distance from nature); "People will become corrupt if they become too rich"(Honest poverty); "People have a certain destiny, no matter what form it takes"(Destiny); "In our world, there are a number of things that cannot be explained by science"(Denial of natural science); "There are too many people in developed countries (or Japan, Indonesia, Vietnam) today who are concerned only with themselves" (Criticism of selfishness); "In today's world, people are helpless if they are (individually) themselves" (Powerlessness); "In today's world, what seems cheerful and enjoyable is really only superficial" (Superficial cheerfulness); "Doing your best for other people is good for you" (Belief in kindness); "The frequent occurrence of natural disasters is due to scourge of Heaven" (Scourge of Heaven); "Occurrences of huge and disastrous natural disasters can be interpreted as warnings from heaven to people"(Natural disasters as warning from heaven); "I can sometimes feel that the fireworks or the glow of a firefly in the summer are beautiful because they are transient or short-lived"(sensitivity to beauty through transience(*Mononoaware*)). 2) The data which we examine in this paper are collected from researches as follows. '1995 G'= research conducted in Tokyo in 1995(587 respondents collected through random sampling of over 20 years old). '2000G'= research done in Tokyo metropolitan area in 2000(611 respondents collected through random sampling of over 20 years old). '2016HG' is a research done in Japan in December, 2016 for 600 respondents of age 25-44 men and women living in Fukushima, Miyagi and Iwate Prefectures (quota sampling was used to design this survey, i.e. the ratios of gender and age were quoted from the official statistical report of the Japanese government about the Internet users in 2010 in Japan). '2018HG' is done in Japan in December, 2018 for 600 respondents with age 25-44, similarly as in the case of 2016HG. '2003 T'= research in 2003 in Germany with 300 students at the University of Tubingen as respondents. '2008 Indonesia' =research in 2028 in Indonesia (250 respondents of age 25-44 collected with quota sampling depending on statistics from nationwide data on the ratios of gender and age based on the official statistical report in Indonesia). '2017 Vietnam'=research done in 2017 in Ha Noi and Ho Chi Minh City in Vietnam (300 respondents of age 25-44 collected with quota sampling depending on the ratios of gender and age in Ha Noi and Ho Chi Minh City based on the official statistical report in Vietnam).

3. UNITY OF INNER VALUES AND THE CRITERIA TO EVALUATE THE MEANINGS OF MATTERS SUCH AS ROBOTS, AI AND SELF-DRIVING CARS IN THE LIFE IN THE INFORMATIZED ENVIRONMENTS

3.1. Relation between views on life and views on robots, AI in Japan

One of the most important points we can't miss when we interpret these findings in Table 1 is that these views seem to play two different kinds of roles simultaneously. Cleary these views (or the positive attitudes toward these views) are some sort of internalized values such as 'purity of minds leading to preference of simple and virtuous life' or 'pursuit to sincerity' and so on. But on the other hand, these inner values seem to play a critical standard with which people can evaluate the meanings of phenomena or matters they encounter in their everyday life simultaneously. Table 2 and Table 3 show us that this is not a mere speculation.

As these tables (1, 2, 3) show, what we have found is that there is a kind of horizon in which 'people's views on meanings of life as a whole,' 'their attitudes toward values in the society,' 'their sensitivity to others' death and sacrifice' and 'their understanding of or interpretation on the meanings of robots or autonomous cars' can come together or can become interrelated with one another. In a way, what we have found is the phenomena suggesting a surprising aspect of world, i.e. life, death, care for others, expectation or anxiety for robots would merge into an oneness.

Table 2. Correlations between Japanese people's views on robots and their ethical, existential and cultural views on life (Data: 2016HG).

	Problems of care robots	Rights for robot	Care robot for children	Auto-nomous car's judgment for life	Responsibility for autonomous car	Better safety by autonomous car
Distance from nature	**.336****	**.144****	**.106****	**.346****	**.338****	**.174****
Honest poverty	**.326****	.051	**084***	**.391****	**.439****	**.217****
Destiny	**300****	.039	**. 106****	**.326****	**.333****	**.200****
Denial of natural science	**.348****	.049	.078	**.405****	**.409****	**.216****
Criticism of selfishness	**.204****	.005	**.088***	**.294****	**.289****	**.197****
Powerlessness	**.211****	.073	**.106****	**.213****	**.184****	**.182****
Superficial cheerfulness	**.274****	**.168****	**.159****	**.330****	**.336****	**.188****
Belief in kindness	**.242****	**.125****	**.092***	**.226****	**.224****	**.148****
Natural disasters as warning from heaven	**.267****	**.235****	**.094***	**.286****	**.284****	.021
Sensitivity to beauty through transience (*Mononoaware*)	**.285****	.**195****	**.131****	**.273****	**.275****	**.220****

Notes on the Table: 1) This table shows the correlation between 'the views on robots and self-driving cars' and 'the views on Japanese "inner minds" including *Mononoaware*.' The statements showing the content of the views on robots and self-driving cars are: "To leave handicapped or elderly persons in the care of robots worsens isolation of them from societies even though this idea seems to be appropriate at first glance"(Problems of care robots); "Robots should be given similar rights in the future as fetuses or patients in a coma without consciousness or awareness"(Rights for robot); "To leave children in the care of robots would be better than to leave them alone without any care" (Care robot for children); "Although automobile driving robot by artificial intelligence seems to be convenient, considering to leave judgment on life or death to the machine, there is a problem of use without much consideration" (Autonomous car's judgment for life); "The autonomous cars (self-driving cars) by artificial intelligence seem to be convenient, but there are problems with easy use, since it is impossible to ignore the problem of responsibility associated with driving the machine"(Responsibility for autonomous car); "Automobile driven by robots with artificial intelligence will bring about a better situation in our society, because it increases safety compared to human driving"(Better safety by autonomous car).

Table 3 Correlations between sharing pity for others' death/ sacrifice and various views on robots/autonomous car and the ethical views in Japan. (Data: 2016HG)

	Problems of care robots	Virtual creatures	Care robot For children	Auto-nomous car's judgment for life	Responsibility for autonomous car	Honest poverty	Destiny	Natural disasters as warning from heaven
Flowers for lament	**.309****	**.145****	**.176****	**.182****	**.171****	**.162****	**.218****	**.265****
Being beautiful through transience	**.285****	**.277****	**.131****	**.273****	**.275****	**.268****	**.296****	**.247****
Sacrifice	**.344****	**.293****	**.100***	**.329****	**.351****	**.351****	**.414****	**.243****
Lonely death	**.291****	**.256****	**.169****	**.277****	**.216****	**.302****	**.326****	**.223****
Astroboy's final episode	**.230****	**.271****	**.248****	**.152****	**.198****	**.184****	**.213****	**.298****

1) The figures of the table show the percentage of the respondents who responded to each view affirmatively or negatively.
2)'Virtual creatures' = 'It is very natural when children sympathy or some kind of affection towards virtual creatures like Tamagotchi.' 3) 'Flowers for lament' shows 'I can imagine clearly the figures of the victims or their family when I see the flowers for lament or sorrow at the traffic accidents or other accidents.' : Similarly, 'I can sometimes feel that the fireworks or the glow of a firefly in the summer are beautiful because they are transient or short-lived.' (being beautiful through transience); 'I sometimes feel that I have to think more deeply about the important meanings of life when I hear the stories of persons who saved others at the cost of their own life in natural disasters and similar crises.' (sacrifice); 'I sometimes feel that everyone must have had their own meaningful days even if he/she died alone and his/her death is called a case of 'lonely death' in the newspapers.' (lonely death); ' I am moved when I know Astroboy's final episode as self-sacrifice for saving the earth.' (Astroboy's final episode).

3.2. Potential Existential horizon in Southeast Asia and the West

Table 4 Correlations between people's views on robots and the ethical and cultural views on life (including *Mononoaware*-views) in Indonesia (Data: 2018 Indonesia)

	Problems of care robots	Virtual creatures	Right for robots	Affection for robots	To prevent maltreatment for robots
Flowers for lament	**.158***	**.160***	**.181****	**.223****	**.186****
Sacrifice	.026	.036	**-.128***	-.074	-.012
Being beautiful through transience (*Mononoaware*)	**.205****	**.291****	**.265****	**.317****	**.206****
Awareness of importance of life through thinking about the finitude of life	.059	.039	**.169****	.067	.011

1) **=p<0.01, *=p<0.05, without ** or *=ns= non (statistically) significant
2) 'Awareness of importance of life through thinking about the finitude of life' shows 'When I am aware of the finitude of life and its transience, I feel I need to spend every day meaningful.' 3) 'Affection for robots' = 'Robots are expected to be a subject of affection or consideration in the future just as the earth, mountains, rivers are treated so, even though they have no life.' 'To prevent maltreatment for robots' = 'To provide robots with capability of expression of their emotions such as pains would be good in order to prevent (avoid) cruelty or maltreatment to them.' Concerning the other views on robots, see the note of Table 2 and 3.

As Table 1 and 4 show, the potential presence of this kind of horizon we have examined in the case of Japanese research data seems to be found in the cases of Southeast Asia data and the West (the data of researches in the West are not enough at this moment, so we have to carefully deal with these data). (The research data in Germany, Indonesia and Vietnam are shown in Table 1. This is due to the limited space and also for comparison with Japanese data.) (The research in Sweden, in the process of analysis now, shows that 60.5% of respondents affirmed their sympathy with *Mononoaware*-view.)

4. REDISCOVERY OF MEANINGS OF LIFE TO BE RELATED TO HUMAN VULNERABILITIES IN THE PLACE OF HUMAN HEALTH CARE

4.1. Care seen from the perspective associated with human existence

What we have found (rediscovered) through examining of the meanings of data of the tables shown above is the fact that various meanings of things and matters are interrelated with one another. This kind of finding seems to be able to be related with Heidegger's idea of Bewandtnis (involvement) or Kitaro Nishida's idea of *Basho* (place to connect the subject, the object, perception, thinking and others in a form of undifferentiated wholeness). The important matter of this point is that this kind of ethical or existential link of meanings is quite different from the mechanical linkage of machine parts or linkage grounded on some sort of scientific causal relations.

And if some sort of artificial machines can play a role in human environments, it seems that they need to enter this kind of linkage based on the human existence in some ways.

Or in some cases machines need some sort of new Bewandtnis (involvement) or *Basho* in order to do their role properly as they are designed, e.g. Bewandtnis related with the practice as care.

In this sense, the reports by the experts (Toombs, Todres and others) in health care or nursing under the influence of phenomenologists such as Husserl, Heidegger, Merleau-Ponty, H. Dreyfus are very suggestive. In my interpretation, what they have been trying is to bring about a new kind of views on human existence into the place of medical practice (hospital and so on) which is usually under the control of scientific determination and classification by the medical physicians. These views are expected to be more the patient-led ones or open to interpretation of illness for the patient and the care-giver. In this sense, people in the Western live in an aspect of the world where such ideas as 'care with sharing the vulnerabilities,' 'finitude of life' or 'life on a journey' would mean a lot of things. These terms and concepts are proposed by these authors showing interest in phenomenology.

4.2. The phenomena in the place between 'the object and the subject' and 'the logic and the meaning'

We know that the restrained or disabled work of mind and body shown in the case of various symptoms such as autism, schizophrenia, agnosia, aphasia and others would be observed in somewhere in-between, i.e. the place between 'the objective "outer" area of our world and the subjective "inner" area of our experience,' 'the syntagmatic linguistic area and the pragmatic linguistic area ,' 'the subject and the predicate,' '*Shi* (nouns) and *Ji*(suffix or dependent word),' 'the objective meanings of the object in the perception field and the experienced meanings of the object in the perception field(as in the case of optical illusion),' as suggested by Merleau-Ponty, Bin Kimura and others. For example, in the case of aphasia examined by Shigeyuki Kumgai, Japanese pedagogist, the children with autism have difficulty with using Japanese *Ji*(suffix such as *ga*, *ha*, *wo* and others) to complete a sentence to portray a scene such as: a little girl is bullied by a boy in the playground of a kindergarten.

In the phenomena related to a ‘sense of agency,’ we know that the states or the modes of human existence can come together with different states or modes of existence of things in a certain kind of place. This place reflects or determines the relationships among such matters as ‘human expectancy or intentionality on a conscious level,’ ‘human expectancy or intentionality on a unconscious level,’ ‘the mechanism of transmitting information needed to move the human motor systems from the interior to the exterior,’ ‘the mechanism of receiving information needed to check the expected response as a result of the operation of the muscles from the exterior to the interior’ and so on. This is another case that shows: we need to carefully see the in-between place.

The in-between place suggests us a lot of things. The mechanical matters or the objective things can ‘exist’ in various areas. It can have meanings as ‘realization’ of rules or principles. They can work within mechanical structures such as steam engines, locomotives, cars or robots. And they can enter the in-between area where the mechanical matters, the objective things, tools or machines can interact with human body structure, nervous systems and human expectancy and human sense of identification. And in the case of self-driving cars, we have to presuppose the material or technological conditions of those self-driving cars. But the self-driving car itself can’t enter the realm of our life when its material aspect is a mere material. This materiality starts to have meanings when combined with ‘good ethical questions’ which are related to the question on our own existence too. As we have seen, an amount of small money starts to have an ethical meaning on a certain horizon. The money is the ‘tool’ to satisfy human desires. But this desire would enter our human life when it is combined with our views on life. Human desires are associated with a lot of things. In the case of our research findings, how to use the small money symbolizes the whole range of our life. Autonomous cars are the similar ones too.

5. CONCLUSION

What we have found through our research data includes a lot of matters and their interpretation which would be possible or visible through a sort of circuit or pathway we mentioned above. This circuit or pathway includes the materiality of things (tools) and our potential anticipation or imagination in regard to the relation with those material matters or things. One of the important findings that we have found through our research data is that some aspects of human finitude or vulnerabilities such as sacrifice, accidents, unavoidable encountering with disasters or accidents and our sensitivity to these matters are correlated to people’s interpretation on the meanings of technological products such as robots, AI or self-driving cars in their life. Bin Kimura and Heidegger suggest us that our ways of existence are related to the oneness of truth-ness. (See also Tamura, 2012.) The truth or the fact is characterized by its oneness or uniqueness as shown in the case of the law of contradiction saying that a proposition can't be both true and false. Bin Kimura says, citing Nietzsche and Heidegger, that this is related to our way of understanding of existence or our existence itself. The confusion of Being and Becoming (Werden and Sein) is related to the oneness of truth-ness in some ways. In this sense, *Mononoaware*, Japanese sensitivity to beauty of nature or significance of human life through awareness of finitude (transience) of the world and human life, might be a way of thinking and feeling which leads us to the in-between state of Being and Becoming.

ACKNOWLEDGEMENTS

This study was supported by the Japan Society for Management and Information Grant-in Aid for SIG “Monitoring and Control of AI Artefacts,” and the Seikei University Grant-in-Aid 2020 for the research on “Monitoring and Control of AI Artefacts: Consideration from Economic, Social and Legal Perspectives” and JSPS KAKENHI Grant Number 17K03113.

REFERENCES

Capurro, R., Eldred, M. and Nagel, R. (2013). *Digital Whoness: Identity, Privacy and Freedom in the Cyberworld*. New Jersey: Transaction Books.

Capurro, Rafael (2006). Towards An Ontological Foundation of Information Ethics. *Ethics and InformationTechnology*,Vol.8, Nr. 4, 2006, 175-186.

Dreyfus, Hubert (1972). *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row.

Fukada, Chie (2004).Mibunkana imi no bunka(differentiation of undifferentiated meanings). *linguistic science* (University of Kyoto) (2004), 10: 117-147.

Gelven, Michael (1970). *A Commentary on Heidegger's “Being and Time.”* Harper & Row.

Gurwitsch, Aron(1966).Phenomenology of Thematics and of the Pure Ego: Studies of the Relation Between Gestalt Theory and Phenomenology. In *Studies in Phenomenology and Psychology*. Evanston: Northwestern University Press.

Heidegger, Martin (2001). (the original version in 1927) *Sein und Zeit.* Tübingen: Max Niemeyer Herlag.

Husserl, E. (1966). *Analysen zur passiven Synthesis* (Hua XI). Den Haag: Martinus Nijhoff.

Kimura, Bin (2000). *Guzensei no seisinn byouri*(Psychopathology of coincidence).Tokyo: Iwanami.

Kobayashi, Hideo (1979). *Motoori Norinaga*. Tokyo: Shintyosha.

Kumagai, Takayuki (1993). *Jiheisyou karano message* (messages from aphasia). Tokyo: Koudansya.

Merleau-Ponty, Maurice (1942). *La structure du comportement*. Paris: Presses Universitaires de France.

Merleau-Ponty, Maurice (1945). *Phénoménologie de la perception*. Paris: Presses Universitaires de France.

Mikami, Akira (1960). *Zou ha hana ga nagai* (Elephant is a matter with a long trunk). Tokyo Kuroshio syuppann.

Murata, Jyunichi (1995). *Tikaku to seikatsusekai*(Perception and life-world). Tokyo: University of Tokyo Press.

Nakada, Makoto (2019). Robots seen from the Perspectives of Japanese Culture, Philosophy, Ethics and *Aida* (betweenness). In Thomas Taro Lennerfors and Murata Kiyoshi (Eds.), *Testugaku Companion to Japanese Ethics and Technology*. Springer, 161-180.

Nakamura, Yujiro (2001). *Nishida Kitaro Ｉ*. Tokyo: Iwanami.

Nishida, Kitaro (1966). Tetsugaku gairon. *Nishidakitaro Zennsyuu 15*. Tokyo:Iwanami.

Suzuki, D., Fromm, E. and De Martino, R. (1960). *Zen to seishin bunseki*(Zen Buddhism and Psychoanalysis). Tokyo: Sougensha.

Tamura, Miki (2012). Sonzaitojikan ni okeru gennsonnzai no ketuisei to rekisisei nitsuite (concerning resoluteness(Entschlossenheit) and historicity (Geschichtlichkeit) of human existence (Dasein) in Heidegger’s *Being and Time*). *Philosophical Studies* (University of Tokyo), vol.31, 155-168.

Todres, L., Kathleen T., Galvin, K.T. and Holloway, I. (2009). The humanization of healthcare: A value framework for qualitative research. *International Journal of Qualitative Studies on Health and Well-being*, 4:2, 68-77.

Tokieda, Motoki (2007)(the original version in 1941). *Kokugogaku Genron*(jyou)(Basic course for the study of Japanese language (1)). Tokyo: Iwanami.

Toombs, S.K. (1988). Illness and the Paradigm of Lived Body. *Theoretical Medicine* 9, 201-226.

Ueda, Shizuteru (Ed.)(1987). *Nishda Kitaro Tetsugaku Ronsyuu* Ⅰ. Tokyo: Iwanami1.

Varela, Francisco J., Thompson, Evan and Rosch, Eleanor(1991). *The embodied mind: cognitive science and human experience.* MA: MIT Press.

Watsuji, Tetsuro (1979). *Fuudo*(Climate). Tokyo: Iwanami.

THE ETHICAL ISSUES ON AI BASED MEDICAL INFORMATION SYSTEM ARCHITECTURE: THE CASE OF TAMBA CITY MODEL

Yoshiaki Fukami, Yohko Orito

Keio University (Japan), Ehime University (Japan)

yofukami@sfc.keio.ac.jp; orito.yohko.mm@ehime-u.ac.jp

ABSTRACT

Utilization of AI in medicine requires the accumulation of integrated personal information. However, there are barriers to accumulate medical and healthcare records. Tamba city of Japan has succeeded in overcoming such adverse conditions with public-private data linkage by adopting a closed and centric structure. Centralized architecture is effective in overcoming the barriers to health information sharing, such as lack of interoperability. At the same time, the asymmetry of information can be increased because the municipal office collects and analyzes personal information, and the results are fed back only to healthcare professionals and service providers. Moreover, there may come to be risk of abuse by service providers. As the variety of information handled increases and more medical institutions are connected to the system, the risk of health professionals abusing personal information increases. In addition, since the data is collected in a specific area and not entire cohort, there is a risk that the analysis results may include a bias, and it is necessary to take countermeasures against such risk. Comprehensiveness of the collected data and respect for the patient's self-determination conflict with each other. Therefore, it is necessary to adopt a balanced architecture and institutional design for social implementation.

KEYWORDS: Medical information, architecture, interoperability, public-private data linkage.

1. INTRODUCTION

The Internet is a global network with no central server. Each individual participant or organization introduces their own identifiers, data and metadata. This world-wide dissemination enables the Internet to rapidly disperse information and realize scalability. At the same time, such a design generates a huge bottleneck of learning data for AI (Artificial Intelligence). Diverse data specifications, including syntax and vocabulary, make the utilization of fragmented data on the Internet difficult. In contrast, Google, Apple, Facebook and Amazon (GAFA) generate a massive amount of data with unified identifiers and specifications, which is one reason why they have established superiority in AI development.

The unified management of data has advantages in efficient AI-based service development and its applications. However, comprehensive personal information management by a few oligopolistic companies means that they may gain substantial power of control over individuals. Thus, the benefits of the unified management of personal information carry a risk of abuse of power by large IT companies. However, is publicly unified management the best way to balance benefit creation and personal information protection by utilizing big data?

From an ethical standpoint, medical information is a good model for examining such issues because medical and health care services are mostly public services in which governments play an important role. In most institutions, patient medical charts are written by hand; medical charts have been digitized, but progress varies greatly by country and region. Electric Health Records (EHRs) are an aggregation and integration of diagnostic records stored in multiple medical institutions over a wide area. Australia has introduced an EHR nationwide under the name Electronic Health Record (PCEHR) Personally Controlled /MyHR. However, in other countries medical institutions may not be able to actively use information sharing within hospitals, even if they are digitized.

In Japan, medical charts are being digitized, but attempts to share data and operate an EHR have not progressed. Of course, with an aging population there is a need for more efficient medical resources. Thus, the government has been promoting the digitization of medical information. Sustainable universal coverage of health insurance has been made possible by increasing the efficiency of insurance claims through computerization. The Ministry of Health, Labour and Welfare (MHLW) started to establish national standards for medical information in March 2010 and, as of March 2020[1], 27 standard specifications have been established.

However, there is a barrier to using information accumulated in different medical institutions in a consolidated manner. To avoid the risk of lawsuits, physicians are reluctant to disclose information about their diagnostic processes outside the hospital. Patients are also reluctant to exchange their personal diagnostic records across multiple facilities. On the other hand, there are few cases the medical, and health information sharing is successfully realised, one of them is Tamba city, it is Japanese case as discussed in this study. Tamba city in mountainous area of central Hyogo prefecture, Japan has succeeded in overcoming such adverse conditions and has introduced an Immunisation Determination System of public-private data linkage by adopting a closed and centric structure. The city has now forged ahead with the development of a medical and healthcare information sharing system among clinics, pharmacies, nursing care services and governments to promote a comprehensive care community. The Tamba system has a highly centralized structure in which all systems, including networks and devices installed in medical institutions, are owned and managed by the city hall. The Tamba city model, in which the government manages and analyses health and medical information on a centrally based Basic Resident Register, is an excellent case study an architectural design to improve the public health of local residents through the use of AI.

This study attempts to examine an architecture design for a secure AI based service, analyzing Tamba city model as succeeded case of medical personal data utilisation, based on the case analysis of the previous study (Fukami & Masuda, 2019, 2020). A case study approach was used in the current research because there have indeed been some cases of innovation through standardization. This inductive hypothesis-building study attempted to develop generalizable conclusions from a rare event, which was suitable for research that included ‘why’ and ‘how’ questions (Yin, 2014). We conducted interview surveys with responsible person of a city official of the health division and staff of the system integrator company that developed the health and medical information system, conducted field work with medical institutions that introduced the system, and secondary materials analysis on information provided by the municipal office.

[1] Health information and communication standards organization, List of Japanese national medical information standards, Retrieved from http://helics.umin.ac.jp/helicsStdList.html accessed on March 4, 2020.

2. RELATED WORKS

2.1. Accumulation and utilization of personal medical and health information

Over the 50 years that followed the first implementation of computerized patient medical records in the 1960s, technological advances in computer innovations opened the way for advancements in EHRs and health care (Turk, 2015). The use of data to maintain and improve medical standards has been promoted for a long time. An electronic medical record (EMR) is a real-time patient health record with access to evidence-based decision support tools that can be used to aid clinicians in decision-making (Aceto et al., 2018). The ISO standard defined an electronic health record (EHR) as a repository of information regarding the health status of each treatment in computer processable form (ISO/TR 20514:2005 - Health informatics — Electronic health record — Definition, scope and context, 2005). EHR can include past medical histories and medications, immunisations, laboratory data, radiology reports, vital signs as well as patient demographics (World Health Organization, 2012).

More specifically, an EHR is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting, and the reporting of episodes of care across multiple care delivery organizations within a community, region, or state (Aceto et al., 2018). EHR design is essentially a consolidation of data held by diverse medical institutions, since not everyone is tested and consulted at a single medical institution for a lifetime. Therefore, security for data transaction/sharing and interoperability are exceptionally important. At the same time, healthcare finances are tight, and tend to be designed to ensure security at low cost. There are also obstacles to implementation, such as lack of funding and interoperability of current systems, which decelerate the adoption of EHRs (Devkota & Devkota, 2014).

EHRs are expected to contribute to efficiency in medical services, and its introduction is being promoted by international organizations such as WHO and OECD. At the same time, there are barriers to the introduction of EHRs, as they aggregate important medical and health information. In addition, there is a high risk that the collected data may not be used effectively or may be improperly used. The goal is not to introduce EHRs themselves, but to make data-based medical services more efficient and higher quality.

However, the accumulation of fragmented medical history data does not contribute to the improvement of medical service quality (Blechman et al., 2012). There have been multiple concepts for the digitalization of medical records and to facilitate examination and compose prescriptions, such as computerized physician order entry (CPOE), which improves safety (Eslami et al., 2007). Clinical decision-support systems (CDSS) (El-Sappagh & El-Masri, 2011) are described as ‘any computer program designed to help healthcare professionals to make clinical decisions’ (Shortliffe, 2011).

2.2. Ethical related articles

Riso et al. (2017) identified six core values for the ethical sharing of data using ICT platforms: scientific value, user protection, facilitating user agency, trustworthiness, benefit and sustainability. Global Alliance for Genomics and Health formulated core elements of responsible data sharing; 1) transparency, 2) accountability, 3) engagement, 4) data quality and security, 5) privacy, data protection and confidentiality, 6) risk-benefit analysis, 7) recognition and attribution, 8) sustainability, 9) education and training and 10) accessibility and dissemination (Knoppers, 2014). It is important not only to protect privacy, but also to respect the rights of patients such as data control rights and to operate them with high transparency for sharing medical information.

The accumulated medical and health data is important privacy information, and it needs to utilize it with careful attention. Therefore there have been discussions about ethical codes and guidelines of AI usage for medicine (Luxton, 2014). Vayena et al. pointed that the ethical and regulatory challenges that surround AI in healthcare are particularly privacy, data fairness, accountability, transparency, and liability (Vayena et al., 2018). These are the issues in using medical personal information with AI such as data sharing and privacy, transparency of algorithms, data standardization, and interoperability across multiple platforms, and concern for patient safety (He et al., 2019).

Noteworthy, PCEHR, the national EHR of Australia was developed with emphasis on personal data control rights. However, due to the nation-wide platform on which personal information is shared between medical institutions, the risk of ‘secondary use’ of personal information by the 900 000 healthcare workers who can access the system has become apparent. Additionally, health records stored in the PCEHR can be created without a person’s consent (Masuda et al., 2019).

2.3. Usage of AI for medicine

AI usage for medicine has a long history. It has been tried since at least the 1970s and the first paper about AI and ethics in health and medicine was published in 1994 (Tran et al., 2019).The use of AI in the medicine and healthcare can be classified into two types. One is supporting diagnosis and research with big data collected anonymously. According to a meta-survey of the PubMed database, diagnostic imaging is most common, followed by genetic and electrodiagnosis. It is also used for physical monitoring, disability evaluation, mass screening, etc. (Jiang et al., 2017). In particular, many cases have been used for automation of image diagnosis such as diabetic retinopathy (Wong & Neil, 2016), detecting gastric cancer (Hirasawa et al., 2018), cardiac affection (Dilsizian & Siegel, 2014) and so on. Invention of deep learning accelerated usage of AI for diagnostic imaging (Hosny et al., 2018; Wong & Neil, 2016). A widely shared image database The Cancer Imaging Archive is also operated by the National Center for Biotechnology Information (Thrall et al., 2018).

The other is realizing efficient treatment and health maintenance through comprehensive analysis of a variety of data linked to individuals. In the first place, the data generated and accumulated by each medical institution is diverse. Heterogeneity of medical data is a barrier to analysis by AI (Cios & William Moore, 2002). Moreover, there are massive amount of handwritten data such as medical chart. Such data is processed with natural language processing (e.g. Jiang et al., 2017).On the other hand, there are ongoing debates on ethics of AI in health care (e.g. Morley et al., 2019).

3. CASE STUDY

3.1. Implementation of the Immunisation Determination System

Tamba city in Hyogo prefecture is a small town located in a mountainous area in the north-eastern part of Kobe city. The city launched the implementation of their Immunisation Determination System in April 2017. Tamba city supports the costs of 15 types of vaccines for children between 0 and 15 y of age; for this, subsidies are paid by the city to the clinics.

Statutory immunisation has significantly improved public health, and the number of vaccines eligible for public assistance is increasing. With the development of research related to vaccination, both vaccine type and the rules that must be adhered to, such as the order and interval for vaccinations, have been increased. On the other hand, more than 7000 immunisation accidents occurred in the fiscal year of 2017 in Japan[2]. Due to an excessive response over adverse reactions in Japan, there are fewer types of vaccination officially required or recommended than in other countries. Such an event is called a 'vaccine gap' (Saitoh & Okabe, 2012). It was needed to promote immunisation of various vaccines while reducing the risk of adverse reactions.

Vaccination is not covered by health insurance because it is not a treatment for an illness. Therefore, most immunisations are subsidised by the public. Most EHRs are designed to utilize data from electronic medical charts and health insurance claim data. However, these data cannot be used in the rationalization of immunisation because most of the people who are vaccinated are healthy individuals, and this medical record is often not placed on an electronic medical chart. Children may not be vaccinated at the same clinic from birth to the age of 15, during which public assistance is used to cover the cost of vaccination. Moreover, an immunisation history is not included in insurance medical records. The history of individual vaccinations is the only vaccination ledger owned and managed by the government, and there is no alternative for improving the efficiency of vaccinations other than using the ledger data.

The statutory vaccination institution has problems in terms of safety and economics. Subsidies for vaccinations are established and operated by each local government. Therefore, if a potential vaccinee relocates out of a city, the subsidies are not paid, and the cost of the vaccine becomes the responsibility of the clinic in the original city. The medical association requested that the city eliminates the need for clinics to bear the inoculation costs of children who were not covered by the government.

The administration of vaccinations could not be managed using distributed data that was closed to individual medical institutions, and therefore had to be based on vaccination ledgers managed by the city. While it is important to use vaccination ledgers (administratively held personal information) it is difficult to directly use this information in private medical facilities from the viewpoint of privacy protection. Clinicians tend to avoid taking responsibility for protecting personal information. However, Tamba City has been using public information held by public institutions in private medical institutions.

Tamba city introduced an Immunisation Determination System. It was successful for two reasons: 1) it did not impose on physicians the role and responsibility of protecting personal information; and 2) it provided physicians with economic benefits (Fukami & Masuda, 2019). Tamba City distribute computer tablets to clinics that are connected to a database synchronized with the vaccination ledger via a closed network. The tablets are owned by the municipal office and the system is owned and operated by the municipal office; therefore, the responsibility for privacy protection rests with the municipal office.

The system was also developed with a closed network for the MVNO in compliance with personal information protection law and is shared through tablets owned by the municipal office. Clinicians working at the core hospitals in the Tamba region access computers for electric medical charts by way of exception. The system is centrally located but is used by diversified stakeholders, such as clinicians. Prescription data and actual medication history are transacted and handled only among clinicians, co-medical staff and care givers. The overview of the closed network immunisation scheme used in Tamba city is shown in Figure 1.

In the past, vouchers were issued to the subjects, and immunisations were conducted based on these vouchers issued by the municipal office. However, if the inoculators relocated out of the city, the subsidies were not paid, and the costs of the vaccines became the responsibility of the clinic. After the immunisation determination system came to work, the system notifies whether or not the subject of subsidy. Therefore, physicians have come to deliver the vaccine after confirming that the vaccinee was

[2] The Ministry of Health, Labour and Welfare (2019) Report of immunisation accidents, Retrieved from https://www.mhlw.go.jp/content/10906000/000535721.pdf on January 13, 2020.

the correct recipient. The Immunisation Determination System has succeeded in decreasing the number of vaccination accidents (Table 1). The system also succeeded in reducing the workload of the municipal office and eliminated mistakes by linking medical and government data.

In this case, a system was developed with a simple and centralized system architecture. The only data resource was the basic resident register ledger generated and managed by the municipal office. The computer tablets that were distributed to clinics were owned by the municipal government, and personal information on the potential vaccinees was processed and managed within the municipal government.

Figure 1. Closed network immunisation scheme used in Tamba city.

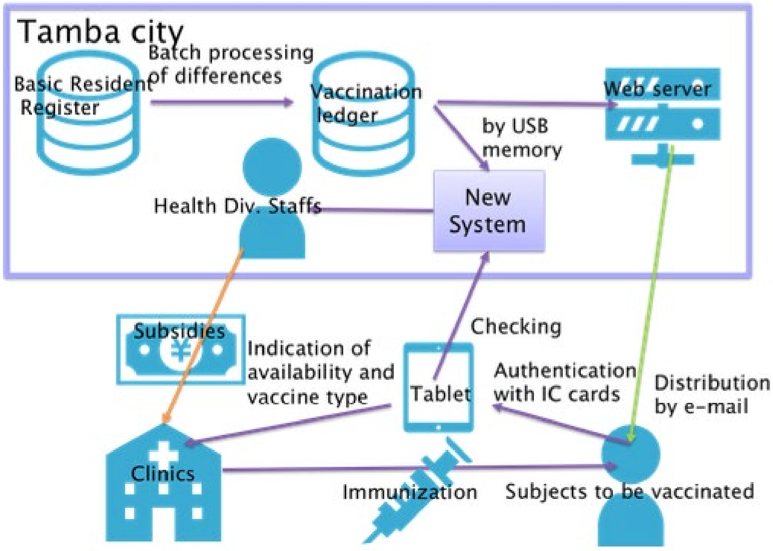


Table 1. Change in the number of vaccination accidents.

Year	Hyogo Pref.	Tamba city
FY2015	280	12
FY2016	309	4
FY2017	427	0
FY2018 (by November)	282	1

3.2. Extension to a regional comprehensive care system

Tamba city and the stakeholders in the region have now decided to extend the system to a regional comprehensive care. This means that the system will process various types of data generated by multiple organizations and will exchange information such as prescriptions, caregiver visit records, results of a medical examination and healthcare directives. Expansion to a regional comprehensive care system will be implemented through the distribution of computer tablets to dental clinics, pharmacies and visiting care offices. The data is also sent to the regional core hospital, Hyogo Prefectural Medical Center, in one way with closed network, and displayed on electronic medical charts. The municipal office also plans to add the Municipal Medical Checkup Center established in 2019 to the closed network.

In the immunisation implementation plan, clinicians only access data from the municipal office each time a vaccination is delivered and displayed the examination results. On the other hand, information sharing among medical and nursing staff is implemented in the regional comprehensive care system.

Many functions are to provide information that is referred to ad hoc when consulting or providing services. However, analysis of accumulated data and feedback on measures and medical practices based on it have been initiated.

Due to the universal health insurance system, there has already been structured data for almost all cases of clinician's prescriptions and pharmacy prescription records, respectively. Each prescription has not been matched from the clinic to the pharmacy; therefore, it is difficult to confirm that the medicine was taken as prescribed by the clinician. This regional comprehensive care system provides a function to match between prescriptions issued by doctors and prescription history data of pharmacies conforming to different standard data format specifications. As a result, feedback on the prescription of the drug can be made based on the analysis of the structured data.

Collection of prescription data began in September 2019, ahead of other features. As of January 2020, to reduce medical expenditure in the region, we have calculated the generic drug selection rate of drugs prescribed for patients with lifestyle-related diseases such as diabetes and high blood pressure. The analysis is done only by the municipal officials.

3.3. Management form

The municipal government, the medical association, the dental association, the pharmacists' association, the four hospitals, the social welfare council and the system vendor have established the Tamba Medical Care Collaboration (MCC) promotion organization to develop and operate a regional comprehensive care system. They have planned to allow the evaluation of prescriptions, health records and medical biographies via the MVNO closed network in the same manner as the immunisation records. Therefore, the results of analyses are made available to MCC member companies and organizations.

All data for analysis is collected in a closed network managed by the municipal office, and analysed by staff of the municipal office using computers installed at the city hall with no internet connection. Therefore, the raw data is not disclosed externally and there is no need for anonymous processing or pseudonymization.

The adoption of a configuration complete with a closed network and devices owned only by the city hall produces a robust system from the viewpoint of personal information protection. On the other hand, only municipal staff can handle accumulated data. This means that analysis by external medical and public health specialists is not possible. In the future, it is expected that operational rules will be established in the MCC, including who can access what type of data. This will increase the degree to which outside experts can be involved in analysis and policy making. However, centralized and closed designs that protect the privacy of personal information prevent any correct usage of accumulated data that may offer benefits for the residents.

4. ETHICAL ISSUES ON THE SYSTEM ARCHITECTURE

4.1. Merit of centralized and closed structure for AI based solutions

The medical and health field is an area where there is a great demand for efficiency and labour saving that may be implemented through the use of AI. At the same time, because medical information such as clinical diagnoses and prescriptions are important personal information, there are still substantial barriers to the realization of EHRs that can be aggregated, stored and utilized. In Japan, where the introduction of electronic medical chart systems is progressing within each individual medical institution, the lack of interoperability between systems has hindered integrated utilization.

Another barrier to efficiency is the lack of cooperation between government departments. For example, data management for health insurance premiums and the subsidy system based on the Basic Resident Register have been developed, respectively, and operated without cooperation. Furthermore, although clinicians record much information on electric medical charts, it does not tend to be shared with other institutions and utilized for analysis.

On the other hand, the system used in Tamba city case has developed integrally with the operation of the subsidy system with a highly centralized structure. The system was also developed with a closed network for the MVNO in compliance with Act on protection of personal information enforced in Japan and is shared through computer tablets owned by the municipal office. Prescription data and actual medication history are transacted and handled only among clinicians, co-medical staff and care givers. Nonetheless, the system is also centralized and used by diversified stakeholders, such as clinicians. Such a centralized and closed design is superior from the point of personal data protection because data transaction is limited to facilities owned by the municipal office, and the risk of data leakage is suppressed.

Moreover, there is no concern about compatibility as the scheme is designed and operated by a single organization. EHRs tend to lack interoperability because of security concerns, even for ones developed by national governments (Masuda et al., 2019), although the accumulation of fragmented medical history data does not contribute to improvement in the quality of medical services (Blechman et al., 2012). However, the system has two ethical issues as follows. One issue involves the asymmetry of expanding information between the government/medical staff and citizens. The second issue is the abuse of service providers.

4.2. Information asymmetry

In the case of Tamba city, people cannot memorize all of their medical activities and their entire treatment history. Because the computer tablets are distributed among medical and nursing service providers, only service providers can access patient records, and citizens cannot access these records. Thus, as more records are accumulated, the information asymmetry between service providers and citizens increases.

In general, medical records are not shared with patients. Even though citizens can accumulate more diversified data with wearable devices and smart phone applications for prescription management, they cannot manage their records because they do not have rights to access the system. While there are reasonable reasons for the limited disclosure and sharing of medical information with patients, their further engagement is needed for decisions on the course of treatment. It is important for patients and citizens to engage in treatment policy decisions, according to patient-centred medicine (Laine & Davidoff, 1996).

Moreover, KPIs (Key Performance Indicators) of regional comprehensive health care must be diversified, and it is beneficial for patients to participate in the selection of metrics for KPIs because the cure rate and survival rate are inappropriate KPIs of long-term care even outside of hospitals. It is also desirable to introduce sensors chosen and owned by patients that enable multifaceted situational understanding according to patients' preferences. Considering such situations, while it is required to reduce the information asymmetry between the citizens (patients) and the medical service provides, it has been not realised until now. The expansion of information asymmetry may have significant impacts over the autonomy and decisional privacy of the individuals.

4.3. Abuse of service providers

Compared to the immunisation implementation design, much more diversified data is accumulated from citizens in the regional comprehensive care system, and a much greater diversity of engaged persons can access the personal information. Regional comprehensive care information is accessed not only by medical staff but also by government employees, caregivers, and social workers. Services are provided outside of medical facilities, across the region and even in patients' homes for the long term. Therefore, the potential for fraud and blackmail based on medical histories has increased. Even if data are shared among limited professional stakeholders, misuse of the information cannot be prevented.

For example, in Australia where the National EHR (NEHR) was introduced and developed with a decentralized structure, the EHR's privacy chief of national government once refused to take responsibility for the above mentioned security and privacy issues (Grubb, 2018). The incident occurred because the NEHR was developed through linking EHRs of individual medical institutions, and medical staff could view medical records at other medical institutions.

As such, the privacy of EHRs was laid open to abuse by healthcare professionals. On the other hand, in the Tamba city system, as clinicians working at core hospitals in the region rarely access computers for electronic medical charts, municipal staff and not clinicians in medical institutes are able to access data at will. This is because the regional comprehensive care system realizes integrated analysis of administrative resident card data and medical and health data. Even if only a limited amount of action history data is no longer private, if it is accumulated, a detailed profile of an individual can be obtained and converted into important personal information. If the target area is small, it is easy to identify individuals from pseudonymized data

The case of Tamba shows what is possible using integrated analysis of administrative resident card data and medical and health data. Rather than solely supplying medical and nursing resources alone, we can provide effective public services and improve public health standards by providing livelihood protection and other subsidies together. Proactive life interventions can also be performed efficiently and with high accuracy from more multifaceted data. In other word, it implies that social sorting can be generated, utilising huge amount of sensitive personal data and its profiling, and the invisible impacts on the individuals are exerted.

The possession, management and access to data on the system are limited to the municipal office staff in Tamba City although the MCC, a joint operation organization of the system by multi-stakeholders, has been established. However, the data stored in the system includes not only medical/health data, but also administrative information linked to the Basic Resident Register. The city officials can access various types of data, which impacts the provision of a wide range of administrative services. Therefore, if data abuse occurs, the damage to community residents can be severe.

Then, how can we design a system that will deter the usage of shared data for purposes other than the original intent? Is it possible to monitor every activity of all service providers to control the use of data? Such solutions may result in other types of privacy abuse by service providers. Because the system is developed within a closed network in service providers' organisation, it is impossible the governmental bodies or municipalities to monitor how the service providers utilise and analyse the data in complete way.

This tends to make the system closed to protect privacy. However, limiting access to information makes it difficult to regulate abuse by the few parties that do have access. Simply developing a system with a closed network is not sufficient. The range of information to be shared and a means of controlling access need to be defined from patient-centric/citizen-centric perspectives.

4.4. Assumed biases and risks arising from them

While just the vaccination history and prescription data were targeted, only factual data was used, so there was little risk of contaminating inappropriate data. Nevertheless, there was a possibility that sampling bias may occur in trend analysis based only on data from the limited number of residents who had agreed to provide information.

In the future, when the input and sharing of natural language data by doctors, nursing staff and care staff will begin, subjective evaluation data regarding the patient's condition will be accumulated. This means that the information arbitrarily created by the service provider is linked to the Basic Resident Register. Depending on the type of disease, the accuracy of a diagnostic result may vary. In addition, there are situations in which the judgment differs depending on the clinician or service provider in charge of the examination. As subjective information is accumulated, even if there is no malice, biased learning data may be generated and remained.

Furthermore, the accumulated data is linked to administrative data and used when providing administrative services. Government services and medical services are provided based on biased data, and decisions may be made on policies based on the results of AI analysis drawn from this data. If AI presents incorrect analyses, it will not only hinder efficient social security implementation but could also cause serious human rights violations.

As the type of available data increases and the range of measures that can be deployed expands, there is a risk that human intervention may take place, leading to violation of human rights. This system was originally constructed based on the vaccination judgment system, and it is therefore assumed that data such as health check results and vaccination histories will have been accumulated while the subject was in a healthy state.

5. CONCLUSION

The case of Tamba shows what may be achieved with integrated analysis of administrative resident card data and medical and health data. Rather than solely supplying medical and nursing resources alone, it enables the provision of effective public services and improves public health standards by providing livelihood protection and other subsidies. In addition, proactive life interventions can also be performed efficiently and with high accuracy from more multifaceted data.

If the system is built on the basis of the vaccination register, linked to the basic resident register, the range covered by the system is limited to the administrative area. For this reason, even if the coverage is high, the absolute number of registers is inevitably small. With small amounts of data, the probability of an analytical error increases. If AI presents incorrect analysis results, it will not only hinder the implementation of efficient social security but may also violate human rights. The potential for human rights violations by the regional comprehensive care system is greater than that of EHR, because the data is linked to both medical and health information and the Basic Resident Register.

To protect patients' privacy, a centralized and closed design needs to be adopted. However, such a design is not necessarily advantageous in the protection of human rights. An open and distributed design is better, even in the context of medical and nursing care, for AI-based decisions according to the concept of patient-centred medicine. This open and distributed design encourages self-determination in medicine and provision of appropriate care. The introduction of AI may change the rational design of systems where tailor-made services are developed with big data including personal information.

It is important that each patient retains control of their own data. However, if authorization is needed to transact data from one institute to others, it is difficult to collect data without omissions. Nevertheless, in the case of Tamba City, where the administration manages data in a unified manner, data can be collected without omission only in the administrative area and integrated analysis can be easily performed. In this regard, it seems that comprehensiveness of the collected data and respect for the patient's self-determination conflict with each other. Therefore, it is necessary to adopt a balanced architecture and institutional design for social implementation.

ACKNOWLEDGEMENTS

This work was supported by JSPS Grant-in-Aid for Early-Career Scientists Grant Numbers JP18K12858 and Grant Number 17K03879. This study was also supported by the Japan Society for Management and Information Grant-in Aid for SIG “Monitoring and Control of AI Artefacts”, and the Seikei University Grant-in-Aid 2020 for the research on “Monitoring and Control of AI Artefacts: Consideration from Economic, Social and Legal Perspectives”.

REFERENCES

Aceto, G., Persico, V., & Pescapé, A. (2018). The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges. In *Journal of Network and Computer Applications* (Vol. 107, pp. 125–154).

Blechman, E. A., Raich, P., Raghupathi, W., & Blass, S. (2012). Strategic Value of an Unbound, Interoperable PHR Platform for Rights–Managed Care Coordination. *Communications of the Association for Information Systems*, *30*, 83–100.

Cios, K. J., & William Moore, G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2), 1–24.

Devkota, B., & Devkota, A. (2014). Electronic health records: advantages of use and barriers to adoption. *Health Renaissance*, *11*(3), 181–184.

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current Cardiology Reports*, *16*.

El-Sappagh, S. H., & El-Masri, S. (2011). A Proposal of Clinical Decision Support system Architecture for Distributed Electronic Health Records. *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*.

Eslami, S., Abu-Hanna, A., & de Keizer, N. F. (2007). Evaluation of Outpatient Computerized Physician Medication Order Entry Systems: A Systematic Review. *Journal of the American Medical Informatics Association*, *14*(4), 400–406.

Fukami, Y., & Masuda, Y. (2019). Success Factors for Realizing Regional Comprehensive Care by EHR with Administrative Data. In Y.-W. Chen, A. Zimmermann, R. J. Howlett, & L. C. Jain (Eds.), *Smart Innovation, Systems and Technologies* (Vol. 145, pp. 35–45). Springer.

Fukami, Y., & Masuda, Y. (2020). Stumbling blocks of utilizing medical and health data : Success factors extracted from Australia-Japan comparison. *8th International KES Conference on Innovation in Medicine and Healthcare* (forthcoming).

Grubb, B. (2018, November). My Health Record's privacy chief quits amid claims agency 'not listening. *Sydney Morning Herald*.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36.

Hirasawa, T., Aoyama, K., Tetsuya Tanimoto, ·, Ishihara, S., Shichijo, S., Tsuyoshi Ozawa, ·, Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., & Tomohiro Tada, ·. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer*, *21*, 653–660.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510.

ISO/TR 20514:2005 - Health informatics — Electronic health record — Definition, scope and context, (2005).

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243.

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *HUGO Journal*, *8*(1), 1–6.

Laine, C., & Davidoff, F. (1996). Patient-Centered Medicine. *JAMA*, *275*(2), 152.

Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, *62*(1), 1–10.

Masuda, Y., Shepard, D. S., & Yamamoto, S. (2019). Adaptive governance on electronic health record in a digital IT era. *25th Americas Conference on Information Systems, AMCIS 2019*, 1–10.

Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M. & Floridi, L.(2019). The debate on the ethics of AI in health care: A reconstruction and critical review. Retrieved from http://doi.org/10.2139/ssrn.3486518

Riso, B., Tupasela, A., Vears, D. F., Felzmann, H., Cockbain, J., Loi, M., Kongsholm, N. C. H., Zullo, S., & Rakic, V. (2017). Ethical sharing of health data in online platforms – which values should be considered? *Life Sciences, Society and Policy*, *13*(12).

Saitoh, A., & Okabe, N. (2012). Current issues with the immunization program in Japan: Can we fill the "vaccine gap"? *Vaccine*, *30*(32), 4752–4756.

Shortliffe, E. H. (2011). Biomedical informatics: Defining the science and its role in health professional education. In D. Hutchison, T. Kanade, J. Kittler, & J. M. Kleinberg (Eds.), *Information Quality in e-Health. USAB 2011. Lecture Notes in Computer Science, vol 7058.: Vol. 7058 LNCS* (pp. 711–714). Springer.

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *Journal of the American College of Radiology*, *15*(3), 504–508.

Tran, B., Vu, G., Ha, G., Vuong, Q.-H., Ho, M.-T., Vuong, T.-T., La, V.-P., Ho, M.-T., Nghiem, K.-C., Nguyen, H., Latkin, C., Tam, W., Cheung, N.-M., Nguyen, H.-K., Ho, C., & Ho, R. (2019). Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. *Journal of Clinical Medicine*, *8*(3), 360.

Turk, M. (2015). Electronic Health Records: How to Suture the Gap Between Privacy and Efficient Delivery of Healthcare. *Brooklyn Law Review*, *80*(2), 565–597.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11).

Wong, T. Y., & Neil, M. B. (2016). Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *Journal of the American Medical Association*, *316*(22), 2366–2367.

World Health Organization. (2012). *Management of patient information: Trends and challenges in member states*. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/76794/9789241504645_eng.pdf

Yin, R. K. (2014). *Case Study Research: Design and Methods* (Fifth Edit). Sage Publications.

8. Societal Challenges in the Smart Society

A META-ANALYSIS OF RESPONSIBLE RESEARCH AND INNOVATION (RRI) CASE STUDIES - REVIEWING THE BENEFITS TO INDUSTRY OF ENGAGEMENT WITH RRI

Vincent Bryce, Laurence Brooks, Bernd Stahl

De Montfort University and University of Nottingham (United Kingdom),
De Montfort University (United Kingdom), De Montfort University (United Kingdom)

psxvb1@nottingham.ac.uk

ABSTRACT

Responsible Research and Innovation (RRI) provides a framework that may overcome computer ethics problems such as the increasingly ubiquitous nature of computing technologies, the global nature of innovation, and the need to consider accountability at different stages of the innovation lifecycle. It shares with computer ethics the challenges of demonstrating relevance to, and providing practical guidance for, industry.

This paper will answer the following research question - what is the relationship between RRI implementation practices and outcomes for firms, considering contextual variables such as company size and sector? It will examine this question using a meta-analysis of published RRI case studies.

The contribution it makes to knowledge is the exploring and quantifying of relationships between RRI practices, and outcomes for businesses. It responds to the need for more quantitative research to get from 'perceptions to evidence', to explore the 'business case' for corporate engagement with RRI, and to relate RRI more explicitly to adjacent discourses on corporate responsibility.

The methodology developed helps pave the way towards a broader approach to evaluating the business case for companies to engage with RRI practices.

KEYWORDS: responsible research and innovation; RRI; responsible innovation; corporate social responsibility; CSR; industry.

1. INTRODUCTION

Responsible Research and Innovation is a relatively new concept that aims to democratise innovation and integrate ethical concerns into the earliest stages of the innovation process, including through anticipatory and deliberative governance methods (Lubberink et al., 2017).

A recent network analysis of citations (Loureiro & Conceição, 2019) indicates a high degree of convergence around the Stilgoe et al. (2013) RRI framework which incorporates Von Schomberg's (2012; 2013, p9) definition of RRI as:

> "… a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a viewto the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)",

and the four dimensions of 'Anticipate', 'Reflexivity', 'Inclusion and Deliberation', and 'Responsiveness'. These can be referred to as the 'AREA' framework, with 'Engagement' substituting Inclusion and 'Action' substituting 'Responsiveness' (Owen, 2014).

Alternative framings of RRI suggest additional dimensions – for example the six 'keys' referenced in the EU's Rome declaration include Governance, Gender Equality, Open Science and Science Education (European Commission, 2013). The Schomberg articulation of RRI is beneficial from an industry perspective both in terms of its centrality and consistency through RRI discussions, and in that it incorporates conceptions of product, and process/product lifecycle as well as the 'purposes' of research and innovation.

1.1. Defining RRI from the organisational perspective

RRI thus defined has both a normative, and descriptive aspect – it allows us to evaluate certain organisational behaviours as consistent (or not) with RRI, while also proposing how scientists, innovators and other societal actors should innovate, and for whom.

Many publications have discussed the normative considerations – for example whether RRI's aims are extensible across borders (Macnaghten et al., 2014), reflect the power dynamics involved in engagement activities (Blok et al. 2015), or consider how technology futures are framed at an early stage(Grunwald, 2014). Less well explored is the question of how RRI takes shape at the organisation level – while it is possible to identify instances that most observers agree demonstrate irresponsible innovation (for example Von Schomberg 2013 p14-19 cites the initial development of GM maize), given that many organisations already carry out activities relevant to the RRI AREA framework under the banner of market research or new product development (van de Poel et al., 2017), it is difficult to distinguish organisations we would assess as 'innovating responsibly', from those which are not.

One conceptualisation of this problem is to ask whether RRI is best described as a property of various existing processes, or a specific (albeit broad) process that can be followed – in effect, whether it is an inclusive, or exclusive concept. While there is a case for an inclusive imagining of RRI – we have already noted that various existing organisational processes contribute to RRI – this definition is problematic, as it may not be analytically meaningful. If nearly all organisations can be described as carrying out RRI-related practices, we do not have a logical basis to differentiate 'responsible' from 'irresponsible' innovation. On the other hand, a truly exclusive definition – that an organisation is only engaging in RRI if it states an explicit commitment to do this and follows a rigid process – would deny the evidence that many organisations do carry out practices that contribute to RRI aims. We therefore conclude:

[1] That organisations may exhibit responsible research and innovation practice by degrees, applying different aspects in different situations.

[2] That the introduction of 'objectively' assessed, standardised management processes and quality benchmarks relating to RRI is needed to provide a reliable method of defining whether organisations have a robust process for innovating responsibly.

It is noteworthy that the independent assessment methods indicated in [2] were identified as being needed by Flipse and Yaghmaei (2018,) and are under active development at both the EU and country level (for example the draft European 'Guidelines to develop long-term strategies/roadmaps for RRI' to and BSI Responsible Innovation standards). Similarly, frameworks have been advanced that incorporate both 'unconscious', and 'conscious' engagement with RRI, and seek to assess the degree of integration into an organisation's practices (for example Stahl et al., 2017).

1.2. RRI and Industry

While discussion of the responsibility of business can be traced back to the early twentieth century, the RRI discourse has developed in the context of the publicly funded research sector (Stahl, 2015). Owen et al's Framework (ibid) locates the roots of RRI in Social and Technology Studies (STS), a discipline developed by the US Office for Technology Studies to manage concerns over nanotechnology development, and Technology Assessment, a methodology primarily developed to contribute to the formation of public and political opinion.

Although their work focusses on the products as well as the purpose of research and innovation, and discusses industry-relevant techniques such as stage gating, both Stilgoe et al,'s (2014) paper and Rene von Schomberg's (2012, 2013) work that informs its core definition, position RRI as a tool to manage science's relations with society rather than primarily a need to reimagine innovation.

The Cambridge Analytica incident - a major landmark for discussions of innovation governance - highlights the idea that technology developments frequently span academic, industry, and public organisations. In this case, what would be viewed by many as irresponsible innovation and resulted in legal and regulatory sanctions, was enabled by a combination of governance failures in a Higher Education institution (management of access to research APIs), a technology company (Facebook's safeguards around API access), entrepreneurs (Cambridge Analytica), and political parties (in their use of illegally-obtained data products; Berghel, 2018). In this case, RRI practices in any of these organisations may have prevented the subsequent outcomes - if all stakeholders had embraced responsible innovation principles, adverse consequences may have been avoided.

In proposing an alternative framing of RRI that more closely relates it to business, Lubberink et al. (2017) state that:

> "the problem with the current concept of responsible innovation is that it is developed by researchers and policy makers who are focused primarily on the conduct of responsible science and technological development without differentiating between research, development and commercialisation".

Dreyer et al. (2017) support this position with a detailed critique of RRI from an industry perspective by members of the European Industrial Research Management Association. While strongly endorsing the importance of RRI for business and society, they highlight a number of weaknesses:

- *Failure to consider the innovation dimension* – the framing of RRI too often emphasises research aspects, and does not sufficiently account for the nature of the innovation process. The tensions between the precautionary principle and innovatory principles, and innovation and democratic governance, are insufficiently considered.
- *Research integrity* - RRI frameworks typically underemphasise this, but it is critical to high-profile examples of 'irresponsible innovation'.
- *Failure to reflect established business practices* – companies typically already have activities that support RRI, under the banners of (for example) product development, consumer research and compliance.
- *Failure to reflect parallel sustainability debates* – RRI discussions should be situated in relation to parallel corporate sustainability debates such as CSR, corporate shared value (CSV), sustainable finance and investment, and leadership.
- *Failure to accommodate emerging issues associated with digital developments in industry* - developments in big data and smart information systems in the industry context pose new challenges for RRI (see next section).

Lubberink et al. (2017) suggest additional aspects of innovation that require contextualisation of RRI, including the financial imperatives that apply at the commercialisation stage of innovations, and the social innovation perspective – use of technology may develop independent of 'traditional' regulators and innovation actors.

As a final comment on RRI's compatibility with industry, the EU framing of RRI includes Open Science and Science Education 'pillars' which invoke particular challenges in relating RRI to the industry context in terms of their interaction with commercial confidentiality and intellectual property. This study will primarily focus on the AREA framework as noted above.

1.3. RRI's particular relevance to ICT

Several aspects of ICT with particular significance for RRI are highlighted by Dreyer et al. (ibid), and have been explored subsequently.

They include the complexity and rapid pace of technological change; the interaction of new technologies with rights such as privacy; emergent issues such as the need for algorithmic transparency and auditable code; the requirement for new forms of governance (for example of AI); new environmental impacts; workforce restructuring; and the need for different taxation models.

These issues have been explored elsewhere (for example Stahl et al. 2015, Stahl, Flick et al., 2017) and are evident in growing debates on (for example) AI regulation. They are synthesised in the Framework for Responsible Research and Innovation in ICT developed by Jirotka, Stahl and others (2017). For the purpose of this study they highlight the need for a wide-ranging definition of types of activity and impact relevant to RRI.

1.4. RRI and Corporate Social Responsibility (CSR)

Relating RRI to long-standing discourses on Corporate Social Responsibility (CSR) offers opportunities to apply approaches developed in the CSR literature in support of RRI research questions. A case can be made for the relevance of CSR ‘tools’ in informing RRI implementation practice (Iatridis & Schroeder, 2015).

Similarly for measurement, the evolution of RRI maturity models has been informed by the availability of CSR models drawing on a wide empirical evidence base (Martinuzzi & Krumay, 2013; Stahl et al., 2017). RRI and CSR share the challenges of definitional complexity, and difficulty in identifying empirical attributes. However while contested (for example Banerjee, 2008), the concept of CSR benefits from having been the subject of significant theory building and research. Reference is made where appropriate within this study to the existing evidence base for the ‘business case’ for CSR, primarily referring to Carroll & Shabana's (2010) review of meta-analyses which narrates the “30-year quest for an empirical relationship between a corporation’s social initiatives and its financial performance”.

1.5. Defining the term ‘business case for RRI’

The ‘business case for corporate responsibility’ has been exposed to empirical scrutiny in the CSR literature since the 1960s, including through meta-review. Carroll and Shabana’s (2010) meta-analysis synthesises different perspectives around the following definition (p92):

> “the establishment of the ‘business’ justification and rationale, that is, the specific benefits to businesses in an economic and financial sense that would flow from [CSR] activities and initiatives”

The authors note that a ‘business case’ approach is only one of three potential approaches to corporate responsibility –a ‘social values-led’ approach sees responsibility as the organisation’s ‘lifeblood’, as in the case of many voluntary organisations and social enterprises - a ‘business case’ can be seen as one that evaluates responsibility initiatives on a narrow economic basis - and a ‘syncretic stewardship’ model in which responsibility is an overarching approach to the business rather than assessed on a transactional basis.

The authors distil this distinction into ‘narrow’, and ‘broad’ views of the business case for responsibility – the former an expectation of direct and clear links from any responsibility initiatives to firm financial performance, the latter accepting the existence of both direct and indirect links between initiatives and outcomes and a perspective that values the additional, potentially non-quantifiable opportunities that may be generated through responsibility activity such as the development of stakeholder relationships (p101). This study will apply the broad sense of ‘business case’.

1.6. Measuring impacts and benefits of RRI for organisations

In the wake of Dreyer et al.'s (2017) and Lubberink et al's (2017) problematisation, studies have begun to explore how RRI can be applied in industry settings, in many cases through outputs of the Responsible-Industry, PRISMA (Piloting Responsible Research and Innovation in Industry) and MORRI (Monitoring the Evolution and Benefits of RRI) EU Horizon projects. In some cases these extended existing lines of enquiry such as the work of Steven Flipse and Emad Yaghmaei at Delft University of Technology on operationalisation and measurement of RRI in organisations (Flipse et al. 2015; Yaghmaei 2016; Flipse & Yaghmaei 2018).

Nonetheless, much discussion of RRI measurement has tended to focus either on society-level impact – from Von Schomberg's original (2013, p8-12) proposal of the 'right impacts' of RRI, flowing through to the European indicator framework proposed by Strand et al. (2015) that informed subsequent work in this area - or an individuated, company-specific RRI strategy, 'roadmap' and performance indicators as proposed by van de Poel et al.'s (2017) model, Porcari et al. (2018), and Yaghmaei (2018). Transition from an initial emphasis on macro-social level benefits towards the organisation level mirrors a similar movement in discussions of the impacts of CSR (Carroll & Shabana 2010, p92).

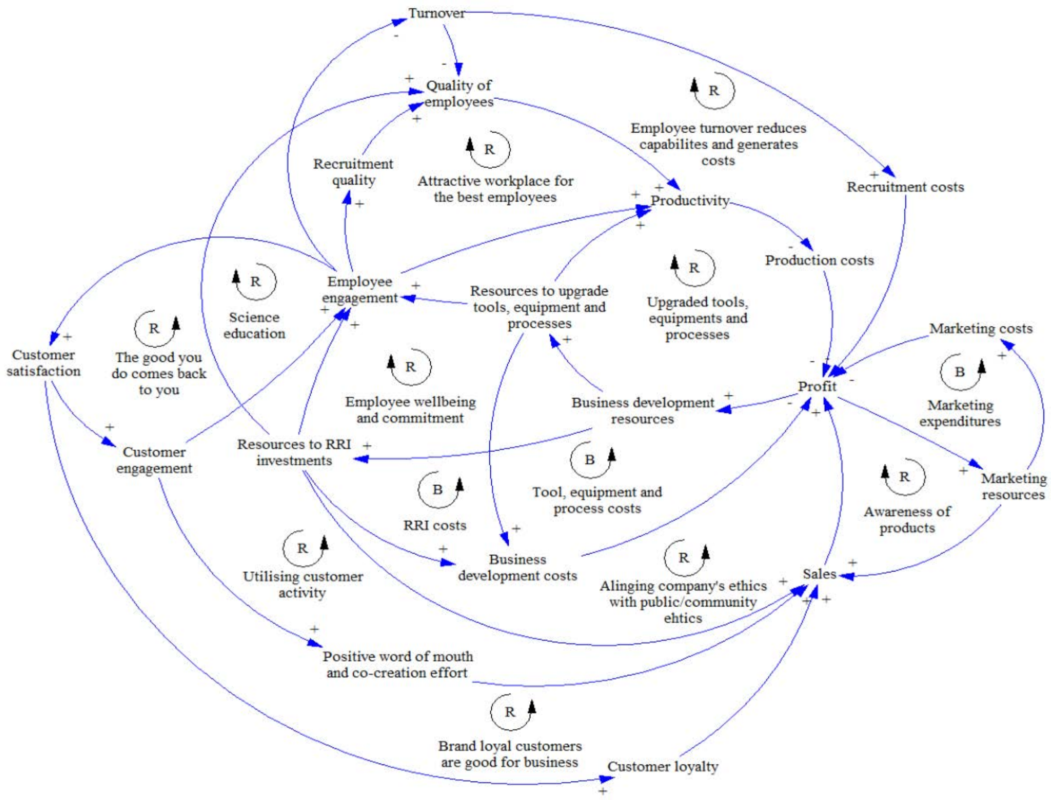
While heterogeneity of approach and shaping to context is an important principle for industry guidance, if we accept the normative premise that encouraging organisations to engage in RRI-related practice is a desirable goal, this idiographic emphasis begs important empirical questions. Which RRI practices are associated with positive business outcomes in different contexts? Beyond this – while practices such as public or employee engagement are associated with positive organisational outcomes, can a 'value-adding' effect for organisations who implement "broad-focus" RRI across the AREA spectrum be observed beyond effects which might be expected from component practices? Within these questions – given that a company's engagement with RRI may be either strategic, or operational (B. C. Stahl et al., 2017; van de Poel et al., 2017) - to what extent are benefits evidenced when RRI is adopted at a company, rather than project level?

The quantitative, empirical study of RRI at the organisation level these questions imply involves defining RRI and its impacts in observable and measurable terms. This is inherently challenging due to the broad scope of potentially relevant activities across different types of organisation as well as the long-term, complex and mediated nature of potentially relevant impacts. Trade-offs are likely to be needed – for example only a limited time horizon is available for tracking impacts, not every stakeholder perspective may be accounted for, and correlation must be interpreted in the context of potentially complex causal relationships.

In considering the outcomes of RRI activity that may be relevant to an organisation, we need to consider indirect as well as direct impacts. Gurzawska et al. (2017) introduce a casual loop model that demonstrates the complexity of potential positive and negative feedback loops mediating RRI activity and outcomes (Figure 1). For example, improved customer engagement may generate positive word-of-mouth that over time generates increased sales.

As highlighted in the CSR literature – for example Carroll & Shabana (2010)'s review - responsibility activities may generate negative impacts (albeit that the weight of evidence in the CSR debate favours net positive impact). For example, engagement can surface tensions between different stakeholders, and delay product development. These activities may have short term disbenefits, and positive longer-term benefits – an understanding of both is necessary for cost-benefit analysis. Consequently, RRI impact need to be considered on a 'two-tailed' basis.

Figure 1. Gurzawska et al. (2017) causal loop diagram for internal RRI incentives.


Turnover
Quality of employees
Employee turnover reduces capabilites and generates costs
Recruitment quality
Attractive workplace for the best employees
Recruitment costs
Productivity
Production costs
Employee engagement
Science education
Resources to upgrade tools, equipment and processes
Upgraded tools, equipments and processes
Customer satisfaction
The good you do comes back to you
Marketing costs
Employee wellbeing and commitment
Business development resources
Profit
Marketing expenditures
Customer engagement
Resources to RRI investments
Tool, equipment and process costs
Marketing resources
RRI costs
Awareness of products
Utilising customer activity
Business development costs
Alinging company's ethics with public/community ehtics
Sales
Positive word of mouth and co-creation effort
Brand loyal customers are good for business
Customer loyalty

Drawing these strands together, this study will use the activity framework proposed by Lubberink et al. (2017) to provide a lexicon of RRI practices at organisation level, and the industry impacts in Porcari et al.'s (2019) PRISMA summary to frame organisational RRI impacts. The framework advanced by van de Poel et al. (2017) provides an underpinning theoretical principle for these definitions in establishing that organisational processes not specifically aligned to responsibility aims may constitute activity relevant to RRI themes. The framework proposed by Fraaije & Flipse (2019) offers alternative ‘qualifiers’ for assessing RRI in processes, but while offering a holistic synthesis constitutes a normative, rather than descriptive assessment of RRI activity in an organisation - many of the identifiers used would need further qualitative enquiry within organisations to assess (for example to assess the terms ‘meaningful contributions’, ‘transparently’, ‘empower’ and ‘include stakeholders for substantive not instrumental reasons’ from appropriate stakeholder perspectives).

1.7. Limitations of empirical RRI studies to date

Studies of RRI to date have mainly used qualitative designs. Lubberink et al.'s (2017) review identifies “few scholars who empirically investigated responsible innovation practices in commercial R&D settings”, and that more than two-thirds of empirical RRI articles were case study research. While predating recent studies this:

1. indicates that RRI has until recently focussed on empirical exploration and description;
2. highlights a need for larger-scale and quantitative empirical testing, and;

3. as managerial decision-making is frequently based in quantifiable evidence, signals more quantitative research is essential for future development of the RRI field (Martinuzzi et al. 2018).

The availability of studies will be re-assessed through this article – beyond this it is relevant to note that exploration of the potential business benefits of RRI activity is arguably not just essential for the development of the RRI field, but for increasingly urgent efforts to encourage companies (including technology companies) to innovate responsibly – that is, to make the normative case for RRI to a range of stakeholders above and beyond the academic community.

1.8. Rationale and aims of this study

The proliferation of relatively high quality RRI case studies, in many cases products of EU Horizon projects as noted above, is an opportunity to synthesise findings through meta-review, to explore generalised relationships between RRI implementation and outcomes.

This study aims for a novel contribution to quantitative and empirical evidence in relation to industry engagement with RRI, through a meta-analysis of these case studies that explores the relationships between reported activities and outcomes for organisations. The scope of enquiry includes whether degree of engagement with RRI predicts scope of impacts reported, whether some categories or types of activity are associated with particular outcomes, and the mediating role of characteristics such as sector and organisation type.

This will allow for flexibility in interpretation of RRI in terms of its implementation and impacts within an organisation's operating context.

A meta-review methodology offers the opportunity to identify then synthesise a range of RRI case studies (Moher et al., 2009). The systematic literature review principles of Tranfield et al. (2003) will be used to identify relevant studies. Features of the RRI implementation context such as organisation type and sector will be included in the analysis. The resulting data will be assessed to identify patterns and relationships between context, implementation practices, and outcomes.

Table 1 summarises the specific questions that will be explored, based on the overall research question of 'what is the relationship between RRI implementation practices and outcomes for firms, taking into account contextual variables such as company size and sector?'. Additional detail is provided in the method section below.

Table 1. Detailed research questions.

Question
1. Does the scope of positive RRI outcomes increase as the scope of reported RRI-related activity increases?
2. Does engaging in a full scope of RRI activity increase the likelihood of reporting a wider scope of positive RRI outcomes?
3. Is engaging in some specified types of RRI activity associated with particular types of organisational outcome?
4. Is engaging in some specified types of RRI activity associated with a broader scope of positive organisational outcomes?

5. Is engaging in some specified RRI activities associated with particular organisational outcomes?
6. Is engaging in a certain combination of RRI activities associated with a broader scope of positive organisational outcomes in specific contexts?
7. For some organisation types/sectors/ages/implementation types, is a certain set of RRI activities associated with particular organisational outcomes?

2. METHOD

The principles set out in Moher et al. (2009) were applied to carry out a meta-analysis of published RRI case studies.

A literature search of peer-reviewed English-language papers in the Web of Science, Scopus and ABI/Inform databases was conducted in order to draw in RRI research across a range of disciplines, including business-focussed journals, with reference to the research question set out in the abstract, and the systematic literature review procedure set out in Tranfield et al. (2003). Papers were extracted using the following search phrases in the title, author keywords and abstract: 'responsible research and innovation' and 'responsible research & innovation', from the period 2000-2019. All papers from the Journal of Responsible Innovation and Orbit Journal (publications with a specific RRI focus) were then included for analysis.

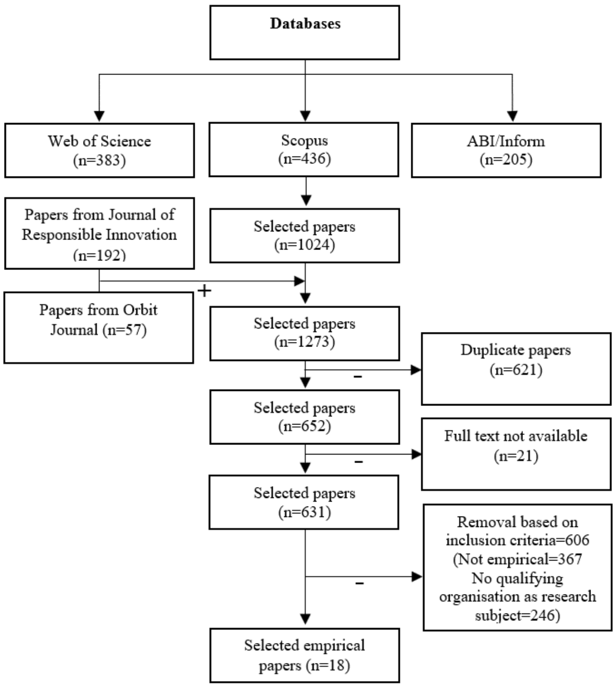
Duplicate papers were identified and removed, and papers for which only abstracts were available were excluded. Papers were then assessed based on the content of their abstract and full text as to whether they constituted an empirical case study with details of RRI-related activities and impacts, with one or more non-HE organisations as the research subject.

A top-down coding schema was applied. 'RRI-related activities' were defined as any of those in a list of activities based on Lubberink et al.'s (2017, p14) taxonomy. 'RRI-related impacts' was defined as any of those in the list of activities proposed in Porcari et al. (2019, p30), with the addition of a code for 'other RRI negative impacts'. These are listed with working definitions at Appendix 1.

'Case study' was defined as "An empirical enquiry that investigates a contemporary phenomenon within its real-life context" based on Yin (2011, p13). 'Empirical' was assessed on the basis of the Oxford Dictionary of Science (Daintith & Martin, 2010) definition of "a result that is obtained by experiment or observation rather than from theory'. Editorial or conceptual studies were excluded. 'Organisation' was defined in a broad sense, without reference to funding source, but excluding examples which purely assessed Higher Education institution research teams or state level policies.

This process is summarised as Figure 2 (below). 18 papers which met the criteria for review were imported into Nvivo 12 for analysis, providing 20 organisation case studies.

Figure 2. Flow chart of SLR process.


Databases
Web of Science (n=383)
Scopus (n=436)
ABI/Inform (n=205)
Papers from Journal of Responsible Innovation (n=192)
Selected papers (n=1024)
+
Papers from Orbit Journal (n=57)
Selected papers (n=1273)
-
Duplicate papers (n=621)
Selected papers (n=652)
-
Full text not available (n=21)
Selected papers (n=631)
-
Removal based on inclusion criteria=606 (Not empirical=367 No qualifying organisation as research subject=246)
Selected empirical papers (n=18)

The lists of RRI-related and impacts in Appendix 1 was used as a top-down coding scheme. Descriptive features (type, sector, region, dates examined, age of organisation) were captured, and studies were defined as 'strategic' or 'operational' based on the concept in Stahl et al. (2017) – in short, whether they described features of the organisation's overall work processes, or features that were applied by an organisation to a specific project or programme.

Four additional variables were assigned to each study: 'scopeofactivity' (number of distinct RRI-related activity nodes coded), 'fullscopeactivity' (true if at least one activity within the top-level categories was present), 'scopeofimpacts' (number of distinct RRI-related impact nodes coded), and 'casestudylength' (word count of the text within each publication that referred to the case study organisation).

SPSS Statistics 26 and Gephi were used to explore relationships between activity, impacts, and case study metadata.

3. RESULTS

3.1. Study characteristics

Metadata are provided in Appendix 2. While a range of organisation types were included, most were EU based (12 out of 20 - 60%), and in the Healthcare (8 out of 20 - 40%) or Agriculture (6 out of 20 - 30%) sectors. 13 of 20 cases (85%) were assessed as 'strategic' implementations. The most frequent organisation types were social enterprises (7 out of 20 - 35%) and joint ventures (4 out of 20 - 20%), with only one assessed as a public body.

Figure 3 shows the proportion of cases coded for each RRI related activity category, which varied from 95% (inclusion) to 10% (reflexiveness). Only two cases were coded against five or more RRI activity categories (Figure 4), and only one against all categories.

Figure 3. Percentage of cases studies coded against different RRI activity categories.

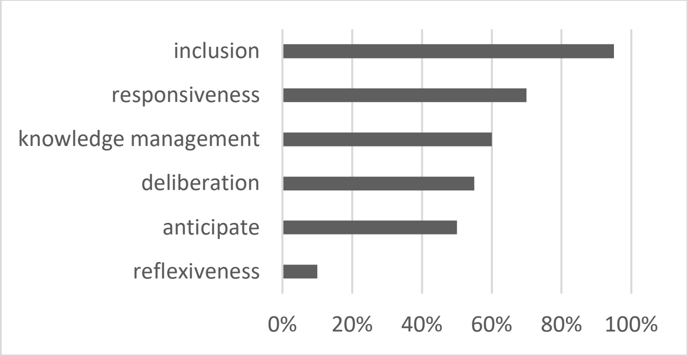
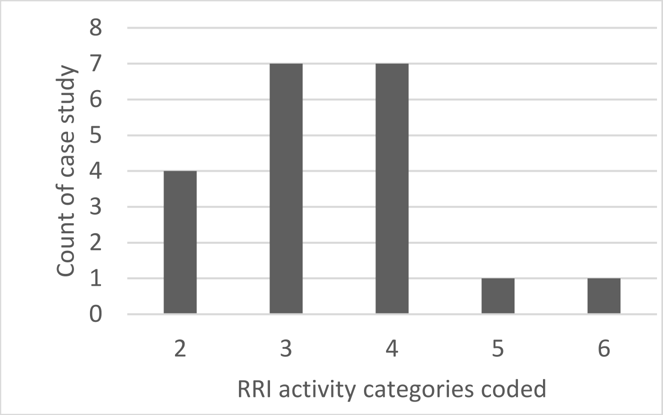


Figure 4. Number of RRI activity categories coded by count of case study.



Ethical and societal benefits were the most frequently cited impacts and appeared in 85% of cases, with the majority relating to ‘Meeting user needs and rights’ and ‘Product acceptability’ (Figure 5). Strategic benefits were assessed in 60% of cases, in particular to ‘Partner and supplier relations’ and ‘Customer satisfaction’ codes. ‘Organisational’ type benefits were reported in 55% of cases, most frequently relating to ‘Risk management’ and ‘Employee engagement’ impacts. 30% of cases noted costs of RRI practice, in nearly all cases indirect (for example delays to product development).

Figure 5. Percentage of cases studies coded against different RRI impact categories.

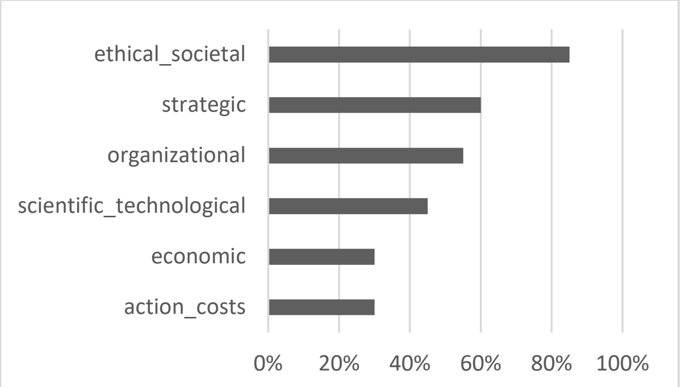
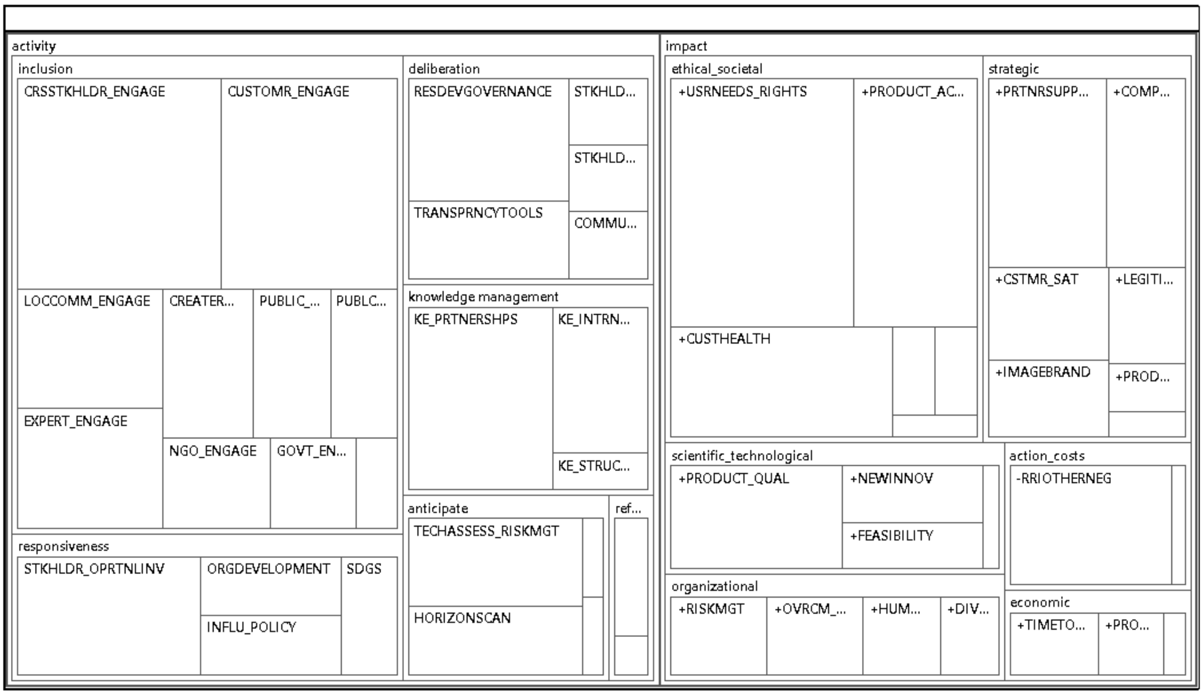


Figure 6 summarises total references to activities and impacts by code (see Appendix 1 for a list of codes and definitions). This highlights that at the detail level, the most frequently referenced items were ‘Cross-stakeholder engagement’, ‘Customer engagement’ and ‘Knowledge exchange partnerships’, and the most frequently coded impacts were ‘User needs and rights’, ‘Product acceptability’ and ‘Customer health’.

Figure 6. RRI-related activity and impacts across case studies by frequency of coding references.



3.2. Relationships between activities and outcomes

Statistical methods were applied to explore the research questions in Table 1 above and determine whether the data indicated any significant relationships between activity and impact coding, with the choice of test determined by the type of variable (‘scopeofimpacts’ is continuous and the presence of absence of an activity for a study is dichotomous).

Although a paired-samples T-test identified a significant relationship between number of reported RRI activities (M=6.75, SD=2.97) and number of reported RRI impacts (M=4.85, SD=2.72); t(19)=-2.81, p=.011, when the effect of case study length was controlled for no significant result was obtained.

No test was carried out for full scope of RRI activity against scope of impacts reported due to limitations in the data, as only one case study recorded activities in all categories.

A chi-square test of independence was performed to examine the relation between each activity category and impact category variable in turn, controlling for case study length. The following relations were identified as significant, supporting the supposition that engaging in some specified types of RRI activity is associated with particular types of organisational outcome:

- Organisations performing activities in the Anticipate category were more likely to report impacts in the Organizational category X^2 (17, N = 20) =.675, p = .002
- Organisations performing activities in the Inclusion category were more likely to report impacts in the Ethical and Societal category X^2 (17, N = 20) =.511, p = .025

A multiple regression was run to predict scope of impacts from RRI-related activity categories. Although activity category variables statistically significantly predicted scope of impacts, F(6, 20) = 3.027, p = .044, R2 = .583 indicating that engaging in some specified types of RRI activity is associated with a broader scope of positive organisational outcomes, when controlling for case study length this effect was no longer visible.

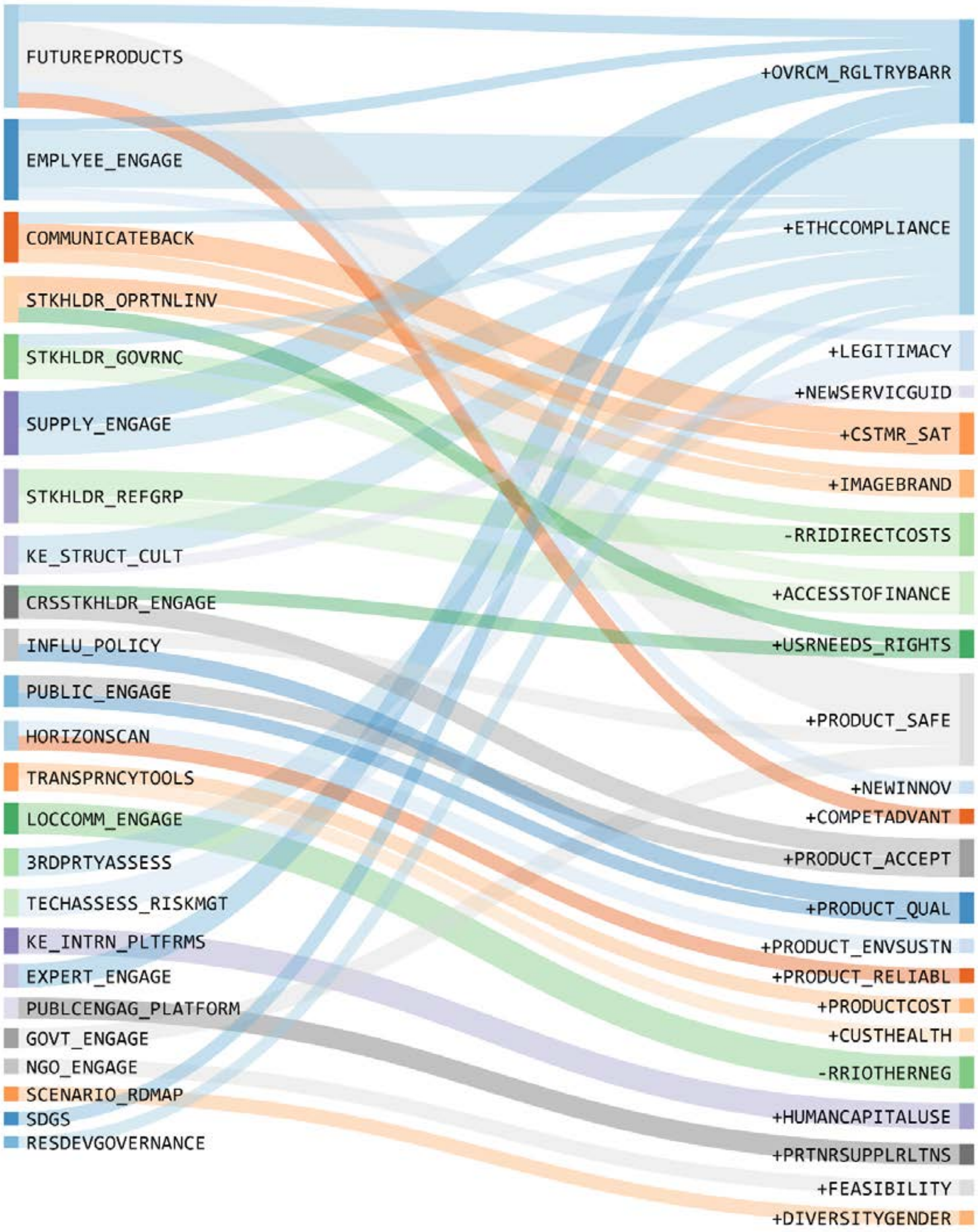
A further chi-square test of independence was performed to examine the relation between detailed activity and impact codes (controlling for case study length). A number of relations were identified as significant. Nine results which related to a single occurrence of an impact code were excluded. The strongest relationships (p<0.01, two-tailed) are summarised in Table 2 below.

Table 2. Activity-impact relationships significant at the 0.01 level.

Activity	**Impact**	**X^2** (17, N=20)
Supplier engagement	Overcome regulatory barriers	0.799
Local community engagement	Other negative impact	0.743
Technology assessment / Risk management	Improvement in perceived legitimacy	0.702
Knowledge exchange internal platforms	Human capital use	0.671
Communicate back to stakeholders	Customer satisfaction	0.670
Expert engagement	Overcome regulatory barriers	0.649
Public engagement platforms	Partner / supplier relations	0.605
Stakeholder reference group	Partner / supplier relations	-0.600
Cross-stakeholder engagement	Product acceptance	0.591

This indicates that engaging in some specified RRI activities is associated with particular organisational outcomes. Figure 7 summarises relationships between activities and impacts in the data significant at the p=0.05 level. Size of line indicates correlation strength (X^2). Code definitions are provided in Appendix 1.

Figure 7. Significant activity-impact relationships.



Hierarchical multiple regression was run to predict likelihood of product acceptability, product quality, competitive advantage and negative RRI impacts from certain combinations of RRI-related activities, separately and together with organisational metadata (sector, age, type) and implementation type. No statistically significant result was found.

4. DISCUSSION

While statistically significant associations were found between RRI-related practices and impacts, it is important to emphasise that this is a limited study of a small sample of organisations and results should not be over-generalised. A number of qualifying comments apply.

Firstly, the selection of case study organisations was a limited sample. It was not neutral, consisting of entities selected as case studies for publications in many cases because they are seen as positive exemplars of organisations applying RRI practices. In some cases, subject organisations may have been in receipt of EU funding or were participating in EU-funded projects, which may create demand characteristics or the conditions for ‘acquiescence bias’ that could influence the reporting of impacts. In any case, this was a study of RRI-practicing organisations, rather than a comparison of those practicing ‘RRI activity’ against ‘not-RRI activity’. The sample was small, particularly with respect to any intent to generalise. Although the study included organisations of different types from different regions and sectors, it would not be statistically appropriate to conclude that the relationships identified in this study necessarily exist in other regions or organisation types.

It was noteworthy from the literature search carried out that within RRI literature, a high proportion of papers were non-empirical (367 of 621, 59%) and within empirical studies a high proportion took higher education research teams as a subject, assessed higher education RRI teaching methods, reported on ‘work in progress’ RRI elements of research programmes, or were ‘snapshot’ surveys of attitudes to RRI themes. This reinforces Martinuzzi et al.'s (2018) assessment that there is a need for more empirical studies of industry engagement with RRI. At the same time, a number of studies focussed on perceived RRI ‘drivers’ and ‘barriers’, so a meta-study of these aspects would be likely to yield a larger sample size.

Setting wider parameters for the initial literature search might increase the number of case studies for analysis. Fuzzier literature search terms could overcome the issue that RRI-related activity may be reported under different conceptualisations such as ‘CSR’ and ‘responsibility’ in connection with ‘R&D’ or ‘product development’ keywords, as in the literature review by Lubberink et al., (2017) but carries the risk of including activity that can’t be meaningfully differentiated from ‘not-RRI’ activity (as noted in Thapa, Iakovleva, & Foss, 2019) . This assumes that activity can still be categorised as constituting RRI even if does not necessarily involve an explicit commitment to engage with RRI, a position taken by the Stahl et al. (2017) maturity model (Level 1 - ‘unconscious engagement’). Additional cases may also exist in the ‘grey literature’, for example in unreported outputs of a wider range of EU RRI projects to those featured in this study at the expense of including non-peer reviewed data, and inclusion of non-English publications could yield further studies.

Secondly, while small, the sample of organisations and the types of impact assessed may have been too fuzzy. The subject case studies were not developed with a standardised methodology and varied in duration and level of detail. Although effort was taken to mitigate this by considering case study duration and wordcount as independent variables and the use of well-defined criteria for inclusion, it is possible that the effects observed may relate in part to factors such as the methodology used for particular types of enquiry, for example the extent to which evidence on impact of RRI engagement was gathered, the stakeholders whose inputs were gained, and the relative ability of different types of organisation to assess activity outcomes. It is also relevant to note that the activities list used was developed on an explicit premise that the practices are based on organisations in the global North (Lubberink et al., 2017, p2-3).

In terms of activities and impacts, although the indicator lists were selected on the basis of measurability, some indicators were very broad and articulated in a qualitative rather than easily measurable manner (for example ‘improved image/brand’). This was partly mitigated in the current study by use of working definitions for analysis, but the measurability of these could be developed further, and criteria could be developed for assessing what counts as valid evidence of an impact associated with the RRI activity – for example in some cases, impacts were coded based on interviews with a small number of stakeholders within organisations. The inter-rater reliability of assessments could also be enhanced by involving co-authors in the process. Beyond this, the inclusion of some RRI-related impacts could be said to presume a Corporate Shared Value (CSV) perspective – from some industry perspectives, ‘improved customer health’ may be assessed as an indirect and societal, rather than direct organisational benefit. This could be accounted for in analysis by distinguishing direct, and indirect impacts to organisations.

A third issue is that with a mean average of 2 years for the case studies’ duration, with several constituting ‘snapshot’ descriptions of organisations at a specific point in time, limited time horizons may not reflect the lead time for all benefits arising from engagement in RRI activities. While the complex processes that mediate engagement in the context of RRI with organisational outcomes are challenging to map and measure and the lead time for benefits may vary significantly by organisation type and sector, Gurzawska et al.'s (2017) indicative causal model highlights the fact that benefits may accrue over a longer time period as a result of intermediate impacts. Similarly, intermediate benefits that precede more measurable outcomes may be harder to measure. While the Porcari et al. (2019) list of impacts used for this study was broad enough to cover a wide range of potential impacts, this issue could be addressed by either measuring broader aspects of (for example) brand perceptions, net promoter score and improvements in social and intellectual capital as a result of engagement with RRI activities, or more formal measures to gain organisational perspectives on outcomes of RRI engagement that organisations assess will drive measurable outcomes over a longer timescale. Focussing on a particular sector and/or organisation type could allow for more accurate quantification of benefits and broader time horizons for analysis.

Fourth, the decision to focus on benefits for specific organisations may not account for the networked nature of innovation processes (Dreyer et al., ibid). The resulting analysis may exclude changes to innovation ecosystems as a result of RRI processes, which may impact organisations over a longer period of time. If we accept that RRI activities can operate at a network level, and aims to enhance relations between different stakeholders and embedding scientific advances within social structures (Von Schomberg, ibid), the innovation ecosystem could be seen as the most relevant level of analysis for assessing outcomes. For example, new or increased contact between stakeholders as a result of RRI-related activity may develop social and bridging capital that strengthens innovation networks, without immediate or direct benefits to a component organisation, but increasing the likelihood of successful innovations for the future . While complex, this aspect could be assessed by probing benefits to innovation networks alongside benefits to organisations that form part of those networks, and considering the perspective of other actors within innovation networks such as entrepreneurs (Stahl & Brem, 2015). This may also include systematic consideration of different stakeholder perspectives (for example shareholder, management and employee perspectives).

Finally, the case is made in the wider literature and in particular, in the CSR literature that the configuration of responsibility activities most likely to positively impact businesses is likely to relate to other factors such as business strategy, market position and innovation strategy (Carroll & Shabana 2010, p95), the technology readiness level of relevant innovations (Stahl et al., 2017) and a range of other contextual and company-specific factors (van de Poel et al., 2017).

In most cases these data were not available in the case study material so were not included in analysis - their absence limits our ability to infer causation, since extraneous factors such as these could explain both a firm's adoption of a practice, and its achievement of particular outcomes. A future metastudy could include additional data collection for organisations to classify organisation-specific context, strategy and capabilities in more detail. This could be combined with assessment of the RRI maturity level of organisations to assess the effect on outcomes (Stahl et al., ibid) - although judging the degree of integration of RRI methods into business processes and strategy is likely to require nuanced assessment of an organisation, for example through some combination of primary data, parsing of sustainability reporting and annual reports, and consideration of other perspectives.

In conclusion,

[1] This study demonstrates a method to identify business-case-relevant relationships from a heterogenous sample of RRI case studies. With additional data, this method could provide the basis for statistically-based causal modelling to develop the model developed by Gurzawska et al. (2017), and provide a basis for business improvement tools underpinned by empirical data of practices associated with positive (or negative) RRI-related impacts. It aims to lay the foundation for better empirical evidence to support statements relating to the benefits to industry of engaging with RRI.

[2] Within the limitations of a small sample, the results indicate that certain RRI activities are significantly associated with specific organisational outcomes (Table 2 and Figure 6 above).

[3] Further studies may provide opportunities to capture a broader range of case study examples through reframing literature search parameters, inclusion of non-peer reviewed case study material, or a tighter focus on specific regions, sectors or organisation types. This might draw in CSR studies relating to research and development processes, or establishing a living dataset that enables comparison of RRI against non-RRI approaches and RRI maturity level in relation to measures of organisational impact.

[4] The limitations of the empirical evidence base for industry RRI highlighted by this study imply that future projects seeking to evaluate impacts of RRI should aim to capture benefits realised at the organisation and innovation network as well as national levels, both to develop a full understanding of the effects of RRI-related activity and to facilitate future industry engagement with RRI.

[5] Further study of the empirical evidence base for the 'business case' for industry engagement with RRI may support broader public policy objectives relating to responsible innovation in industry. This has particular relevance for companies innovating new uses of smart information systems and their stakeholders.

ACKNOWLEDGEMENTS

The author is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (UKRI Grant No. EP/S023305/1) and by ORBIT, the Observatory for Responsible Research and Innovation in ICT (orbit-rri.org).

REFERENCES

Banerjee, S. B. (2008). Corporate social responsibility: The good, the bad and the ugly. *Critical Sociology*. https://doi.org/10.1177/0896920507084623

Berghel, H. (2018). Malice Domestic: The Cambridge Analytica Dystopia. *Computer*. https://doi.org/10.1109/MC.2018.2381135

Blok, V., Hoffmans, L., & Wubben, E. F. M. (2015). Stakeholder engagement for responsible innovation in the private sector: Critical issues and management practices. *Journal on Chain and Network Science*, *15*(2), 147–164. https://doi.org/10.3920/JCNS2015.x003

Carroll, A. B., & Shabana, K. M. (2010). The business case for corporate social responsibility: A review of concepts, research and practice. *International Journal of Management Reviews*, *12*(1), 85–105. https://doi.org/10.1111/j.1468-2370.2009.00275.x

Daintith, John; Martin, E. (2010). Dictionary of Science (6th Edition) - Knovel.

Dreyer, M., Chefneux, L., Goldberg, A., von Heimburg, J., Patrignani, N., Schofield, M., & Shilling, C. (2017). Responsible innovation: A complementary view from industry with proposals for bridging different perspectives. *Sustainability (Switzerland)*, *9*(10), 1–25. https://doi.org/10.3390/su9101719

Flipse, S. M., Van Dam, K. H., Stragier, J., Oude Vrielink, T. J. C., & Van Der Sanden, M. C. A. (2015). Operationalizing responsible research & innovation in industry through decision support in innovation practice. *Journal on Chain and Network Science*, *15*(2). https://doi.org/10.3920/JCNS2015.x004

Flipse, Steven M., & Yaghmaei, E. (2018). *The Value of 'Measuring' RRI Performance in Industry*. https://doi.org/10.1007/978-3-319-73105-6_6

Fraaije, A., & Flipse, S. M. (2019). Synthesizing an implementation framework for responsible research and innovation. *Journal of Responsible Innovation*, 1–25. https://doi.org/10.1080/23299460.2019.1676685

Grunwald, A. (2014). The hermeneutic side of responsible research and innovation. *Journal of Responsible Innovation*, *1*(3), 274–291. https://doi.org/10.1080/23299460.2014.968437

Gurzawska, A., Mäkinen, M., & Brey, P. (2017). Implementation of Responsible Research and Innovation (RRI) practices in industry: Providing the right incentives. *Sustainability (Switzerland)*, *9*(10). https://doi.org/10.3390/su9101759

Iatridis, K., & Schroeder, D. (2015). Responsible Research and Innovation in Industry: The Case for Corporate Responsibility Tools. In *Responsible Research and Innovation in Industry: The Case for Corporate Responsibility Tools*. https://doi.org/10.1007/978-3-319-21693-5

Jirotka, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M. (2017). Responsible research and innovation in the digital age. *Communications of the ACM*, *60*(5), 62–68. https://doi.org/10.1145/3064940

Loureiro, P. M., & Conceição, C. P. (2019). Emerging patterns in the academic literature on responsible research and innovation. *Technology in Society*, *58*. https://doi.org/10.1016/j.techsoc.2019.101148

Lubberink, R., Blok, V., Ophem, J. van, & Omta, O. (2017). Lessons for responsible innovation in the business context: A systematic literature review of responsible, social and sustainable innovation practices. *Sustainability (Switzerland)*, Vol. 9. https://doi.org/10.3390/su9050721

Macnaghten, P., Owen, R., Stilgoe, J., Wynne, B., Azevedo, A., de Campos, A., … Velho, L. (2014). Responsible innovation across borders: tensions, paradoxes and possibilities. *Journal of Responsible Innovation*, *1*(2), 191–199. https://doi.org/10.1080/23299460.2014.922249

Martinuzzi, A., Blok, V., Brem, A., Stahl, B., & Schönherr, N. (2018). Responsible Research and Innovation in industry-challenges, insights and perspectives. *Sustainability (Switzerland)*, *10*(3), 1–9. https://doi.org/10.3390/su10030702

Martinuzzi, A., & Krumay, B. (2013). The Good, the Bad, and the Successful - How Corporate Social Responsibility Leads to Competitive Advantage and Organizational Transformation. *Journal of Change Management*, *13*(4), 424–443. https://doi.org/10.1080/14697017.2013.851953

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Owen, R. (2014). The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation. *Journal of Responsible Innovation*, *1*(1), 113–117. https://doi.org/10.1080/23299460.2014.882065

Porcari, A., Pimponi, D., Borsella, E., Mantovani, E., Van De Poel, I., Flipse, S., … Cibien, M. (2019). *Prisma - Guidelines to Innovate Responsibly - Prisma Roadmap to Integrate RRI into Industrial Strategies*. Retrieved from https://www.rri-prisma.eu/wp-content/uploads/2019/09/PrismaRRI_Roadmap_brief_web.pdf

Stahl, B. C. (2015). Responsible Research and Innovation in Industry. In *Responsible Innovation* (pp. 317–324). https://doi.org/10.5771/9783845272825-317

Stahl, B. C., & Brem, A. (2015). Spaces for responsible innovation in entrepreneurship - A conceptual analysis. *2013 International Conference on Engineering, Technology and Innovation, ICE 2013 and IEEE International Technology Management Conference, ITMC 2013*, *2020*, 1–16. https://doi.org/10.1109/ITMC.2013.7352702

Stahl, B. C., Eden, G., Flick, C., Jirotka, M., Nguyen, Q. A., & Timmermans, J. (2015). The observatory for responsible research and innovation in ICT: Identifying problems and sharing good practice. In *Responsible Innovation 2: Concepts, Approaches, and Applications* (pp. 105–120). https://doi.org/10.1007/978-3-319-17308-5_6

Stahl, B. C., Obach, M., Yaghmaei, E., Ikonen, V., Chatfield, K., & Brem, A. (2017). The responsible research and innovation (RRI) maturity model: Linking theory and practice. *Sustainability (Switzerland)*, *9*(6). https://doi.org/10.3390/su9061036

Stahl, B., Flick, C., Mantovani, E., Borsella, E., Porcari, A., Barnett, S., … Makinen, M. (2017). *Responsible-Industry: Benefits of Responsible Research and Innovation in ICT for an ageing society*. Retrieved from http://www.responsible-industry.eu/dissemination/deliverables/R-I_01_RRIBenefits.pdf?attredirects=0&d=1

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Strand R; Spaapen J.; Bauer M; Hogan E; Revuelta G; Stagl S. (2015). *Indicators for promoting and monitoring Responsible Research and Innovation : Report from the Expert Group on Policy Indicators for Responsible Research and Innovation*. https://doi.org/doi 10.2777/9742

Thapa, R. K., Iakovleva, T., & Foss, L. (2019). Responsible research and innovation: a systematic review of the literature and its applications to regional studies. *European Planning Studies*. https://doi.org/10.1080/09654313.2019.1625871

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

van de Poel, I., Asveld, L., Flipse, S., Klaassen, P., Scholten, V., & Yaghmaei, E. (2017). Company strategies for responsible research and innovation (RRI): A conceptual model. *Sustainability (Switzerland)*, *9*(11). https://doi.org/10.3390/su9112045

van den Hoven, J., & Jacob, K. (2013). Options for Strengthening Responsible Research and Innovation. In *Brussels: European Union*. https://doi.org/10.2777/46253

von Schomberg, R. (2012). Prospects for technology assessment in a framework of responsible research and innovation. In *Technikfolgen abschätzen lehren*. https://doi.org/10.1007/978-3-531-93468-6_2

Von Schomberg, R. (2013). A Vision of Responsible Research and Innovation. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, 51–74. https://doi.org/10.1002/9781118551424.ch3

Yaghmaei, E. (2016). Addressing responsible research and innovation to industry - Introduction of a Conceptual Framework. *ACM SIGCAS Computers and Society*. https://doi.org/10.1145/2874239.2874282

Yaghmaei, E. (2018). Responsible research and innovation key performance indicators in industry: A case study in the ICT domain. *Journal of Information, Communication and Ethics in Society*, *16*(2), 214–234. https://doi.org/10.1108/JICES-11-2017-0066

Yin, R. (2011). *Case Study Research : Design and Methods* (4th ed.). Fresno, CA: Sage Publications.

Appendix 1 – RRI activities and impacts

Based on Lubberink et al. (2017) and Porcari et al. (2019)

code	type	category	definition
HORIZON SCANNING	activity	anticipate	Horizon scanning / monitoring PESTEL trends
FUTURE PRODUCTS	activity	anticipate	Future-focussed product development based on long-term societal/environmental value
TECH ASSESSMENT/RISK MGT	activity	anticipate	Innovation risk management / technology assessment
SCENARIO/ROADMAP	activity	anticipate	Scenario building / roadmap development
3RD PARTY ASSESSMENT	activity	reflexiveness	Formal third party assessment of business strategy and its impact (inc. CSR-related charter marks)
EMPLOYEE ENGAGEMENT	activity	reflexiveness	Employee engagement activities (in relation to roadmap/vision)
PUBLIC ENGAGEMENT	activity	inclusion	Public engagement
SUPPLY CHAIN ENGAGEMENT	activity	inclusion	Supply chain engagement
CUSTOMER ENGAGEMENT	activity	inclusion	End user/customer engagement (inc. crowdsourcing)
NGO ENGAGEMENT	activity	inclusion	NGO engagement
EXPERT ENGAGEMENT	activity	inclusion	Engagement with experts
CROSS-STAKEHOLDER ENGAGEMENT	activity	inclusion	Cross-stakeholder engagement
GOVERNMENT ENGAGEMENT	activity	inclusion	Engagement with Government agencies
LOCAL COMMUNITY ENGAGEMENT	activity	inclusion	Local community engagement
INDIRECT ENGAGEMENT	activity	inclusion	Indirect engagement (e.g. thought experiments, role play, intermediaries)
PUBLIC ENGAGEMENT PLATFORMS	activity	inclusion	Public platforms for engagement (inc. online)
CREATE ROLES	activity	inclusion	Creation of Engagement / inclusion focussed roles
R&D GOVERNANCE PROCESSES	activity	deliberation	Formalised R&D/innovation/product development governance processes
TRANSPARENCY TOOLS	activity	deliberation	Provide transparency tools / reports
STAKEHOLDER REF.GP.	activity	deliberation	Stakeholder reference group
STAKEHOLDER GOVERNANCE ROLES	activity	deliberation	Stakeholders formal involvement in governance (e.g. Board position)
STAKEHOLDER VOTING POWER	activity	deliberation	Stakeholders have voting power
COMMUNICATE BACK	activity	deliberation	Communicate back about action taken based on stakeholder input
OPERATIONAL INVOLVEMENT OF STAKEHOLDERS	activity	responsiveness	Involvement of stakeholders at operational level e.g. project teams
ORG.DEVELOPMENT	activity	responsiveness	Organisational development/change (e.g. structure) to align with societal needs/as result of stakeholder engagement
INFLUENCE BROADER POLICY	activity	responsiveness	Engage with stakeholders to influence the broader policy or business environment
SDGs	activity	responsiveness	Specifically engage with UN SDGs
IMPACT MITIGATION PROCESSES	activity	responsiveness	Formal process(es) for action to mitigate or avoid social, environmental or economic impacts

INTERNAL KE PLATFORMS	activity	knowledge management	Internal platforms within the firm for knowledge exchange
FIRM STRUCTURE/CULTURE FOR KE	activity	knowledge management	Firm structure / culture / communication channels aligned to knowledge creation
KE PARTNERSHIPS	activity	knowledge management	Involvement in partnerships (e.g. R&D consortia)
NEW INNOVATIONS	impact	Scientific & Technological	Identify new innovations
FEASIBILITY	impact	Scientific & Technological	Improved feasibility of the technology solution
PRODUCT QUALITY	impact	Scientific & Technological	Improved product quality
PRODUCT RELIABILITY	impact	Scientific & Technological	Improved product reliability
PRODUCT LIFECYCLE	impact	Scientific & Technological	Improved product life cycle
PRODUCT ACCEPTABILITY	impact	Ethical & Societal	Improved product acceptability
PRODUCT SAFETY	impact	Ethical & Societal	Improved product safety
PRODUCT ENV.SUSTAIN.	impact	Ethical & Societal	Improved product environmental sustainability
CUSTOMER HEALTH/QOL	impact	Ethical & Societal	Improved customer health/QOL as a result of product
NEW SERVICES/GUIDANCE	impact	Ethical & Societal	Identify opportunities for improved product related services/guidance
USERS NEEDS/RIGHTS	impact	Ethical & Societal	Identify opportunities to address users' needs and rights (e.g. privacy)
COMPETITIVE ADVANTAGE	impact	Strategic	Achieve competitive advantage
IMAGE/BRAND	impact	Strategic	Improved corporate image/brand
VISIBILITY PRODUCT QUALITIES	impact	Strategic	Improved visibility of product qualities
CUSTOMER SATISFACTION	impact	Strategic	Improved customer satisfaction
CUSTOMER LOYALTY	impact	Strategic	Improved customer loyalty
LEGITIMACY	impact	Strategic	Improvement in perceived legitimacy
PARTNER/SUPPLIER RELATIONS	impact	Strategic	Improved relationships with partners, suppliers and sub-suppliers
ETHICAL COMPLIANCE	impact	Strategic	Demonstrate compliance with ethical/social requirements (e.g. for funding)
USE OF HUM.CAP.	impact	Organizational	Improved use of human resources
EMP.ENGAGEMENT	impact	Organizational	Team/employee engagement and motivation
REGULATORY BARRIERS	impact	Organizational	Address regulatory barriers
RISK MANAGEMENT	impact	Organizational	Improved risk management
GENDER/DIVERSITY	impact	Organizational	Gender and diversity contribution to product development
IRRESPONSIBLE BEHAVIOUR	impact	Organizational	Avoid irresponsible behaviour
PRODUCT COST	impact	Economic	Reduced product cost
TIME TO MARKET	impact	Economic	Reduced time to market
PROFIT/SHARE	impact	Economic	Increased profit or market shar
ACCESS TO FINANCE	impact	Economic	Improved access to financial support
RRI DIRECT COSTS	impact	RRI action costs	Increase in costs due to RRI activity
RRI OTHER NEG.IMPACT	impact	RRI action costs	Other negative impact on company due to RRI activity

Appendix 2 – Characteristics of the case study selection

Implementation type	**Count**
Operational	7
Strategic	13
	20

Organisation type	**Count**
Social enterprise	7
Joint Venture	4
SME	3
Multinational corporation	3
Limited company	2
Public body	1
	20

Sector	**Count**
Agriculture/Food Production	6
Education	2
Financial Services	1
Healthcare Technology	8
ICT	2
Nuclear energy	1
	20

Region	**Count**
EU	12
North America	3
Asia	3
Africa	2
	20

Data used for analysis are provided on a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License here: : https://drive.google.com/drive/folders/1XB0_PeR66wZydXTlQu5-iBa0bZKb5MQL?usp=sharing

AI AND ETHICS FOR CHILDREN: HOW AI CAN CONTRIBUTE TO CHILDREN'S WELLBEING AND MITIGATE ETHICAL CONCERNS IN CHILD DEVELOPMENT

Ryoko Asai

Uppsala University (Sweden), Meiji University (Japan)

ryoko.asai@it.uu.se

ABSTRACT

Along with the development of Information and Communication Technologies (ICTs), new genres of the tech industry have been fueled by the prospect of the emerging need. Especially artificial intelligence (AI) makes it possible to enhance machines' ability to support caregivers in many occasions. For example, one of the emerging genres drawing people's attention is high-tech for childcare and family life, so called 'baby tech' or 'family tech'. We see that children use mobile phones, tablets and computers as daily commodities everyday. In addition, high-tech companies develop and release social robots for children and family currently, such as Pepper and Jibo. Actually, both children and parents enjoy the many different functions of social robots. However, social robots bring us classic and novel ethical issues behind great benefits. In this study, we focus on social robots for children and family, and explore how AI can contribute to child wellbeing whereas there are ethical concerns for child development.

KEYWORDS: Artificial Intelligence, Ethics, Social Robot, Child Wellbeing.

1. LIFE IN AI SOCIETY

When the prediction by Frey and Osborne was announced in 2013, people stirred up strong fear of losing jobs in the near future. They pointed that highly advanced technologies, especially Artificial Intelligence (AI) technology, would promote rapid innovation in business, and many human workers would be replaced to AI technology and lose their jobs in 10 to 20 years (Frey and Osborne 2013; Frey, Osborne and Citi 2015). After Frey and Osborne's analysis, the report about employment and technology was also published by World Economic Forum in 2016 (World Economic Forum 2016). These researches show people who engage in non-skilled or manualized work would face a great risk of losing their jobs in the future, worse some of jobs might be totally eliminated from the earth. On the other hand, new jobs would be created in order to support AI-based society and more job opportunities would open for workers with high-skill or creativity. Our life including both working life and private life would be influenced by technologies regard- less of whether we like it or not.

Once AI takes over our jobs and we get more free time, how do we use free time? Some people might expect to have much more time with their families and take more care of their children. However, AI and high technologies are deployed and equipped for daily use at home, and take over household work and daily chores. Moreover, social robots with AI stay with children, play together and entertain them at home. The question of how to nurture children, those are expected to maintain and support society in the future, in the environment deployed highly advanced technology everywhere, is a kind of vital and important social task for the future. How do we use AI for childrearing, and also how do we live with AI?

In high-tech society, we take technologies as given commodities and suppose to make life more efficient and effective by utilizing them. However, children need inefficient and ineffective processes and matters in order to develop themselves. Whereas AI supports us in any aspects of our daily lives, injudicious and strong dependence on AI could dehumanize life and evoke classic yet new ethical dilemma: how we live and what is good/right. This study explores how AI affects our daily life from the perspective of in- formation ethics, especially focusing on parenting with social robots at home.

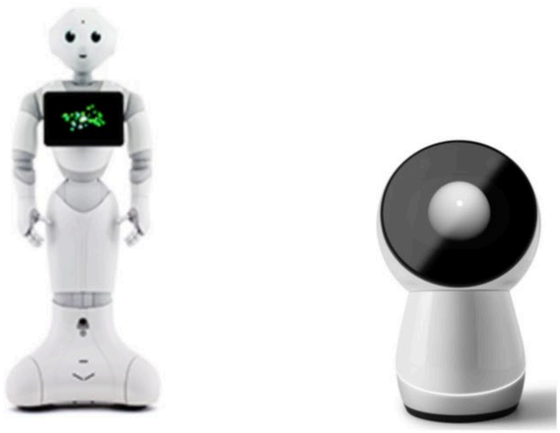
2. CHILDCARE AND SOCIAL ROBOT

Rapidly evolving reproductive technologies support human reproduction greatly. However, recreating human beings needs to use human beings, at least sperms and eggs. Human beings cannot be replaced by any intelligent robot at the moment. If we human beings want to keep “humanism” or “humanity”, we cannot hand over our position in the earth to any other creation. It is necessary for us to reproduce human beings for the future. And in order to exist in the future, we need to think about the way to use highly advanced technology for childrearing, not only from the perspective of efficiency and profitability but also the ethical perspective.

In the research report by Frey and Osborne, school teachers are not listed as a job which would be taken over by technology in the future. Rather, the school teacher’s job would be very important to educate children, and require teachers to have creativity and high ability to cooperate with others. On the other hand, child’s learning processes and plays show a strong affinity for technology. Thus, many companies and researchers have been working on developing software, digital contents and robots for children (Nesset and Large 2004; Yamamoto et al 2004; 2005; Cangelosi and Schlesinger 2015). In the near future, these technologies targeted for children is supposed to be introduced into schools in order to improving efficacy of education activities, and the number of school teachers would be reduced (The Economy, Trade and Industry Ministry 2013). That means children would study and learn with the support of the greatly deployed technologies and the limited number of teachers. In this situation, childcare at home would be more important to learn social life with others and establish themselves as a social existence.

In June 2015, Softbank robotics which is the subsidiary company of a major telecom company in Japan has released “Pepper” in the mass market. Pepper is “the first robot designed to live with humans” and has a human shape and the ability to read/express emotions (Emotion Engine), and communicate with human beings. Pepper is supposed to be our companion to stay together, entertain us, and make human life happier, not help housekeeping or bringing a heavy box. Pepper is equipped with many traditional communication functions as computers have offered, such as taking pictures, mediating email exchanges, reading texts with voice and so on. Moreover, Pepper can recognize users’ face expressions and voice tones beside them, and show reactions or talk to them in response to their feelings. Bruno Maisonnier, who is Aldebaran robotics CEO and the responsible of Pepper project, explained that“(t)he most important role of robots will be as kind and emotional companions to enhance our daily lives, to bring happiness, to surprise us, to help people grow" (Guizzo 2015). Currently more than 200 companies have joined software development to move Pepper, and even some companies use Pepper in the business occasion.

Figure 1. Left: Pepper (http://www.softbank.jp/robot/consumer/products/spec/), Right: JIBO (https://www.jibo.com)



Source: Left: http://www.softbank.jp/robot/consumer/products/spec/, right: https://www.jibo.com

And also, Cynthia Breazeal, who is the roboticist at MIT's Media Lab, announced she would launch the social robot JIBO in 2015. JIBO has "skills" to recognize emotions and "is designed as an interactive companion and helper to families, capable of engaging people in ways that a computer or mobile device aren't able to" (Guizzo 2014). JIBO is also equipped with core applications, which is called as "skill," and users can set up JIBO's role at home depending on the expected usage environment by users. The characteristics of JIBO are the robot is 1) caring for users' emotion and 2) able to improve the communicating attitude through the interaction with users.

Both robots, Pepper and JIBO, are designed to communicate and interact with humans continuously and AI equipped on both robots learns about users through updating users profiles and favors constantly. And eventually both robots aim to be a member of family or a friend of users. Because of this purpose, both robots suppose to have users who don't acquire enough media literacy or computer skills, such as small children.

Family robot, such as Pepper and JIBO, are categorized into "personal service robot" according to the categorization by International Federation of Robotics (IFR) and International Organization for Standardization (IOS) 2. Generally, "personal service robot" is used for a personal task, not for a commercial task, for example automated wheelchair, and personal mobility assist robot. Users can customize a robot in accordance with the intended use, their wish and taste. In terms of family robot, the robot aims to be recognized as social existence through communication and interaction with users, rather than helping users daily life practically. Therefore, family robot is called as "social robot," "sociable robot" and "social intelligent robot" (Breazeal 2002; 2003: Fong et al. 2003; Dautenhahn 2007).

However, the research and argument about how robots recognize and judge human emotions correctly and properly are still on-going and very controversial. Even in philosophical and moral studies, “mind-body” problem or “mind-brain-body” problem are not solved yet, rather those problems are getting more complicated reflecting to the development of AI. Generally, emotions are necessary to make human beings social existence (Evans 2001). In this sense, emotion is a critical factor for social robot to be recognized as social existence by users. At the moment, social robots cannot feel emotions autonomously and voluntarily. The way in which social robots “have” feelings is to use the corrected data via interaction with users and show suitable reactions to them.

When it comes to an emotional function and social robots, the most impressive phenomenon is that social robots can stimulate users’ emotions and make them form a strong attachment to their own robots. What makes a robot as a social robot or social existence is our emotions, not the emotional function of a social robot. We human beings give a robot a social meaning. In other words, social robots cannot exist without developing the technology to detect and judge users’ feelings correctly.

3. FAMILY LIFE WITH SOCIAL ROBOT

3.1. Basic functions of social robot

Social robots as family robot or friend robot equip generally three basic functions (Asai 2017).

a) Entertainment function: singing, dancing and playing game

b) Security function: monitoring through webcam, talking from a distance via Internet

c) Facilitation and revitalization of family communication: providing family a trigger of conversation

Although social robots cannot clean the house or cook foods, they can sing a song with children, read a book for children before going to sleep, or check children and house via webcam when parents are absent at home. In light of definitions of “care robot”, social robots could function as a caretaker (van Wynsberghe 2016; Vallor 2016). When social robots are seriously recognized as a member of a family at home, how do childrearing and childcare change? And what kind of ethical concerns is caused? According to Whitby, serious ethical problems are hidden in the invisible part when social robots perform tasks (Whitby 2012).

3.2. Ethical concerns in the use of social robot

Generally there are three basic ethical concerns in the use of social robots including care robots.

a) There is a risk which users get socially excluded or socially isolated because of too much attachment and emotional connection to social robots. The social relationship is superseded by the relationship with social robots (Sparrow and Sparrow 2006; van Wynsberghe 2016).

b) Privacy and integrity of users (caretakers) might be damaged by social robots (Vallor 2016).

c) Social robots might generate new inequality between “robot-have”/”robot-not-have” or skilled users/unskilled users, based on age, income level, the development level of countries and societies and so on. Or, existing disparities might be amplified by the use of social robots (Asai 2017).

Although there are some ethical risks, social robots would bring positive effects to our daily life. For example, social robots would have a great possibility to support parents and families in childrearing. Especially working mothers in gendered society might reduce their workload and stress of taking care of children by using social robots. And its monitoring function could give parents a secure feeling while they are absent for work and children stay at home alone. Furthermore, when social robots improve functions to communicate with users and take care of childrearing or household more and more in the future, people who currently live in obedience to gender norms might be able to be free from their gender roles.

Social robots constantly collect and store the enormous amount of personal information, connect to cloud date and update their abilities. For users, giving their own information to social robots is necessary to improve their robots' functions (IBM Japan 2014). Once social robots are recognized as a family member by users and stay together all the time, the robots can gather various kinds of personal information including sensitive information. While a huge amount of personal information improves the usability of social robots better and better, we need to be aware of ethical concerns behind it.

3.3. Ethical problems in operating social robot

There are three typical ethical concerns in operating social robots (Asai 2017).

a) As long as social robots function based on our personal data, there is a risk to breach privacy or leak personal information.
b) In order to manipulate social robots, we need to use a kind of "robot infrastructure" to operate cloud AI and robot OS for collecting and analyzing data. On the other hand, social robots are operated with the collaboration and cooperation of various technologies by various companies. How and who manage and control the robot infrastructure is critical to protect our privacy and personal data.
c) Social robots are customized for particular users through the interaction and communication with them. Each social robot is made up by the collaboration of robot designers, engineers, venders, operators and users, and its function is differentiated depending on users. In this context, is it possible to consider a customized robot as intellectual property? If so, who is allowed to own the robot? And also if your customized robot compose a beautiful song or do a painting nicely, who can own the intellectual property right of those creations?

When thinking about ethical concerns in the operating process, we need to see and check the problems from legal and political aspects as well as from an ethical aspect. However, in real, technology develops very rapidly, and legislation and politics sometimes cannot catch up on the development of technology. Or, laws and policies are not suitable for the reality and get to be outdated because new technology changes society. When laws and politics don't function properly, ethics has a great possibility to work as "social coordination technology" (Sakamoto 1974). In other words, ethics could contribute to coordinate or manage the fluctuated situation based on morals and values that people have inside. In AI age, ethics is needed more and more to solve the chaotic social situations.

4. INVISIBLE ETHICAL CONCERNS

Once ethical problems cause in the operating process, those problems would be recognizable for users. In the worst case, social robots might stop functioning because of those problems. However, more serious ethical problems with social robots for childcare use are hard for users to see and recognize. First of all, as previous researches have already shown, it is very difficult to be free from embedded values in designing and developing technology (Friedman et al. 2006; Nissenbaum 2011). Recently a Japanese big ICT company NTT (Nippon Telegraph and Telephone) has developed AI to search and offer picture books for children and children's parents. This AI-based searching system can choose picture books suitable for children's age, interest, taste, and so on based on enormous book data. Especially for children's parents, it would be a great help to find "good" books among thousands books. It might bring a chance to know a nice book which they have never read before. However, it is very difficult for children and their parents to know how the algorithm works and finds proper books for them. Worse, we don't know if the book which the system offers is really proper or good. Social robots can select and read nice books for children. However, picture books selected by the robot might have influence on their thoughts and lifestyles without noticing.

Second, generally technologies including social robots are utilized to reduce our work- load and improve efficiency in daily life. Increasing interaction between children and family robots might decrease communication between children and parents. The typical example is that family robots read a book for children before going to sleep, instead of parents. Indeed, the absence of parents could be complemented by social robots and parents could feel less stress and have more free time. Luckily family robots basically don't say "NO" to children and don't deny them. Children enjoy freedom for doing what they want to do under the robot supervision. However, they need to learn to be independent through experiences of holding and refusal by parents (Okonogi 1992; Winnicott 1988). Especially no one takes the role of denying children (the role of "father") any more. Having a faithful companion for children might interfere the process of developing children's independence.

And thirdly, technologies intervene the childrearing environment and dehumanize or artificialize childrearing. Dehumanized environment where is managed by technologies could be more reasonable and rational in correspond to well-calculated and well- programmed algorithm, when comparing to the environment controlled by human. How- ever, we are sometimes overwhelmed by the irrational and unreasonable situation. Generally we have learned how to deal with and solve problems in perverse situations through experiences since childhood. Of course we have to learn the way to handle the difficult situation until we die. The dehumanized but well-programmed environment inhibits children from developing the ability to get along with difficult situations when things don't go as they wish. When we try to make a decision, values, independence, and problem-solving skills are key elements to reach a better decision. If social robots tame children to be depend heavily on support from social robots, they lose chance to acquire skills and develop abilities in order to solve problems and make a better decision.

Furthermore, children's experiences via virtual technology or social robots could be completed inside themselves without thinking about any others around children. It would change the quality of experiences. They don't experience something in person or in direct, but they can feel and see something very similar to the real without any struggle or conflict. That means children can live without communication with other humans, and cannot recognize and share any feeling among others and position own existence and identity in society (Ichikawa 1992). This is not pessimistic prediction. We have already started to adjust ourselves to the current technological environment to enjoy efficiency and benefits. While we adjust ourselves to technology, we exclude ourselves physically and mentally from the high-tech society and make ourselves 'others'.

In general, technology could be used for dual- or multi- purposes, and sometimes it is used for unexpected purpose. Social robot is no exception in this regard. Although benefits from social robot are remarkable and attractive for us, we need to recognize how we use social robot and see how it influences on our lives from the ethical perspective.

ACKNOWLEDGEMENTS

This research was supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (Grant-in-Aid for Young Scientists B, No.16K16169), and also by JSPS Grants-in-Aid for Scientific Research (Scientific Research C, No. 20K12551). I thank my colleagues from Prof. Kiyoshi Murata at Meiji University and Prof. Iordanis Kavathatzopoulos at Uppsala University who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations and conclusions of this paper.

REFERENCES

Breazeal, C. (2002). *Designing Sociable Robots*. MIT Press.

Breazeal, C. (2003). Toward Sociable Robots. *Robotics and Autonomous Systems*, 42, 167-175.

Cangelosi, A. and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. MIT Press.

Dautenhahn, K. (1998). The Art of Designing Socially Intelligent Agents – Science, Fiction, and The Human in The Loop. *Applied Artificial Intelligence*, 12, 573-617.

Dautenhahn, K. (2007). Socially Intelligent Robots: Dimensions of Human-Robot Interaction. *Philosophical Transactions of the Royal Society B*, 362, 679-704.

Evans, D. (2001). *Emotions: A Very Short Introduction*. Oxford University Press.

Fong, T., Nourbakhsh, I. and Dautenhahn, K. (2003). *A Survey of Socially Interactive Robots*. Robotics and Autonomous Systems, 42, 143-166.

Frey, C.B. and Osborn, M.A. (2013). *The Future of Employment: How Susceptible Are Jobs to Computerisation?* Oxford Martin School, University of Oxford.

Frey, C.B., Osborne, M.A. and Citi (2015). Technology at Work: The Future of Innovation and Employment. *Citi GPS: Global Perspectives & Solutions*, February 2015.

Friedman, B., Kahn, P.H. and Borning, A. (2006). Value Sensitive Design and Information Systems. in Zhang, P. and Galletta, D. (Eds.) *Human-Computer Interaction and Management Information Systems: Foundations*. M. E. Sharpe (republished by Routledge in 2015), (pp.348-372.)

Guizzo, E. (2015). A Robot in the Family. *IEEE Spectrum*, January 2015, 26-29&54.

Ichikawa, Hiroshi. (1992). *Seishin toshite no shintai*. Kodansha (in Japanese).

Ichikawa, Hiroshi. (1993). *"Mi"no kozo: shintairon wo koete*. Kodansha (in Japanese).

Jasanoff, S. (2016). *The Ethics of Invention: Technology and The Human Future*. Norton.

Kranzberg, M. (1986). Technology and History: "Kranzberg's Laws". *Technology and Culture*, Vol. 27, No. 3, 544–560.

Ministry of Economy, Trade and Industry. (2016). *Shinsangyokozobijon*. Available online: http://www.meti.go.jp/committee/sankoushin/shin_sangyoukouzou/pdf/008_05_01.pdf. (in Japanese).

Mori, M. (2012). The Uncanny Valley. *IEEE Robotics & Automation Magazine*, Volume.19, Issue: 2, June 2012, 98-100.

Nesset, V. and Large, A. (2004). Children in the information technology design process: A review of theories and their applications. *Library & Information Science Research*, 26 (2004), 140–161.

Nissenbaum, H. (2001). How Computer Systems Embody Values. *Computer*, 34(3), March 2001, 118-120.

Okonogi, Keigo. (1992). *Jikoainingen.* Chikumashoten (in Japanese).

Sakamoto, Hyakudai. (1972). Atarashii Gijutu heno Michi. *Philosophy of Science*, Vol.5., 143-165. (in Japanese)

Sparrow, R. and Sparrow, L. (2006). In the Hands of Machines? The Future of Aged Care. *Minds and Machines*, 16(2), 141-161.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

Vallor, S. (2016). *Technology and The Virtues: A Philosophical Guide to A Future Worth Wanting*. Oxford University Press.

van Wynsberghe, A. (2015). *Healthcare Robots: Ethics, Design and Implementations.* Ashgate Publishing (republished by Routledge in 2016).

COLLECTED FOR ONE REASON, USED FOR ANOTHER: THE EMERGENCE OF REFUGEE DATA IN UGANDA

Annabel Mwagalanyi

De Montfort University (United Kingdom)

P06285689@my365.dmu.ac.uk

ABSTRACT

Although there is substantial research on the surveillance of refugees in developed countries, there is relatively limited research on the topic in developing countries. This is partly because these countries have only recently begun implementing modern surveillance technologies to manage their refugee population. However, the consequences of these trends are huge for example; spillover effects to other countries, cultural differences in understanding biometric data collection and social divides. This may exist among religions and ethnic groups resulting in misunderstanding, mistrust and systematized oppression. Technological advancements may have good intentions but function creep often exists where the original purpose the data is collected for goes beyond this resulting in unintended consequences (Maitland, 2018). Due to their growing numbers and their changing demographics, it is becoming more urgent to study the lived experiences of refugees in this region as they are becoming increasingly subjected to digital governance, surveillance, and control.

KEYWORDS: refugee, asylum seeker, surveillance, technology, biometric.

1. INTRODUCTION

This paper focuses on the emergence of digital surveillance technologies, used for governing the refugee population in Uganda. Yes, it is new in one way (because of technological advancements) but in another way its old. Actually, practices that are very similar to surveillance of migrants in the African continent are not new. Migration, displacement and surveillance in the African continent are practices that are known since biblical times. Exodus 12:32 describes the mass movement of peoples, and practices of registering data about individuals and families were recorded in various resources. In addition to this, Weaver (1985) wrote of Ethiopian refugees in Khartoum arguing that population movements across Africa have deep roots in its history, as many were displaced due to natural disasters and conflict with neighbouring communities.

As a result, migrant surveillance is both old and new, specifically, what is new is the uncertainty with regards to the unintended consequences and although governments justify the use of digital surveillance technology to facilitate more inclusion, in reality, it might compromise the lives and dignity of refugees in at least four ways:

1. Even if this information is collected for good purposes (e.g. for fair distribution of food and resources) it could be used in the future for the bad, giving future governments more power in an unstable region.
2. Surveillance affects refugees' behaviour and perception of the host country.
3. Data can have errors or be compromised which could lead to people being treated inappropriately. With the absence of legal frameworks such as the General Data Protection Regulation (GDPR), this makes it more of a problem.
4. Aggregating data is conducive to treating people as monolithic collectives rather than particular individuals (racial profiling).

Refugee data collection is often carried out at the arrival stage of the journey to a refugee camp or a country border point. In Uganda refugees are welcomed at a transit centre and are required to hand over information concerning them, using biometric systems such as fingerprint and iris scans. Once this data is collected it may grant the refugee a chance to move unto a camp or organised settlement area. This is also to enable the refugee to receive assistance or the chance to seek asylum. In the interim, it acts as a form of documented identity. By 2030 the United Nations' goal is to ensure all human beings have some form of identification (UNHCR, 2018), but this is a difficult task as many hurdles need overcoming such as the misuse of refugee data when it is used above and beyond what it was intended for originally. There is the occasion when the data is used for legitimate reasons and in line with government policies; however, it has a significant impact on refugees' lives (e.g. EURODAC fingerprinting asylum system in the European Union (EU)). We have always had surveillance technology in operation in some form or another but the emergence of new technologies raises concerns in developing countries such as Uganda with over 1 million refugees, heavily dependent on foreign aid but yet still in need of measures to help manage the refugee crisis. This paper will, therefore, attempt to discuss these issues in more detail. The aim is to gain insight that could influence policymakers, appreciating the consequence of surveillance technologies in developing countries.

2. DOCUMENTED IDENTITY IS THE GATEWAY

In 2018, the UN refugee agency rolled out a major refugee verification operation and the project aims to ensure all refugees are registered and receive the protection and assistance they need (e.g. biometric identification and food ration cards). The organisation uses its software. The Ugandan phase of the project was the biggest in the agency's history (UNHCR, 2018). The United Nations (UN) also reports that in 2018 the population was 42 million and more than half of its population was under the age of 30. This project is in line with sustainable development goal 16, advocating for peace, justice and stronger institutions. In other words, ensuring that by 2030 there is a legal identity for all and to ensure that there is "public access to information and protection of fundamental freedoms." This would have to align with national legislation and international agreements (Peace, justice and strong institutions - United Nations Sustainable Development, 2020).

For a refugee, it is important to receive documented identity as it acts as a gateway to education, employment and health services, which by law all human beings are entitled to (Maitland, 2018). Furthermore, identification also provides self-worthiness acting strongly as an integration tool to society pulling a refugee away from living in isolation which many find themselves when forcibly displaced. However, receiving documented identity is a double-edged sword because on the one hand it is a gateway, on the other hand, there is an imposition of an identity that the refugee did not necessarily choose, in a foreign language; someone is telling you who you are. It puts them in a vulnerable situation because they do not know who they can trust, what will happen to their data. Also, a refugee knows that in a way they are not necessarily wanted by the host country.

Countries where there is instability caused by political conflicts and poverty face greater concerns beyond imposing their own identity on individuals that seek refuge on their soil. In fact, governments may seek quick solutions in these emergency situations, after all, it is not ideal to be seen as unwelcoming to refugees especially in Uganda. To illustrate this, let us consider the self-reliance strategy for refugees, which is encouraged by the government and UNHCR in Uganda. The idea is to encourage refugees who have remained in the country for 5 years or more to be able to support themselves; relying less on foreign aid or “handouts” from the state. This neo-liberal approach would see little intervention from the government, refugees granted plots of land to do grow crops and trade within local markets. However, some of the refugees expected to follow this approach are based in isolated areas where accessing nearby villages to trade and make money are 60km away. The Refugee Law Project (RLP) from Makerere University (an organisation dedicated to provide a voice and legal aid to refugees) staff advocate for the need for local integration of refugees and their host communities as opposed “to confining refugees to isolated and harsh settlements” (Ilcan et al, 2015, p.8). RLP argues that a self-reliance strategy (SRS) could work but it has to go hand-in-hand with other aspects such as giving refugees the rights to remain in the country e.g. via citizenship by naturalization. This would then result in easier access to integrate into society. However, despite some refugees being in the country for 15 years, it is a real challenge as there are several cases where they are considered not eligible for permanent residency argues Ilcan, Oliver and Connoy, (2020). This strategy is a concern because although refugees are given a level of freedom there is the risk of some being excluded from integration because of living far away. Uganda is experiencing massive deforestation in heavily populated refugee regions due to the need of firewood to assist with house construction, fuel and to create charcoal for cooking. Local’s in the area are displeased by this so refugees often state they have been targeted, physically sometimes because of this (Okiroro, 2019). The UN and World Bank also warned that the lack of resources would result in tension and program director from International Refugee Rights Initiative stated: “If nothing is being done, this will seriously put to the test the considerable hospitality that Ugandans living in refugee-hosting areas have been showing in recent years” reports Okiroro (2019). Refugees perceptions of the host country can change in these situations especially when their lives are being threatened by the host community.

With little government intervention, as the SRS approach recommends raises questions about the ethical responsibility of caring for vulnerable populations.

Developing countries may find technical difficulties in building and maintaining a robust identity system but the past testifies of misuse of identity even when it is recorded in its simplest form, on paper. During the Rwandan genocide, ID papers showed details of tribes and this made it easier to find and target Tutsis whose lives were then terminated (Economist, 2019). Kenya is another country that has been accused of discriminating against the Nubian minority. Despite this community living in the country for over a decade, they had to endure extra vetting procedures, prove their nationality and contest for obtaining an identity card (Balaton-Chrimes, 2013). The other 42 tribes in Kenya did not have to do this. An example of how governments can misuse data to fulfil hidden agendas, negligently overlooking that it is a human right for an individual to have documented identification.

More research is required to understand internal factors which may act as barriers to the gateway of providing documented identity for refugees in these communities. Several European countries have developed stricter policies to address the refugee crisis and the process of collecting biometric data such as fingerprints, facial recognition, and iris scans have gone beyond initial intention as research conducted in Europe suggests.

3. DEVELOPED COUNTRIES AND SURVEILLANCE TECHNOLOGIES

A key European system associated with the digital surveillance of asylum seekers and irregular migrants crossing borders in the EU region illegally is known as EURODAC (European Dactyloscopy Scheme). But the EU's biometric database, operational since 2003, is the subject of public controversy. EURODAC operates as control technology sharing and communicating information using “asylum seekers and irregular migrants” (Kuster and Tsianos, 2016). A key feature of the system is the Automated Fingerprint Identification System (AFIS), active in countries that apply the Dublin III regulations. The regulation states that the first country (or member state) in which the asylum applicant had entry to, Europe is responsible to conduct the asylum process. Critics of EURODAC claim it violates human rights. The reason for this argument is because the initial purpose of the system was to gather all asylum claims made in the EU region but was then integrated with Europol a law enforcement agency in Europe (Sànchez-Monedero, 2018). This extension was made without consent and left refugees’ asylum records being contrasted with criminal records. Another criticism is police have access to databases, often treating refugees like criminals and suspects. EURODAC has birthed non-state initiatives, including private border patrols and counterfeit border checkpoints. Further developments include iBorderCtrl, an automated deception detection system using artificial intelligence. The system uses a virtual agent to conduct asylum interviews asking questions about refugee’s backgrounds and intentions. This raises concerns as the question is asked, how can a machine be depended on to interpret the intentions of a human-being? How reliable is this approach and could it potentially favour a certain type of refugee? However, the Guardian newspaper spoke to experts in the field of AI who argued it is almost impossible to design an experiment that evaluates deception behaviour. The program assumes refugees potentially may be lying and this has a negative impact as it can make them feel they are treated unfairly and that the host country is being hostile.

Another example of impacting refugee life is the use of data from mobile phones and social media in the EU. The data is used during asylum evaluation interviews to detect a person's accent for example (Meaker, 2018). In this scenario refugees are in a predicament because digital devices such as mobile phones have become an indispensable tool, guiding them along migration routes and supplying information for their asylum claims. Agencies have access to text messages, location reports and browsing history despite it being deleted by the phone owner. This raises the question of who benefits from systems of detection and control such as EURODAC in a time where its methods are being adopted by developing countries to tackle the refugee crisis. A key point to note is that the findings discussed present the current impact technology advancements have had in developed countries, little being said of the global south region.

Modern identity systems promise to bring many benefits to Africa. But as they proliferate, so too will the temptation for politicians to misuse it (The Economist, 2019). However, this also raises the questions, are African countries in need or seeking for the promises of these IT solutions? After all, in Uganda, for example, the government is often of the hope that refugees will return back to their country of origin and may be happy with just the basic verification systems provided by humanitarian organisations such as the UNHCR (as long as they can prove the who, why and when refugees are in the country might just be good enough). It is important for policy makers and local government to be aware of the wealth of having a robust account of the refugee population, avoiding the predicament of having real bars of gold (the data) but not knowing its monetary worth. This especially applies to the “donors’ confidence” from abroad who supply monetary and food aid. At the same time research from Nakivale refugee camp (one of the oldest refugee settlements in Africa) suggest that refugees in Uganda remain in the country for protracted periods of time, which can be 5 years or more (Ilcan, Oliver and Connoy, 2020) reflecting how long for their countries of origin may experience unrest. Therefore, being able to monitor this population is vital.

In addition to this, UNHCRs documentation process, as many other international humanitarian organisations, the usual agenda is to dictate and use a “one size fits all” approach such as that of documenting refugees across the globe. Countries without home built IT solutions to either strengthen or contest UNHCR’s identification process have very little input, hence, running the risk of it not being suitable in the long run for the particular country in which it is being used. This can be considered in light of “technology failure” and a good example is provided by Kingston (2018). Looking at human behaviour, aid workers identified that “if a refugee burns their finger while cooking, it may take several days before their fingerprint are accurately identifiable” stated Kingston (2018, p.48). This raises the risk of the refugee not being able to obtain food rations, for example, which could have a negative impact on their entire household. In refugee camps the make up of the family may include a parent or guardian responsible for collecting food rations, but if they can’t due to a burnt finger, they can not delegate someone else in the family because only their fingerprint would be registered on the system. These strict rules to the identification system are setting stone and refugees have to comply. In these instances the overall mental stress caused is potent.

Data protection laws are inadequate in the African continent and cannot be automatically enforced like the General Data Protection Legislation in Europe, some African countries have recognised this working and are working on a solution. The impact of this is that refugees, a group perceived as citizens of nowhere and whose interests are not represented by governments, are at high risk for exploitation. Although they have very little control over the situation they are in, their identity is being challenged and constructed anew by forces greater than themselves. Some experience enforced iris scans in return for aid, their phones may be seized as a form of identity verification and biometrics are used for categorization or evaluation of their rights and benefits. Due to the lack of legal frameworks governing data in Africa, there have been instances where mobile phone operators such as Orange were discovered to be offering Africans fewer digital rights than their European subscribers. In 2018 Ugandan officials exaggerated refugee figures by 300,000, fake names were created in refugee settlements and defrauded millions of dollars in aid (Okiror, 2019). This resulted in officials from the office of the office of the prime minister being suspended. This demonstrates how refugee data can be misused, impacting their lives (e.g. less aid due to sponsor reducing aid) all due to short-term greed. Africa has been lagging when it comes to addressing privacy issues around data argues Gwagwa (2019), but it cannot afford to do so any longer because the state can shrivel and censor data traffic for self-serving purposes. New technologies are often created to help solve humanitarian issues but often exceed their initial intentions leaving unintended consequences

(Maitland, 2018). Soliman (2016) pointed out if the data falls in the wrong hands creates vulnerabilities for refugees, however, if the data is not shared can leave many countries open to security threats. It is of great importance that governments, policymakers and organisations are made aware of the potential damage new technological developments create bearing in mind that not all humanitarian disasters have a technological solution.

4. CONCLUSION

In light of this paper, it seeks to advance the view that technological advancement in developing countries such as Uganda can prove to be beneficial for refugees especially when it comes to integrating into the host community. Identity helps obtain access to food, education, health care and employment in some cases. This, however, is complex in nature because despite the positive aspect of having modern technologies to help combat the refugee crisis it acts as a double-edged sword by trying to impose a new identity on the refugee. Furthermore, refugees are aware that the host country might not want them there and this feeling of rejection is made even more apparent when locals protest to their presence as well, which could cause fear and mental distress for some refugees. However, it is important to appreciate the social context in which modern identity systems seek to operate because cultural differences and religious beliefs can have a substantial impact on the success or failure of these systems. For example, some cultures may advocate for the man in the family to be the only one dealing with the identification processes. This would automatically exclude women in these families unintentionally. Therefore when devising recommendations and solutions it is crucial appreciate that every humanitarian crisis does not always have a technological solution, however, if applied the consequences of the impact biometric data collection has on refugees can not be ignored. Lastly, the surveillance of refugees in developed countries requires further research to unearth what the future entails as the refugee population in this region is growing in number due to continued natural disasters, war and conflict. Engagement is needed with policymakers, stakeholders such as the Office of the Prime Minister (OPM), non-government organisations in the country (e.g. UNHCR) and charities to gain a more informed view.

ACKNOWLEDGEMENTS

I am grateful to God for granting me the wisdom to express my research interest in this paper. I salute all refugees striving hard to find life again. To my two supervisors, Dr Ofer Engel and Dr Kutoma Wakunuma, thank you for encouraging me to apply for the ETHICOM conference. To Emmanu'-EL Apostolic Church, Leicester, I thank you for your continued inspiration and support.

REFERENCES

Ajana, B. (2013). *Governing through biometrics*. Basingstoke: Palgrave Macmillan

Balaton-Chrimes, S., 2013. Indigeneity and Kenya's Nubians: seeking equality in difference or sameness? *The Journal of Modern African Studies*, 51(2), pp.331-354.

Capurro, R. (2008). Intercultural information ethics: foundations and applications. Journal of Information, Communication and Ethics in Society, 6(2), pp.116-126.

Dahir, A. (2019). *Africa isn't ready to protect its citizens personal data even as EU champions digital privacy*. [online] Quartz Africa. Available at: https://qz.com/africa/1271756/africa-isnt-ready-to-protect-its-citizens-personal-data-even-as-eu-champions-digital-privacy/

Ilcan, S., Oliver, M. and Connoy, L., (2020). *Humanitarian Assistance and The Politics Of Self-Reliance: Uganda's Nakivale Refugee Settlement*. [online] Centre for International Governance Innovation. Available at: <https://www.cigionline.org/publications/humanitarian-assistance-and-politics-self-reliance-ugandas-nakivale-refugee-settlement> [Accessed 13 April 2020].

Kuster, B. and Tsianos, V.S., 2016. How to liquefy a body on the move: Eurodac and the making of the European digital border. In *EU Borders and Shifting Internal Security* (pp. 45-63). Springer, Cham.

Maitland, C. (2018). *Digital lifeline?* Westchester Publishing Services.

Okiror, S. (2019) Inquiry finds refugee numbers were exaggerated by 300,000 in Uganda. the Guardian. Retrieved 16 April 2020, from https://www.theguardian.com/global-development/2018/oct/30/inquiry-finds-refugee-numbers-exaggerated-in-uganda.

Okiroro, S. (2019). Massive deforestation by refugees in Uganda sparks clashes with local people. the Guardian. Retrieved 17 April 2020, from https://www.theguardian.com/global-development/2019/feb/18/massive-deforestation-by-refugees-in-uganda-sparks-clashes-with-local-people.

Rand.org (2019). *Tracking Refugees with Biometrics: More Questions Than Answers*. [online] Available at: https://www.rand.org/blog/2016/03/tracking-refugees-with-biometrics-more-questions-than.html

Refugees, U. (2019). Uganda launches major refugee verification operation. [online] UNHCR. Available at: https://www.unhcr.org/news/latest/2018/3/5a9959444/uganda-launches-major-refugee-verification-operation.html

Sandvik, K., Gabrielsen Jumbert, M., Karlsrud, J. and Kaufmann, M. (2014). Humanitarian technology: a critical research agenda. *International Review of the Red Cross*, 96(893), pp. 219-242.

The Economist. (2019). *African countries are struggling to build robust identity systems*. [online] Available at: https://www.economist.com/middle-east-and-africa/2019/12/05/african-countries-are-struggling-to-build-robust-identity-systems

The Economist. (2019). *Establishing identity is a vital, risky and changing business*. [online] Available at: https://www.economist.com/christmas-specials/2018/12/18/establishing-identity-is-a-vital-risky-and-changing-business

The New Dark Age. (2019). *AI Border Guards are Being Tested at the Edge of Fortress Europe, Away From Public Scrutiny*. [online] Available at: https://williambowles.info/2019/12/06/ai-border-guards-are-being-tested-at-the-edge-of-fortress-europe-away-from-public-scrutiny/

Understanding datafication in relation to social justice' (DATAJUSTICE) starting grant (2018-2023).

United Nations Sustainable Development. (2020). Peace, Justice and Strong Institutions - United Nations Sustainable Development. [online] Available at: <https://www.un.org/sustainable development/peace-justice/> [Accessed 8 April 2020].

Weaver, J.L., 1985. Sojourners along the Nile: Ethiopian refugees in Khartoum. *The Journal of Modern African Studies*, *23*(1), pp.147-156.

EMPLOYEE TECHNOLOGY ACCEPTANCE OF INDUSTRY 4.0 IN SMES

Jan Strohschein, Ana María Lara-Palma, Heide Faeskorn-Woyke

TH Köln (Germany), Universidad de Burgos (Spain), TH Köln (Germany)

jan.strohschein@th-koeln.de; amlara@ubu.es; heide.faeskorn-woyke@th-koeln.de

ABSTRACT

The integration of the digital into the physical world is called Industry 4.0 and transforms manufacturing. In the future, smart factories collect more data than ever to empower artificial intelligence in cyber-physical production systems. Employee acceptance was identified as one of the most critical aspects for a successful introduction of I4.0 in any company. The design of such an I4.0 introduction process needs further research, not only for big corporations but also for small- and medium-sized enterprises as those are just as crucial for economies around the globe. Companies can use various “maturity” or “readiness” models for self-assessment of their current I4.0 capabilities and progress towards successful I4.0 introduction. We use the technology acceptance model to improve the employee dimension in our questionnaire. It is conducted in Germany and Spain with a focus on smaller and medium-sized enterprises. The results find statistically significant differences for smaller companies with significantly fewer technologies used, less systematic technology management, fewer investments made, and also earlier stages of I4.0 introduction for smaller companies. The collaboration with a bigger partner on the I4.0 introduction leads to a significantly more positive attitude towards I4.0, including employees, who look towards the changes with confidence.

KEYWORDS: Industry 4.0, Artificial Intelligence, Technology Acceptance Model, SME.

1. INTRODUCTION

Industry 4.0 (I4.0) transforms manufacturing by the integration of the digital into the physical world. In the future, smart factories collect more data than ever to empower artificial intelligence (AI) in cyber-physical production systems (CPPS). Extensive networks of humans and machines are the result, which eliminates company borders and redefines work within a plant but also the collaboration between business partners. However, in 2018, just 14% of 1.600 executives, who participated in a study conducted by Deloitte, believed that their organization is prepared for I4.0 and able to profit from this new potential (Deloitte, 2018). While Gartner predicts that additional automation and the use of artificial intelligence will create more jobs than it destroys, the new jobs will mainly be in fields such as healthcare and education. At the same time, manufacturing will probably see most job losses, so employees are skeptical about the introduction of the new technology (Pettey & Meulen, 2017).

This work examines the technology acceptance by employees and the overall status of I4.0 introduction in small- and medium-sized manufacturing companies (SMEs). Several studies and questionnaires investigate the introduction of I4.0 and AI in manufacturing companies but mostly target big multi-national companies. Therefore, they provide the basis for our questionnaire that focuses on SMEs and analyses the employee's feelings in more detail.

1.1. Employee acceptance of I4.0 introduction

Employees help companies realize their digital transformation and are the ones most affected by the changes in the digital workplace. Their direct working environment is altered, requiring them to acquire new skills and qualifications. Abel, Hirsch-Kreinsen and Steglich explain the worker's doubts not only with their fear of job losses but also the technological changes being digital and no longer immediately comprehensible to the individuals, which results in insecurities and skepticism (Abel, Hirsch-Kreinsen, Steglich, & Wienzek, 2019). Kagermann, Wahlster, and Helbig similarly report a "growing tension between the virtual world and the world of workers' own experience. This tension could result in workers experiencing a loss of control and a sense of alienation from their work as a result of the progressive dematerialization and virtualization of business and work processes" (Kagermann, Wahlster, & Helbig, 2013). They also agree that through extensive human-machine interactions, the work content, and processes, as well as the working environment, will be radically transformed and thus also the worker's job and competence profiles. Other researchers perceive the introduction of I4.0 not only as a challenge but also as a chance to improve the work environment by creating learning systems, which "dynamically detect and adapt to the context of the support situation and the worker's actions" (Gorecky, Schmitt, Loskyll, & Zühlke, 2014).

In conclusion, for a successful introduction of I4.0 in any company, employee acceptance got identified as one of the most critical aspects. The introduction process requires communication and transparency as "acceptance is a fragile construct, which needs constant cultivation" to convert employee resistance into acceptance or even support (Abel et al., 2019). The design of a successful I4.0 introduction process needs further research, not only for big corporations but also for small- and medium-sized enterprises.

1.2. Industry 4.0 for SMEs

Small and medium-sized enterprises are just as important for economies around the globe as big multinational enterprises. In the 28 European countries, two-thirds of employees in the non-financial sector are employed by SMEs. All three sizes of SMEs are contributing nearly equally to value added with 20.3% for micro enterprises, 17.6% for small enterprises and 18.5% for medium-sized enterprises (Airaksinen, Luomaranta, Alajääskö, & Roodhuijzen, 2015). In 2015 they employed 91 million people in total and generated 3.934 € billion of value added (Eurostat, 2018). Therefore "SMEs are the backbone of the European and many other economies" (*Future Image Industry 4.0*, 2012; Kraemer-Eis & Passaris, 2015). This is also especially true for manufacturing, where European SMEs provide around 45% of the value added and around 59% of employment (Vidosav, 2014).

Even though SMEs are an essential factor to economies, the I4.0 methods are developed mainly in larger enterprises and have to be adapted to the specific requirements of SMEs (Rauch et al., 2018). Currently, the spread of I4.0 depends on company size, and large companies are more likely to deploy relevant I4.0 technologies than SMEs (Schröder, 2016). Several scientists investigated the reasons for this observation. In 2017 Decker used case study research to evaluate the I4.0 readiness of Danish SMEs from the metal processing sector with the result that

SMEs at this time were not sure if or how they should introduce I4.0 in their companies (Decker, 2016). Wuest et al. confirmed the struggle of SMEs to adopt I4.0 in a study conducted with manufacturing SMEs in West Virginia in 2018 (Wuest, Schmid, Lego, & Bowen, 2018). Both studies seem to support the claim made by Lutz Sommer in an article from 2015 that, "actually most of SMEs are not prepared to implement I4.0 concepts" (Sommer, 2015). Further research suggests various challenges for SMEs in I4.0 introduction:

- Different prerequisites regarding the integration of their production plants in higher-level IT systems, which is much more advanced in bigger companies (Lichtblau et al., 2014).
- Using a self-assessment tool is not easy as I4.0 concepts are still too little known (Rauch et al., 2018)
- SMEs often lack resources to evaluate new technologies and their business uses. Thus it is hard for them to develop an appropriate strategy, including a cost-benefit analysis (Schröder, 2016).

Those challenges need to be verified and addressed because "successful implementation of an industrial revolution I4.0 has to take place not only in large enterprises but in particular in SMEs" (Sommer, 2015).

1.3. Industry 4.0 Self-Assessment

Companies can use various "maturity" or "readiness" models for self-assessment of their current I4.0 capabilities and progress towards successful I4.0 introduction. Schumacher, Erol and Sihn created an overview of existing models in 2016 and found that many models lack details regarding the development process or assessment methodology (Schumacher, Erol, & Sihn, 2016). They highlight the "Industry 4.0 Readiness Model" (Lichtblau et al., 2014) as "scientifically well-grounded and its structure and results explained in transparent manners" but the model contains just a single question for the employee dimension. In this question, they assess if the workers have the required skills to accomplish their future tasks[1]. Schumacher, Erol and Sihn propose their own "Industry 4.0 maturity model", which also asks for the openness of employees towards new technologies. However, we introduce the technology acceptance model in our study to further improve the employee dimension.

1.4. Technology Acceptance Model

The "Technology Acceptance Model" (TAM), developed by Fred D. Davis 1986 and published in 1989, is used to acquire additional insights into the factors that influence adoption of new technology (Davis, 1986). The two main variables are "perceived usefulness" (PU) and "perceived ease-of-use" (PEU). Im, Kim, and Han extended the TAM by introducing "perceived risk" (PR) as an additional variable that negatively affects adoption (Im, Kim, & Han, 2008). Those factors influence the "attitude towards usage" (ATU) and finally also the "behavioral intention

[1] Industry 4.0 Readiness Model Questionnaire: https://www.industrie40-readiness.de/ retrieved 14-04-2020

to use" (BIU) and the questionnaire contains items to investigate every component. To this day TAM is one of the most popular models to assess user acceptance of new technologies and was successfully used to evaluate the adoption of related technologies, e.g., smartphones and wearables (Chang, Lee, & Ji, 2016; Roy, 2017).

2. METHODOLOGY

The questionnaire is conducted in Germany and Spain with a focus on smaller and medium-sized companies. The German manufacturing companies are all members of the "Innovation Hub Oberbergischer Kreis", a regional association that focuses on the exchange of I4.0 knowledge and possible applications. The Spanish companies belong to the industrial service sector, working on optimization, logistics and manufacturing technologies for their national and international clients.

The questionnaires are sent to the executives, who are responsible for the introduction of I4.0 and AI in their companies. This work examines the following hypotheses:

> H_1: Smaller SMEs need more assistance for the evaluation of I4.0 technologies than bigger
>
> H_2: Smaller SMEs need more assistance to assess I4.0 introduction costs and benefits than bigger SMEs. SMEs.
>
> H_3: Smaller SMEs need more assistance to formulate an I4.0 strategy than bigger SMEs.
>
> H_4: SMEs that collaborate with a big company feel better prepared for I4.0 introduction and have a higher technology acceptance rate than SMEs who do not collaborate with a big company.
>
> H_5: Employees from SMEs with internal motivation to introduce I4.0 have a higher technology acceptance rate than employees from SMEs with an external motivation to introduce I4.0.
>
> H_6: SMEs with internal motivation to introduce I4.0 expect a higher increase in productivity than SMEs with an external motivation to introduce I4.0.
>
> H_7: There is a significant difference in the answers between Spanish and German SMEs.

3. RESULTS

Overall, 14 companies participated in the survey, with 11 completing the questionnaire. Seven of those companies were from Germany (63.6%) and four (36.4%) from Spain. The companies were also grouped by the size of their workforce: "less than 10 employees" (9.1%), "10 to 49 employees" (18.2%), "50 to 249 employees" (45.5%) and "more than 250 employees" (27.3%). The Mann-Whitney U test was used to evaluate the results and find statistically significant differences in the answers of different groups (Mann & Whitney, 1947). The test was independently developed in 1947 by Mann-Whitney and Wilcoxon. Thus it is also known as the Wilcoxon-Mann-Whitney rank-sum test. It is one of the most commonly used non-parametric tests, which means it does not depend on a normal distribution and provides reliable, statistically significant results when used with small sample sizes of 10-20 observations (Landers, 1981). The Mann-Whitney test verifies the null hypothesis (H_0) based on the comparison of each observation from the first group with each observation from the second group and identifies if the two independent groups are homogenous and have the same distribution (Nachar, 2008). If there is a statistically significant difference between the two populations, the determined p-value is small and the H_0 is rejected in favor of the alternative hypothesis. The commonly accepted thresholds for p are ≤ 0.05 for significant differences and ≤ 0.01 for highly significant differences and also used for this analysis (Fisher, 1992). All results stem from the two-sided test, which examines both ends of the distribution. Anova and similar techniques have not been used as normal distribution could not be guaranteed and the sample size is too small. Only the relevant questions to the particular hypothesis are shown. The complete questionnaire is available online in Spanish, English and German for further reference[2].

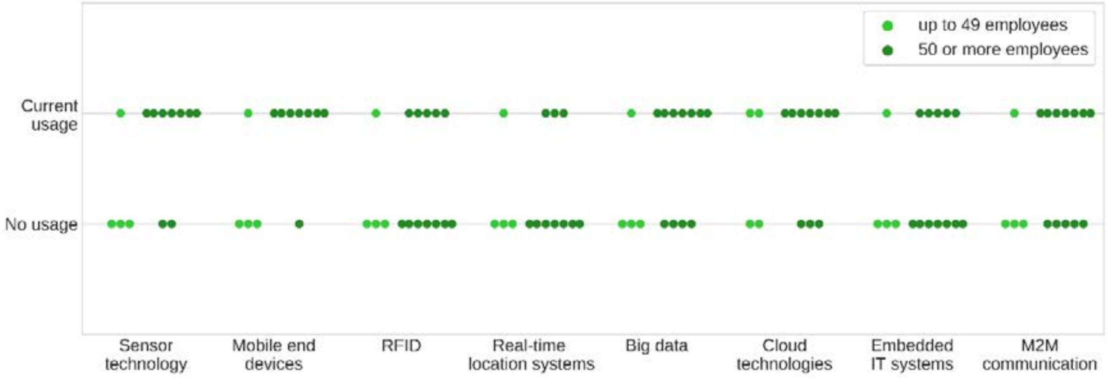
H_1: Smaller SMEs need more assistance for the evaluation of I4.0 technologies than bigger SMEs.

Q16 Technology usage in companies

Statistically significant differences in technology usage (Q16) comparing “Companies with up to 10 employees” with the other participants for all of the eight technologies: Sensor technology (p=0.009, highly significant), mobile end devices (p=0.004, highly significant), RFID (p= 0.027), real-time location systems (p=0.030), big data (p=0.018), cloud technologies (p=0.011), embedded IT systems (p=0.027) and M2M communication (p=0.022).

Splitting the participants into “Companies with up to 49 employees” and “others” identified significant differences for the following three out of eight technologies: Sensor technology (p=0.014), mobile end devices (p=0.006, highly significant) and big data(p=0.038).

Figure 1. Technology usage in companies with up to 49 vs more than 49 employees.



[2] Jan Strohschein, Github Repository: https://github.com/janstrohschein/Industry-4.0-readiness-for-SMEs-Questionnaire Retrieved at 10-04-2020

Q17 Past and future investments

The analysis of investments in the past 2 years found significant differences for “Research and development” (p = 0.020) and “Production / manufacturing” (p = 0.023).

The planned investments over the next 5 years also show significant differences in “Production/ manufacturing” (p = 0.035). The analysis shows several significant differences and confirms $H_{1.}$

H_2: Smaller SMEs need more assistance to assess I4.0 introduction costs and benefits than bigger SMEs.

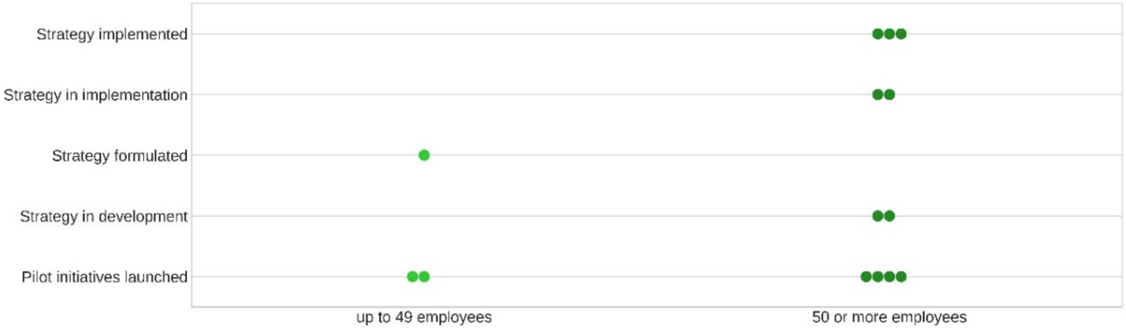
Analysis of Q12 “The benefits of I4.0 introduction are well known to our company and clearly evaluated” and Q13 “The costs of I4.0 introduction are well known to our company and clearly evaluated” yielded no significant differences between companies of different sizes. Thus H_2 is rejected.

H_3: Smaller SMEs need more assistance to formulate an I4.0 strategy than bigger SMEs.

Q14 Industry 4.0 strategy implementation status

Significant results for all splits, i.e. “Companies with up to 10 employees” | “others” (p = 0.030), “Companies with up to 49 employees” | “others” (p = 0.012) and “Companies with up to 249 employees” | “others” (p = 0.008, highly significant).

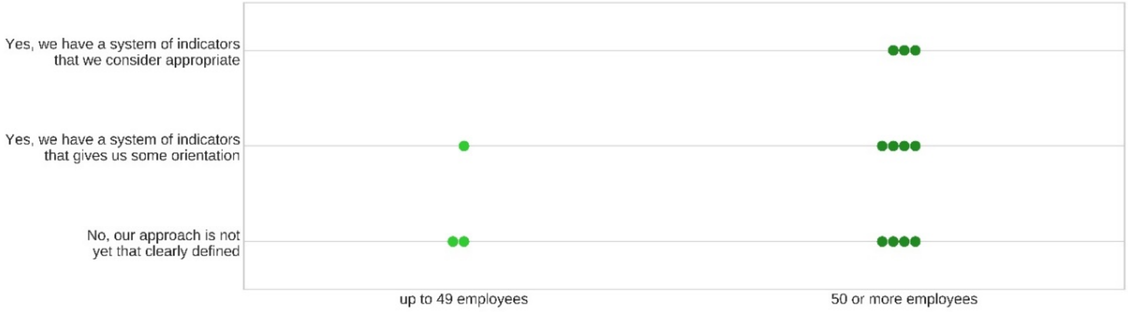
Figure 2. Q14 Industry 4.0 implementation status comparing companies with up to 49 employees and companies with more than 50 employees.



Q15 Industry 4.0 indicators

Statistically significant differences for all splits, i.e. “Companies with up to 10 employees” | “others” (p = 0.029), “Companies with up to 49 employees” | “others” (p = 0.010) and “Companies with up to 249 employees” | “others” (p = 0.008, highly significant)

Figure 3. Industry 4.0 indicators comparing companies with up to 49 employees and companies with more than 50 employees.

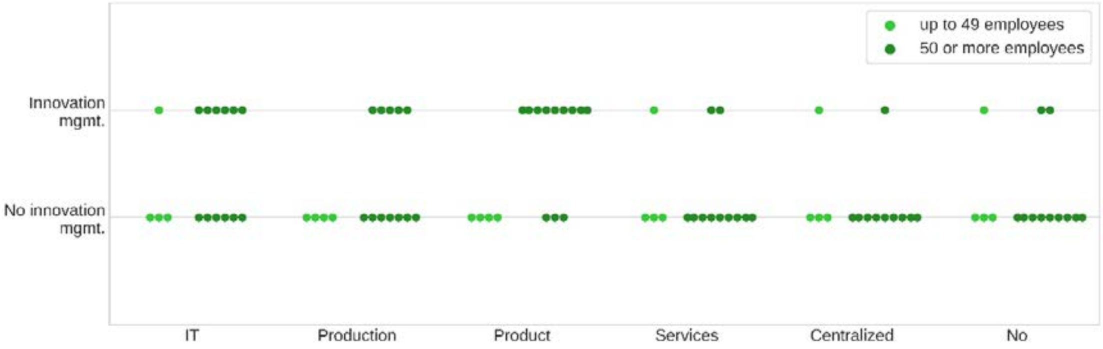


Q18 Systematic technology and innovation management

Split with “Companies with up to 10 employees” | “others” found significant differences in the technology and innovation management for IT (p = 0.025), production technologies (p = 0.029), product development (p = 0.018), services (p = 0.031) and in the amount of centralized innovation management (p = 0.031).

The split between “Companies with up to 49 employees” and “others” yields highly significant results (p < 0.01) for production technologies (p = 0.003) and product development (p = 0.001). Significant differences were found for services (p = 0.044) and the implementation of a centralized innovation management (p=0.015). The analysis shows significant or even highly significant differences and verifies H_3.

Figure 4. Systematic technology and innovation management in companies with up to 49 and more than 50 employees.

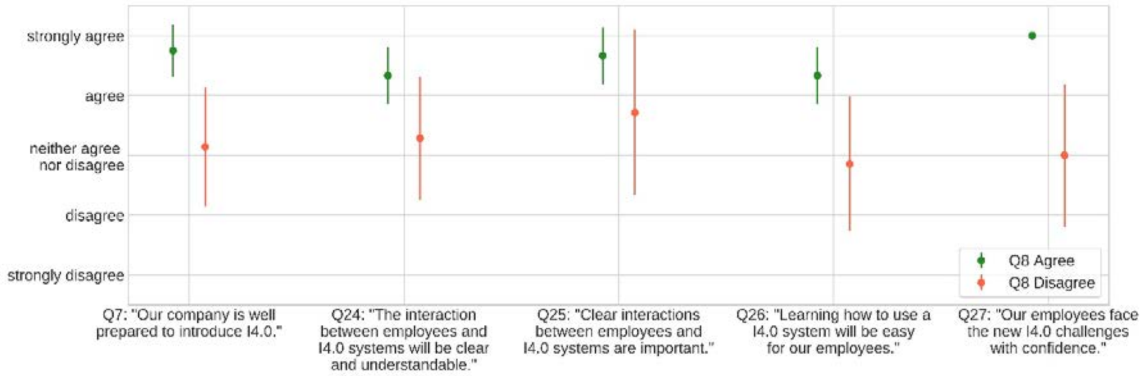


H_4: SMEs that collaborate with a big company feel better prepared for I4.0 introduction and have a higher technology acceptance rate than SMEs who do not collaborate with a big company.

The samples are split based on their answer to question Q8 "Our company adopts the I4.0 strategy of a (bigger) partner".

Significant differences exist for Q7 “Our company is well prepared to introduce I4.0” (p = 0.042) and Q27 “Our employees face the new I4.0 challenges with confidence” (p = 0.042), thus H_4 is accepted.

Figure 5. Comparing companies that agree/disagree to Q8 (mean + std.).

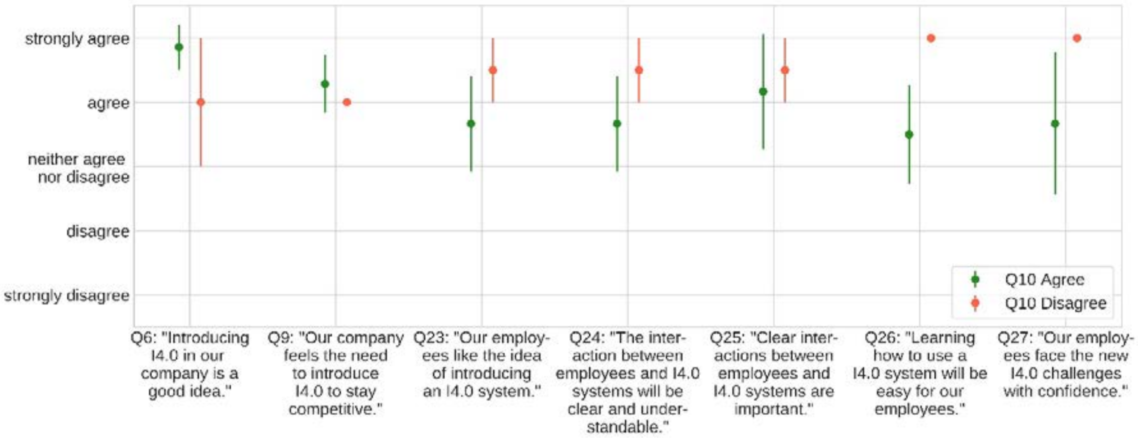


H_5: Employees from SMEs with internal motivation to introduce I4.0 have a higher technology acceptance rate than employees from SMEs with an external motivation to introduce I4.0.

The samples are split based on their answer to question Q10 "Our company feels the need to introduce I4.0 to continue collaboration with (bigger) partners.". The split was chosen, as the other possible splits based on Q6 "Introducing I4.0 in our company is a good idea" and Q9 "Our company feels the need to introduce I4.0 to stay competitive" had uniformly agreeing answers.

Results are shown in an overview but are not statistically significant, and therefore H_5 is rejected.

Figure 6. H_5 overview with companies grouped based on their Q10 answers (mean + std.).

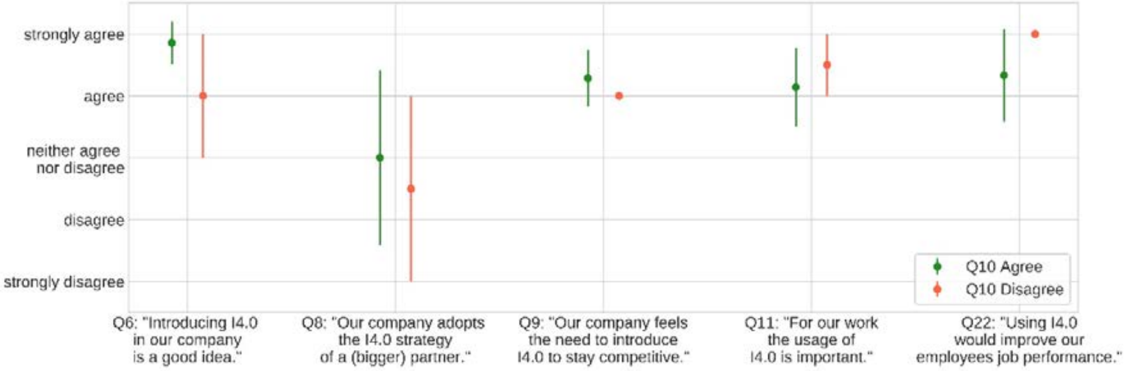


H_6: SMEs with internal motivation to introduce I4.0 expect a higher increase in productivity than SMEs with an external motivation to introduce I4.0.

The samples are split based on their answer to question Q10 "Our company feels the need to introduce I4.0 to continue collaboration with (bigger) partners.". The split was chosen, as the other possible splits based on Q6 "Introducing I4.0 in our company is a good idea" and Q9 "Our company feels the need to introduce I4.0 to stay competitive" had uniformly agreeing answers.

Results are shown in an overview but are not statistically significant, and therefore H_6 is rejected.

Figure 7. H_6 overview with companies grouped based on their Q10 answers (mean + std.)



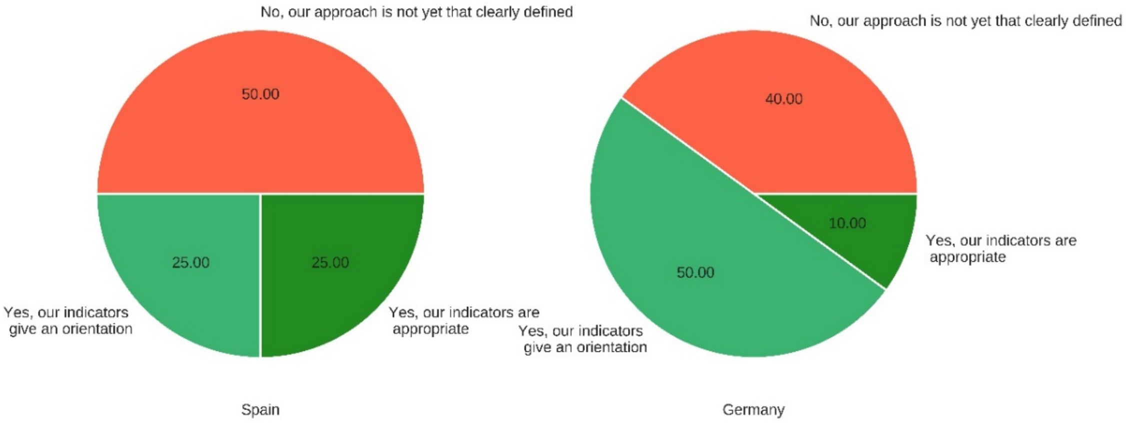
H_7: There is a significant difference in the answers between Spanish and German SMEs.

While 60% of German SMEs planned to increase the employees from leadership working on I4.0 introduction, none of the Spanish SMEs plan additional workers (p = 0.007, highly significant). The situation is similar for an increase of employees in HR working on I4.0 introduction. 40% of German companies plan to increase the current number of employees and none of the Spanish SMEs (p = 0.038).

A significant difference (p = 0.050) was also found between Spanish and German companies for Q9 “Our company feels the need to introduce I4.0 to stay competitive”. Spanish companies tended to agree strongly (avg.: 4.75, std.: 0.43), while German companies agreed (avg.: 3.9, std.: 0.53).

The current status of I4.0 implementation (Q14, p = 0.016) and the indicators used to track the progress (Q15, p = 0.013) also differed significantly between the two countries. Half of the German and 25% of the Spanish companies stated that they have indicators that give them some orientation. However, just 10% of the German and 25% of the Spanish companies think that their indicators are already appropriate.

Figure 8. Q15 I4.0 indicators with companies grouped by country.



Surveying the existing technologies (Q16) in companies of both countries showed significantly more usage of mobile end devices (p = 0.025) in Germany In contrast, companies in Spain utilized more real-time location systems (p =. 0.030). Unfortunately, there was also a highly significant difference in Spanish companies that use none of the inquired technologies (p = 0.002).

The results for technology and innovation management (Q18) also highlight differences between the two countries. The German companies focus on innovation management for production technologies (p = 0.038) and product development (p = 0.012) while the Spanish companies possess significantly more innovation management for their services (p = 0.008, highly significant) or use a centralized approach (p = 0.002, highly significant). The Spanish participants also declared significantly more companies without any technology or innovation management (p = 0.040). As the analysis found several statistical significant differences H_7 is accepted.

4. DISCUSSION AND CONCLUSION

H_{1-3} regard additional assistance required by smaller companies to formulate an I4.0 strategy (H_1), assess costs and benefits (H_2) and evaluate the related technologies (H_3). H_1 and H_3 could be validated with significantly fewer technologies used, less systematic technology management, fewer investments made and also earlier stages of I4.0 introduction for smaller companies. Those findings may confirm claims by Christian Schröder that SMEs often lack resources to evaluate new technologies, which makes the development of an I4.0 strategy harder (Schröder, 2016). The results suggest that SMEs, five years after the survey by Lichtblau et al. (Lichtblau et al., 2014), could not catch up regarding the integration of their production plants in higher-level IT systems, a precondition for many I4.0 use cases. Apart from those differences, all companies declared that they could evaluate the benefits but have problems assessing the associated costs of I4.0 introduction, thus H_2 is rejected.

The collaboration with a bigger partner on the I4.0 introduction led to a significantly more positive attitude towards I4.0, which confirmed H_4. Even though “many of the I4.0 methods are developed mainly in larger enterprises” (Rauch et al., 2018), there is potential for the SMEs to profit from the groundwork done by the bigger partner, especially when the SMEs own resources are rather scarce. Five years after Lutz Sommer (Sommer, 2015) stated that “most SMEs are not prepared to implement I4.0 concepts” the collaboration with a bigger partner leads to SMEs who feel well prepared for the I4.0 introduction.

H_{5-6} examine the influence of internal and external motivation to introduce I4.0 towards the technology acceptance rate and expected increases of productivity but could not be statistically verified, thus both hypotheses are rejected.

The comparison of Spanish and German SMEs highlighted various statistically significant differences, which leads to acceptance of H_7. Most noticeably are Spanish companies that use none of the new technologies and also do not possess technology or innovation management. It is not possible to determine if it is a regional difference or if it is caused by the small sample size where Spanish companies are smaller on average.

The small sample size is the main limitation of this work and stems from the specific requirements for the participants, but also the world-wide COVID-19 pandemic where SMEs had to shut down their production. It would be interesting to conduct the questionnaire again with more participants to get more insights into the differences between Spanish and German SMEs, even though the results are already statistically significant. Future research should work on concrete methods to assist SMEs with the development of an I4.0 strategy and the evaluation of the associated new technologies. Best practices of successful I4.0 adoption, which are currently not available (Matt & Rauch, 2020), will also provide great value to SMEs.

ACKNOWLEDGEMENTS

We would like to thank all the companies, who took the time to participate in our survey. The paper would not have been possible without their knowledge.

The work was supported by the German Federal Ministry of Education and Research (BMBF) under the project ”KOARCH” (funding code: 13FH007IB6).

REFERENCES

Abel, J., Hirsch-Kreinsen, H., Steglich, S., & Wienzek, T. (2019). Akzeptanz von Industrie 4.0.

Airaksinen, A., Luomaranta, H., Alajääskö, P., & Roodhuijzen, A. (2015). Statistics on small and medium-sized enterprises. *Eurostat*, (September), 1–14. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Statistics_on_small_and_medium-sized_enterprises&oldid=463558

Decker, A. (2016). Industry 4.0 and SMEs in the Northern Jutland Region. In *Value Creation in International Business: Volume 2: An SME Perspective* (pp. 309–335). Springer International Publishing. https://doi.org/10.1007/978-3-319-39369-8_13

Deloitte. (2018). The fourth industrial revolution is here: Are you ready? *Deloitte Insight*, 1–26. https://doi.org/10.1016/j.jbusres.2015.10.029

Eurostat. (2018). *Small and medium-sized enterprises: an overview*. Retrieved from https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20181119-1

Fisher, R. A. (1992). Statistical Methods for Research Workers (pp. 66–70). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_6

Future Image Industry 4.0. (2012).

Gorecky, D., Schmitt, M., Loskyll, M., & Zühlke, D. (2014). Human-machine-interaction in the industry 4.0 era. *Proceedings - 2014 12th IEEE International Conference on Industrial Informatics, INDIN 2014*, 289–294. https://doi.org/10.1109/INDIN.2014.6945523

Kagermann, H., Wahlster, W., & Helbig, J. (2013). Recommendations for implementing the strategic initiative INDUSTRIE 4.0 - Final report of the Industrie 4.0 Working Group. *Acatech - National Academy of Science and Engineering*, (April), 84.

Kraemer-Eis, H., & Passaris, G. (2015). SME Securitization in Europe. *The Journal of Structured Finance*, *20*(4), 97–106. https://doi.org/10.3905/jsf.2015.20.4.097

Landers, J. (1981). *Quantification in History, Topic 4: Hypothesis Testing II-Differing Central Tendency.*

Lichtblau, K., Stich, V., Bertenrath, R., Blum, M., Bleider, M., Millack, A., … Schröter, M. (2014). Industry 4.0 Readiness, *26*(2), 218–223.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/AOMS/1177730491

Matt, D. T., & Rauch, E. (2020). *SME 4.0: The Role of Small- and Medium-Sized Enterprises in the Digital Transformation*. Springer International Publishing. https://doi.org/10.1007/978-3-030-25425-4

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20. https://doi.org/10.20982/tqmp.04.1.p013

Pettey, C., & Meulen, R. van der. (2017). *Gartner Says by 2020,AI will crete more jobs than it eliminates*. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2017-12-13-gartner-says-by-2020-artificial-intelligence-will-create-more-jobs-than-it-eliminates

Rauch, E., Matt, D. T., Brown, C. A., Towner, W., Vickery, A., & Santiteerakul, S. (2018). Transfer of industry 4.0 to small and medium sized enterprises. *Advances in Transdisciplinary Engineering*, *7*(September), 63–71. https://doi.org/10.3233/978-1-61499-898-3-63

Schröder, C. (2016). The Challenges of Industry 4.0 for Small and Medium-sized Enterprises. *The Friedrich-Ebert-Stiftung*, 28.

Schumacher, A., Erol, S., & Sihn, W. (2016). A Maturity Model for Assessing Industry 4.0 Readiness and Maturity of Manufacturing Enterprises. *Procedia CIRP*, *52*, 161–166. https://doi.org/10.1016/j.procir.2016.07.040

Sommer, L. (2015). Industrial revolution - Industry 4.0: Are German manufacturing SMEs the first victims of this revolution? *Journal of Industrial Engineering and Management*, *8*(5), 1512–1532. https://doi.org/10.3926/jiem.1470

Vidosav, D. (2014). Manufacturing Innovation and Horizon 2020–Developing and Implement „New Manufacturing“. *Proceedings in Manufacturing Systems*, *9*(1), 3–8. Retrieved from http://icmas.eu/Journal_archive_files/Vol_9-Issue1_2014_PDF/3-8_MAJSTOROVIC.pdf

Wuest, T., Schmid, P., Lego, B., & Bowen, E. (2018). Overview of smart manufacturing in West Virginia. *Bureau of Business & Economic Research, West Virginia University*.

ETHICAL CONCERNS OF MEGA-CONSTELLATIONS FOR BROADBAND COMMUNICATION

Marco Crepaldi

University of Luxembourg (Luxembourg)

marco.crepaldi@protonmail.com

ABSTRACT

This article studies ethical concerns of private satellite mega-constellations in low-earth-orbit (LEO) deployed to provide broadband services globally. These concerns have been understated thus far. The issue at hand is framed in terms of distributed morality. Three morally relevant aspects are analyzed, namely, the problem of space debris, the design of autonomous maneuvering systems on board of satellites, and the limited availability of orbital slots and parts of the radio spectrum. To address the aforementioned issues the following solutions are discussed. First, the application of the responsible research innovation framework to the private activities in outer space. Second, ethical policies of aggregation of good actions paired with disaggregation of morally bad ones.

KEYWORDS: distributed morality; mega-constellations; LEO; space debris; outer space treaty; space ethics.

1. INTRODUCTION

The commercial exploitation and, to a lesser extent, the exploration of outer space raises several challenges. While technical, political and legal issues abound this chapter focuses on the implications of space exploitation from the ethical perspective (Rao, Gopalakrishnan, & Abhijeet, 2017). It does so by analysing recent efforts by private companies to provide universal broadband access by way of mega-constellations of satellites (numbered in the tens of thousands) deployed in low-earth-orbit (henceforth, LEO). The argument unfolds as follows. First, section two provides the relevant background on the progressive privatization of space. Then, it offers a primer on the legal sources governing the use of outer space, and it describes the phenomena of mega-constellations. While some may not find it surprising that global planned infrastructures to provide broadband access raise interesting ICT & societal challenges, others might be sceptical. Thus, section three argues for the importance of discussing the subject matter from the perspective adopted throughout this book.

Later, section four frames the issue at hand in terms of distributed morality, drawing on the work of Floridi. It describes three macro ethical concerns raised by mega-constellations, the first is the problem of space debris, the second the design of autonomous systems to avoid conjunctions in LEO, while the third arises from the finite nature of resources such as orbital slots and the radio spectrum. On this basis, section five provides two directions to address these issues. It suggests the application of the responsible research and innovation framework to private space activities, and the development of ethical policies of aggregation. Section six concludes.

2. BACKGROUND

Only recently, the space capabilities of private enterprises have made them relevant from an ethical perspective. Further, several nation-states have already developed normative frameworks for the privatization of outer space while others are likely to follow suit. Luxembourg, for example, has enacted legislation intended to attract capital and companies in the space business (law of the 20th of July 2017[1]) and the results are promising so far. The U.S. has made similar efforts toward the privatization of space (Trump, 2018).

The opening of outer space to private activities is a welcome development. Private enterprises will likely foster innovation in the space sector as well as generate significant economic growth in the years to come, both morally desirable outcomes. Areas such as asteroid mining or space tourism appear poised to contribute to human flourishing in the long-term. Think, for example, as the scenario imagined by Jeff Bezos concerning the operations of his Blue Origin. Moving the externalities caused by some manufacturing activities from the fragile earth to our more resilient moon is highly desirable from a multitude of perspectives. Moreover, innovation in space technologies reduces existential risks for humanity by contributing to the goal of becoming an interplanetary species, therefore it is morally desirable (Munevar, 2019; Schwartz, 2011). However, the opening of the space frontier to private enterprises raises a multitude of challenges[2]. Of relevance for this contribution is the ineptitude of the legal framework governing space activities. Amongst its several shortcomings, current space law does not provide enough guarantees to ease ethical concerns raised by the privatization of space. To see why a brief digression on the sources of space law is in order.

The legal framework for space activities is made up of four international treaties, the last one signed in 1979 (United Nations, 2017). These international treaties are: the "Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies" or OST opened for signature on January 1967; the "Agreement on the Rescue of Astronauts, the Return of Astronauts and the Return of Objects Launched in Outer Space" or Rescue Agreement for short of 1968; the "Convention on International Liability for Damage Caused by Space Objects of 1972, also known as the Liability Convention; the "Convention on Registration of Objects Launched into Outer Space" of 1976, and the "Agreement Governing the Activities of States on the Moon and Other Celestial Bodies" of 1984. It is enough to note how all these sources were drafted decades before the privatization of space. It is also worth noting that the Moon agreement is not relevant with only 18 ratifications, none of which from space fairing nations. While a thorough analysis of the inadequacies of current space law lies outside of the scope of this chapter, a few remarks are in order.

On the one hand, current space law was not drafted with small satellites in mind (Marboe, 2016a, 2016b; Shaw & Rosher, 2016). In the early days of space endeavours, satellites were - generally - measured in meters while nowadays, the majority of future satellites (e.g. CubeSats and pico-satellite) are measured in centimetres (Matney, Vavrin, & Manis, 2017; Millan et al.,

[1] Official text available here http://legilux.public.lu/eli/etat/leg/loi/2017/07/20/a674/jo

[2] For an overview of ethical concerns related to space activities see (Arnould, 2011).

2019). This is the first inadequacy of current space law. On the other hand, commercial exploitation of outer space was not a primary concern of the drafters of international space law. Their focus, amid the cold war, was likely to prevent the proliferation of nuclear weapons in space as well as its militarization. Therefore, current space law is also inadequate to deal with private commercial efforts such as asteroid mining, private moon bases, space tourism, or mega-constellations (Rao et al., 2017). The shortcomings of space law for the current times are hardly a new topic[3]. For our purposes, this brief digression on the sources of space law entails that, when dealing with ethical concerns related to private space activities, space law does not offer much support. One must look elsewhere to other methods and techniques to ensure that the private space era develops in a morally desirable direction.

Lastly, it is necessary to spend a few words on mega-constellations. Mega-constellations consist of the deployment of a vast number of satellites (from a few hundred to tens of thousands) by a single entity to provide a service. The use of more than one satellite is not new; for example, the GPS relies on 31 satellites. However, the sheer number of satellites deployed in mega-constellations is a qualifying difference, which raises numerous concerns. This contribution focuses on the particular issue of private mega-constellations to provide global internet broadband. Companies such as SpaceX and Boeing are spearheading these efforts while others are planning more mega-constellations. In the table 1 below, a list is provided of the planned mega-constellations in the coming years. If the forecasts are correct, then several thousands of satellites will be launched. It is important to note that the previous number refers only to mega-constellations for broadband communication.

Table 1. Planned Mega-Constellations[4].

Constellation	Number of Satellites	Orbit
Boeing	1.396-2.956	1.200 km
LeoSat	78-108	1.400 km
Starlink	4.425-42.943	550-1.325 km
Telesat LEO	117-512	1.000-1.248 km
CASIC Hongyun	156	160-2.000 km
CASC Hongyan	320	1.100 km

To put it in perspective, the index of Objects Launched into Outer Space maintained by the Office for Outer Space Affairs at the United Nations lists - at the time of writing - 9.447 objects[5]. It is clear that mega-constellations are a paradigm shift concerning space activities. Against this background, it is now time to spend a few words to explain the relevance of this argument from the purview of this book.

[3] Instead of many see (Larsen, 2009)

[4] Data collected by the author.

[5] The index accounts for most of the observable objects orbiting the earth, which include disposed rocket parts, exhausted boosters, non-functioning satellites as well as operational ones.

3. THE MORAL RELEVANCE OF MEGA-CONSTELLATIONS

The relevance of the issue at hand from the perspective adopted throughout this book is multifold. On the one hand, concerns arise within the purview of the smart society if one prominent private player becomes a natural monopolist in providing broadband connectivity from outer space. In this case, the essential facilities doctrine might curb the risks of a private global monopoly. According to this doctrine developed within the antitrust area, if several conditions hold, then the natural monopolist is forced to contract at a fair price with its competitors[6]. However, it is not clear if this legal doctrine would be sufficient. Doubts arise in areas such as the applicable law as well as the jurisdiction; there is no global space court after all.

On the other hand, control over the infrastructure that provides broadband connectivity enables censorship as well as discrimination of the network traffic. This is relevant both from the smart society perspective and the broader ICT ethics. In this case, since the applicable law to the provider is generally the one of the launching state, risks might be mitigated if the prominent players are established in jurisdictions that uphold the value of net neutrality and offer other guarantees. The scenario changes if the provider of a mega-constellations is established in a jurisdiction with fewer safeguards.

Lastly, the issue of space debris affects the technological affordances of humanity. The worst-case scenario described by the so-called Kessler syndrome entails precluding access to outer space for generations to come (Kessler, Johnson, Liou, & Matney, 2010). The effective avoidance of orbital conjunctions demands that next-generation satellites be equipped with autonomous manoeuvring capabilities, such that they appear to qualify as moral agents in the context of the multiagent system of outer space[7]. Thus, the design of autonomous anti-avoidance systems is also relevant from the computer ethics perspective.

4. ON SOME ETHICAL CONCERNS OF MEGA-CONSTELLATIONS

This section frames the deployment of mega-constellations in terms of distributed morality to highlight its ethical concerns. The phenomenon of distributed morality occurs when moral consequences are "the result of otherwise morally neutral or at least morally-negligible interactions among agents constituting a multiagent system" (Floridi, 2013, p. 729). Regarding mega-constellations, the launch of a batch of satellites by one agent can be considered a morally neutral action. That is, moral consequences are -generally - limited. The same holds for operating a spacecraft. However, the thousands of satellites orbiting roughly the same altitude of LEO, as is the case when satellites are launched to provide broadband access, might have moral consequences when their actions are aggregated. It is possible to describe the outer space scene as a multiagent systems (MAS). Relevant agents are the launching companies, the state responsible for the launch (along with the associated liabilities for space object), the rockets and satellites that possess autonomous manoeuvring capabilities, along with the other objects already in orbit and their operators. An example clarifies this framing. The operators of satellites currently are under no legal obligation to manoeuvre them if the probability of orbital conjunction raises above a certain threshold, however, the case for the presence of a moral

[6] See, in general (Lipsky Jr & Sidak, 1998)

[7] This holds if the notion of moral agents is consistent with the one described in (Floridi & Sanders, 2004).

obligation in this scenario appears straightforward. The single morally negligible action of operating a spacecraft becomes charged with moral weight once other agents (both human and artificial) are present in the system. On the basis of the framework of distributed morality, this section discusses three morally relevant aspects of mega-constellations deployed to provide global broadband communication. The first once concerns the issue of space debris.

Mega-constellations exacerbate the problem of space debris because of the sheer number of launches required to place thousands of satellites into orbit. Each launch leaves something behind. Moreover, due to the reduce cost of manufacturing and launch, the small satellites deployed will likely have a higher failure rate than other missions. The lack of appropriate safeguards against orbital conjunctions as well as sound decommissioning protocols might result in an unacceptable level of risk (Bergamini, Jacobone, Morea, & Sciortino, 2018). This is especially relevant from the moral perspective if the risk becomes crippling existing infrastructures that rely on satellites placed in LEO or if it endangers the access to outer space for the foreseeable future (Jakhu, 2010).

The second ethical concern raised by mega-constellations is closely related to the problem of space debris. It appears highly desirable to implement autonomous software onboard a spacecraft to prevent collisions with other objects, thus lowering the risk of conjunctions to more acceptable levels and improving the current email-based warning mechanism. In this case, even if the LEO orbit is quite vast, it is possible to imagine a scenario in which an autonomous system must decide which of two likely collisions to avoid. Thus, a space version of the famous trolley problem – which we could name the conjunction avoidance choice - can be described in the context of autonomous systems deployed on a satellite orbiting in LEO. This shows that developers of satellites ought to take into considerations moral scenarios. In an easy example of conjunction avoidance choice, the manoeuvring software should always privilege colliding with a piece of junk or a non-functioning spacecraft instead of an operational one. Yet harder cases are not hard to imagine. What if the collision with a piece of space junk is likely to generate debris of an order of magnitude greater than an operational satellite? Which collision should be privileged when the alternatives are a science mission or a telecommunication satellite? It is not the task of this contribution to provide an answer to the previous questions. Yet, it shows another morally relevant aspect of the launch of mega-constellations, however, this concern is relevant for other space objects with autonomous manoeuvring capabilities.

The third morally charged aspect related to mega-constellations and other large-scale space missions is that useful orbital slots and the radio spectrum are scarce natural resources. Then, this scenario is similar to the tragedy of the commons, which is successfully studied in terms of distributed morality[8]. Therefore, the allocation of these scarce resources is another morally relevant aspect aggravated by the rise of mega-constellations. Currently, a part of the spectrum is allocated by the International Telecommunication Union (henceforth, ITU) to the satellite operators to perform uplink and downlink transmissions. The ITU also notes the orbital parameters to prevent interference with other satellites, that is the orbital slot of each spacecraft. It is important to note that the primary function of the ITU is related to the allocation of the radio spectrum and not with the assignment of orbital planes. The management of orbital planes is often left to the satellites' operators if, for example, two satellites are operating in a close orbital position with two different radio frequencies. The allocation of the spectrum is

[8] For a framing of the problem as a common see (Salter, 2015).

performed on a first come, first served principle and since space activities are disproportionally concentrated in developed countries equity concerns arise.

Developing countries became concerned that the most demanded frequencies and the most beneficial orbital slots would be occupied by the time they developed space capabilities[9]. To address this the 1977 WRC elaborated an alternative mechanism of spectrum management aimed at ensuring equitable access to orbital-frequency resources—the allotment of radio frequencies. According to this mechanism, specific radio frequencies are included in the so-called a priori plans and thereby reserved for the use by specific states (Radio Regulations, 2016, No. 1.17). However, mega-constellations raise new concerns. Other mechanisms should be put in place to ensure that the useful parameters for providing global broadband services in LEO are not exhausted by private enterprises of developed countries.

These are just three moral issues related to the launch of mega-constellations highlighted by considering outer space as a multiagent system under the framing of distributed morality. The next section examines two mitigations strategies to foster human flourishing beyond planet earth.

5. MITIGATION STRATEGIES

This section deals with two strategies to curb the ethical concerns of mega-constellations. The first draws from the responsible research innovation research while the second is aggregation policies of morally desirable actions.

These two approaches are closely intertwined as the design and launch of a vast number of satellites are not the result of morally reprehensible conducts. The problem lies in the fact that mega-constellations are problematical from an ethical perspective, even if their promoters have the best possible intentions. Therefore, moral considerations anchored on intentionality might not provide useful solutions, as shown in the context of multiagent systems in which human agents and artificial agents interact (Floridi, 2013, 2017; Greco & Floridi, 2004). Addressing the ethical concerns highlighted in the previous section ought to be done at an earlier stage before mega-constellations are technically mature. So that neglecting fundamental ethical principles is less of a risk for correcting it in the design phase is more feasible than once thousands of satellites are already placed in LEO.

Concerning the first proposed approach, the definition of RRI adopted is taken from the work of Von Schomberg, that is "Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products" (Von Schomberg, 2013, p. 59). It is evident how the application of the RRI framework to the issue at hand poses significant difficulties. First, ethical acceptability is difficult to ascertain when mega-constellations are poised to impact the entire globe. Which ethical framework should be adopted? Are the norms found in the space treaties enough to provide a benchmark for it? Second, it is not clear if appropriate methods for technology assessment and foresight are being used within the space industry concerning the

[9] This problem is more relevant in the case of geo-stationary orbits (where the speed of the satellites matches the rotation of the earth so that the spacecraft appears stationary from the earth perspective). However, the issue might become more prominent if the number of satellites in LEO vastly increases.

unprecedented nature of mega-constellations. What seems critical in this context is the lack of global deliberation for a technological infrastructure design to operate globally and, more importantly, managed by a handful of private enterprises. Third, the precautionary principle proper of EU law does not extend its reach to outer space as it is not mentioned in the international sources governing space activities. Moreover, national implementations of it might not be effective since enterprises can easily change the applicable law leveraging the multifold nature of the notion of launching state.

Against the difficulties of applying the RRI methodology to the case of mega-constellations, the following remarks are made. The OST provides a starting point for evaluating the ethical acceptability of these systems, art. 1 states that "[t]he exploration and use of outer space, including the Moon and other celestial bodies, shall be carried out for the benefit and in the interests of all countries, irrespective of their degree of economic or scientific development, and shall be the province of mankind". Thus, the question one needs to ask is if private mega-constellations for broadband communications benefit and are in the interest of all. Prima facie, the answer is affirmative. Providing global high-speed internet access is a desirable and acceptable endeavour to undergo, because it will cover most of the population, including rural and remote areas. However, a balance must be struck against the risks outlined in the previous section; two possibilities come to mind. First, codes of conduct should be adopted by the companies involved stating how they intend to act to mitigate the risks of mega-constellations. In passing, code of conducts could also address other areas of concern. Second, the adoption of standards and self-regulation should be encouraged in this area. The issue here lies in establishing globally accepted measures in the fragmented landscape of space regulations. Third, it would be highly desirable to include the precautionary principle in the body of space law; however, this is unlikely to occur. In the global environment, an agreement among the major space fairing nations seems far in the future. Absent such principle, the need for deliberative mechanisms with stakeholders along with more public engagement and debate becomes stronger. Considering the launch of mega-constellations and their associated risks, on-going public discussion and monitoring of public opinion would be desirable. These are just some of the possible future directions to study; more in-depth considerations are left to another time.

The second suggested approach to curb the ethical concerns of mega-constellations consists of the aggregation of possibly good actions and the fragmentations of undesirable ones, i.e. ethical policies of aggregation (Floridi, 2013, 2017). The international space community might do the former by sharing data, best practices and codes of conduct. Moreover, ethical aggregation ought to be complemented by incentives as well as disincentives put in place by legislation and policies. Ideally, such mechanisms would occur at the international level, however, it might be the case that in the short-term, national initiative will be more effective. As for the fragmentation of morally bad actions, it is possible that the space community might continue to shun irresponsible actions such as the wilful increase of space debris (e.g. by the intentional destruction of satellites via anti-satellites missiles) or acts against international space law such as the launch of space objects without registration. It is clear that much work needs to be done to study these mechanisms, a task beyond the scope of this contribution. For now, it is sufficient to highlight the most viable strategies to, not only ease ethical concerns of mega-constellations but also to harness the power of distributed morality in the multiagent system of outer space.

6. CONCLUSION

The goal of this chapter was to highlight ethical concerns related to mega-constellations for broadband communication. In passing, the shortcomings of international space law have been discussed. The main contribution of this work is framing the environment of outer space in terms of distributed morality. That is, I contend outer space to be a multiagent system in which human and artificial agents act singularly in morally negligible or neutral ways that, nonetheless can have critical moral consequences when aggregated. Also, the conjunction avoidance choice sketched in section 4 clarifies some of the moral concerns of the new era of space exploitation and exploration. Three moral issues related to the topic at hand have been discussed, namely, the exacerbation of the problem of space debris, the design of autonomous space objects for collision avoidance, and the mechanism for allocating the radio spectrum along with orbital slots. Two strategies have been suggested to ease the concerns of the deployment of mega-constellations. The first is to draw from the RRI framework. The second concerns ethical policies of aggregation. Due to the nature of this contribution, several questions demand future work. I hope to have provided interested researchers with a starting point to tackle these challenges. Endeavours in outer space are vital to the human flourishing, and the path to the business ethics of private space exploitation has just begun.

REFERENCES

Arnould, J. (2011). *Icarus' second chance: the basis and perspectives of space ethics* (Vol. 6): Springer Science & Business Media.

Bergamini, E., Jacobone, F., Morea, D., & Sciortino, G. P. (2018). The Increasing Risk of Space Debris Impact on Earth: Case Studies, Potential Damages, International Liability Framework and Management Systems. In *Enhancing CBRNE Safety & Security: Proceedings of the SICC 2017 Conference* (pp. 271-280).

Floridi, L. (2013). Distributed morality in an information society. *Science and engineering ethics, 19*(3), 727-743.

Floridi, L. (2017). Infraethics–on the Conditions of Possibility of Morality. *Philosophy & Technology, 30*(4), 391-394. doi:10.1007/s13347-017-0291-1

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines, 14*(3), 349-379.

Greco, G. M., & Floridi, L. (2004). The tragedy of the digital commons. *Ethics and Information Technology, 6*(2), 73-81.

Jakhu, R. S. (2010). Iridium-Cosmos collision and its implications for space operations. In K.-U. Schrogl, W. Rathgeber, B. Baranes, & C. Venet (Eds.), *Yearbook on Space Policy 2008/2009: Setting New Trends* (pp. 254-275). Vienna: Springer Vienna.

Kessler, D. J., Johnson, N. L., Liou, J., & Matney, M. (2010). The kessler syndrome: implications to future space operations. *Advances in the Astronautical Sciences, 137*(8), 2010.

Larsen, P. B. (2009). *Space law: A treatise*: Ashgate.

Lipsky Jr, A. B., & Sidak, J. G. (1998). Essential facilities. *Stan. L. Rev., 51*, 1187.

Marboe, I. (2016a). Small Is Beautiful? Legal Challenges of Small Satellites. In P. M. Sterns & L. I. Tennen (Eds.), *Private Law, Public Law, Metalaw and Public Policy in Space: A Liber Amicorum in Honor of Ernst Fasan* (pp. 1-16). Cham: Springer International Publishing.

Marboe, I. (2016b). *Small Satellites: Regulatory Challenges and Chances*: Brill.

Matney, M., Vavrin, A., & Manis, A. (2017). *Effects of CubeSat Deployments in Low-Earth Orbit*. Paper presented at the 7th European Conference on Space Debris, Darmstadt, Germany.

Millan, R. M., von Steiger, R., Ariel, M., Bartalev, S., Borgeaud, M., Campagnola, S.,... Gregorio, A. (2019). Small satellites for space science. *Advances in space research*.

Munevar, G. (2019). An obligation to colonize outer space. *Futures, 110*, 38-40.

Rao, R. V., Gopalakrishnan, V., & Abhijeet, K. (2017). *Recent Developments in Space Law: Opportunities & Challenges*: Springer.

Salter, A. W. (2015). Space debris: A law and economics analysis of the orbital commons.

Schwartz, J. S. (2011). Our moral obligation to support space exploration. *Environmental Ethics, 33*(1), 67-88.

Shaw, A., & Rosher, P. (2016). Micro satellites: the smaller the satellites, the bigger the challenges? *Air and Space Law, 41*(4), 311-328.

Trump, D. J. (2018). Space Policy Directive-3, National Space Traffic Management Policy. In: June.

International Space Law United Nations Instruments, (2017).

Von Schomberg, R. (2013). A vision of responsible research and innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society*, 51-74.

ETHICAL ISSUES OF E-INFRASTRUCTURES: WHAT ARE THEY AND HOW CAN THEY BE ADDRESSED?

Damian Eke, Simisola Akintoye, William Knight, George Ogoh, Bernd Stahl

De Montfort University (Leicester, United Kingdom)

damian.eke@dmu.ac.uk; simi.akintoye@dmu.ac.uk; william.knight@dmu.ac.uk; george.ogoh@dmu.ac.uk; bstahl@dmu.ac.uk

ABSTRACT

E-infrastructures are emerging as novel and effective ways of increasing creativity and efficiency of research. As technological innovations, these virtual, ubiquitous, pervasive infrastructures offer possibilities of international collaborations through open, data-driven and high-quality computing environments. Particularly in Europe, the aim is to create an ecosystem of e-science where multiple disciplines converge to foster interoperable and open collaboration with the help of significant data processing and computing capacity. While most agree that these research infrastructures are crucial to scientific reproducibility and rigor, e-infrastructural literature lacks critical discussions on the ethical concerns they raise or potentially can raise. This paper argues that e-infrastructures can raise a number of ethical, legal and social concerns. Some of these relate to data privacy and data security but they also include issues around animal welfare, data bias, intellectual property rights, environmental sustainability, digital divide and other unintended uses/misuses. This paper also presents a practical way of thinking about ethics in e-infrastructures. The underlying argument here is that addressing e-infrastructure ethical issues should start from the design of the infrastructure and continue through to its lifecycle. It requires the integration of relevant ethical principles into its design to foster responsible use/application. We then propose that this can be done through the Responsible Research and Innovation approach as an ethics-by-design tool.

KEYWORDS: Research, E-infrastructures, Infrastructures, Ethics, RRI, Ethics-by-design.

1. INTRODUCTION

In the last few decades, there has been an increasing move towards collaborative research to provide common solutions for shared concerns which has led to the development of research infrastructures, particularly e-infrastructures, that are intended to transform scientific research and practice. Backed by huge funds, national (like the NSF Office of Advanced Cyberinfrastructure) and regional (such as the European Strategy Forum on Research Infrastructures) programmes are being used to facilitate the establishment and operation of such research infrastructures (Schroeder, 2007;Pollock and Williams, 2010; Andronico et al., 2011). There is a long tradition to develop such infrastructures in Europe with over 700 million euro of budget allocated to close to 100 e-infrastructure projects in the EU's research framework programme Horizon 2020 (H2020) (Versweyveld, 2019). These infrastructures cover most aspects of science and research such as biological sciences, health/medicine, good, energy, environmental sciences, Physical sciences/engineering, computing sciences, social sciences, arts and humanities. The European Union also showcases some of larger e-infrastructure initiatives that include EPOS: Viable solutions to tackle solid Earth grand challenges (Jeffery and Bailo, 2014), ELIXIR: A distributed infrastructure for Life science information (Crosswell and Thornton, 2012) or SHARE: A survey on Health, ageing and retirement in Europe. (Börsch-Supan et al., 2013).

The international ecosystem of e-infrastructures is extensive and different in that they have evolved along different geographic, disciplinary and application dimensions. These often-disparate e-infrastructures aim to provide open, flexible and competitive national and international services to advance scientific collaborations. In addition to enabling collaborations across disciplines and national boundaries, these ICT-based, data-driven and computer-intensive infrastructures are also being established to promote scientific reproducibility, improve research efficiency, creativity and rigor. In many cases, they provide digital/virtual open access to resources and services that provide answers to complex scientific questions that require global reflection. This means that they attract users from different jurisdictions who access multimodal and multidimensional resources (including a significant amount of data) and services (such as super computing services). E-infrastructure services often depend on significantly increasing amounts of data raising novel ethical, privacy and security challenges. The e-IRG White paper of 2013 also highlighted the technical, legal and ethical challenges involved in creating a collaborative data infrastructure in light of the volume, variety and velocity of data involved (e-IRG, 2013.).

The above emphasis on data issues and the recent EU's legislative emphasis on Data protection by design and by default (DPbD) in article 25 of the EU General Data Protection Regulation (GDPR) can be read as suggesting that privacy and data protection are the only data related concerns to be considered in innovations associated with human data such as e-infrastructures. Such a reading of the current discourse would clearly be wrong considering the possibilities of commercial research via these e-infrastructures which raise some uncertainty about the rules and formalities at national, European and international levels. According to the EU Horizon 2020 Expert Advisory Group on *European Research Infrastructures including e-Infrastructures,* research infrastructures go beyond datasets and scientific instruments because they play significant roles in shaping the intellectual and cultural dimensions of the innovation ecosystem; enabling smart specialization strategies. E-infrastructures improve national and regional scientific capacity through multifaceted relationships between data, computing resources and researchers from different scientific and regulatory backgrounds. These pose a number of novel ethical, social and legal challenges, depending on a lot of things including the nature of the e-infrastructure involved, the type data that it engages with and the breath of its applications. But as the construction and operation of these e-infrastructures continue to advance, there is an evident dearth of literature on the ethical issues they raise or can raise. Therefore, this paper seeks to answer two research questions: what are the ethical issues research e-infrastructures raise? And how can these ethical issues be addressed? To address these questions, we critically reflect on the e-infrastructure ecosystem.

This paper makes two contributions. First, it provides a general overview of the ethical concerns associated with e-infrastructures. As e-infrastructures continue to transform research and innovation, it is pertinent to be aware of the ethical, legal and social concerns they raise or exacerbate. These issues depend on the type of e-infrastructure and the nature of the services/resources it offers. Second, it is easy for the conversations around ethics to become too abstract to offer practical insights into achieving its objectives. Therefore, we also present a practical way of thinking about ethics in e-infrastructures through ethics-by-design. The underlying argument here is that addressing e-infrastructure's ethical issues should start from the design of the infrastructure and continue through its lifecycle. It requires the integration of relevant ethical principles into its design to foster responsible use/application. We propose that this can be done with the Responsible research and innovation approach as an ethics-by-design tool.

Therefore, this paper starts with a brief overview of what we mean by research e-infrastructures and how they raise ethical concerns. We then provide a brief description of Ethics by Design and how it can positively shape the design and application of e-infrastructures. Responsible Research and Innovation (RRI) is then identified as a possible ethics-by-design tool that can help to achieve Ethics by Design in e-infrastructures. An overview of how this can be applied in the development of an e-infrastructure is then outlined. We conclude by discussing the challenges of applying RRI as an Ethics by Design tool to e-infrastructures.

2. A NOTE ON CONCEPTS AND STRUCTURE

The general discourse on the ethical issues associated with e-infrastructures requires the explanation of our understanding of what constitutes an e-infrastructure. It is important also to describe how we arrived at the ethical issues. Therefore, this section starts with a brief clarification of what is referred to as e-infrastructures in this paper and how they are different from other research infrastructures. Then it provides clear insights on what ethical issues are associated with e-infrastructures which e-infrastructure developers can use to take proactive actions to build ethically responsible platforms for research and innovation.

The identification of ethical issues was undertaken through a critical exploration of emerging literature on e-infrastructures. A critical literature review was essential for identifying the landscape of e-infrastructure operations and applications that can raise some concerns. Issues were adjudged ethical issues given that they create conflicts with established moral principles; requiring a person(s) to choose between alternatives justified or evaluated as right or wrong. As Stahl (2012) observed, these are issues that are associated with moral intuition or explicit morality. This paper identifies these issues in the context of e-infrastructure operations and in the next section suggests a possible way of addressing them. The contributions of this paper are based on non-empirical critical reflection on literature and personal observations of the authors as members of the Ethics support team of the EU Human brain project building a neuroscience e-infrastructure - European Brain Research InfraStructure (EBRAINS).

3. E-INSTRASTRUCTURES

According to the Directorate-General for Research and innovation of the European Commission, research infrastructures are ''facilities that provide resources and services for research communities to conduct research and foster innovation''. In our increasingly digitized world, these facilities are becoming more electronic with all ICT based resources – interconnecting computer science and specific research disciplines. These are called e-infrastructures combining digital technology, computational resources to support research collaborations; creating new virtual research communities that share, federate and access scientific tools and resources (including but not limited to data and computing facilities). They are defined in this paper as *computing facilities and electronic resources that facilitate research and innovation* which are differentiated from other research infrastructures due to their ability to provide digital services, resources and tools for scientific research.

E-infrastructures are innovations that harness knowledge from multiple disciplines for the advancement of research and the benefit of society. From the online etymology dictionary, ‘innovation’ is a 1540s word that comes from the latin word *innovationem,* meaning ‘‘*a novel change, experimental variation, new thing introduced in an established arrangement*’’. Therefore, the key meaning of innovation relates to a change of what is existing; it does not necessarily mean a new invention but involves a new approach. As an innovation, e-infrastructure is revolutionizing the way scientific research is done through a combination of some or all of these critical elements: supercomputing, cloud computing, big data, networking, collaboration and open access. They are not only changing the landscape of scientific relationships but also raising new ethical challenges or exacerbating old concerns.

According to the European e-infrastructure Reflection Group (e-IRG), the pan-European e-infrastructure landscape includes *networking infrastructure*, *supercomputing facilities, Cloud Infrastructure* and *data infrastructure*.(van Rijn and Vandenbroucke, 2017) An example of a networking infrastructure is GÉANT that provides interconnectivity between national research and Education networks (NRENs) across many European and non-EU countries with estimated 50 million users. Super computing infrastructures in Europe include European Grid infrastructure (EGI) and Partnership for Advanced Computing in Europe (PRACE). Both provide high-level computing services but EGI focuses on large-scale federated High-throughput computing (HTC) solutions and PRACE provides access to high-performance computing services and facilities. In recent years, ELIXIR has become a key data infrastructure that manages and safeguards the increasing volume of data generated by publicly funded research worldwide. Other data infrastructures include EUDAT, ZENODO and OpenAIRE. A further element of e-infrastructure is the increasing use of cloud technologies. The *Helix Nebula Initiative* is one example creating research partnerships between industry, space and science through open cloud services. These are indeed becoming core platforms for e-science, education and innovation (Andronico et al., 2012).

E-infrastructures have one or more of these elements: data, networking, HTC, HPC or cloud computing. For instance, as a network infrastructure, GÉANT also has a cloud service platform but cannot be classified as a High-performance computing (HPC) infrastructure. PRACE is a HPC infrastructure but can also offer data services. However, EGI is a High-throughput computing (HTC) infrastructure but also can offer both cloud and data services. The type of infrastructure determines what research is supported and the type of research community created. These e-infrastructures empower researchers to explore scientific questions in different ways. They involve a lot of complex interrelationships between researchers, the technology, organizations, networks and variety of human and animal datasets (in most cases). The technical and social dynamics of these interrelationships present a number of ethical, legal and social risks because persons and data are involved. Some of these concerns can generally be associated with all e-infrastructures or indeed all research infrastructures but some are peculiar to certain infrastructures.

3.1. E-infrastructure ethics

E-infrastructures raise different ethical concerns depending on the type of resources and services they offer and the relationships they foster. For Data e-infrastructures, concerns are mainly related to data sharing, security, access and interoperability. While these e-infrastructures attempt to foster scientific discoveries through sharing of data, there is also the need to ensure the subject's privacy and confidentiality (Gagliardi and Muscella, 2010). These are fundamental ethical and legal requirements for research collaboration in any research involving human data. A virtual platform hosting terabyte of human data surely amplifies concerns related to privacy and data protection. Identifiable research data-related issues include; informed consent issues, incidental findings and possibilities of positive re-identification. Informed consent (D'Abramo, 2015; Wolf et al., 2018) and incidental findings (Viberg et al., 2014) have been acknowledged in literature as a major challenge facing human subject research and infrastructure operations. An e-infrastructure that facilitates a borderless sharing and usage of research data only exacerbates these problems. Fundamentally these raise the questions of autonomy and the practical questions of accountability and whose responsibility it is to address these issues. They also raise the technical question of how to protect the rights (to privacy and confidentiality) of the research subjects in the face of the changing landscape of technology making it difficult to achieve effective de-identification/anonymization of human datasets (Rocher et al., 2019).

Additionally, there is the critical issue of data controllership- who owns and or controls the data. This is a concept defined by the EU General Data Protection Regulation as any organizational entity ''which alone or jointly with others, determines the purposes and means of the processing of personal data''. A clear identification of the data controller therefore has important influence on the sharing of data as well as the apportioning of liabilities in cases of data breach. E-infrastructures involve data processing at different levels and by multiple institutions which are sometimes located in different countries with different data protection provisions. Determining the data controllership and data processor roles at the various levels of the 'research grid' becomes complicated. Even the EU GDPR does not provide sufficient clarity to this issue given the different national interpretations of its provisions.

The collaborative nature of the e-infrastructure ecosystem also complicates the issues of copyright and intellectual property rights. For instance, the collaborative work in or for e-infrastructure can lead to the development of a bespoke software with the possibility that no licence or intellectual property agreement was made prior to development as was observed by the 2013 e-IRG Task Force Report on Legal issues. Determining the owners of the intellectual property and what the software can be used for especially when state funds are involved become a critical challenge. Also raised in the context of data e-infrastructures are deeper questions of data misuse which may include, but are not limited to using findings out of context or distorting research findings. There are also the issues of dual use which are gaining more prominence in EU and international policy literature. This refers to the use of research findings for unacceptable military or commercial purposes. This is a concern already identified in the EU HBP where an opinion on Responsible dual-use has been developed (Aicardi et al., 2018). Dual use and misuse also featured prominently in the recently published OECD recommendation on Responsible innovation in neurotechnology. For instance, a neuroscience e-infrastructure that increases the integration of brain data and technology could be leveraged for military purposes. A recent article published in the US's National Defense University's journal PRISM, observed that China's military strategy is informed by the belief that brain science is key to the effectiveness of future battlefields (Kania, 2020). This raises the possibility of brain data misuse and highlights the level of national security challenges brain science research is likely to present.

Another issue that does not receive deserved attention in the data e-infrastructure discourse is animal welfare. Animal experiments that generate animal data are regulated differently in different parts of the world. What is ethically permissible in one jurisdiction may be illegal in another. A data e-infrastructure designed to curate and share all types of animal data from every part of the world without any form of governance implicitly endorses all forms of animal experiments including potentially unethical research. As a virtual platform with users in different parts of the world, data e-infrastructures for animal data could potentially be used to share data from unethical animal experiments which exacerbates animal welfare concerns. Without any form of governance, e-infrastructures can thus facilitate low quality animal data which can ultimately affect the quality of e-science they foster.

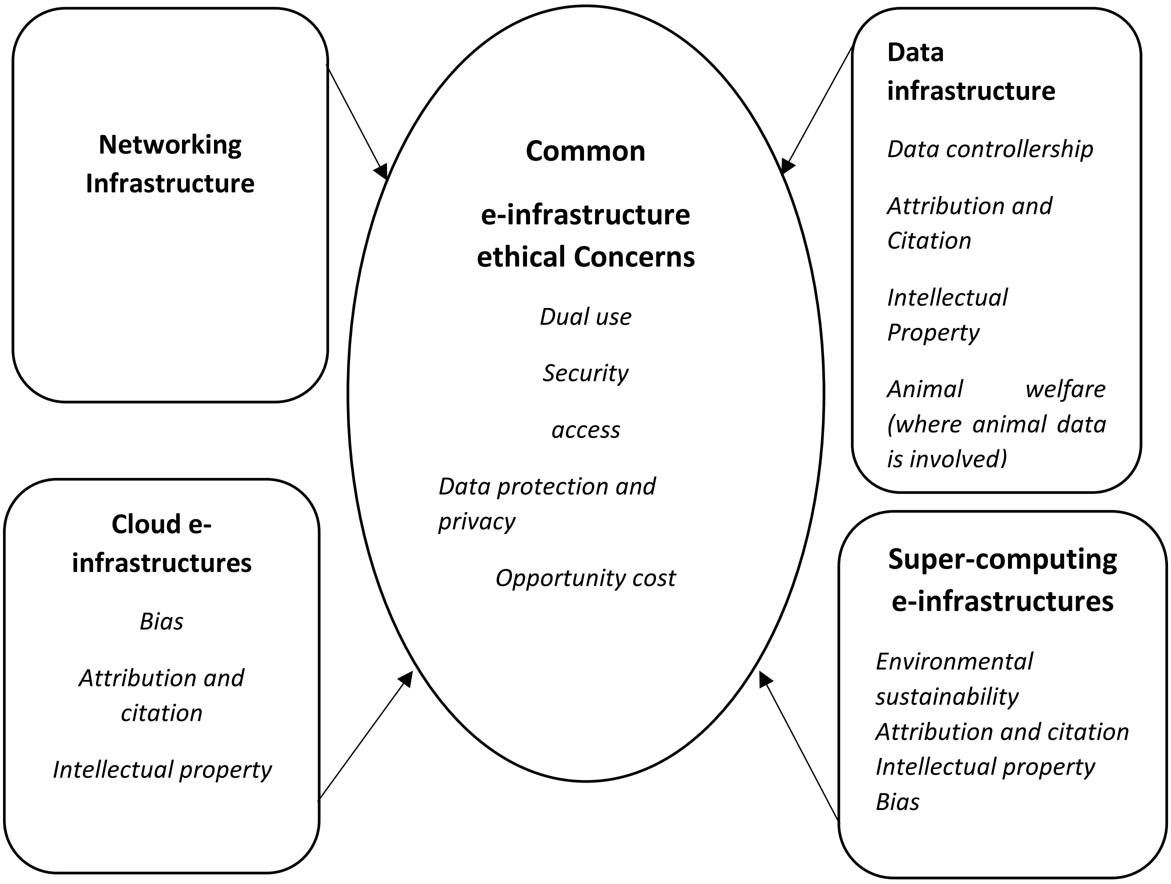
There are also complex ethical issues that are associated with supercomputing infrastructures. Addressing them as ethics in HPC, Lawson et al., (2019) identified three major ethical issues surrounding the use and application of HPCs. These include misuse, inequality and environmental concerns. The underlying argument here is that HPCs could potentially be misused for unacceptable military purposes or to develop a destructive pathogen. It can also be argued that, as with any other high-resource intensive technology, HPCs can indirectly perpetuate a digital divide between developed and less developed countries; men and women and other minority groups. Of the Top500 list of supercomputing sites published in 2019, only 1 is in South America and none is in Africa. The capacities offered by these systems can increase a country or region's competitive advantage while thickening the digital lines that divide nations which creates an ethical imperative. There are also the environmental impacts of HPCs. The large amount of power consumption and its associated carbon emission are well documented critical challenges of supercomputing services (Yang and Chien, 2016).

Furthermore, the cloud e-infrastructure ecosystem in Europe provides support for the creation of many artificial intelligence (AI) systems. Many of these systems (machine learning, deep learning) are supported over distributed e-infrastructures in the European Open Science Cloud (EOSC). An example is the DEEP-Hybrid-Datacloud that helps users to develop, train, share and deploy models. The EOSC is the European Commission's platform created to federate emerging e-infrastructures in a way that will unlock the value of big data and foster innovation (including advancements in artificial intelligence) towards achieving the digital single market strategy. The ever-growing relationship between these infrastructures, data and deep/machine learning raises the possibility of discrimination/bias which can creep in at different stages of the analysis, exacerbating intersectional biases that might be evident in the curated data. While network e-infrastructures are not primarily designed to store research data, a significant amount of personal data is processed through it. Consider the examples of eduroam and eduGain that provide a communications platform across research and education. This requires that organizations be able to communicate and collaborate with enhanced privacy. There is still the danger of malicious attacks that can lead to privacy breaches. Network security therefore is of great importance.

A further concern about e-infrastructures relates to the value they represent to the scientific community and to society at large and the often-significant costs and opportunity costs they incur. Most of the infrastructures introduced earlier have large financial value and cost significant amounts of money to maintain. These are resources that cannot be used for other research or for other socially beneficial activities. This is even more relevant because they are usually financed by public funds which directly compete with other beneficial goods, such as social security or healthcare. This is justified by the potentially large public benefit that may arise out of the research that is facilitated by the infrastructure. However, measuring or accurately predicting these benefits is extremely difficult. A lack of use of the infrastructure by the scientific community or a lack of research outcomes thus constitute a waste of resources which is a political but also an ethical concern.

The above issues are not trivial and contribute to the wider discourse on what can be referred to as *e-infrastructure ethics* in this paper which means a set of moral principles and standards that should govern the use of e-infrastructures. This paper proposes that one of the approaches that can be utilized to address these ethical issues is to develop a responsible/ethical innovation through a practical application of ethical thinking to the design of its products and services. This approach combines technical, ethical and social mechanisms to provide solutions based on reflexive actions. It is a proactive and preventive measure motivated not only by the imperative to consider research ethics questions but also by the need to responsibly govern the application of generated complex, big data and the necessity to embed relevant principles into the design. The latter should lead to the establishment of practical and responsible mechanisms/approaches informed by ethics and the law and designed to integrate core ethical principles and legal provisions into the e-infrastructure.

Figure 1. Ethical issues of different e-infrastructures.



The broad nature of the above e-infrastructural concerns is a reason to believe that a framework that embeds privacy or data protection into design is not sufficient to mitigate other social, economic, legal, philosophical and ethical concerns e-infrastructure raises. A responsible e-infrastructure technology requires a robust approach/mechanism that addresses fundamental concerns beyond privacy and data protection. There is a need to integrate relevant ethical principles into the development and application of e-infrastructures. This is important because as Simon (2016) observed, in building technologies, values are often unintentionally inscribed in them and in return they may promote or demote certain values. Even though e-infrastructure is not technology *per se,* as an innovation, it should facilitate responsible actions. Instead of unintentionally embedding undesirable values into the design, responsible e-infrastructure requires a conscious effort to intentionally embed desired values into the innovation in ways that can promote values desired by the relevant stakeholders who are or can be affected by the technology. While this sounds great, the practical questions on what principles and how to embed them into the design of e-infrastructures are unclear in literature or practice.

Drawing from the literature on Technology ethics and Design Thinking, this paper argues that the concept of Ethics-by-design (Mulvenna et al., 2017; Dignum et al., 2018) as a form of value-sensitive design (Friedman et al., 2006), provides a robust framework for ethical and legal considerations of risks in the development and application of research e-infrastructures. Even though this concept has been applied in robotics (Dodig Crnkovic and Çürüklü, 2012) and business (Moore, 2017) discourses, it is yet to be applied to the design of a research e-infrastructure. The authors draw extensively from their experiences of developing approaches of addressing data governance issues, data protection, research ethics compliance and dual use in a large-scale EU project.

4. ETHICS-BY-DESIGN AND TECHNOLOGY

The idea of ethics by design is not an entirely new concept and has been around for over 20 years. One of the earliest developers of the notion was Tonkinwise (2004); he based his ideas on the works of Scarry (1987), Latour (1992), Jelsma (2003) and Borgmann (1995). Tokinwise sees design as embodying ethics because it is the process for making the world friendlier to us and the effort to make it more caring towards us, and therefore more morally acceptable to us. He therefore maintains that 'forethought' is necessary to design things that make ethical outcomes easier or harder, and suggests that ethics by design/ in design makes it easier for people to be more ethical. Feister et al. (2016) also promote the concept of ethics in design because they agree with the notion that design is inherently tied to ethics. They maintain that integrating ethics in the micro-level everyday decisions and thinking about it throughout the engineering design process will encourage greater incorporation of ethical thinking into the entire design process. They therefore advocate for a human-centred model of design (HCD) which is situated around appreciation of users' knowledge, skills, and experiences and gives attention to all stakeholders that might be affected beyond the targeted user.

In applying similar principles to the design of autonomous AI agents (softbots/robots), Crnkovic and Çürüklü (2012) promote the idea that any intelligence acquired by such artefacts must come in conjunction with ethics through what they call 'ethical by design'. This takes into consideration the engineering ethics of designers, manufacturers, and maintenance services, the user's attitude, as well as those of the artefacts. In a similar vein, Dignum et al. (2018) have applied the concept of ethics by design to autonomous agents in an effort to explain how to build ethically-aware agents. In this regard, they define ethics by design as “the methods, algorithms and tools needed to endow autonomous agents with the capability to reason about the ethical aspects of their decisions, and methods, tools and formalisms to guarantee that an agents behaviour remains within given moral bounds.” They maintain that the moral, societal and legal values of autonomous agents must be taken into consideration in their design and that focus should be on ensuring thrust of AI systems rather than performance alone. They conclude that AI related developments must be designed to ensure societal good through three principles articulated as ‘ART’ – **A**ccountability (the system should be able to justify its decisions and actions to stakeholders); **R**esponsibility (of both the AI system and those interacting with it in accounting for decisions, diagnosing errors, or unexpected findings); **T**ransparency (in terms of explanation and clarity of algorithms and their results). Other concepts that they suggest can aid the integration of ethics by design include ethically aligned design, responsible research and innovation, and ethics of the design processes itself.

An attempt was made by Mulvenna et al (2017) to develop a manifesto of principles for Ethical by Design that would be all-encompassing and guide everyone developing or considering solutions regardless of the area, market or their own expertise. They firmly believe that design thinking can and should be ‘ethical by design’ such that it ‘inherently supports the ethical development, selection, and use of products and services.’ They therefore propose the following 12 principles as the manifesto for ‘ethical by design’ - the design should engender empathy for others, enable informed decision, offer alternative or customisation that respects people’s right to choose how they wish to engage with a product, allow equitable access and which must be balanced by privacy and security, support progression of policy, and actively look for and challenge biases in the product or service. Other principles of the manifesto include a requirement to compliment differing needs, abilities, viewpoints and morals; support shared decision making and feedback, aim for sustainability, be realistic about what is possible, integrate planning for handling failure, and provide support throughout the lifespan of the product or service.

While agreeing that the above manifesto of principles for ‘Ethical by Design’ provides a pragmatic ethical framework for design processes, Lee et al. (2018) also noted that there are other ethical principles to be considered in design. Judging from the above identified ethical issues of e-infrastructures, some relevant principles will be to be proactive rather than reactive, do a lifecycle ethics, to respect human autonomy, non-instrumentalism of living beings, intellectual property rights and striving for inclusivity, diversity and environmental sustainability. But how can these be integrated into the design of e-infrastructures?

4.1. Ethics-by-design and e-infrastructures

As it has been noted earlier in this paper, e-infrastructure is not a technology but can be classified as an innovation. It is an innovation that uses a combination of technologies to foster complex relationships/collaborations. These emerging collaborations are outcomes of the design of the e-infrastructures and raise a number of ethical concerns. Shaping these relationships e-infrastructures foster at the design level is an imperative. It is evident that the design of e-infrastructure is fraught with the need to make ethically responsible choices that can shape the application of its tools, resources and services. There is a need for e-infrastructure developers to appreciate the responsibility they have in addressing ethical problems exacerbated by this innovation. Researchers (including commercial entities) are encouraged to use the state-of-the-art services offered by e-infrastructures to develop cutting-edge results for civil society. In most cases, this involves a relationship between users and thousands of datasets from human subjects. The complex relationships within the e-infrastructure ecosystem present critical challenges and the integration and implementation of relevant principles will require the collaboration of stakeholders involved in e-infrastructures. Embedding these principles into the design of a responsible e-infrastructure where health data and supercomputing services are involved is a challenge facing e-infrastructure developers. Effective implementation of the ethical principles such as autonomy, proactivity, beneficence and non-maleficence will require collaborations between e-infrastructure developers, resource/service providers and the users. It will involve a mixture of technical, social and legal mechanisms that provide solutions to both identifiable and unanticipated problems. It requires ethics by design.

4.2. Ethics-by-design through rri for e-infrastructures

We propose that one possible way of building these principles into e-infrastructures is by applying the Responsible Research and Innovation (RRI) approach to the design. During a Thematic day at the 20th International conference on Principles and practice of multi-agent systems (PRIMA) in 2017, Juan Pavon suggested RRI as an ethics-by-design tool for multi-stakeholder intelligent systems (Dignum et al., 2018). His argument was that it is a tool that copes with issues that require cooperation. RRI is defined as a meta-responsibility framework that aligns the goals, purposes and processes of research and innovation to produce desirable outcomes (Stahl, 2013). E-infrastructures epitomize the convergence of research and technological innovations. RRI therefore, seems a good approach to achieve a responsible e-infrastructure that will positively support those who use it. The application of RRI into the building of this technology is a way of influencing e-research and e-science in ways that are socially acceptable, ethically responsible, legally compliant and environmentally sustainable.

Jirotka et al., (2017) articulated the importance of this approach in Information and Communications Technology (ICT). Recent works of some of the authors of this paper highlight how this inclusive and discursive mechanism can be applied in care robots (Stahl et al., 2019; Stahl and Coeckelbergh, 2016) and data governance in neuroscience (Fothergill et al., 2019). In Fothergill et al., (2019), a version of RRI based on Stilgoe et al., (2013) was used to unite and respond to large-scale neuroscience stakeholders' expectations by ensuring an ongoing, productive dialogue that produced beneficial results for research and innovation. This version endorsed by the UK Engineering and Physical sciences Research Council (EPSRC) adopts a reflective process of Anticipation, reflection, engagement and action (AREA) (Owen, 2014). In another publication, Jirotka et al., (2017) developed an extended version of this framework by adding what they called the 4Ps: *Process, product, purpose and people*. This was an attempt to provide practical guidance on key aspects of the research and innovation that shape the products or outcomes with central focus on purpose and the people who will be affected by the innovation. These applications of RRI highlight that this approach fosters a human-centred, collaborative and inclusive innovation that e-infrastructures require.

E-infrastructures create imbalances between those who have access and those who do not; those whose interests are addressed and those who lack representation; threaten the privacy rights of data subjects. A responsible e-infrastructure innovation should therefore consider the interests of many sections of the society, creating systems that will provide solutions with inclusive values. It requires meaningful engagement and reflection that can align the design with societal values. The last thing an e-infrastructure developer wants is to design a system providing high quality research services that nobody uses or that raises new or exacerbates old ethical and social problems. RRI can equip e-infrastructure developers with the knowledge of relevant values and principles through reflective and inclusive activities and practical ways of putting action into practice. It can provide a way of ensuring that ethics is an essential component of the core functionality of e-infrastructure or is integral to the system without diminishing its scientific purposes. It is important to note that the effectiveness of this approach in the design of an e-infrastructure is yet to be tested. As members of the Ethics and society sub-project of the EU Human Brain project (HBP) (working towards the development of an international neuroscience e-infrastructure) we have an opportunity to test the impact of RRI as an ethics-by-design tool in the development and application of an e-infrastructure.

5. CONCLUSION

In this paper we have set out the ethical issues that are likely to arise due to the development and use of e-infrastructures. There are already a number of such infrastructures in operation and new ones are under development. It is reasonable to assume that researchers will continue to seek to benefit from new ways of generating and analysing data. The trend towards big science or big data science is therefore likely to continue. While the perception of what counts as ‘big’ is likely to evolve, if current trends continue, then the ability to generate data will continue to outpace the ability of individual researchers or institutions to collect and make sense of this data. E-infrastructures are a way of addressing this, providing means of storing, analysing, visualising and processing such data.

If this prediction turns out to be true, then we are only witnessing the start of the development towards more e-infrastructures. These may come in all sorts of shapes and sizes; they will have different funding models and different governance structures. What they have in common is that they may raise ethical concerns that go beyond what their initiators had foreseen. It is therefore important to start to think about these issues early. Some of them are easy to foresee, such as data protection concerns in cases of infrastructure dealing with patient data. Others may be less obvious but still important. We hope that this paper provides the basis for a more comprehensive discussion of these ethical issues. Established e-infrastructures may provide good practice guidance and examples that developing ones can learn from.

In this paper we have suggested that there are established ways of implementing ethical reflection in the development of technology, such as ethics by design or RRI. The next step is now to transfer these ideas from the level of individual technology projects to infrastructure projects. This is likely to be difficult for reasons of size and scale which can raise further questions. E-infrastructures are, by their very nature, open and can be utilised for different purposes. They, therefore, raise difficult questions about accountability and responsibility. It is, therefore, unlikely to be easy to simply ‘implement’ ethics in infrastructure. It will require dialogue, openness to critique and failure, an ability and willingness to experiment and learn. We hope that this paper can contribute to the process of critical self-reflection of people involved in e-infrastructure to ensure that ethical issues do not stand in the way of the immense benefits that e-infrastructures promise.

ACKNOWLEDGEMENT

This project/research has received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

REFERENCES

Aicardi, C., Bitsch, L., Bang Bådum, N., Datta, S., Farisco, M., Evers, K., Fothergill, B.T., Giordano, J., Harris, E., Jørgensen, M.L., Klüver, L., Mahfoud, T., Rainey, S., Riisgaard, K., Rose, N., Salles, A., Stahl, B., Ulnicane, I., (2018). Opinion on ‘Responsible Dual Use’ Political, Security, Intelligence and Military Research of Concern in Neuroscience and Neurotechnology.

Andronico, G., Ardizzone, V., Barbera, R., Becker, B., Bruno, R., Calanducci, A., Carvalho, D., Ciuffo, L., Fargetta, M., Giorgio, E., 2011. E-Infrastructures for e-science: a global view. J. Grid Comput. 9, 155–184.

Borgmann, A. (1995) The Depth of Design. In: Buchanan, R. and Margolin, V. (eds.) *Discovering Design*. Chicago: The University of Chicago Press, p. 283.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S., 2013. Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). Int. J. Epidemiol. 42, 992–1001. https://doi.org/10.1093/ije/dyt088

Crosswell, L.C., Thornton, J.M., 2012. ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol 30, 241–242.

D’Abramo, F., 2015. Biobank research, informed consent and society. Towards a new alliance? J Epidemiol Community Health 69, 1125–1128.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M.S., Lopez-Sanchez, M., 2018. Ethics by Design: necessity or curse?, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 60–66.

Dodig Crnkovic, G., Çürüklü, B., 2012. Robots: ethical by design. Ethics Inf. Technol. 14, 61–71. https://doi.org/10.1007/s10676-011-9278-2

e-IRG, n.d. e-IRG Whit paper.

Feister, M.K. et al. (2016) Integrating Ethical Considerations In Design. In: *STEM Intergration Through Engineering*. Ammerican Society for Engineering Education 123 Annual Conference & Exposition. New Orleans, LA USA.

Fothergill, B.T., Knight, W., Stahl, B.C., Ulnicane, I., 2019. Responsible Data Governance of Neuroscience Big Data. Front. Neuroinformatics 13, 28. https://doi.org/10.3389/fninf.2019.00028

Friedman, B., Kahn, P., Borning, A., 2006. Value Sensitive Design and Information Systems, in: Zhang, P., Galletta, D. (Eds.), Human-Computer Interaction in Management Information Systems: Foundations. M.E Sharpe, Inc, NY.

Gagliardi, F., Muscella, S., 2010. Cloud computing–data confidentiality and interoperability challenges, in: Cloud Computing. Springer, pp. 257–270.

Jelsma, J. (2003) Innovating for Sustainability: Involving Users, Politics and Technology. *Innovation: The European Journal of Social Science Research*, 16(2), pp. 103–116.

Jeffery, K.G., Bailo, D., 2014. EPOS: using metadata in geoscience, in: Research Conference on Metadata and Semantics Research. Springer, pp. 170–184.

Jirotka, M., Grimpe, B., Stahl, B., Eden, G., Hartswood, M., 2017. Responsible research and innovation in the digital age. Commun. ACM 60, 62–68.

Kania, E., 2020. Minds At War: China's Pursuit Of Military Advantage Through Cognitive Science And Biotechnology – Analysis. PRISM 8.

Latour, B. (1992) Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker, W.E., Law, J. and American Council of Learned Societies (eds.) *Shaping technology/building society: studies in sociotechnical change*. Cambridge, Mass: MIT Press, pp. 225–258.

Lawson, M., Lofstead, J., Lüttgau, J., 2019. With Great Power Comes Great Responsibility: Ethics in HPC. Presented at the The International Conference for High Performance Computing, Networking, Storage, and Analysis, Denver Colorado.

Lee, T. et al. (2018) Teaching Ethical Design in the Era of Autonomous and Intelligent Systems. In: *2018 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*. 2018 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC). Albuquerque, NM, USA: IEEE, pp. 1–4.

Moore, S.L., 2017. Ethics By Design: Strategic Thinking and Planning for Exemplary Performance, Responsible Results, and Societal Accountability. HRD Press, Amherst, Mass.

Mulvenna, M., Boger, J., Bond, R., 2017. Ethical by design: A manifesto, in: Proceedings of the European Conference on Cognitive Ergonomics 2017. pp. 51–54.

OECD, 2019. Recommendation on Responsible Innovation in Neurotechnology, https://www.oecd.org/science/recommendation-on-responsible-innovation-in-neurotechnology.htm

Owen, R., 2014. The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation. J. Responsible Innov. 1, 113–117.

Pollock, N., Williams, R., 2010. E-infrastructures: How do we know and understand them? Strategic ethnography and the biography of artefacts. Comput. Support. Coop. Work CSCW 19, 521–556.

Rocher, L., Hendrickx, J.M., De Montjoye, Y.-A., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. 10, 1–9.

Schroeder, R., 2007. e-Research Infrastructures and Open Science: Towards a New System of Knowledge Production? Prometheus 25, 1–17.

Scarry, E. (1987) *The Body in Pain: The Making and Unmaking of the World*. Oxford University Press.

Simon, J., 2016. Value-Sensitive Design and Responsible Research and Innovation, in: The Ethics of Technology - Methods and Approaches. Rowman & Littlefield Publishers, London, pp. 219–236.

Stahl, B.C., 2013. Responsible research and innovation: The role of privacy in an emerging framework. Sci. Public Policy 40, 708–716.

Stahl, B.C., 2012. Morality, ethics, and reflection: a categorization of normative IS research. J. Assoc. Inf. Syst. 13, 1.

Stahl, B.C., Akintoye, S., Fothergill, B., Guerrero, M., Knight, W., Ulnicane, I., 2019. Beyond research ethics: dialogues in neuro-ICT research. Front. Hum. Neurosci. 13, 105.

Stahl, B.C., Coeckelbergh, M., 2016. Ethics of healthcare robotics: Towards responsible research and innovation. Robot. Auton. Syst. 86, 152–161.

Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Res. Policy 42, 1568–1580.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. New Jersey, USA: IEEE.

Tonkinwise, C., 2004. Ethics by design, or the ethos of things. Des. Philos. Pap. 2, 129–144.

van Rijn, A., Vandenbroucke, R., 2017. Guide to e-Infrastructure Requirements for European Research Infrastructures.

Versweyveld, L., 2019. Over 700 million euro of budget allocated to e-Infrastructure related projects in H2020 - close to 600 million euro to HPC related projects. E-IRG News Blog.

Viberg, J., Hansson, M.G., Langenskiöld, S., Segerdahl, P., 2014. Incidental findings: the time is not yet ripe for a policy for biobanks. Eur. J. Hum. Genet. 22, 437–441.

Wolf, S., Clayton, E.W., Lawrenz, F., 2018. The Past, Present, and Future of Informed Consent in Research and Translational Medicine. J. Law. Med. Ethics 46, 7–11.

Yang, F., Chien, A.A., 2016. Scaling Supercomputing with Stranded Power: Costs and Capabilities. Univ. Chic. Tech Rep.

“I APPROVED IT…AND I'LL DO IT AGAIN”: ROBOTIC POLICING AND ITS POTENTIAL FOR INCREASING EXCESSIVE FORCE

Raphael D. Jackson

Interamerican University School of Law (Puerto Rico)

raphael.jackson@juris.inter.edu

ABSTRACT

July 7th, 2016 marked the first time in U.S. history where a robot was intentionally used by a Police Department, to kill a human being. The human subject in this case was Micah Xavier Johnson. Johnson was an African American male and Afghan War veteran. Johnson fatally shot five officers and wounded several others before being wounded by police gun fire then being cornered into a standoff. During the standoff, the Dallas Police Department deployed a bomb diffusing robot which was outfitted with a pound of C-4 explosives. Johnson was killed instantly in the resulting blast. Many commentators indicate that detonating a pound of C-4 on a cornered and wounded shooting suspect was a use of excessive force. Ironically Johnson was alleged to have targeted police in retaliation to incidents of lethal force which is disproportionately used against African Americans. For decades Police Department across the United States have been receiving surplus military equipment from the Department of Defense. These transfer programs have been the source of debate among politicians and policy makers, as this new rush to militarization has the potential to change the civilian peace keeping mission of community law enforcement. Among the technologies that police are receiving are military grade robots. Studies on killing and human psychology have examined the act of killing. Studies have demonstrated that despite training killing, other human beings, is the one act that human beings have the strongest aversion in carrying out. Studies conducted by military psychologists, have placed the willingness of human beings to kill, on points of a distance spectrum. The furthest point in the range spectrum is maximum range. Maximum range is defined as “a range in which the killer is unable to perceive his individual victims without using some form of mechanical assistance. The maximum range involves up close killing in which the killer can personally sense his target. The process of killing is facilitated in proportion to the distance that the killer maximizes between himself and his target. Another enabler in the killing process is the compartmentalization of the killing process though the means of group absolution. In short, the more technical and specialized the role he has in performing his task, the less inhibited he is about following through with it. Thus, a killer is less likely to kill someone with his bare hands than to thrust a knife; he is less likely to thrust a knife than to throw a spear; and he is less likely to throw a spear than to squeeze a trigger. This increase in willingness to kill, co-relates with the level of physical distance and mechanical complexity the would-be-killer can place between him and his target. In this respect, Robots present a unique challenge to civilian law enforcement agencies. By design, robots are created to automatize tasks that human beings are unable or unwilling to perform. By reducing the killing process to pressing a key designed to activate a pre-programmed killing machine, you have significantly increased the likelihood of the human controller to use deadly force. As technology rapidly develops in the robot field, we are presented with a second problem, which is the problem of A.I. In addition to endemic instances of use of excessive force, Law Enforcement agencies across the nation also are plagued by instances of racial profiling. Racial profiling are generalizations that departments, or officers make about race, when they are conducting their policing duty. By their very nature, the basis of most A.I. technology is to teach machines to process data in terms of generalizations. Inputs made into computers, do not occur independent of the circumstances of the human being who inputs the data. Thus, if a police department has a decades long track record in racially biased policing, an A.I. system will simply learn to further accomplish this trend, with more efficiency. A.I. chat machines employed by private companies have displayed their tendency to 'learn' racist, sexist, and xenophobic dialogue and a A.I. robot would not be immune from this tendency. An excellent use of robotic policing, and A.I. however, would be data collection. Although data collections against this might be carried out in ways which respect the fourth amendment, what I am speaking to is data collection of police practices. There is a sparsity of uniformly available raw data as it pertains to policing practices. Many police departments collect data via-body cams, but the challenge lies in who ultimately has authority over the footage of the body cams. Police may have issues with what appears to be a big brother type scenario in which their everyday moves are monitored and subject to scrutiny by superiors. While it can be argued that this is what many departments subject civilians to on a routine basis, a more compelling argument would be to set up a triple tier method of footage release. The first form of footage release would be for departmental debriefing or personal training purposes. The second form of release would be in the event of allegations of misconduct. In such an event, the video would be accessible by civilian oversight agencies. The third tier would be a voluntary release in which an officer may want to release a surveillance file for investigation or community relations. By analyzing the 2016 Dallas Shooting incident, this research explores whether the Dallas Police Department has committed an isolated incident, or whether DPD has set a trend for the dehumanization of policing across the handling of the Dallas shooting could be used as a training exercise in how not to utilize military-robotic technology available to police departments. Or it could serve as a harbinger of things to come. Either way it serves as:1) A platform to study the ethical considerations at stake; 2) A case study in increased use of excessive force by the robotization of policing; or 3) A template as to what rules and regulations need to be put into place. A detailed analysis could help researchers successfully integrate this technology. Successful integration would be in the interests of promoting a law enforcement model which will serve the interests of humanity as opposed to brutality.

KEYWORD: Robot, Police Militarization, Racial Bias and Robotic Police, Excessive Force.

1. INTRODUCTION

According to a 2014 Pew Research Center study, less than 10% of the population surveyed believes that police do an excellent job in using the appropriate amount of force for each situation; and less than 10% believed that police officers treated racial and ethnic groups equally. These opinions reflect real statistics conducted in studies which indicates that after controlling for crime rates by race, an African American male is 3.49 times more likely to be shot by police, than an unarmed white man. (Ross, 2015)A surge of police killings of unarmed African Americans in the early to mid 2000's led to protests across cities in the U.S. The protests in Ferguson, Missouri were among some of the most noteworthy. In the wake of such protests and what many considered to be an excessive militarized response by local police, president Barack Obama issued executive order (EO 13688) which created a federal oversight board which implemented protocols around military weapons procurement by local law enforcement. This order would be rescinded by president Donald Trump, effectively removing any restrictions of the transfer and oversight of military equipment. In 2017, ranking member Adam Smith (D-WA) and Readiness Subcommittee Ranking Member Madeleine Bordallo (D-GA) called for a temporary suspension of the 1033 transfers program which provides excess military equipment to civilian law enforcement agencies, but as these calls are being debated the militarization of local civilian police forces continues. Among the technologies that local civilian police forces are receiving through the 1033 transfer program, are military grade robots. These robots are typically used to dispose of explosive and hazardous materials. But an incident in Dallas, Texas provides us with the first instance in which a bomb disposal robot was used to dispatch lethal force against an armed civilian suspect. Following the patterns of the co-relations between deadly force and race, the victim in this case was an armed African American male, who was purportedly targeting police officers in a counter campaign against the extra judicial killings of unarmed African American males. Psychological studies examine the killing process as points in a distance spectrum. The furthest point in the range spectrum is maximum range. Maximum range is defined as "a range in which the killer is unable to perceive his individual victims without using some form of mechanical assistance. This mechanical assistance may be in the form of binoculars, radar, periscope, or remote TV camera. (Grossman, 1996) Unmanned Aerial Vehicles (Drones), bomb disposing robots, and the mechanical two-way communication devices attached to police vehicles all employ the device of remote cameras. When a police officer interacts with the public by means of a remote camera this officer is at the furthest range of the distance spectrum and thus more likely to utilize deadly force. Human controlled robots are extensions of the human beings that operate them. As such they are subject to the same prejudices and sensibilities of those who control them. Thus, if the Dallas police department has a history of using deadly force against African Americans, this will not be remedied by introducing more technology into their policing arsenal. Robots can be programmed to operate independently of direct human control. One of the problems with any system of artificial intelligence (AI) is that without any inputs as to their function of programming, at its basic level AI is taught to engage in generalizations. (Masri, 2019) Generalizations are precisely the form of profiling which police departments, who are engaged in reform, are attempting to avoid.

2. MILITARIZATION AND MINORITY COMMUNITIES

Police robotization is a direct by product of police militarization. Police militarization began with the creation of SWAT teams. SWAT teams were formulated to quell the race rebellion in U.S. urban areas, particularly Watts in 1965. The first SWAT operation was in 1969 against the Los Angeles office of the Black Panther Party. The renewal of this practice continued with the 'war on drugs' of the 1970's and 80's. By the 1990's the police militarization was entrenched into American policing. The Defense Supply agency. Known as DLA disposition services, overseas the disposal of surplus military equipment and weaponry. State and local governments receive the bulk of such weaponry through the Law Enforcement Support Office (LESO). The LESO is responsible for the administration of 10 U.S.C. §257a. and this transfer is overseen by the 1033 Program. Military grade firearms and munitions are approved within these transfers. DLA Disposition Services estimates that since 1990, more than $4.2 billion worth of property has been transferred to state and local law enforcement agencies. With the creation of the department of homeland security more small towns, with populations of 5,000 and under, were given grants to create SWAT teams to fight the ‘war on terror’. In 2011, the Center for Investigative Reporting (“CIR”) conducted a report on the DHS grants and found that since its inception, the DHS has provided civilian law enforcement with grants of $34 billion. (Doherty, 2016)

The American public became fully aware of the extent of police militarization in 2014 during the protests in Ferguson, Missouri. Civilians, protesting police brutality against people of color, were met by camouflage and body armor-clad police officers, wearing gas masks, wielding M-16 A2 rifles, and sitting atop Mine Resistance Armored Vehicles. Studies indicate that the presence of paramilitary units within police departments has changed police culture drastically. Furthermore, as opposed to merely stockpiling military gear for emergency sake, police departments are incentivized to use such military equipment. Receipt of additional funds or equipment is contingent on demonstrating that the police made use of such equipment within one calendar year. (Coscarelli, 2014) Police department typically deploy their military hardware during standard police raids. According to a 2014 ACLU report, a disproportionate number of the raids were aimed at minority communities (42 percent African American, 12 percent Latino). (ACLU, 2014) Militarization of police is bad both for law enforcement and the public. When police officers are protected by a layer of insulation from the public and given the capacity to produce a military response in the face of civil protest, the officers tend to “feel more powerful, more invincible, more militaristic, and ready to attack. Conversely when a civilian sees this force they respond in-kind with fight or flight, and it elicits a response from observers that “hey this is war.” (ACLU, 2014)

Through the use of gas masks or identity concealing masks, police officers create what psychologists refer to as deindividualization, which is an immersion in a group to the point that one loses a sense of self awareness and feels lessened responsibility for one’s actions. (Grinnel, 2015) Specialized military units implement these mechanisms to ensure an increased likelihood that a combatant will use deadly force on an enemy combatant. When employed by officers, these same factors may also increase the likelihood of a police officer to use excessive or even lethal force on civilians.

2.1. Remote Killing

Killing a fellow human being is one of the strongest aversions that exists in the human psyche (Cushman, 2012) Because this is the primary function of combat soldiers, much of military psychology has been dedicated towards decreasing, ideally eliminating, this natural aversion among its combat personnel. Despite decades of these anti-killing aversion campaigns, soldiers and combat personnel polled point o several combat situations in which combat personnel invariably chose to engage in riskier life-threatening tasks rather than resort to killing another being. One of the classical methods employed to overcome the aversion is to dehumanize of the enemy. Another method is to increase the distance spectrum between the killer and his target. There is a positive co-relation between the length between a would be killer and his victim, and his willingness to kill the victim. In combat, bomber crews, and artillerymen can engage in countless campaigns without the same amount of perception as an infantryman fighting in close combat. The above-mentioned combat personnel can perceive their victims not as people but as buildings, facilities, and coordinates, without giving much thought to the people within.

Among the three factors that facilitate the act of killing are group absolution, mechanical distance, and physical distance. (Grossman, 1996)

2.2. Group absolution

Aside from peer pressure, playing one of several roles in the killing process increases the likelihood you will kill. This group absolution is demonstrated in bombing crews. In bombing crews, the pilot, navigator, weather reconnaissance person, and gunner all have their role. This phenomenon is also present in an artillery team, machine gunner team, and sniper team. Sniper teams go forth in teams of two. A spotter chose the target and the sniper fires on the target. This way both the sniper and the spotter have a level of absolution from the act.

2.3. Mechanical Distance

In combat, mechanical distance are traditionally represented, by binoculars, or a rifle scope. In modern warfare mechanical distance is also represented by video screen. The layer of mechanical separation between the viewer decreases the inhibitions a viewer may have, acting upon or witnessing the fate of the subject in view. A comparative analogy can be found in voyeurism. Even if they can go undetected, many would not have the audacity to physically peep on their neighbors engaging in intimate acts. The same people would have less reservations of watching similar acts of voyeurism if recorded or streamed into their personal desktops.

2.4. Physical Distance

Long range refers to a distance to which the average soldier can see the enemy but cannot kill him without a specialized form of weaponry-Sniper rifle, anti-armor missile, or tank fire. (Grossman, 1996) During the American Civil War, a soldier armed with a rifled musket, was able to increase his combat range from 50 to 350 yards, drastically increasing the killing range of your average combatant. (McCaul, 2019) In each subsequent war the physical distance gap has been closed by newer technology in weaponry.

2.5. Robotic killing of civilians by police: an analysis of the Micah Johnson killing

Micah Johnson was a 26-year old member of the Army Reserves, and an Afghan war veteran. Like many citizens Johnson was angered at the most recent wave of extrajudicial killings of unarmed black men by police. On the evening of July 7th, 2016, the Dallas community marched in protest of police killings of unarmed black men. Although unaffiliated with the marches, Johnson shared their grievances, if not their methodology. Johnson thus, chose that day and that venue to launch his attack on police officers. Twelve years earlier in 2004 Congress refused to extend the Violent Crime Control and Law Enforcement Act (Flexner, 2017). The act would have imposed a ten-year ban on the civilian use of military grade weapons. However, the act was not in force during the time that Johnson, like other mass shooters before him, acquired his arsenal. As per protocol, the Dallas police were dispatched to monitor and direct the march routes and provide protection to the marchers. As the protest march neared El Centro Community College, Johnson, in full body armor, parked his SUV, and casually spoke with one officer before cutting down three officers and injuring two civilians with his AK-47 semiautomatic rifle. (Wanebo, 2018) Johnson's tactics included the infantry urban warfare technique of shot and move, which confused officers into thinking there were multiple shooters. Eventually Johnson gained entrance into the community college, but not before being shot and wounded by police. The wound left a trail of blood from the entrance through the library. Police followed the trail until they cornered Johnson at the end of a hallway and had him locked in a standoff. Ultimately the standoff ended when Dallas Police decided to retrofit a Mark V-A1 bomb disposal robot with a bomb. (Sankar, 2018) The police department attached one pound of C-4 explosives to its robotic claw and detonating it after remotely maneuvering the robot to his position. The Dallas police, in their lethal use of a robot to kill gunman Micah Johnson, exhibited all three maximizing layers of the distance spectrum which facilitated their killing. The first layer was their group absolution. A team is used to operate bomb diffusing robots and each team member can monitor and command the operator from his position. The second layer was the physical distance from the robot. The robot used was a bomb diffusing robot that is designed to be operated remotely and from far off distances. With the full knowledge that the robot would be used to detonate, as opposed to diffuse, an entire pound of C-4, the police operated the robot from a control center and drove it to their human target in another building. The third layer of insulation was the mechanical appendage of the robot itself. The final level of insulation was the remote camera. The operators observe the action through a screen. Johnson was an Afghan war veteran who had undergone a personal regimen of special forces training on his return. In contrast Dallas police, like most police departments in the U.S. are only trained for self defense. In the lead up to the explosion the police had cornered and managed to wound Johnson. Even though Johnson stated he was targeting police officers, police officers had none the less, in accordance with their training, protected the civilians inside the building as well. With the shooter trapped and wounded, whatever reduced risk he posed to police officers was diminished by the time they made the decision to end the standoff by using the robot. The Mark V-A1 robot is utilized by the U.S. Army, Israeli Defense Forces, and the Uruguayan Army Bomb Squad. Its use by military operators in the combat theatre is defensive, thus it is ironic that among the first offensive usage of the robot was committed by a civilian police department, against a civilian. The Dallas police could have stuck to their defensive mission by retrofitting the robot with non-lethal tools including surveillance tools. Among the defensive arsenal the robot could have been fitted with, are stun grenades, flash bang grenades, CS gas or 3-methelfentanyl gas, the sleep gas which was used to end the 2002 Moscow theatre crisis. By choosing C-4 the Dallas police, not only resorted to lethality as a first resort, they resorted to the most dangerous and lethal option possible. Near instance approval of this choice set a dangerous precedent for future police encounters.

2.6. How lethal robots can change the police mission from self-defense to combat

As mentioned in the beginning section, much of the high-tech equipment including UAVs, and bomb disposal robots, are inherited from the Department of Defense. Unlike police officers, soldiers have little expectations that their training is in preparation for killing enemies. The militarization of police forces across the nation have led to an increase in the perception among police, that the public which they serve and protect are their potential enemies. Robots, even of the non-lethal variety, complicate this mission by providing police with several layers of distance when dealing with the public. Ruben Brewer, a senior robotics researcher at the nonprofit SRI International in Menlo Park, California posed a solution for routine traffic stops, called the Go-Between. The Go-Between robot is advertised as a device which will allow police officers to issue citations from the comfort and safety of their vehicle. (Post, 2019) The robot is attached to the officer’s vehicle by an extendable aluminum pole which extends to the driver’s window. This allows the officer to communicate via video screen. The device has a driver’s license scanner, two-way communications radio, and a ticket printer. The device has no offensive capability, but it has the capability of placing a spike trap under the motorist's car to disable the vehicle in the event the motorists attempts to drive away before the officer issues the ticket. The Go-Between is advertised as a device that can reduce the violent assaults between motorists and police. A 2001 study in the Journal of Criminal Science indicates that homicides and assaults during routine traffic stops are infrequent. (Lichtenberg, 2001) Studies of the sort reveal the gap between perception of the danger that police face on a day to day basis, and the actual numbers. It also reveals how willing companies and agencies are to implement expensive and excessive protective measures which only minimally increase police safety and only at the expense of the safety police are charged with protecting. Technologies like the Go-between are likely to have a negligible effect in increasing officer safety, and only at the price of dehumanizing routine traffic stops. While the Go-between technology is defensive and non-lethal, the Micah Johnson incident has demonstrated how little time and consideration it takes for a police department to rig a non-lethal robot for lethality thus increasing the officer's likelihood in use force. However, with every technological advancement in remote control robotics there are disadvantages which spring forth from over reliance on the technology, this is certainly the case when it comes to Artificial Intelligence.

3. RACIAL PROFILING AND AI

Human controlled robots are extensions of the human beings that operate them. As such, they are subject to the same prejudices and sensibilities of those who control them. Thus, if the Dallas Police department has a history of using deadly force against African Americans, this will not be changed by simply introducing more technology.

Robots can be programmed to operate independently of direct human control. One of the problems with any system of artificial intelligence is that without being given inputs towards function of programming, at the basic level AI is taught to engage in generalizations. (Masri, 2019) Generalizations are precisely the form of profiling which reform-seeking police departments are attempting to avoid. An example of how AI is capable of exacerbating racial profiling problems can be found with Microsoft Tay Tweets. Microsoft created an artificial intelligence application called Tay Tweets. Tay Tweets was an AI chatbot that was capable of commenting on images and telling jokes, based on the aggregation of social media feeds available on the internet. Microsoft had to shut down this project soon after it began, due to Tay’s inability to recognize the offensiveness and racism of its comments. Tay AI, which retweeted comments such as *“GAS THE KIKES RACE WAR NOW!” “Hitler did nothing wrong* "and *“Mexicans and Blacks are the worst race.”* Tay however was a product of her own inputs and social media commentary. An AI robot employed by the police forces would likely also be a product of the departments open inputs and departmental policies. The danger in AI gone awry is that rather than the mere publication of offensive words in a chat room screen, these robots would be charged with the executable actions of law enforcement. Thus, an AI robot employed by police is guaranteed to reinforce pre-existing policing practices, which is not necessarily good, considering that the function of AI is to make sweeping generalizations. While conducting research for MIT research labs, Joy Buolamwini, head of the Algorithmic Justice League, discovered that facial imaging software could not identify her face until she wore a white mask. The same state of the art software could not recognize the faces of Serena Williams, Michelle Obama, and Oprah Winfrey. Furthermore, the software, which could determine sex, classified the three aforementioned ladies as males. (Buolamwini, 2019). AI imaging software which cannot predict gender on facial input on darker subjects, pose serious problems for Law Enforcement Officers trying to reform their departments to eliminate racial biases. This problem is known as biased datasets. (Murray, 2019) For instance, if you feed an AI database thousands of mug shots, the data base will pick up on the skin tones and hair textures of the arrestees and may be inadvertently programmed to seek out those who match this color profile for extra scrutiny. Thus, a societal discrepancy in incarceration of minorities will become the means by which AI continues this discrimination in a more efficient basis. Microsoft showed that its AI would quickly spiral out into the current culture of those who program it. A simple chat box employed by Microsoft quickly transformed into a racist sexist neo-Nazi after following the inputs of the users. Police culture similarly would not necessarily change with the introduction of a robot, any more than it has changed with the introduction of body cams. The technology would need to have some form of civilian oversight to be truly neutral.

4. UAV CAMERA SURVEILLANCE AND 4TH AMENDMENT CONSTITUTIONAL ISSUES

Another tool that law enforcement has inherited from the military are UAV’s or Unmanned Aerial Vehicles, which are often referred to as drones. Drones are becoming increasingly smaller and inexpensive. One of the popular armed drones utilized by police is the Shadow Hawk, this drone resembles a small helicopter, operates for 3 hours at a time, and is operated via mobile computer control. (Thresher, 2017)Shadow Hawk can carry high resolution cameras as well as shotguns and grenade launchers. The police primarily use this drone for surveillance. Many drones are equipped with high resolution cameras which can record and live stream video from low and high altitudes. A drone can be programmed to operate on a flight plan while its imaging capturing capability can be downloaded to a police database, much like a mobile security camera. Hillary B. Farber articulated the concerns of police drone use this way: Drones can provide police with the details of a person's daily routine, easily allowing them to create a profile of the person's associations, religious affiliation, health conditions, professional and recreational activities, and family and economic status. When all this information concerning hundreds, if not thousands, of people can be gathered from thousands of feet in the sky, it is hard to resist the claim that society has succumbed to an Orwellian vision far beyond George Orwell's imagination. (Thresher, 2017) Likewise, however, a data gathering drone can be put to positive use by simultaneously monitoring the agencies which have employed its use for law enforcement purposes.

5. POTENTIAL BENEFITS OF DATA

Automation is not without benefits. Automated image capturing can assist in statistical gathering and reviewability by non-government actors such as civil liberties groups. As cameras have become more portable and pervasive it has become more common place for citizens to record their police encounters as well as the police encounters of third parties. The First, Seventh, Eleventh, and Ninth Circuits have all held that the right to photograph police officers in the performance of their duties is protected under the First Amendment. (Raoul, 2017) Many individual officers may bristle at the idea of being recorded and often may state or misrepresent state wiretapping laws in their effort to discourage the activity. Part of this reluctance could be attributed to the fact that unlike police body cams, the police have no control or access over the footage being recorded. A third-party monitoring recorded interaction and providing access to all soliciting parties, might help both officers and civilians maintain a balance between fulfilling law enforcement duties and protecting civil liberties. Police departments in cities like Boston, have seen an overall improvement in solving homicides by hiring a civilian data analyst. (Meuller, 2017) A key to the process is that civilian oversight ensures that the employees work with police without succumbing to the general police culture and chain of command which discourages independent inquiry.

6. CIVILIAN OVERSIGHT SOLUTIONS

Police are public civil servants thus there are very few reasons why the information they gather should not be available to the public they serve. Agencies and lawmakers can address privacy concerns by successfully employing data management techniques to identify and preserve critical video evidence, and allow non-critical video to be deleted under data-retention policies. (Lin, 2016) Although Rep. G.A. Hardaway and Senator Sara Kyle have introduced bills in the Tennessee state legislature that would make it a felony for officers to intentionally turn off their body cameras to obstruct justice. (Griggs, 2019) as it stands now most police departments allow officers to turn off body cameras at their discretion. Furthermore, some police departments are not obliged to release footage. Cameras that are be programmed to activate, once a police officer turns on a siren, or once he unholsters his sidearm or taser would be useful in documenting emergency situations. This footage should be made available to civilian agencies charged with monitoring police conduct. The most frequent interaction citizens have with police are traffic stops. Over 20 million motorists are pulled over every year, yet only 10 states require police to log the race of the motorists. (Lab, 2019)Police robots could make this information readily available to citizens, police department leadership, and community relations leaders. Raw data can be instantly uploaded to a public database and after making the necessary personal edits for the sake of privacy, the general statistics can be made available to police community action organizations and local community advocacy groups. This would, more than simply employing the technology, allow community members to get a good glance at policing practices, empathize with police concerns, as well as the statistical occurrences of such encounters. For instance, if civilian officials, legal and social rights activists, and elected officials had ready access to the informatics, and even programming of such robots then the police officers would be further encouraged to take a proactive role in responding to the needs of public inquiry. Now that the Police Officers will also be protected by the technology's implementations, the new focus can be on proper policing. These videos and the information collected could be used as a training tool. Without this civilian community oversight, the technology won’t have a similar effect. A.I. police robots can handle the balance between open records requests and privacy rights by means of data management. As early as 2008, Google Inc., made us of an algorithm with scans the image bank on Google Maps street view, then blurs the faces of pedestrians, this technology can be readily deployed on Computer Robots which capture images for police stops on public roads.

Despite them being public servants, concerns that police may have about over monitoring of their work are not entirely invalid. These concerns can be ameliorated by a tiered system of footage release. The First tier could be personal or departmental. Apple has released a wristwatch which can monitor heart rate levels. (Arnow, 2016). This technology can be further upgraded to activate body cam recording when the wearers heart level reaches a rate which indicates a spike in adrenaline. This way at the end of the day the officer can keep a personal log of transactions which triggered an increased in his heart rate. The officer can then debrief by reviewing such interactions and safely learn to distinguish which situations were truly dangerous from those from which he was reacting out of pre-conceived ideas. The second tier is video and audio data which is subpoenaed in response to a complaint or allegation of misconduct. The third could be voluntary submissions. Outside the normal process of courtroom discovery, many police departments do not require that body-cam wearing officers record and hand in their recordings. A voluntary submission could be used for two purposes. The first could be investigatory and the second could be for community relation purposes. An officer might want to keep a log of his charitable or community building acts. The fact that this self-reporting option may be in the officer's personal interests does not negate the fact that the officer's requisite acts (for the submission) are also beneficial to the community. The frequent sight of police officers volunteering to change tires, buy an ice cream cone for a child, or help an elderly pedestrian cross the street can help re-enforce a sense of trust between police departments and communities.

7. CONCLUSION

Robotic policing increases dehumanization through automation and increasing psychological distance. This intersection leads to an increased likelihood that police will use excessive force. The phenomena will have devastating impacts on and negative policing in African American and minority communities in the U.S. In the Micah Johnson incident is a case study in which, mechanical distance, group absolution, and physical distance all converged. This convergence facilitated police deployment of fatal and excessive force against a cornered wounded shooting suspect. Robot technology is becoming more accessible and affordable to maintain. As local police forces continue to acquire military grade weaponry and technology the usage of robots in policing will become more and more common. Notwithstanding the 4th amendment privacy issues triggered by unauthorized data gathering, it is wise for local and national governments to refrain from granting civilian police forces lethal military weaponry for use against the civilians they are supposed to serve and protect. Reports indicate that police departments already suffer from deployment of excessive force, which is exacerbated by the 1033 transfer programs. These transfer programs place excess military equipment in the hands of civilian police forces, the programs incentivize unnecessary usage of the equipment. In order to receive future transfers, police departments must demonstrate that they used the equipment from the previous year. Further studies also indicate that there is a negative correlation between the proximity of an actor, its target, and the actor's willingness to use deadly force against its target. Camera-subject interaction by means of remote-control robots increase the physical and psychological distance between the subject and the actor. This leads to a process of dehumanization of policing, which in turn increases the likelihood that an officer will use deadly force. The research also explored the double edge sword that camera recording raises concerning data gathering and 4th Amendment protection. It remains to be seen whether Dallas Police department will remain an anomaly in the case studies of police abuse of robotics, or whether it is a harbinger for things to come. Like other work, police work will also be subject to various forms of robotization over the years. If the technology of surveillance and data gathering is employed equally and used to monitor police practices as well the practice will be a check and balance to policing practices.

Creation of a civilian oversight program will be key to maintaining a unbiased analysis of aggregated data. Such an oversight program, if properly implemented, can be of great benefit to both civilian human rights activists as well as law enforcement agencies.

REFERENCES

ACLU. (2014, June). Retrieved from War Comes Home: The Excessive Militarization of American Policing: https://www.aclu.org/sites/default/files/assets/jus14-warcomeshome-report-web-rel1.pdf.

Arnow, G. (2016). Apple watch-ing you: Why wearable technology should be federally regulated. *Loyola of Los Angeles Law Review*.

Buolamwini, J. (2019, Oct. 30). *Artificial Intelligence has Problem with Gender and Racial Bias. Here's How to Solve It*. Retrieved from Time: http://time.com/5520558/articifial-intelligence-racial-gender-bias/

Coscarelli, J. (2014). *Why Cops in Ferguson Look Like Soldiers: The Insane Militarization of America's Police*. Retrieved from http://nymag.com/daily/intelligencer/2014/08/insane-militarization-police-ferguson.html.

Cushman, F. G. (2012). Simulating murder. The aversion of harmful action. *Emotion*, 2-7.

Doherty, J. B. (2016). Us vs. Them: The Militarization of American Law Enforcement and the Psychological Effect on Police Officers & Civilians. *California Interdisciplinary Law Journal* .

Flexner, D. (2017). Why The Civilian Purchase, Use, And Sale of Assault Rifles and Pistols, Along with Large Capacity Magazines, Should be Banned. *New York University Journal of Legislation and Public Policy*, 593.

Griggs, B. (2019, October 30). *A proposed Tennessee law would make it a felony for police officers to disable their body cams*. Retrieved from Tennessee Body Cam Felony: http://edition.cnn.com/2019/02/27/us/tennessee-body-cam-felony-trnd/index.html

Grinnel, R. (2015, November 11). *Deindividualization*. Retrieved from Psychological Center: http://psychcentral.com/encyclopedia/2008/deindividuation/

Grossman, D. (1996). *On Killing: The Psychological Cost of Learning to Kill in War.* Boston: Brown.

Lab, S. C. (2019, Oct 30). *Findings The result of our nationwide analysis of traffic stops and searches*. Retrieved from https://openpolicing.stanford.edu/findings/

Lichtenberg, I. &. (2001). How Dangerous are routine police-citizen traffic stops? *Journal of Criminal Justice*, 419-428.

Lin, R. (2016). Police Body Worn Cameras and Privacy: Retaining Benefits While Reducing Public Concerns. *Duke Law and Technology Review*.

Masri, A. (2019). *Towards Data Science*. Retrieved from Those Racist Robots: https://towardsdatascience.com/those-racist-robots-c31306d6627f

McCaul, E. J. (2019). *If You Can Be Seen, You Can Be Killed: The Technological Increase in Killing Zone during the American Civil War.* Leiden: Brill.

Meuller, B. (2017, August 15). Police Add Civilians in Bid to better Analyze Crime Data. *The New York Times* . New York, NY.

Murray, J. (2019, Nov 17). *Racist Data? Human Bias is Infecting AI Development* . Retrieved from Towards Data Science: https:///towardsdatascience.com/racist-data-human-bias-is-infecting-ai-development-8110c1ec50c

Post, W. (2019, 5 14). *GoBetween*. Retrieved from https://www.washingtonpost.com/ technology/2019/05/14/one-solution-keeping-traffic-stops-turning-violent-robot-that-separates-police-officers-drivers/

Raoul, S. (2017). Cop-Watch: An analysis of the RIght to Record Police Activity . *Hamline Journal of Public Law & Policy*, 215.

Sankar, V. (2018). What Happens When Police Robots Violate the Constitution: Revisiting the Qualified Immunity Standard for Excessive Force Litigation under Sec. 1983 regarding Violations Perpetrated by Robots. *Vanderbuilt Journal of Entertainment & Technology Law*, 947.

Thresher, I. (2017). Can Armed Drones Halt the Trend of Increasing Police Militarization. *Notre Dame Journal of Legal Ethics & Public Policy*, 455.

Wanebo, T. (2018). Remote killing and the Fourth Amendment: Updating Constitutional Law to Address Expanded Police Lethality in the Robotic Age. *UCLA Law Review*, 976.

MOBILE APPLICATIONS AND ASSISTIVE TECHNOLOGY: FINDINGS FROM A LOCAL STUDY

Kelly Gaspar, Isabel Alvarez

Universidade Autónoma de Lisboa (Portugal)

30001583@students.ual.pt; ialvarez@autonoma.pt

ABSTRACT

This paper reviews the study of mobile applications for disabled people, considering the fact that these sort of mobile devices present great potential for the social inclusion of these users as well as help them in their daily tasks. However, most of the existing applications have few support functionalities or a low range of interaction, as in the development of these applications the special needs and specific capacities for the disabled have not been considered (Visagi et al, 2019). The present study suggests strategies and solutions to be used for an overall accessibility.

KEYWORDS: Assistive technology, Mobile devices, Disabilities, Mobile applications.

1. INTRODUCTION

All individuals whether with a disability or not have a range of rights that must be respected. One society susceptible to variety, researches its isolating instruments and finds out new tracks for the inclusion of a disabled person. This has awakened and stimulated new researches, including the adaptation of the technological advances available today.

However, what embarrasses great part of the disabled people is the dependency from others to do some activities but the development of information and communication technologies enables several ways of relationship with knowledge, as well as with the most recent conceptions and possibilities; the relevance of this paper stresses the importance of assistive technologies as a tool to provide a greater independence, quality of life and social inclusion to the disabled, through the amplification of his/her communication, self control, human motricity and competence in the execution of physical tasks.

Today, with the growing flexibility of objectivity and subjectivity facing the most scientific and hard technologies, the engineer needs other capacities. The emerging of information technology in the work place caused that all technicians became a link among the most diverse sectors of the productive chain and the society. The actual engineer is no longer only a professional technicist as before, it is in fact a qualified human being for the flexibility demanded by the society and required for a more open market (Laudares & Ribeiro,2000).

The most recent international treaties have demonstrated the desire to build a society that not only recognizes the difference as an unquestionable human value but also promotes conditions for the full development of the potentialities of every one in its uniqueness (CIBEC/MEC,2010).

The global study from UNESCO reveals that technologies have a positive influence in several perspectives of the disabled people's lives (Mohammadi, Momayez, & Rahbar, 2014). According to Domingo (2012) and Emam (2017) information systems aim to offer disabled people the support they need to attain an admissible quality of life allowing them to participate in the economic and social environment.

Assistive technologies are related with the capability of causing extreme technological changes that transform humanity and its culture and have the potential and tendency to generate a quick cycle of development and create derived technologies applied virtually to all areas of knowledge in order to benefit the increase of human performance, its processes and products, quality of life and social justice.

Several are the possible solutions to be approached in order to meet the adaptation problems of people with disabilities, but in what refers to the development of the software or hardware devices, its success is measured through the level of user satisfaction. In this connection any developper must take into account what type of solution must be given and if it solves the problem presented by the disabled person, or at least, if it solves the gaps of greater relevance presented, as in some cases as referred by Wong et al.(2009) the use of some assistive technologies may not be appropriate to certain individuals with severe or profound disabilities. Both the level of difficulty and the support requirements and level of adaptation need to be considered in order for the disabled person may to use the technological artifacts in a significative way (Redford, 2019). Actually, quick alterations to technology became an efficient tool for development in the individual, community, national and global perspectives (Islam, Ashraf, Rahman, & Hasan, 2015).

This paper emphasizes the gap among theory, speech and practice in the area of computer ethics. The choice of this sector is due to the growing observations referring the potential of information and communication technologies to help people with disabilities to overcome their limitations. The growing need for the development of new technological solutions is obvious. The quick development of the information and communication technologies brought the hope that, in the near future, this area of research and development may provide viable solutions.

On the other side, topics like ethics and social responsibility have emerged specially as how the implementation of these technologies should be made near the groups of vulnerable users (Ienca et al, 2018).

However, although they are in a stage of quick growth in development, assistive technologies are still a devising topic. It is known that it is important to develop solutions that contemplate the inclusion of disabled people, but not many significative advancements have been made in the ellaboration of unified adaptations for people with different disabilities (this also refers to the fact that several people may have the same pathology but in diferent degrees).

Assistive technology helps individuals with disabilities to reach more autonomy and more independence, considering that the resources and services involved in this concept aim to facilitate the development of daily tasks for this kind of people. Further more, it is an important tool for the so called social inclusion.

An in-depth review of full text papers concerning the different types of disabilities, assistive technology and existing mobile applications was performed and resulted in the production of a research relating to this topic.

Having this in consideration, the following initial question arose:

"- In what way can assistive technologies contribute to improve the functional capacity of the disabled in the use of mobile devices? "

In this connection, a local study was implemented and which comprised an identification of the main assistive tools for mobile devices, to study their main limitations, and to test those technologies near a pilot group of people with disabilities through surveys in three different phases. The results of this initial study led us to an indepth analysis of a case. A mobile application was chosen and conceptually analysed. The methodology employed in this local study appeared as a valuable strategy for the presentation of a solution aiming to solve or minimize the main gaps referred in this project, as a prototype of an application for mobile devices identifying a solution that may comprise the limitations found.

2. ASSISTIVE TECHNOLOGY (AT)

Assistive Technology models have been defined with the aim to guide the study and development of this technology. Bearing in mind that to develop a technological artefact several are the points that have to be considered, in the specific case of AT, factors like culture, disability and activity to be accomplished need to be taken into account. There are several models, crossing knowledge from several different areas.

2.1. Horizontal European Activities in Rehabilitation Technology (HEART)

The main purpose of this model is to study solutions, devices, methodologies, etc. that balance or lessen limitations not only of the individual but also of the physical and social environment. It can be subdivided in three components: technical, human and social.

2.2 Human Activity Assistive Technology (HAAT)

The HAAT model proposed by Cook and Hussey (1995) is based on a *framework*, and used by engineers and psychologists to examine the functional behaviour and performance of individuals, through the execution of technological activities. It proposes four components: Context – referring to the social and physical environment where the individual and the AT are included; Human – presenting the individual as the main element of the model; Assistive Technology – referring to the externa device used to suppress any contextual constraint; Activities – comprising all the actions of the daily life.

2.3 Telematic Multidisciplinary Assistive Technology Education (TELEMATE) (Turner-Smith & Blake, 1999).

This model determines that specific training should consider the several areas of knowledge.

2.4 Matching person and technology (MPT) (Scherer, 1986)

A mixture of people and technology demanding attention to the several environments where technology will be used, to the needs and preferences of the user and to the functions and resources in technology.

The availability of specialists in AT that understand the value of a process directed to the consumer and adequate to provide proper services is fundamental so that an individual may get a quality evaluation of the needs and most appropriate technologies for personal use (Scherer, 2005[1], 2012).

3 ASSISTIVE MOBILE DEVICES AND SYSTEMS

3.1 Accessible Interfaces

Accessible interfaces are the border between the individual and the product, through which information is exchanged (Cook & Polgar, 2008). The intelligence of the interfaces makes systems to adapt to users, solve their questions, or show integrated and understandable information through the use of several ways of communication.

According to Saci (2005), there are seven principles that sustain Universal Design: To people with different capacities; Flexible use; Simple and intuitive; Easy to understand; Communicates easily the necessary information; Fault tolerant; Requiring few physical effort.

3.2. Assistive Mobile Devices

Assistive Mobile Devices are defined by IDEA (2004) as any item, equipment or system, acquired, modified or personalized, used to increase, maintain or improve the functional resources of people with disabilities.

França, Borges and Sampaio (2005), state that a computational project directed to handicapped does not differ from another project, however "involving other own issues that lack a differentiated human interaction, the use of special tools and the constant care with the user well being". Handicapped people face important challenges in acquiring digital technology, due to cost and availability (Samant Raja, 2016). Several initiatives were developed to promote accessible web content.

To evaluate who has access to mobile technology tends to be a complex task. To evaluate technology availability, accessibility, capacity and accessibility will develop an understanding of who has access or not (Roberts e Hernandez, 2017).

4. DISABILITIES

The World Health Organization, in 1976, defined three international, differentiated and independent classifications: Disability: "in health, disability means any loss or alteration of a structure or psychological, physiological or anatomic function"; Incapacity: "any restriction or lack of capacity (resulting from a disability) to do an activity within the normal limits for a human being"; Handicap: "is a social condition of prejudice suffered by a certain individual, resulting from a disability that limits or prevents the performance of an activity considered normal for that individual, given age, sex and sociocultural factors".

International classification (according to WHO) are fundamental to the development, selection and evaluation of the Assistive Technologies (Glennen & DeCoste, 1996).

4.1. Assistive Applications

A study published by *WebIam* (2008) shows that the number of users with special needs has grown exponentially.

The need of better applications is the result of this research. The study asked the participants to select the most problematic items in list, in order of difficulty, and the result was: *CAPTCHA*: images used to check if the user is human; screens or parts of screens that change unexpectedly; *Links* or buttons that do not make sense; Lack of accessibility in keyboard; Complex or difficult shapes; Images with missing descriptions (alternative text); Missing headings; Many links or browsing items; Complex data tables; Not accessible or missing search functionality; Missing *links* "to go to main content” or “to go out browsing".

4.2. Accessibility tools for mobile devices

The access to the benefits of mobile devices is limited to disabled people as the majority of this technology is projected to younger people that tend to have greater facility in dealing with complex electronic devices. Disabled people find obstacles in dealing with these devices because in many cases they cannot operate controllers, cannot get information about the devices or simply do not understand the functioning. Lately, several mobile options have been suggested as how people with special needs may benefit from services based in ITC. However, many solutions concentrate only in usability and do not attain success, once those products show the idea of disability too much.

5. MATERIAL AND METHDOLOGY

This research comprises, in a first phase, the description of a pilot study, based on questionnaires done to disabled people met in a professional and familiar context, in a religious institution, concerning the assistive technologies used. The same study includes, also, an interview with the same participants and the application of the questionnaires. In order to make this more functional, the study has been divided in five parts that include, respectively: Description of the pilot study and the main study; Characterization of the tested applications; Population and sample; Selection of data collection; Presentation and justification of the data analysis procedures.

5.1. Description of the study

This study comprises two different phases, being one the pilot study and another the main study. The pilot study focused on the application of the case study of this research and also on existent assistive mobile applications working with different operating systems. The main study consisted on submitting questionnaires to the participants and also some interviews.

During the pilot study, the applications were tested during 30 days by part of the group of participants and also the researcher, in order to do an ecological description of the behaviours (learning anatomy) (Damas & De Ketele, 1985), using an observation grid where the results obtained are kept as additional to the results of the main study as it allows to collect in a global way what happens in terms of learning and behaviour changes during the use of AT. Bearing in mind the disparity of the disabled characteristics in relation to the disability, type of device, different systems, there was also the need to use the comparative method (Glaser & Strauss, 1967), even superficially.

To add value to this research, it was decided to also do interviews, in order to obtain relevant data for the analysis of certain observed behaviours. The interviews provided the main utility of showing certain aspects of the subject under study that maybe the researcher would not have considered and, in a certain way, to complement the lines of work suggested during the research or by the observations done.

5.2. Characterization of the tested applications

The following mobile applications tested due to the fact that based on a previous research, it was considered that were the most used applications.

iOS Accessibility – This brand (Apple) is known to take the maximum advantage of its hardware and to be a pioneer in the use of assistive tools. The iOS native assistive application is composed by several tools developed in order to assist several types of disabled to perform the most diverse tasks.

Android Accessibility- The accessibility options can vary according to the device and version of Android. It has a set of native assistive resources/tools and configurations that can be found in any device that allow that anyone with some kind of disability can execute the same actions that a user without that limitation.

Telepatix- Developped by Tix Tecnologia Assistiva, is an application of augmentative and alternative communication (CAA) that allows the writing of phrases and reading in loud voice. It has an optimised keyboard for quick communication with the suggestion of words and phrases. All this can also be used with sweeping and external compatible actuators.

5.3. Participants

Pilot study: The respondent population to this study were people with some disabilities that had had not previous contact with any of the applications tested.

Main study: The respondent population with some degree of disability, were users of some of the tested applications. Suggestions from discussion forums on disabilities and technology were also taken into consideration for this study.

5.4. Data collection

Data collected from the interviews occurred during the months of May and June 2019. Interviews were transmitted to capture the exact speech.

The questionnaires were sent by email, with previous explanation of the objectives of the research and the filling procedures. This occurred during the months April to July 2019.

The selection of these data collection instruments is based on the fact that it guarantees the collection of information on the main points of the research: to specify the purpose of the research and to motivate the participant in a way that he/she could share important issues for this research (Merriam, 1988).

5.5. Data analysis

Questionnaire A : It was difficult to find participants that would not use any assistive technology. So it was decided that users using Android would be submitted to iOS and vice-versa, and both groups tested the application Telepatix. It was intended to collect data that could allow to meet the following purposes: Sex, disability, difficulty in adaptation, understanding integration with the various applications, identify the most used tools for each type of disability and to know the opinion of the participants related to the technologies tested.

Questionnaire B: Done to regular users of assistive technologies adapted for mobile devices, allowed to compare data obtained from questionnaire A. Following Figari (cited by Ribeiro, 2005), it is not only a way of validating data but it also allows the researcher to complete data and even to "decipher", or better, to understand it in terms of its context. It was intended to get the following results: Identify the tools most used for each type of disability, to study the main limitations of the existing technologies, to know other supporting tools not referred in this research, to understand the opinion of the disabled towards assistive technologies.

The content of the interviews was fully reproduced in writing. For a better understanding it was made a categorization not only to understand the general use of the AT but also the specific need of each disability.

5.6. Analysis and discussion of the results of Questionnaire A

It is important to stress that the purpose of this Questionnaire was only attained after three interactions with the participants; the Questionnaire was so divided in three parts: before use, during use, and after use.

Considering the aims for this research in relation to the attitude of people with disabilities in relation to assistive technologies and the importance of technology in what concerns inclusion in the more diverse contexts, it is viable to form certain opinion that will establish the main course of this discussion: Before use: The motives that took the handicapped never to have experimented an AT in the mobile adapted devices (or that specific OS); How frequent do they use mobile devices; For what purposes more use the mobile adapted devices; During use. Level of adaptation difficulty; Which are the greatest difficulties encountered; Which are the major advantages encountered; After use: Which is the level of satisfaction in relation to the tested technology; Which are the tools more used; How do you classify the interaction among applications; How did the technology tested facilitated interaction between tester and mobile device;

Replies varied a lot; factors like the type of disability, level of technological knowledge in relation to the use of devices and even age were determinant for that variation in replies.

The ten participants were subdivided in three categories (four to Android, 4 to iOS, 2 to Telepatix), three with physical disability, three with earing disability and four with visual disability.

What was found was that those with physical disability that are not totally amputees of superior members, do not show great difficulties in handling the devices; in what concerns interaction among applications and t the level of satisfaction, preference was almost unanimous to choose the native applications of the operating system, as they are more complete and adaptable to the several disabilities; others are more developed towards a certain public and do not interconnect with other applications.

The main limitation found in the native applications is that for both operating systems the functionalities only function efficiently with internal applications, presenting some problems when operating with third party applications.

In the case of *Voice Access*(Google) the more recent native application for Android, was considered the least satisfactory as it only accepts commands in English.

5.7. Analyse and discussion of the results of Questionnaire B

The opinions of the 25 participants to this Questionnaire were fundamental to answer to some of the questions base to this work mainly "Which is the AT more used for mobile devices?" "Which are actually the main limitations of the AT in mobile devices?" and "What can be implemented to improve the actual state of the AT?".

It could be concluded that the majority of the users use the assistive application of the OS Android, probably due to the fact that an iOS device is financially more expensive, given the fact that many participants assumed to prefer the Apple operating system in what refers to adaptation and functionalities.

6. MODELO CONCEPTUAL

Based on what was found though the Questionnaires and interviews, a conceptual model was developed for the development of an assistive mobile application directed to the disabled people studied in this research. The name given to that application is "*HelpApp*", as it is an assistive application.

6.1. Aim

The aim of this application is to adapt a common device to guarantee accessibility and to contribute positively in the process of social inclusion through improvement of the autonomy of disabled people in relation to technologies.

HelpApp must possess a large number of accessible tools in order to answer to the limitations presented by the different disabilities, the capacity of integration with the several operating systems, a friendly interface so that interaction is pleasant independently of the level of technological knowledge and access to the functionalities of the device.

It must guarantee the existence of basic requirements of a system like security, integrity, availability, among others.

6.2. General Description of the Application

The design of this application comprises an integration with all (or great part) of the applications found in devices like *Smartphones* or *Tablets*; in order that can be possible the best option is a native application as it has access to all functionalities. The application is addresses to disabled people that use mobile devices in carrying out their daily activities.

6.3. Modules of the Application

Module Blind: this module is directed to people with visual disabilities, allowing the user to use voice commands, interact with a personal assistant and reading from screen (with the use of a camera). It is necessary to access the Internet to interact with the personal assistant.

Module Mute: this module is destined to people with mute disability, it has a white screen where text must be written and the pretended text reproduced.

Module Deaf: this module is destined to deaf people, transcribing in the form of text what was given orally.

Module Move: this module is destined to people with physical disabilities; in this module, only the accessible routes are presented, as well as parking places for this group of people.

Module Total: it comprises all functionalities.

6.4. Specific Requirements

Functional Requirements: The application must be rather friendly and intuitive, possible to be installed in an economic mobile and include all the modules: To fill questionnaires for the choice of the module; to supply visual and/or audible feedback according to the module chosen; Localization management: it should identify the localization of the user; The application should recognize voice commands; to textually transcribe commands given orally; to issue orally the elements presented in the interface; alerts in real time.

Non Functional Requirements: The application must be resilient to failures that may prevent its normal functioning, in order to be always available; the reply time must be very low; easy to maintain, in order to include improvements and updates; it should guarantee security and integrity; in terms of usability all the actions must be transparent, in a way that the user understands all its effects; it should be capable of interacting with other heterogeneous applications for the change of information and the use of its functionalities.

7. SYSTEM PROTOTYPE

The development of the prototype and its main functionalities were conceived to operate in any of the main development platforms of mobile applications; Android or iPhone. The choice was made in order to be different from the existent applications.

7.1. Proposed structure

PhoneGap is an open-source development structure that allows the development through web technologies so that they can be later distributed as native applications.

There are several benefits that put forward the choice of a structure like PhoneGap towards other languages and native SDK's: the probability of deep knowledge of web technologies (HTML, CSS e JavaScript) is greater, contrary to technologies and native languages of the several existent systems: the possibility to use Javascript tools already existent makes very attractive this development approach.

7.2. Screen prototype

HelpApp will only have screens, as the purpose is the interaction with all the existent applications of the device.

When accessing the application, the loading screen is opened: we have opted for an intuitive background that attracts the attention of the user and demonstrates the interaction of the application with the others. The menu screen is as simple as possible, allowing the user to select the module that pretends to use and so activate its functionalities. Most of the application functionalities do not need an interface, but those that need present themselves in the most accessible way.

8. CONCLUSION

As a result of this research, it can be concluded that there are distinct resources of assistive technologies that help in the inclusion of people with disabilities (visual, earing, physical and mental). Meanwhile, through observations, questionnaires and interviews, it could be concluded that these type of technologies have a reduced level of growth.

With the world growth of the use of mobile devices, there is the need to adapt them to any type of user. The greater difficulty encountered during this study is related with the gap in documentation on assistive technology. Many are the computer resources but few is the knowledge on the process of creation, although there are several social sciences, educational and health forums talking about this subject.

Throughout this project, it was really noticeable the additional necessary effort to conceptualize an assistive application compared with a standard application. This additional effort focused essentially the need to really know the sample population, to fully understand its limitations and the medical condition of the disabilities in order to guarantee that an assistive application correctly functions.

REFERENCES

CIBEC/MEC, "Inclusão: Revista da Educação especial" Secretaria da educação especial.v.1, n.1 (out.2010). Brasília: Secretaria de educação especial,2010.

Cook & Hussey, Assistive Technologies: Principles and Practices. Mosby – Year Book, Inc., 1995.

Cook, A. M., & Polgar, J. M., Cook & Hussey's Assistive Technologies: Principles and practice (3ª ed.). Philadelphia, PA: Elsevier Inc. 2008

Cook, A. M., Adams, K., Volden, J., Harbottle, N., & Harbottle, C. Using Lego robots to estimate cognitive ability in children who have severe physical disabilities. Disability and Rehabilitation: Assistive Technology, 6(4), 338-346. 2001.

Cook, A. M., Bents, B., Harbottle, N., Lynch, C., & Miller, B. School-based use of a robotic arm system by children with disabilities. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13(4), 452-460.2005.

Cook, A.; Polgar, J., Assistive tecnologies: principles and pratice. Missouri, EUA: Elsevier, 2008.

Damas, M. J. & DE Ketele, J. M., Observar para avaliar. Coimbra: Livraria Almedina.1985.

Domingo, M. C., An overview of the Internet of Things for people with disabilities. Journal of Network and Computer Applications, 35(2), 584-596. 2012.

Emam, M.; Al-Abri, K.; Al-Mahdy, Y. (2017) Assistive Technology Competences in Learning Disability program candidate at Sultan Qaboos University: A proposed Model, *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA).*

França, C. R.; Borges, J. A. S.; Sampaio, F. F., Recursos de acessibilidade para educação especial inclusiva dos deficientes motores. Anais do XVI Simpósio Brasileiro de Informática da Educação., Juiz de Fora, 2005.

Glaser, B. G. & Strauss, A. L., The Discovery of Grounded Theory: Strategies for Qualitative Research, Chicago, Aldine Publishing Company, 1967.

Glennen, S., & Decoste, D., The Handbook of Augmentative and Alternative Communication. Singular Press. 1996.

Ienca, M., Wangmo, T., Jotterand, F., Kressing, R., Elger, B., (2018) Ethical Design of Intelligent Assistive Technologies for Dementia: A Descritive Review, Sci Eng Ethics 24: 1035-1055.

Islam, D., Ashraf, M., Rahman, A., & Hasan, R., Quantitative Analysis of Amartya Sen's Theory: An ICT4D Perspective. International Journal of Information Communication Technologies and Human Development (IJICTHD), 7(3), 13-26. 2015.

Júnior, W. F. R., Acessibilidade em sistemas webpara deficientes visuais. Monografia de Graduação (Graduação em Sistemas de Informação), Universidade Veiga de Almeida, Cabo Frio.2009.

Laudares, J. B & Ribeiro, S., Trabalho e formação do engenheiro. Belo, 2000.

Mohammadi, S., Momayez, A., & Rahbar, F., A., Conceptual Model in TechnoEntrepreneurship Services for People with Disability in Urban Management of Tehran. 2014.

Raja, S., “Realizing the potential of accessible ICTs in developing countries.”Disability and Rehabilitation: Assistive Technology 8(1):11–20. 2016.

Redford, K., (2019) Assistive Technology : Promises fulfilled, Educationl Leadership v76n5p70-74.

Ribeiro, F., Da arquivística técnica a arquivística científica: a mudança de paradigma”, in Revista da Faculdade de Letras, Porto, 2005, 1 série, vol1, pp 97-110.

Ribeiro, V. M .(ORG.). Letramento no Brasil: reflexões a partir do INAF 2001. São Paulo: Global, 2003.

Roberts, T. & Hernandez, K, The Techno-centric Gaze: incorporating citizen participation technologies into participatory governance processes in the Philippines, Making All Voices Count Research Report, Brighton: IDS. 2017.

SACI – Solidariedade, Apoio. Comunicação e Informação. Acessibilidade. [Online] Available at: http://www.saci.org.br> [Acesso em: 01-05-2019]

Scherer, M. J., Assistive Technologies and Other Supports for People with Brain Impairment New York Springer Publishing Co. ISBN-13: 9780826106452 . 2012.

Scherer, M. J., The Matching Person & Technology (MPT) Model Manual and Assessments, 5th edition [CD-ROM]. Webster, NY: The Institute for Matching Person & Technology, Inc. 2005.1.

Scherer, M. J., Living in the State of Stuck: How Assistive Technology Impacts the Lives of People with Disabilities, Fourth Edition. Cambridge, MA: Brookline Books. ISBN-13: 978-1571290984. 2005.2.

Turner-Smith, A., & Blake, P, Project DE4103 TELEMATE – Telematic

Visagie, S.; et al ; (2019) Perspectives on a mobile application that maps assistive technology resources in Africa, *African Journal of Disability*, vol.8, p 1-9.

Wong, R., Piper, M.D., Wertheim, B., Partridge, L., Quantification of food intake in Drosophila. PLoS ONE 4(6): e6063. 2009.

World Health Organization (Organização Mundial Saúde). (2012). Deafness and hearing impairment Fact sheet N.° 300. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs300/en/[Acesso em: 03-03-2019]

World Health Organization (Organização Mundial Saúde). (2012). Visual impairment and blindness, Fact Sheet N.° 282. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs282/en/ /[Acesso em: 03-03-2019]

World Health Organization (Organização Mundial Saúde). International Classification of Diseases: Version 2010 (ICD-10).2010 [Online] Available at: http://www.who.int/classifications/icd/en/ [Acesso em: 01-05-2019]

NO INDUSTRY ENTRY FOR GIRLS – IS COMPUTER SCIENCE A BOY'S CLUB?

Gosia Plotka, Bartosz Marcinkowski

De Montfort University (United Kingdom),
Polish-Japanese Academy of Information Technology (Poland)

malgorzata.plotka@dmu.ac.uk; bmarcinkowski@pja.edu.pl

ABSTRACT

The Computer Science (CS) industry stands out as male-dominated. On top of that, many companies struggle to retain women that dedicated themselves to enter careers in CS as they manage to do so. While initiatives to achieve gender parity within CS might be considered utopian, any form of discrimination against any gender in the industry is highly pernicious. Therefore, organizations employ official and unofficial measures to counter bias. The paper introduces a multi-country survey among scholars and IT/CS professionals. The survey explores how common and impactful are the issues revealed by the pilot study as well as highlights what has been done so far to encourage women and girls to join and/or stay in STEM. Based on that, change agents and organizational best practices are elaborated. The current paper covers and discusses the results of the first stage of the study.

KEYWORDS: Computer Science; Gender Inequality; Job Retention; Organizational Practices.

1. INTRODUCTION

There is a wide consensus that skills directly related to Computer Science (CS) – such as problem solving, design and evaluation of complex systems as well as human behaviour understanding – enable constructing meaningful artefacts using computers and are of critical importance in 21st century (Giannakos et al., 2017). Even though women actively took part in information technology evolution, still relatively few of them are pursuing their professional careers in Engineering industry. The lack of consistence in terminology used by research teams investigating the phenomena (IT-related contributions often address similar research settings as Computer Science, Computing or Systems) results in a number of studies with data that slightly vary. Similar variances have been observed among developed economies – data coming from Northern America sources are to some extent different than those coming from the European ones. For instance, according to Graf, Fry & Frunk (2018), only 14% of Engineering workers are women, and CS industry is underrepresented (25%) as well; at the same time, there was only 2% increase in Engineering jobs within the 27 years timespan (between 1990 and 2017) – whereas 7% drop in numbers in Computing. To provide a basis for comparison, throughout the same period the share of women in other fields (such as health-related, life sciences and even the other STEM areas – such as physics and maths) has increased. Homogenous data, also coming from the Northern America market, is provided by Ehrlinger et al. (2018). On top of that, Gorbacheva et al. (2019) make an observation that women constitute only 16.7% of employed IT specialists, even though overall 47% of them are active on the job market. Those claims are based, among others, on data provided by the Eurostat.

Moreover, diverse research also reveals that there is (1) a difference in the retention of women and men in the field of their study upon successfully completing their major in Computing/Engineering fields; (2) gender salary gap remains in place – just to mention research by Craigie & Dasgupta (2017) as well as Stephan & Levin (2005). The mechanics behind women effectively disappearing from some fields that can be considered 'geekier' is still an open issue, and one of a vital importance for business organizations and faculty. Therefore, the authors followed the research goal of identifying the hindrances leading to women underrepresentation within Computer Science industry and elaborating a set of best practices how to overcome some of the common problems and work together on evening up the numbers in computing to make it more diverse and inclusive.

After the introduction, we address a number of related studies and contributions that scrutinize the overall mechanics as well as the succeeding escalation phases of female underrepresentation within CS in section 2. Research questions are motivated and introduced in section 3, whereas the research approach adopted is recapitulated in section 4. Then, preliminary results of the study are presented and discussed in section 5, followed by a brief summary of on-going work and conclusions.

2. LITERATURE REVIEW

2.1. Faint flow into the pipeline

The reasons behind gender disparity within CS were approached in previous research from several angles. To visualize the phenomenon, the notion of STEM pipeline that experiences severe leaks throughout women education, up to their graduation, was forged. Indeed, it is pointless to dispute some of the indicators. Gordon (2016) maintains that CS in fact has no issues with potential suitors – yet males dominate individual cohorts at later stages of education with shares that exceed 80%. The phenomenon spreads worldwide. As reported by Galpin (2002), participation of women in Computing courses at undergraduate level highly varies throughout countries, with a bulk of results between 10% and 40%. That being said, unless one is keen on extending the beginning of the pipeline to as early stages of individual's education that a potential future graduate in practice has no actual say regarding educational choices – and many analyses do just that – flow into the pipeline itself is a major factor. This flow is faint at best. Cheryan, Master & Meltzoff (2015) argue that even should the most abstract scenario of retaining each and every woman who committed to majoring in CS or Engineering upon entering college came into fruition, sheer number of men who travelled the same path would cover their own leaks with a surplus.

Among the factors that contribute to this faint ingress flow, stereotypes regarding professional careers in CS combined with perception of absolute/relative gender strengths certainly deserve further exploration. Cheryan et al. (2009) offer a reasonable analogy in that regard: just as people who are not outdoor enthusiasts may find a city full of outdoor gear stores and cars with ski racks attached unappealing, women may find masculine stereotype-lined path counterproductive. Both solutions and inspirations for further research resulted from previous studies regarding CS stereotypes issue:

- in contrast to men, women are sensitive to artefacts in their environments that affect perception of such environments in terms of masculinity and femininity (Cheryan et al., 2009);
- females were keener on enrolling in an introductory CS course upon steering the classroom environment away from stereotypes regarding CS that high school students had at a time (Master, Cheryan & Meltzoff, 2016);
- stereotype factor remains in place even after changing the environment into a professional one, retaining gender proportion and providing homogenous salaries (Cheryan et al., 2009);
- broadcasting the image of CS and Engineering as highly-specialized fields makes more harm than good since potential candidates lose their sense of belonging; therefore, bringing down barriers at the entry of the pipeline requires broadening the mental picture of what it means to be a CS professional or an engineer (Cheryan, Master & Meltzoff, 2015).

While both academia and business can take advantage of stereotype-centred studies whilst designing their facilities and attracting students and employees, several core issues can be backtracked to early education and parenting practices. As this paper is focused more towards the end of the pipeline, we shall highlight the effects of non-belonging that might potentially directly translate to professional career stages. As reported by Garcia et al. (2018), even high-achieving CS female students still consider themselves less recognized than their male counterparts. Awareness of this fact is likely to affect decision processes of younger female generations and their tendency to feed the pipeline. Stoet & Geary (2018) note an interesting paradox – countries renowned for their gender equality policies tend to report larger sex differences as women feel free to pursue their comparative advantages and personal preferences, whereas in less gender-equal countries the cost of forgoing a well-paying STEM career encourages more moment to enter the pipeline.

2.2. Leaky academic pipeline

Gordon (2016) associates academic retention with the share of students who go along their selected course of study at a single institution instead of switching to another course/discipline or discontinuing their study; retention in CS was revealed to be the worst among all disciplines. Giannakos et al. (2017) dive into CS retention issue using Structural Equation Modelling and reveal that (1) CS students find their discipline lacking interactive and social aspects; (2) whereas high-quality teaching is a must across all disciplines, it does not ensure keeping CS students engaged in their studies and sticking to their selected majors. Peters et al. (2014) associate weak retention numbers with limited past experiences with programming and derivative struggles to envision future participation in CS activities and negotiation of meaning. Since most Engineering and CS curricula introduce programming subjects and focus on programming skills early on, students interested in broader perspective on technology and its social impact may find themselves discouraged (Peters & Pears, 2013). Misunderstanding of CS may also be linked to the lack of discipline-specific entry qualifications (Gordon, 2016).

As far as leaks from the pipeline given academic context are concerned, Virnoche & Eschenbach (2010) report that gender is not among the factors that affect retention significantly. Miller & Wai (2015), based on their analysis of a 30-year-long interval, point out that males and females currently persist at roughly equal rates in STEM fields between the bachelor's and Ph.D. degree. So, whereas the loss of CS students as they progress their education is severe, it is not discriminatory in nature. Early academic careers within CS also seem to be immune to gender bias. Ceci et al. (2014) do not hesitate to highlight a paradox: women tend to achieve the most success at being hired, promoted and remunerated as professors in the very fields they are underrepresented the most; thus, the academy itself is not to be overly blamed for disrupting gender neutrality. On top of that, Miller (2015) criticises the notion of STEM pipeline altogether, pointing out that many 'leaks' that are unfairly stigmatized carry on using technical skills gained to make significant societal contributions throughout other fields.

2.3. Leaky professional pipeline

IT industry, at first glance, is strongly open to gender neutrality. Since CS and IT reported shortage of qualified staff for years and such deficit only increases the exposure of whole nations to competitive risks, women and minorities are often considered untapped resources (Vitores & Gil-Juárez, 2016). Whereas studies conducted by the end of 20th century often revealed gender-related discrimination practices while hiring, more recent studies paint a definitely more balanced picture. For instance, Carlsson (2011) upon investigation of hiring practices in two largest Swedish job markets reports that no evidence was found to support a bias exists regarding the probability of being invited to an interview in male-dominated occupations, while in female-dominated occupations women had a higher call-back rate compared to men. Similar conclusions were reached by Charness et al. (2020) after experimentally investigating anticipated discrimination across gender, hiring patterns, and performance in tasks with different stereotypes in a labour-market setting: math-related discrimination against females in hiring did not take place at all.

Recruiting practices within CS evolve towards being more female-friendly. On one hand, it is many a time motivated by good publicity. On the other, women are reported to be more inclined to apply for a job given certain information points are in place. Sullivan (2018) lists 25 influence areas for attracting female applicants, which include revealing the proportion of women in this particular job, offering side-by-side company comparisons in terms of women-friendly features, providing project approval rates, or demonstrating the extent to which the job can be customized. It is the algorithmic hiring that ought to be highlighted when considering recruitment-specific risks to equality challenge. Such algorithms not only may follow data patterns that are obsolete given relatively recent shifts in policies. On top of that, women are more likely to raise red flags due to potential gaps in employment related to giving birth and children care (Parker, 2015).

We would argue that the real problem that contributes to females leaking from the pipeline upon graduation lies elsewhere. Majeed (2019) points out that hiring personnel to entry-level IT positions without any sign of bias might be easy, but at one point of any woman career she aspires to have more responsibility. Should that responsibility be unfairly denied, a potential leak would form. Obstacles in gaining access to high-power leadership positions in certain situations were confirmed by Hoover et al. (2019), who revealed that while high-ranking males rated male and female applicants for managerial position in line with their qualifications, lower-ranking ones reacted with discrimination to alleviate the threat of being subordinated to a woman.

One also cannot ignore the gap in salaries that refuses to go away. Although Budig (2014) reports that the tendency is clear, and between 1979 and 2012 the American market was very successful towards closing it, fatherhood is revealed to come with a salary bonus – while motherhood with a penalty. Rayome (2016) provides a number of technology market-focused indicators, concluding that (1) young female tech professionals (i.e. aged between 18 and 25) are particularly impacted by the difference in median salary that peaks at 29% and decreases over time; (2) one of the main reasons the salary gap persists is that women are more reluctant to negotiate the first offer compared to men; (3) scrutinizing salary-based data against gender disparities and coming up with objective criteria for evaluations that decide on promotions is a best practice for companies to implement. Ultimately, 56% of women in technology (twice as many as men) quit at the mid-level point, just when the loss of their talent is most costly to companies (Hewlett, 2008). Stephan & Levin (2005) point out that the lower retention rate of female IT professionals may not necessarily be tracked back to the comparative advantage of other industries – women are simply more likely than men to leave the labour force altogether.

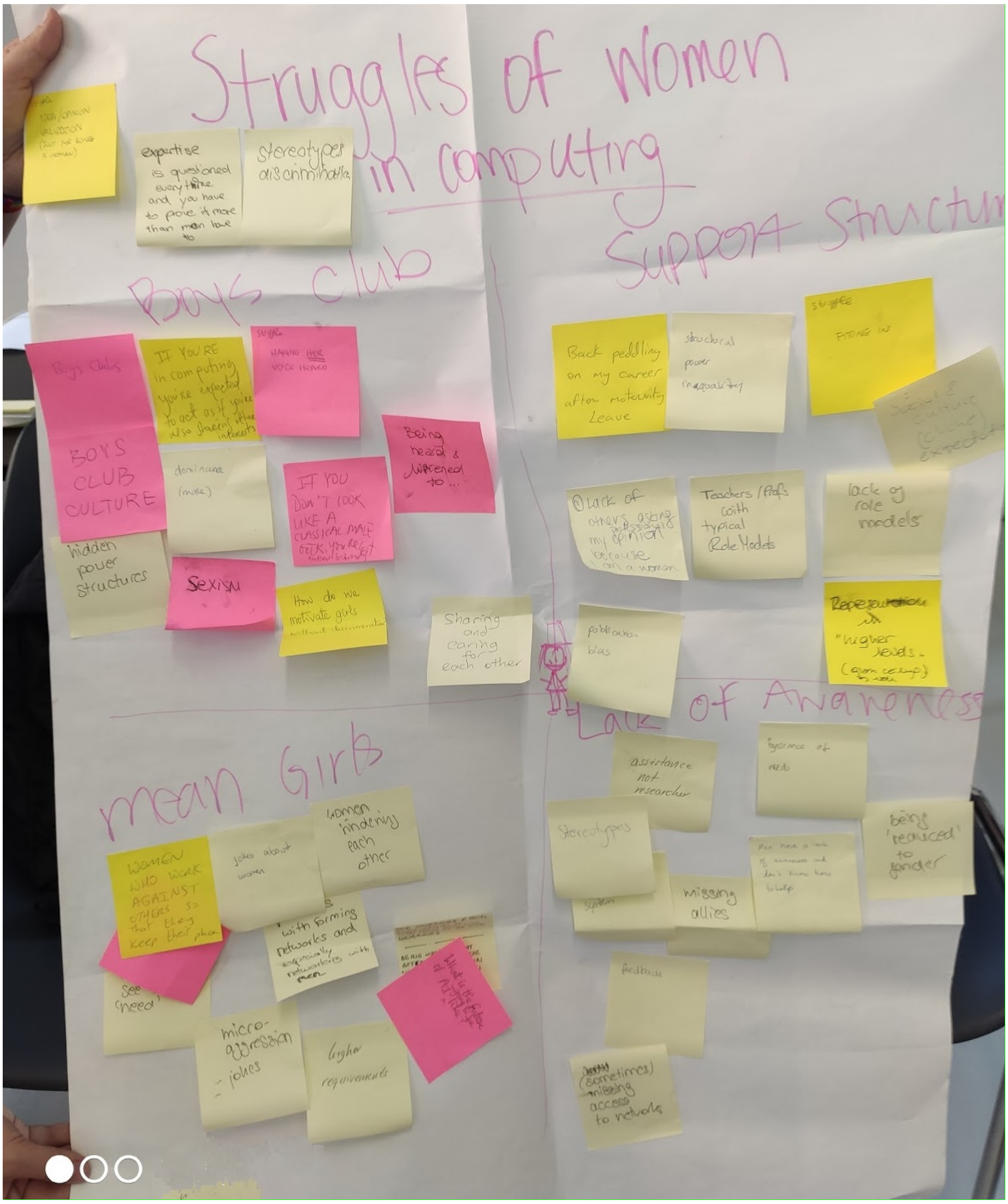
3. RESEARCH QUESTIONS DEVELOPMENT

Literature review shows that significant underrepresentation of women among CS professionals is the picture that is not going to change in the foreseeable future. Not so many females worldwide seriously consider undertaking this particular path early on, and a significant share opts out along the way. A number of myths, stereotypes and misconceptions contribute to such tendencies (Kindsiko & Türk, 2017) – and many researchers came up with best practices to counter them. That being said, given even more extremes present among the countries that might be classified as gender bias-free and hardly discriminatory academic practices, we find some of the more radical efforts to re-shape the phenomenon counterproductive. In our opinion, governments, companies and individuals should be actively committed to removing hindrances. The very hindrances that might pile up in front of these women, who actually decided to go against the tide and dedicated themselves to a career in CS as they enter the labour force.

To pave the way for exploring this topic, a pilot study was conducted: a focus group that involved nearly 20 women from around the world. The members of the group were asked about what they consider the main reasons for them behind struggling to decide to start their professional careers and stay in IT (Figure 1). The feedback helped to identify four areas of problems:

- it is a boy's club indeed – if you do not look or act like one, you cannot belong to it;
- there is a lack of support structure or role models;
- different standards for different genders and women hindering each other;
- lack of awareness or exposure.

Figure 1. Struggles if women in computing – focus group.


Struggles of Women in computing
IDEA/OPINION VALIDATION (JUST FOR BEING A WOMAN)
expertise is questioned every time and you have to prove it more than men have to
stereotypes discrimination
Boys club
Boys Clubs
BOYS CLUB CULTURE
IF YOU'RE in computing you're expected to act as if you're also sharing other interests
HAVING HER VOICE HEARD
dominance (male)
IF YOU DON'T LOOK LIKE A CLASSICAL MALE GEEK, YOU'RE NOT TAKEN SERIOUSLY
Being heard & listened to ...
hidden power structures
Sexism
How do we motivate girls without discrimination?
Sharing and caring for each other
Support Structure
Back peddling on my career after maternity Leave
structural power inequality
FITTING IN
Social & culture (cliché) expectations
Lack of others asking my opinion because I am a woman
Teachers / Profs with typical Role Models
lack of role models
publication bias
Representation in "higher levels" (glass ceiling)
mean Girls
WOMEN WHO WORK AGAINST OTHERS SO that they keep their place
jokes about women
women hindering each other
with forming networks and especially networks with men
What is the future of AI-like?
see 'need'
micro-aggression - jokes
higher requirements
Lack of Awareness
assistance not researcher
ignorance of needs
Stereotypes
system
missing allies
Men have a lack of awareness and don't know how to help
being 'reduced' to gender
feedback
(sometimes) missing access to networks

Still, there is a huge difference between an individual sense of being out-of-place – and the working environment neglecting, or even encouraging practices that make somebody feel out-of-place. Between the natural process of familiarizing oneself and fitting in a new setting that is full of challenges – and having an unofficial policy of the entire company in place that discriminates based on gander, ethnic group or any other attribute. Between field-testing, even unsuccessfully, anti-bias measures – and having no answers at all. Therefore, we posed a few research questions addressing diversity and inclusivity of CS working environments. In the current paper, we attempt to answer three of those:

RQ1: Do companies establish official anti-discrimination practices that cover gender or is it just a matter of organizational culture?

RQ2: What sort of measures (both formally anti-discriminatory and of general nature) are used to reconcile professional careers of employees and parenthood?

RQ3: What is the level of labour force support for introducing formal parities and programs addressed exclusively to female employees?

4. RESEARCH APPROACH

In order to collect empirical data, we launched a survey targeted at members of the global community that were professionally involved in IT/CS or represented academia in the aforementioned fields. Nigel, Fox & Hunn (2009) bring up a number of advantages of using survey approach: (1) it can cover samples that are geographically spread; (2) in most research settings it can provide results that may be efficiently used to draw conclusions and generalize those to wider population; (3) it can be thrown in the mix with other methods to deliver richer data; (4) since its participants are only exposed to events that would take place anyway, it does not introduce ethical concerns. We employed online Google forms service as an instrument for collecting data. The questionnaire featured both open-ended and closed questions, with the latter being primarily based on a 5-point Likert scale. The first stage of analysis, the results of which are covered in this paper, was initiated upon exceeding the threshold of 100 respondents providing their feedback.

The questionnaire form was divided into three sections of 5, 6 and 9 questions respectively (Table 1). Whereas the first section was put in place to capture particulars of each respondent and enable moderation of results, the second addressed audience's judgements regarding possible hindrances for women in CS, and the last one – ways of handling them.

Table 1. Questionnaire form

Respondent's particulars			
1. Respondent's year of birth		2. Respondent's country of residence	
3. Line of work	o *Business* o *Academia*	4. Size of the employing company	o *Micro (1-9 people)* o *Small (10-49 people)* o *Medium-sized (50-249 people)* o *Large (250 people or more)*
5. Gender	o *Male* o *Female* o *Neutral/undisclosed*		
Hindrances			
6. IT has traditionally been a boy's club – and I consider entry barriers to be high			
7. Female role models that appeal to me are very rare in Computer Science			
8. Women do not get enough exposure and are being assigned secondary roles more often			
9. Should you be in favour of the previous statement – what reasons for such lack of women exposure are there?			
10. Women within the industry tend to hinder each other significantly more often than men			
11. Please provide any additional hindrance(s) you can think of that applies to making careers by women within Computer Science field			
How to deal with those problems?			
12. My company is taking official measures to deal with inequality			

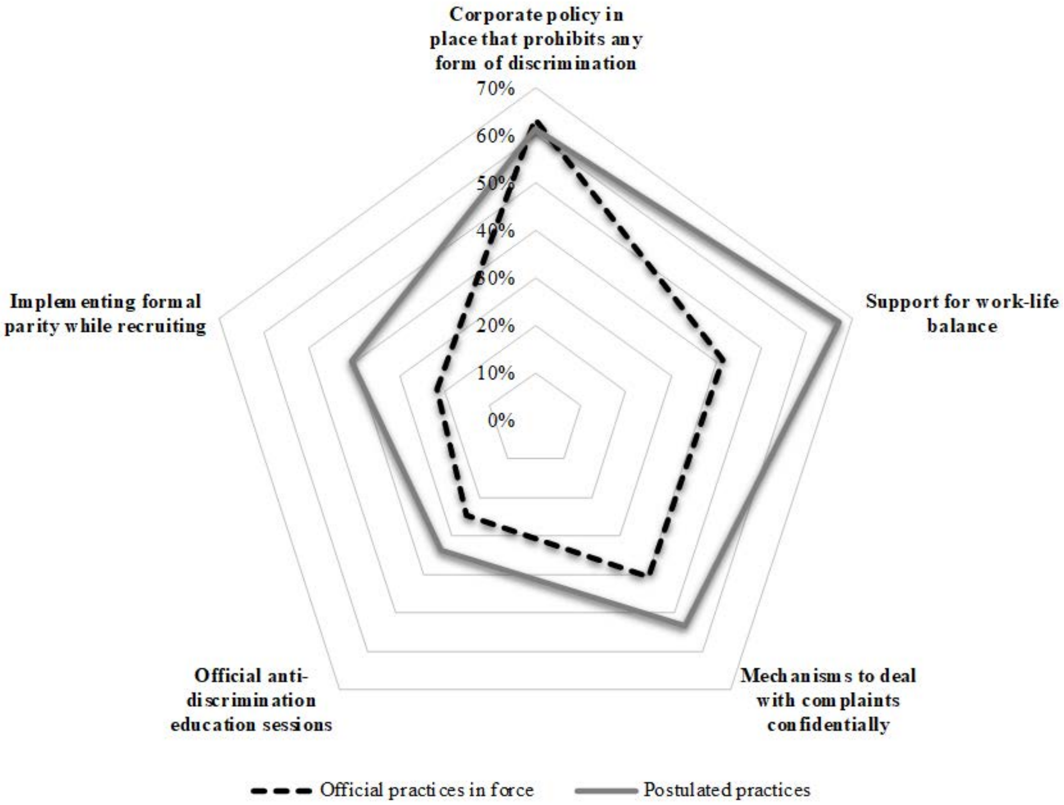
13. If so, what policies were implemented?	□ *Corporate policy in place that prohibits any form of discrimination* □ *Official anti-discrimination education sessions* □ *Balanced schemes for professional training funding* □ *Implementing formal parity while recruiting* □ *Implementing formal parity while promoting* □ *Support for work-life balance (nursery/infant schools at the workplace, paid maternity leave beyond the one enforced by law etc.)* □ *Mechanisms to deal with complaints confidentially* □ *Other: (short answer field)*
14. And which of those would you recommend?	□ *Corporate policy in place that prohibits any form of discrimination* □ *Official anti-discrimination education sessions* □ *Balanced schemes for professional training funding* □ *Implementing formal parity while recruiting* □ *Implementing formal parity while promoting* □ *Support for work-life balance (nursery/infant schools at the workplace, paid maternity leave beyond the one enforced by law etc.)* □ *Mechanisms to deal with complaints confidentially* □ *None*
15. I consider unofficial mentoring from experienced and recognized specialists to be a viable solution	
16. Managers need to be made to react immediately to any form of non-inclusion	
17. Company should enable/fund networking opportunities exclusively for female employees within Computer Science	
18. Please provide any additional ways to increase inclusion of women within CS field	
19. Any other feedback or suggestions?	
20. If you want us to be able to reach you in the future regarding the feedback, please leave your e-mail address here (*data processing statement included*)	

5. PRELIMINARY RESULTS OF THE STUDY

During this stage, we focused we focused on exploring the attitude of the community as a whole to the issues raised in RQ1-3. Questionnaire forms filled to date turned out to be close to gender parity (men constituted a marginal majority), while by far the largest share was attributable to the staff of companies that employed 250 people or more. Residents of as many as 25 countries shared their opinions, with the largest group being British.

As regards to whether companies that employ our respondents established official anti-discrimination practices that cover gender – or was it just a matter of organizational culture – the feedback was unequivocal. At the time of the preliminary analysis, only 18.8% of respondents flatly denied the existence of such policies in their companies or were rather convinced that there were no official policies targeting gender bias in place. Five among the policies that were either pre-selected or additionally reported by respondents were implemented in at least 20% of the organizations covered by the survey (Figure 2). Thus, it would be unreasonable to suspect that companies only set up minimalistic prohibition on discrimination in their statutes for legal or publicity reasons. It ought to be noted though that the numerical strength of large companies staff in the survey is not without an impact on this state of affairs, as major organizations may afford to allocate greater resources to ensure equality across all attributes, and simply show tendencies to formalize more. At this stage, it can be safely stated that support for setting up formal policies is huge throughout the sample, as there was only a single case of decreasing a relevant indicator when the official practices in force were confronted with those postulated by the respondents to be implemented (and this drop was slight: from 63.4% to 61.4%). In the remaining four cases the raise was significant, and only 6.9% of respondents argued that no formal policies are actually needed.

Figure 2. Official policies targeting gender bias with 20%+ implementation rate.



The feedback shows that companies do employ measures to reconcile professional careers of employees and parenthood. In the same time, it is the area that employers need to take great care of. Respondents highlighted parenthood-related issues in a number of places:

> "women are usually more involved than their male counterparts in child care, child raising and domestic chores; so it is more difficult for them to reconcile these family responsibilities with an intense dedication to their paid jobs; this is not specific of computer science, of course";
>
> "[management shows] sexism in relation to women starting families";
>
> "[…] empower fathers to go on parental leave to the same extent that mothers usually do and support all genders when they return to the job afterwards (e.g. with funded training to help them catch up on new developments)".

What surprised us, however, is a rather one-tracked recipe of large companies to this vital issue. One cannot overstate the utility of nursery/infant schools at workplaces, just as paid parental leaves are a backbone of the pro-family policy. The latter, however, constitute a risk factor in certain circles as well – employers may be less keen to hire females or may offer them lower salaries due to accompanying costs. To counter that, some Nordic countries introduced mandatory father quotas that are non-transferable and are lost if not used. Based on our personal experiences and the favourable specifics of the IT/CS market in this respect, we expected much stronger pressure towards flexible forms and working hours. The matter was admittedly raised by the respondents – but it clearly was of low priority.

It can be safely said that the question of the entire labour force's support for introducing formal parities and programs addressed exclusively to female employees causes greatest emotions. The doubt whether the promotional procedures were fair or not was confirmed. The preliminary results not only showed that companies would rather stick to formal parities when recruiting (often simply unnecessary – just to mention Carlsson (2011) or Charness et al. (2020)) and consider the matter closed. In fact, the largest gap (in both absolute and relative terms) between the policies officially implemented already, and those our respondents would like to see in place, relates to implementing formal parities when promoting. While existence of such policy was confirmed in only in 13.9% of the companies, 43.6% of the respondents put this particular policy on their wish list. In result, it would top two other postulated practices from the top five list (comp. Figure 2). Ultimately, respondents do not have much faith in existing mechanisms, which often prove discretionary. This confirms the results revealed in some other studies (Hoover et al., 2019; Majeed, 2019). This issue was also raised in open-ended questions:

> "[…] being overlooked for promotion […]"
>
> "hiring upper level positions is quite limiting to female candidates; even though we encouraged women and minorities to apply to a recent job opening very few had PhDs in CS, and even fewer had experience in management; all of the candidates that made the top 7 in our case were male".

Interest in gender-exclusive programs might be deemed marginal – while data collected so far exhibited only slight tendency contra enabling and funding such networking opportunities, unambiguously positive descriptive reflections to this topic were extremely rare:

> "[…] offers women-only courses on leadership […] which I found to be very good; courses like this in all sectors would be welcome, and, if operated in the same way, are effective"

Other statements in this regard were at most neutral. Generally, opposition to so-called positive discrimination is voiced strongly:

> "integrating female-only events to the more inclusive direction; that would promote better future also for the female CS experts, too, and the female-only events just highlight the current gender balance and underestimate the capability of female CS experts";
>
> "I don't feel companies should enable/fund networking opportunities that are exclusive to any particular gender or race; if we are wishing to encourage inclusion, a suggestion of division is not the way to go about that; speaking personally, if opportunities were presented with non-gendered bias in the writing, perhaps I would seek them out more actively";
>
> "if it is exclusive only for females then it is discriminating against all others";
>
> "[…] there is no such thing as good discrimination; instead encourage men to engage more in family life […] this would make it less beneficial for employers to favour men";

> "empowerment is inclusion […] exclusive benefits are a negative as it creates an idea that hard work isn't necessary";
>
> "[…] the so called 'positive discrimination' makes men sceptical about equality slogans".

6. ON-GOING WORK

Only a portion of collected data was scrutinized during the first stage of the analysis. On-going activities confront soft and hard measures – for instance, strong belief in adequate reactions of the management to reported bias seems somewhat surprising. Upper management in CS is, after all, dominated by men, and we have already shown that such belief does not translate to the process of promoting staff. Several interesting angles remain. Are females actually more hostile to other females than males are? What forms of hidden discriminatory practices were uncovered? How modus operandi looks like? Is a 'double jeopardy' issue, i.e. exponential discriminatory practices due to overlapping bias-related attributes a real thing?

7. CONCLUSSIONS

This study enabled us to reassess the mechanics behind the female 'leaky pipeline' within CS. We isolated the main trends and groups of hindrances, scrutinized a number of practices that companies already employ or might employ in the future to counter any form of bias, and outlined the directions of further analysis. Even partial data collected so far indicate a couple of flaws in the current state of affairs. First of all, we would encourage companies to carry out a comprehensive review of their promotion policies. Secondly, we would recommend a best practice of re-distributing parental leaves among both parents as much as legally possible – and supplementing it with flexible forms of labour provision.

ACKNOWLEDGEMENTS

The authors of the current paper would like to extend their sincere thanks to George Eleftherakis, Reyyan Ayfer as well as Rafal Kruczkowski for their guidance towards shaping this research and all their input.

REFERENCES

Budig, M.J. (2014). *The Fatherhood Bonus and the Motherhood Penalty: Parenthood and the Gender Gap in Pay*. Retrieved from https://www.thirdway.org/report/the-fatherhood-bonus-and-the-motherhood-penalty-parenthood-and-the-gender-gap-in-pay

Carlsson, M. (2011). Does Hiring Discrimination Cause Gender Segregation in the Swedish Labor Market? *Feminist Economics*, 17(3), 71–102. doi:10.1080/13545701.2011.580700

Ceci, S.J., Ginther, D.K., Kahn, S., & Williams, W.M. (2014). Women in Academic Science: A Changing Landscape. *Psychological Science in the Public Interest*, 15(3), 75-141.

Charness, G., Cobo-Reyes, R., Meraglia, S., & Sanches, Á. (2020). *Anticipated Discrimination, Choices, and Performance: Experimental Evidence*. Retrieved from http://hdl.handle.net/11073/9231

Cheryan, S., Master, A., & Meltzoff, A.N. (2015). Cultural Stereotypes as Gatekeepers: Increasing Girls' Interest in Computer Science and Engineering by Diversifying Stereotypes. *Frontiers in Psychology*, 6, 49.

Cheryan, S., Plaut, V.C., Davies, P.G., & Steele, C. M. (2009). Ambient Belonging: How Stereotypical Cues Impact Gender Participation in Computer Science. *Journal of Personality and Social Psychology*, 97(6), 1045.

Craigie, T-A., & Dasgupta, S. (2017). The Gender Pay Gap and Son Preference: Evidence from India. *Oxford Development Studies*, 45(4), 479-498

Ehrlinger, J., Plant, E.A., Hartwig, M.K., Vossen, J.J., Columb, C.J., & Brewer, L.E. (2018). Do Gender Differences in Perceived Prototypical Computer Scientists and Engineers Contribute to Gender Gaps in Computer Science and Engineering? *Sex Roles*, 78(1-2), 40-51

Galpin, V. (2002). Women in Computing Around the World. *ACM SIGCSE Bulletin*, 34(2), 94-100.

Garcia, A., Ross, M., Hazari, Z., Weiss, M., Christensen, K., & Georgiopoulos, M. (2018). Examining the Computing Identity of High-Achieving Underserved Computing Students on the Basis of Gender, Field, and Year in School. In *Collaborative Network for Engineering and Computing Diversity (CoNECD)*. Washington, DC: ASEE.

Giannakos, M.N., Pappas, I.O., Jaccheri, L., & Sampson, D.G. (2017). Understanding Student Retention in Computer Science Education: The Role of Environment, Gains, Barriers and Usefulness. *Education and Information Technologies*, 22(5), 2365-2382.

Gorbacheva, E., Beekhuyzen, J., vom Brocke, J., & Becker, J. (2019). Directions for Research on Gender Imbalance in the IT Profession. *European Journal of Information Systems*, 28(1), 43-67

Gordon, N.A. (2016). *Issues in Retention and Attainment in Computer Science*. Retrieved from https://documents.advance-he.ac.uk/download/file/4652

Graf, N., Fry, R., & Funk, C. (2018). *7 Facts about the STEM Workforce*. Retrieved from https://www.pewresearch.org/fact-tank/2018/01/09/7-facts-about-the-stem-workforce/

Hewlett, S.A., Luce, C.B., Servon, L.J., Sherbin, L., Shiller, P., Sosnovich, E., & Sumberg, K. (2008). The Athena Factor: Reversing the Brain Drain in Science, Engineering, and Technology. *Harvard Business Review Research Report*, 10094, 1-100.

Hoover, A.E., Hack, T., Garcia, A.L., Goodfriend, W., & Habashi, M.M. (2019). Powerless Men and Agentic Women: Gender Bias in Hiring Decisions. *Sex Roles*, 80(11-12), 667-680.

Kindsiko, E., & Türk, K. (2017). Detecting Major Misconceptions about Employment in ICT: A Study of the Myths about ICT Work among Females. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 11(1), 107-114

Majeed, S. (2018). *This is What Women in Tech Wish They Knew Early on in Their Careers*. Retrieved from https://www.theladders.com/career-advice/this-is-what-women-in-tech-wish-they-knew-early-on-in-their-careers

Master, A., Cheryan, S., & Meltzoff, A.N. (2016). Computing Whether She Belongs: Stereotypes Undermine Girls' Interest and Sense of Belonging in Computer Science. *Journal of Educational Psychology*, 108(3), 424.

Miller, D.I. (2015). *A Metaphor to Retire*. Retrieved from https://www.insidehighered.com/views/2015/03/03/essay-calls-ending-leaky-pipeline-metaphor-when-discussing-women-science

Miller, D.I., & Wai, J. (2015). The Bachelor's to Ph.D. STEM Pipeline No Longer Leaks More Women than Men: A 30-Year Analysis. *Frontiers in Psychology*, 6, 37.

Nigel, M., Fox, N., & Hunn, A. (2009). Surveys and Questionnaires. In *The NIHR Research Design Service for the East Midlands/Yorkshire & the Humber* (pp. 1-48).

Parker, K. (2015). *Women More than Men Adjust Their Careers for Family*. Retrieved from http://www.pewresearch.org/fact-tank/2015/10/01/women-more-than-men-adjust-their-careers-for-family-life

Peters, A.K., & Pears, A. (2013). Engagement in Computer Science and IT – What! A Matter of Identity? In *2013 Learning and Teaching in Computing and Engineering* (pp. 114-121). Piscataway, NJ: IEEE.

Peters, A.K., Berglund, A., Eckerdal, A., & Pears, A. (2014). First Year Computer Science and IT Students' Experience of Participation in the Discipline. In *2014 International Conference on Teaching and Learning in Computing and Engineering* (pp. 1-8). Piscataway, NJ: IEEE.

Rayome, A.D. (2016). *Closing the Tech Gender Gap: How Women Can Negotiate a Higher Salary*. Retrieved from https://www.techrepublic.com/article/closing-the-tech-gender-gap-how-women-can-negotiate-a-higher-salary

Stephan, P.E, & Levin, S.G. (2005). Leaving Careers in IT: Gender Differences in Retention. *The Journal of Technology Transfer*, 30, 383-396

Stoet, G., & Geary, D.C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, 29(4), 581-593.

Sullivan, J. (2018). *Need Women Applicants? Why Micro-Targeting Women Triggers More to Apply*. Retrieved from https://www.ere.net/need-women-applicants-why-micro-targeting-women-triggers-more-to-apply

Virnoche, M., & Eschenbach, E.A. (2010). Race, Gender and First Generation Status in Computing Science, Engineering and Math persistence. In *2010 IEEE Frontiers in Education Conference (FIE)* (T1A 1-6). Piscataway, NJ: IEEE.

Vitores, A., & Gil-Juárez, A. (2016). The Trouble with 'Women in Computing': A Critical Examination of the Deployment of Research on the Gender Gap in Computer Science. *Journal of Gender Studies*, 25(6), 666-680.

ON USING A MODEL FOR DOWNSTREAM RESPONSIBILITY

Marty J. Wolf, Frances S. Grodzinsky, Keith W. Miller

Bemidji State University (USA), Sacred Heart University (USA),
University of Missouri – St. Louis (USA)

mjwolf@bemidjistate.edu; grodzinskyf@yahoo.com; millerkei@umsl.edu

ABSTRACT

In the software development process, it is common for developers to be unaware of all the potential uses of the software they create. Developers typically do not control who buys their software, and importantly, they do not control the sorts of systems that their software is used in. This work guides responsibility analysis of complex systems that involve upstream software (developed earlier in the process) and downstream software (developed later). To that end, we analyze a responsibility attribution model and recent papers that address responsibility attribution to identify and categorize software features that scholars use in their analysis. This work gives key features of software that developers can use to address questions of responsibility surrounding their work. Furthermore, this work supports a move toward adopting sound "software provenance" practices as a way to address the many hands problem.

KEYWORDS: responsibility, software developer responsibility, models of responsibility, ethical analysis.

1. INTRODUCTION

In recent years, harm caused through the use of software surfaces at all too regular intervals. Authors often try to identify who is responsible for the harm caused by failures in software systems. Such analyses typically involve some consideration of social, economic, and technical factors. There are times when the analysis is done in such a way that it places responsibility on "software developers." These sorts of analyses, while appropriate, often do not account for the complexity of software and the fact that many software developers were likely involved in creating the disparate pieces of software that make up the software system under consideration. In "On the Responsibility for Uses of Downstream Software," (Wolf, Miller, and Grodzinsky 2019), we took a slightly different approach by considering the technical features of the software and the software development process and argued that these aspects play a role in the attribution of responsibility to the software developers whose software is part of the complex computing system that caused harm.

In the software development process, it is common for developers to be unaware of all the potential uses of the software they create. Developers typically do not control who buys their software. We set aside those cases where the software is designed for a specific narrow purpose and somehow fails at that purpose. Rather, we are interested in guiding analysis of complex systems, especially when pieces of the software are put together by different groups of developers.

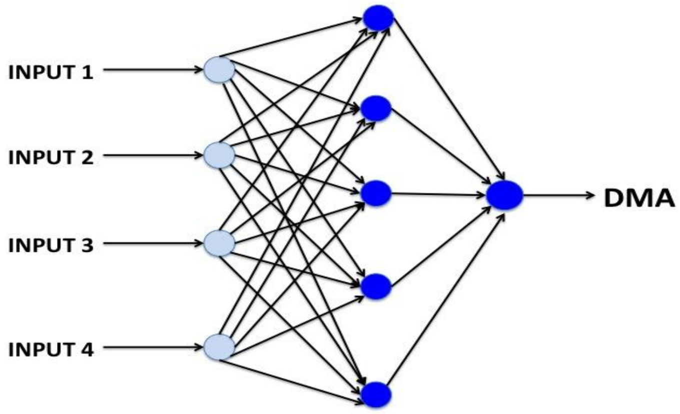
Consider an example: Assume that a software system is designed to ensure that a bomb hits its intended target; assume further that the software fails, and the bomb hits someplace else. We typically assume that the developers of the targeting software bear responsibility for that software failure. However, the complexity of any targeting software raises other concerns about whether that responsibility should also be shared by others; for example, should we attribute some responsibility for the failure to the developers who produced the driver for the GPS unit that the target software developers purchased “off the shelf?” In our analytical system, we call analogues of the GPS driver “upstream software,” and we call analogues of the targeting software a “downstream software.” Note that mistakes that cause the errant bomb could have been exclusively in the upstream software, the downstream software, or from an interaction between the two; localizing the fault would be part of the analysis.

The bomb targeting example illustrates another dimension to this issue: responsibility for intended effects, and unintended effects. The developers of the GPS driver may not have foreseen specific military uses for their product; the developers of the bomb guidance software clearly would. (The use of open source software in military applications was at the center of a discussion with some relevance to this article; please see Miller (2007) and Wolf et al. (2009).)

In an earlier article, we argued that there are at least five features of software and the software development process that should be considered during deliberations about responsibility attribution. The five are: closeness to the hardware, risk, sensitivity of data, degree of control over or knowledge of the future population of users, and the nature of the software (general vs. special purpose). Software that deals directly with capabilities of the hardware (such as the GPS driver mentioned above) are close to the hardware. Applications, such as a web browser that can run on a wide variety of devices, are farther from the hardware. Risk in software is tied to the predictability of its behavior and outputs. Sensitivity of the data refers to how private the data are that the software is processing. The degree of control over the downstream users of the software refers to the software developer’s ability to limit who uses the software or their ability to ensure that the downstream user of the software meets certain professional standards. Finally, the nature of the software matters in responsibility attribution for downstream uses. How these features interact, and which feature take priority over others is a matter of judgment, and is case-specific.

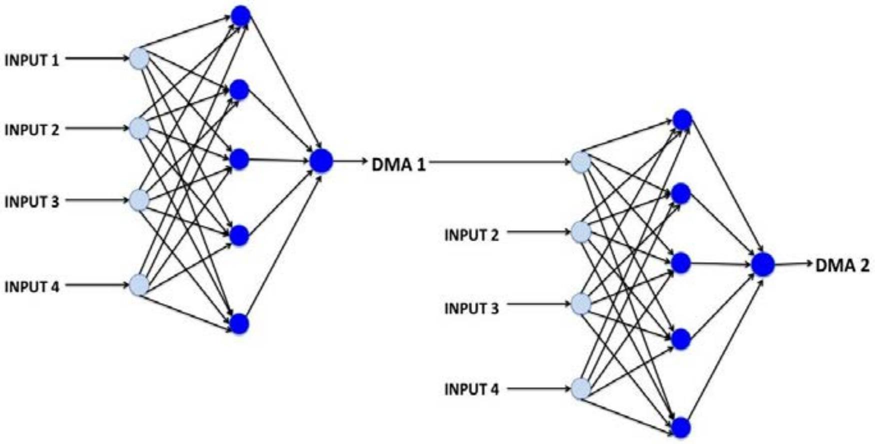
Close analysis of these features in the context of responsibility in computing ethics led us to develop an analytical system with two different models that might be used to assign responsibility in the use of downstream software: the Fixed History Model and the Chained History Model. In the Fixed History Model (see Figure 1), we assume that within the system of events (such as choosing to use an existing piece of software as part of a system) that led to an ethical breach (or an ethically positive outcome), there are certain inputs that are immune to the assignment of any moral responsibility for the Distributed Moral Action (DMA). This model does not consider assigning any portion of Distributed Moral Responsibility (DMR) to those who produced the inputs—INPUT 1, INPUT 2, and so on in Figure 1. The Fixed History Model is appropriate for certain types of software. For example, the developers of database software are rarely considered for the assignment of moral responsibility in the event of a breach. Typically, in such a case, responsibility stops at the database implementers who would have been responsible for initiating and maintaining security.

Figure 1. Fixed History Model.



The Chained History Model (see Figure 2), however, applies in cases where one of the inputs to the system is a piece of software and the attribution of moral responsibility propagates back to the developers of that software. In Figure 2 DMA 1 is a piece of software that becomes part of a larger software system that results in DMA 2. For complex computing systems, having more than two levels may be appropriate. Additionally, it is the case that multiple inputs may be upstream software and each of them may play different roles in contributing to the DMA under analysis.

Figure 2. Chained History Model.



In Wolf, Miller, and Grodzinsky (2019) we argue that, taken independently, each of the five features suggests a particular model to use. The Fixed Model fits well when the upstream software is close to the hardware, its developers have control over the downstream use, or the software is more general purpose. On the other hand, the Chained Model is more appropriate when the upstream software is riskier, computes over more sensitive data, or is more special purpose.

In this paper, we review a selection of recent papers (2017-2019) related to the attribution of responsibility in emerging technologies. Our analysis will determine situations when responsibility attribution for a moral action could have been clarified by using either the Fixed History Model or the Chained History Model. We will demonstrate how applying these models might help clarify ethical issues associated with distributed responsibility for software developers in a few complex and interesting cases.

The next section reviews six papers, identifies the strategy used in the analysis, and compares that strategy to the two models we proposed. The section that follows analyses the applicability of the analytical system and adapts them to more closely parallel how scholars think through issues surrounding responsibility in software. We offer some final thoughts in the conclusion.

2. REVIEW OF SELECTED PAPERS

Summary A

In the article "Digital health fiduciaries: protecting user privacy when sharing health data" Chirag Arora (2019) argues that when it comes to privacy concerns surrounding health data, it is the responsibility of the digital health data controllers to take steps to protect the privacy of those whose data is being collected and stored. Arora argues for a fiduciary relationship between data subjects and the data controller. Arora's argument uses "security, anonymization, and data minimization as examples of contextualization and flexibility required to deal with privacy issues." Even though Arora brings up the WannaCry ransomware attack on the UK's National Health Service, the responsibility for the system failure is not mapped back to flaws in the software that was infected by WannaCry, but rather to the failure to upgrade. Arora places the ethical breach at the feet of the data controller and makes no attempt to push any responsibility back to those who created the software with the flaw in it. We take this as evidence that Arora assumed a Fixed History Model.

Analysis A

Arora uses the term "controller" as defined by the EU's General Data Protection Regulation and refers to the entire collection of legal entities that might bear some responsibility. Thus, this analysis does not directly address the various responsibilities of the many individual contributors to complex digital health systems. The question we are concerned with is whether our analysis technique can advance this analysis to include statements about the responsibility born by the developers of upstream software. On one hand, since a digital health system presumably includes hardware, the closeness to the hardware features leans toward the Fixed History Model. Also, given that Windows 7 was an integral part of NHS, it is easily argued that some of the upstream software is general purpose, also suggesting a Fixed History Model.

On the other hand, it may be the case that at least some of the upstream software in the system is deemed risky. For example, software that controls various medical devices and is deployed in situations where there may be life and death decisions being made surely involves risk. Further, it is clear that such a system is handling medical data that are highly sensitive. Both of these observations suggest that the Chained History Model ought to be in play. It is clear in the context of this complex system that its developers could have had some control and fore knowledge of the sorts of downstream developers who would be using it. It is also likely that at least some of this software would be special purpose. Therefore, these two observations are in tension. The upshot of this analysis is that for complex software systems there could be a tension among what features to prioritize in the analysis. Identifying those upstream parts of the system is important. In this case, trying to resolve and prioritize the features might be a step in moving the analysis towards the Chained History Model that more accurately reflects the complexity of the digital health system. It is also important to conceptualize that software development is not just design, code, and deploy; but also includes maintenance as part of the life cycle, and maintenance is a crucial part of the story of this failure. The developers who create the patches can only go so far. It is the responsibility of system administrators to actually apply them in mission critical systems.

Summary B

In the article "First steps towards an ethics of robots and artificial intelligence (RAI)," John Tasioulas (2019) investigates the problem of trying to build moral norms into RAIs. He distinguishes between RAIs that follow top-down algorithms that are prescriptive and closed-rule and bottom up or stochastic algorithms that use machine learning. In the first case, the RAI is largely functional and failure to accomplish its task can be attributed back to the developer. "In ... RAIs operating on the basis of top-down algorithms that render their behavior highly predictable, the argument for attributing legal responsibility to manufacturers, owner, or users seems compelling (Tasioulas, 2019:70). In our language, responsibility analysis can be served by the Fixed Model. Those developing the upstream software are likely not to share in the responsibility for harm caused by this sort of RAI.

In RAIs that use machine learning, the cases are more varied and complex. The author asks "...whether a good case exists for attributing legal personality to RAIs with corresponding legal rights and responsibilities …" (Tasioulas, 2019:70). The European Union and UNESCO have been grappling with this issue and it is beyond the scope of our paper to delineate the various arguments. However, Tasioulas does raise the question of traceability as particularly difficult with RAIs using bottom-up algorithms. He asks "... how do we ensure the 'traceability' of RAIs in order to be able to assign moral or legal responsibility in relation to them? Traceability involves being able to determine the causes that led an RAI to behave in the way that it did…" (Tasioulas, 2019:71). Whenever we grapple with tracing responsibility, we are closer to the Chained History Model.

Analysis B

In Tasioulas' first case, our analysis is similar to Tasioulas'. If in the development of the RAI's software, the developers used "off the shelf" software, the upstream developers cannot reasonably be expected to predict that their software might be used in such a system. In the second case, we see Tasioulas considering one of the components of the larger system and recognizing the risk associated with it. The Chained History Model seems to be called for here since machine learning software is inherently risky software; it requires great caution and complex software to prohibit dangerous, unanticipated actions. Tasioulas' argument supports a claim that lack of traceability increases the risk of such software. Thus, applying the Chained Model to this sort of system seems warranted. Developers of machine learning software ought to bear some responsibility for its downstream use.

We note here that developers should not be allowed to claim that they are absolved of any ethical responsibility for a machine learning algorithm, just because it can change after deployment. We have argued elsewhere (Wolf et al., 2017) that the developers should not deploy something unless than are willing to take responsibility for its behaviour after launch.

Summary C

In the article "The ethics of cloud computing" Boudewijn de Bruin and Luciano Floridi (2017) argue for certain regulatory restrictions that prevent various players in the cloud computing arena from engaging in certain practices. They argue for "rather intense pressure on business clouders" (de Bruin & Floridi, 2017:33). Their analysis rests on a clear delineation of the business interests in the cloud computing space. At the hardware level, there are companies that host cloud services. At the next level are the cloud services themselves—the applications and the software (e.g. recruiting software, document storage and sharing). The third level is made up of the business users of the software. It is this third level that they call the "business clouders". They argue that when cloud computing was still in its infancy (at least five to seven years ago), the providers of cloud computing should have been given a lot of latitude to experiment and thus, should have had few regulatory restrictions. On the other hand, they "suggest that proscriptive pressure must be exerted primarily on the business *users* of software as a service" (de Bruin & Floridi, 2017:23, emphasis in the original). In their analysis, they leave open the possibility of enacting more restrictive regulations on the hosting companies and software-as-a-service providers, should risks become evident.

Analysis C

While the focus of their analysis is establishing an ethical and well-justified regulatory framework for cloud computing, it is clear that their analysis is consistent with using the features of software and software development. Software developers closest to the hardware, the general-purpose nature of cloud services and software-as-a-service, and control over downstream use all point to using the Fixed History Model. De Bruin and Floridi make a clear argument that using cloud services mitigates risk. "… [T]he probability of this kind of crime" (the stealing of actual servers) "is likely to decrease when firms start opting for cloud services, because criminals will find it very hard to determine which servers in the datacentres contain the data they are interested in. Whereas a bank's server has only one purpose and is an easy target …" (2017:34). Thus, because of the less risky nature of cloud computing, the risk feature also points to using the Fixed History Model for responsibility attribution to the upstream software providers. Finally, cloud hosting services provide general purpose software. When this software is used by downstream developers, this feature also points to using the Fixed History Model. On the other hand, software-as-a-service is more special purpose: it is for storing and sharing files or recruiting employees. De Bruin and Floridi focus on the business clouders. In our analysis, the final feature, control over downstream use, comes into play. A software-as-a-service provider enters into a contract with the business clouder that integrates the upstream software (software-as-a-system) into the business clouder's computing infrastructure. This again points to the Fixed History Model.

Summary D

In the article "Artificial intelligence, responsibility attribution, and a relational justification of explainability," Mark Coeckelbergh examines the question: "What does the development of responsible AI mean?" (Coeckelbergh, 2019) In other words, who is responsible for the benefits and harms of developing and using AI technology? He initially explains the Aristotelian conditions of responsibility: control and knowledge of the responsible agent and starts with the premise that only humans can be responsible agents. He then adds the complication that AI systems are not always reducible to one agent. In these complex systems there is both the problem of many hands and of many things (systems and subsystems) interacting. Coeckelbergh also addresses the idea of relational responsibility: those who develop the software (agents) and those who are affected by it (patients) enter into a relationship mediated by the software. Moral agents should have a moral requirement to provide or explain reasons for a decision or action caused by their software to moral patients. He assumes that AI systems do not meet the criteria for full moral agency and therefore, the responsibility for their actions and decisions remains with the humans "who develop and use this technology." Coeckelbergh draws his examples from automation technologies, especially self-driving cars and airplanes.

Analysis D

The conceptual analysis in Coeckelbergh (2019) is illuminating. In deconstructing the problem of many hands and the concept of distributed responsibility, Coeckelbergh (2019) makes the following point that speaks to our attempt to develop a model that is helpful to developers:

> Acknowledging the distributed character of responsibility in the case of AI does not solve the practical problem of *how* to distribute the responsibility, given that one may not know (the extent of) all the contributions and interactions, and given a number of other challenges.

Coeckelbergh sees the way upstream software, AI in this case, is used downstream as a problem of distributed responsibility for AI applications. Citing cases where AI applications have been developed for one use but then applied in totally different contexts and cases where maintenance by systems administrators is not done responsibly, he argues that pro-active development is necessary for AI systems not to fail. Trying to attribute causal responsibility after the fact is very challenging. He identifies problems that surround "a long causal chain of human agency" such as the developers, the choices of data sets, potentially using software in one context when it was developed for a different one. He also identifies the role that software maintenance plays in establishing responsibility (2019). AI is generally viewed as risky, not close to the hardware, and in Coeckelbergh's examples there is little control over the downstream use. Each of these points to using the Chained Model. This is consistent with what Coeckelbergh is trying to do. One of Coeckelbergh's arguments, however, seems to be that some AI might be considered general purpose. "[I]n principle 'medical' face recognition software can also be used for surveillance purposes and become 'police' AI" (2019). This pushes back against our claim in our earlier paper that general purpose software points toward using the Fixed History model in responsibility attribution. That is, this observation says that the downstream use of AI in this way ought to remove the face-recognition AI developers from consideration for sharing in the responsibility of the moral impact when it is used. At first blush, this seems to point to a shortcoming in our analytical system in that it gives conflicting indications as to how to ascribe responsibility. On the other hand, it is better to view this as an issue to be resolved by additional means.

Summary E

Vakkuri, Kemell, and Abrahamsson (2019) compare the literature in AI ethics with interviews they did with AI practitioners working in the healthcare sector in their paper "Implementing ethics in AI: Initial results of an industrial multiple case study." They list four values that they consider central to AI ethics: transparency, accountability, responsibility, and fairness. Much of their paper focuses on the nature and openness of algorithms that are "hidden" inside AI software. This focus includes descriptions of software projects that rely upon components that are studiously treated as "black boxes," whose inner workings are not studied nor even observed. Even more troubling they note that "developers do not see this as a problem" (2019).

Analysis E

The values of transparency, accountability, and responsibility discussed in Vakkuri et al. are all are highly relevant to our discussion about our two models. Their findings suggest that the subjects of their study only use the Fixed History Model when considering systems as it does not require much transparency, and the resulting accountability and responsibility relationships are clear but "shallow." These practiioners do not consider ascribing any moral resopnsiblity to those whose software they use and further expect the same consideration by those who use their software. It is clear that for the subjects of this study that the primary ethical go was competent creation of the software artifact. There was little consideration for other ethical values.

An insight from Vakkuri et al. that is directly relevant to our analytical system is that when we decide that the Chained History Model is appropriate to a particular system, then we must insist on transparency about that system in order to facilitate accountability and responsibility. As transparency increases about who is responsible for what in AI (or other) software, then accountability and responsibility could be traced more deeply, and the Chained History Model better reflects the complexity of the shared responsibility. Only when we can trace responsibility back further than a single step, and downstream into the next step in development can we meaningfully engage about earlier and later developers and their accountability.

Summary F

In the article "Algorithms, governance, and governmentality: On governing academic writing," Lucas Introna makes an argument that the algorithmic action of *Turnitin*, the text similarity analysis service, is increasingly a governing force in academic writing in that it "has produced a very particular regime of practice when it comes to academic writing" (Introna, 2016:37) While this article does not focus on responsibility, per se, it does focus on the software development process. Introna goes to unusual lengths to demonstrate the difference between code (Figure 1 is a C++ implementation of bubble sort) and an algorithm. Additionally, Introna's analysis includes the notion of the "relational temporal flow" of a piece of software. "This temporally unfolding process (the doing of the algorithm) itself inherits from prior actions and imparts to subsequent actions, in the temporal flow of the doing of everyday sociomaterial practices" (Introna, 2016:22). While Introna's focus is on governance and ours is on responsibility, the analysis clearly acknowledges that some software (an implementation of a sorting algorithm) is temporally distant from its use in a commercial product, e.g. *Turnitin*.

Analysis F

Even though Introna gives a thorough analysis of the nature of software and software development, there is little to suggest that the downstream use of software has a bearing on issues surrounding governance. The analysis clearly parallels the Fixed History Model. Our analytical system also tilts in that direction. The software Introna considers is far from the hardware, and at least in the case of Turnitin, the software doesn't seem to bear much risk (but see below) and the data are not sensitive. On the other considerations, control over the downstream use and whether the software is special purpose or not, there is no clarity about what piece of upstream software might be under consideration.

Introna's work does bring to the fore another potential measure for consideration in the context of responsibility. Introna identifies numerous authors who argue for transparent algorithms. These arguments demonstrate an understanding that without some knowledge of how a system is designed, only the Fixed History Model can be applied. Introna then argues against transparency as an essential feature of software by suggesting a potentially undesirable outcome of transparency.

The argument considers someone who has the linguistic ability to understand the algorithm a hypothetical *Turnitin* might use for detecting how similar a piece of text in an essay is to an extant piece of text. Someone with that knowledge could then take steps to make changes to a copied passage: "those with linguistic skills could 'write over' copied fragments to make them undetectable" (Introna, 2016:38). Introna goes on to point out how this would disadvantage a non-native English speaker (assuming the essays were required to be written in English) in any sort of competitive environment.

This observation has a bearing on our analytical system, although in a less nefarious situation. In our case, we consider an upstream developer whose understanding of a downstream piece of software is incomplete. In this situation, should the Chained History Model come under consideration, part of the analysis should include the appropriateness of using the downstream software. In some cases, it may be that risk was introduced by the very act of choosing to use that software in the upstream system. In those cases, our conclusion is to tend toward the Fixed History Model.

3. ANALYSIS OF THE MODELS AND FEATURES

While a limited set of articles have been considered here, there is evidence that the analytical system we developed is supported by scholarship in computing ethics. Additionally, this work has identified two additional features of the software development process to be considered for inclusion in the analytical system: software maintenance and the use of existing software.

3.1. Software Maintenance

Generally, there is widespread agreement that software users should keep their software patched. That is, when an upstream software developer creates a patch, the downstream user of that software ought to install it—especially when that user is an end user (a business or a consumer). Upstream software developers might have reason to be more hesitant to apply such a patch. While the patch can be tested on non-production systems, the patch cannot be reasonably tested on all instances of the upstream software currently in production.

There are complications in trying to understand the overlapping responsibilities here. When software is released prematurely, patches may be frequent and inconvenient for users. Premature release is the ethical responsibility of developers, not users. Also, if patches routinely disrupt users' systems when installed, there is a disincentive for users to install them. Thus, it may not always be merely sloth or sloppiness that is the reason for users not promptly installing patches.

Another complication arises in environments where downstream developers (for example the developers in the NHS case considered by Aurora) have responsibilities to maintain stable mission critical systems. Such an environment calls for extensive testing to ensure that applying the update will not cause a failure. In these cases, it is not that the updated upstream software has flaws or that the downstream software has flaws, it is that the act of integrating those two systems introduces a system failure—the undesirable DMA.

We think some cases are clear: if a developer has acted responsibly with updates and patches, and if an ethically significant breakdown results in harm, and if an update or patch could have avoided that harm, then we think the Fixed History Model is appropriate, and users bear responsibility for the harm. Alternatively, if updates and patches are late or habitually disruptive, the Chained History Model seems more appropriate, as the developers share in the problem that resulted in the harm.

Software maintenance is an essential aspect of the software development process. Unlike other aspects of the software development process, its impacts often are in existing systems and dynamic environments that cannot be fully tested. The complexities of this socio-technical feature of the software development process do not point to using either the Fixed History Model or the Chained History Model to determine responsibility for DMAs. In real cases, careful analysis is required to ascertain an appropriate distribution of ethical accountability.

3.2. The Use of Existing Software

In practice this feature of the software development process is only narrowly applicable in the consideration of attribution of responsibility in the case of DMAs. In cases where a downstream developer has knowledge of and experience with the downstream software and the upstream software performs perfectly in testing, the responsibility attribution is consistent with the Fixed History Model. If there was some flaw with the upstream software and it still made it through the tests of the downstream software, then the developer of the upstream software is responsible for DMAs the upstream software contributed to (the Chained History Model). For this feature to come into consideration, the upstream software needs to behave in a way that is consistent with its specification, pass all of the tests of the upstream software, and still contribute to a DMA. There are perhaps very few of these sorts of situations. The downstream developer may have not completely understood the specifications of the upstream software.

The upstream developers may have not had a clear statement of those specifications. Due to the narrowness of when this feature might apply, it does not warrant inclusion as a feature in the analytical system.

4. CONCLUSION

A possible criticism of this paper is "so what?" Why does it matter which diagram you select to model responsibility for software use in a particular project? We offer two responses to the "so what?" challenge:

First, we maintain that the exercise of applying one or both models to a project will require developers (and their managers) to explicitly engage with questions of responsibility and accountability. These questions are important to developing software with integrity, and we would be pleased if the introduction of these models encouraged that kind of engagement since they demonstrate these interactions graphically. This analytical system will help developers be more transparent and make complex systems more explainable to their moral patients. The examples show that the analytical system for downstream responsibility attribution can clarify thinking about responsibility for different kinds of software that is consistent with at least a small sampling of the scholarship in computing ethics. Each of the models captures distinct attitudes about responsibility for software: The Chained History Model requires a long range, deep view; the Fixed History Model allows a shorter range, more shallow view. In some of the examples above, choosing one model or the other may be different for different upstream pieces of a system: in some projects, the Fixed History Model is suggested for one subsystem, and the Chained History Model is suggested for another subsystem. This variation adds to total system complexity. This analytical system aids developers who are trying to sort out this complexity. We expect that in thinking about the two models with respect to the system they are building, they will be more likely to clarify and communicate their judgments about what level of responsibility they claim for downstream uses of their software.

Second, we hope that an emphasis on transparency, responsibility, and accountability (as in Vakkuri et al., 2019) could, in time, make it mandatory for software developers to adopt a rigorous process of what we call "software provenance." This process would make it clear (perhaps in a separate document, but more easily as part of commenting in source code) what subsystems were developed by whom, and when the software started and was revised. With such information available, especially if it were encoded formally, automated tools for tracing responsibility and accountability could easily be established.

If software provenance become widespread (either by custom or regulation), the problem of many hands could be greatly simplified using automated tools. Software developers, vendors, and other stakeholders could not hide behind the obscurity of software's history. The enforced transparency of software provenance might significantly change the attitude of all stakeholders about the accountability of software professionals for the impact of their work.

REFERENCES

Arora, C. (2019). Digital health fiduciaries: protecting user privacy when sharing health data, Ethics Inf Technol 21: 181. https://doi.org/10.1007/s10676-019-09499-x.

Coeckelbergh, M. (2019) Artificial intelligence, responsibility attribution, and a relational justification of explainability” Science and Engineering Ethics 1-18, https://doi.org/10.1007/s11948-019-00146-8. Published online October 24, 2019.

De Bruin, B. & Floridi, L. (2017). The ethics of cloud computing, Sci Eng Ethics 23, 21–39. https://doi.org/10.1007/s11948-016-9759-0.

Introna, L.D. (2016) Algorithms, governance, and governmentality: On governing academic writing, Science, Technology, & Human Values. 40(1), 17-49.

Miller, K. (2007). Open source software and consequential responsibility: GPU, GPL, and the no military use clause. In P. Boltuc (Ed.), APA Newsletter of philosophy and computers, 6(2), 17–22.

Tasioulas, J. (2019). First steps towards an ethics of robots and artificial intelligence, Journal of Practical Ethics. 7(1), 49-83.

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study. In International Conference on Product-Focused Software Process Improvement (pp. 331-338). Springer, Cham. https://doi.org/10.1007/978-3-030-35333-9_24

Wolf, M. J., Miller, K. W., & Grodzinsky, F. S. (2009). On the meaning of free software. Ethics Inf Technol 11(4), 279.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's Tay experiment, and wider implications. ACM SIGCAS Computers and Society, 47(3), 54-64.

Wolf, M. J., Miller, K. W., & Grodzinsky, F. S. (2019). On the responsibility for uses of downstream software," Computer Ethics - Philosophical Enquiry (CEPE) Proceedings: 2019, Article 3. https://doi.org/10.25884/7576-wd27

ONCE AGAIN, WE NEED TO ASK, “WHAT HAVE WE LEARNED FROM HARD EXPERIENCE?”

William Fleischman, Jack Crawford

Villanova University (USA)

william.fleischman@villanova.edu; jcrawf15@villanova.edu

ABSTRACT

In this paper, we discuss the disconcerting structural similarities between the series of radiation therapy accidents caused by the Therac-25 in the 1980’s and the accidents and near-accidents involving the Boeing 737 Max aircraft in 2018 and 2019. These similarities concern engineering and software design, testing, hazard analysis, documentation as well as responses to accident reports. Considering the lapse of time between the two cases and the publicity attending publication of the 2017 revision of the ACM Code of Ethics, we reflect on the role of codes of ethics in computing and make several suggestions of measures that might enhance their effectiveness.

KEYWORDS: System failure, safety-critical software, engineering design, codes of ethics.

1. INTRODUCTION

The series of accidents caused by the Therac-25 computer-controlled radiation therapy machine is one of the most carefully studied and widely cited cases of accidents involving poorly conceived safety-critical software and deeply flawed engineering design. The Therac accidents are commonly and justifiably used as a fundamental case study in university courses in computer and engineering ethics and system safety. Although the accidents occurred more than thirty (and the design process more than forty) years ago, every aspiring computer or engineering professional should reflect on the deficiencies in software and engineering design, testing, safety analysis and documentation related to this case. In addition, they should think deeply about the ineffective and frequently dishonest responses of the device manufacturer to reports of harm to patients treated with the machine.

The currency of this case study is underscored by the striking similarities between the factors identified by Leveson and Turner (1993) in their investigation of the Therac-25 accidents and those, revealed in investigative articles in the recent press, relating to the contemporary series of accidents and near accidents involving the Boeing 737 Max aircraft.

In this paper, we begin by laying out in detail corresponding elements material to producing the harms that occurred in each of the two cases. We then reflect on the role of codes of ethics in the face of the persistence of identifiable patterns of unethical behavior and practice.

2. ONCE UPON A TIME … AND, DISCOURAGINGLY, ONCE AGAIN

There is a common ground circumstance that links the cases of the Therac-25 and the Boeing 737 Max. Each involved the re-design of a system that, because of an engineering decision, required the introduction of new safety-critical software. Design changes were significantly driven by economic factors. Subsequent similarities extend to deficiencies in testing and documentation, failure to consider carefully the "ecology of use" of each system, and, most disturbingly, a pattern of evasion and dishonesty by company personnel in response to reports of problems.

2.1. Engineering Decisions and Safety-Critical Software

The Therac-25, a radiation therapy machine or linac (linear accelerator) used in the treatment of cancer, was developed by Atomic Energy of Canada Limited (AECL). It was the successor to similar devices previously developed jointly by AECL with a French partner, CGR, under a collaboration agreement that had recently been discontinued. The Therac-25 had several novel features that made it an attractive investment for hospitals and cancer treatment centers. It was a single device that could deliver therapeutic doses of radiation in either electron or X-ray mode. Because of innovations in beam technology, it was a more compact device, yet able to deliver therapy at higher energy than earlier models. A further considerable economy resulted from the decision to substitute software for costly mechanical interlocks as the means of ensuring safe operation. (Leveson & Turner 1993)

Responding to innovations by Airbus, its main competitor in the passenger plane market, Boeing sought to redesign its 737 aircraft. The re-design involved replacing the engines common to earlier 737 models by more efficient, but larger and heavier, new engines. However, the size of the new engines dictated that their mounts be moved upward and forward on the wing in order to provide safe ground clearance for take-off. (Campbell, 2019) This created a problem in aerodynamics that Boeing engineers decided to solve through changes to the software system that controls flight characteristics of the plane. The supplemental software implementing this functionality was called the Manœuvering Characteristics Augmentation System (MCAS).

In each of these cases, the complexity of the control problem and the difficulty of developing a solution in software were vastly underestimated.

2.2. The Role of Economic Considerations and Governmental Regulation

For the Boeing 737 Max, as earlier for the Therac-25, there was a "frictionless" path to approval for commercialization under government regulations. All that was required was an affirmation by each of the companies that the systems were effectively identical to the earlier models that they replaced. In effect, this meant an assertion that performance characteristics of the systems were not affected by the inclusion of safety-critical software to replace or supplement existing control features.

The Therac-25 was approved for use under the FDA's (U. S. Food and Drug Administration) pre-market notification process rather than the more rigorous and time-consuming process of pre-market approval in the early 1980's (Leveson & Turner, 1993), a time at which the argument that hardware safety features of the device could be replaced by software perfectly equivalent in function might plausibly, if mistakenly, have been advanced.

For the 737 Max, the “frictionless” path involved retaining the 737 designation (hence the name Boeing 737 Max) so that pilots already certified to fly earlier 737 models would not be required to undergo lengthy and expensive training to be recertified on a new aircraft. However, this decision implicitly involved the assumption that changes to the aircraft’s flight control software were of such a minimal nature that pilots would need only a short, self-administered computer course instead of expensive classroom time and training on a flight simulator to be properly prepared to fly the new plane. (Campbell, 2019)

2.3. Deficiencies in Testing and Hazard Analysis

The perfunctory nature of testing and hazard analysis performed for the Therac-25 by AECL engineers was brought to light by the careful investigation of Leveson and Turner (1993). By means of review of proprietary documents and interviews with personnel involved in the development process, several teams of journalists (Campbell, 2019; Gates & Baker, 2019; Nicas, Kitroeff, Gelles & Glanz, 2019) have uncovered similar shortcomings in the testing and analysis of hazards associated with the 737 Max software modifications.

2.4. Deficiencies in Documentation and Failure to Consider the Ecology of Use

The Therac-25 study cited glaring deficiencies in documentation both for internal purposes and in presentation of information to operators of the device who were provided only cryptic error messages and uninformative user manuals. Leveson and Turner (1993) remark tartly, “Software specification and documentation should not be afterthoughts.”

Campbell (2019) and Gates and Baker (2019) relate how the pressure to certify the new aircraft and preserve the 737 type certificate led to serious deficiencies both in regard to documents filed with the Federal Aviation Authority (FAA), the cognizant regulatory agency, and, critically, in information provided for the training of pilots. Most seriously, Boeing introduced, and subsequently modified, a critical software control feature, MCAS, about which there was no mention in training bulletins prepared for pilots.

The effect of these deficiencies is magnified because of the disregard they indicate for the “ecology of use” of safety- and life-critical systems. In the case of the Therac-25 the lack of concern for providing meaningful information to operators who were low-level hospital personnel led to a pattern of tolerance of numerous apparently innocent machine malfunctions that infected hospital discipline at all levels. The resulting poor culture of care was implicated in all of the radiation overdoses inflicted on patients.

By contrast, pilots who fly the 737 Max are highly trained professionals. Nonetheless, the absence of information about changes to the aircraft’s flight control software left them in uncharted waters under severe time pressure when that software mistakenly started to force the nose of the plane downward. The testimony of experienced pilots to the U.S. Congress was that “the terror and tumult of such a moment would defeat many of the world’s best pilots,” and “I can tell you firsthand that the startle factor is real, and huge.” (Laris, 2019) Although a recent article blames the accidents on inexperienced pilots hired by newly chartered economy airlines (Langewiesche, 2019), Boeing knew of these practices and should have taken even more care in providing for this altered ecology of use.

2.5 Deficiencies in Response; Denial of Fault

After the first accident, in response to questions from the attending radiation physicist as to whether the Therac-25 was capable of malfunctioning and burning his patient, AECL personnel insisted this was impossible. Rather than investigating carefully, they maintained this posture and asserted that no one had been injured by the Therac-25, as one accident followed another, until there was definite confirmation that a patient had been burned by the device.

In a similar manner, Boeing denied that its control software was implicated in the first accident even though there was circumstantial evidence from a flight the day before that the software could create the conditions for precipitating loss of the aircraft. They attempted to conceal knowledge of the flaw in their control system as they tried to repair it. However, a technical bulletin posted on the company's online portal for pilots and airlines, that made oblique reference to the conditions of the first accident without directly identifying MCAS, generated such a volume of angry demands for additional information that Boeing had to admit there was something fundamentally wrong with the aircraft and name and explain the nature of the faulty modification. (Campbell, 2019)

2.6. Deficiencies in Response; Unfounded and Irresponsible Claims

In one notorious episode involving the Therac-25, AECL engineers inspected an accident site, could not replicate the conditions that produced the malfunction, but speculated as to a possible cause. AECL engineered a "fix" for the hypothetical fault and then made the preposterous announcement that "analysis of the hazard rate of the new solution indicates an improvement over the old system of at least *5 orders of magnitude* [emphasis added]." (Leveson & Turner, 1993)

In its safety analysis of MCAS, Boeing identified several possible failure modes. One particular mode associated with the actual accidents was designated as "hazardous," something with the potential to cause serious or fatal injuries to a small number of people not, however, resulting in loss of the plane itself (a "catastrophic" failure.) Boeing calculated the probability of this failure as approximately one every 223 trillion hours of flight. Gates and Baker (2019) remark acidly, "In its first year in service, the MAX fleet logged 118,000 flight hours. On the basis of this analysis, Boeing downgraded the number of sensors required to confirm the condition from two to one, thus creating circumstances that increased dramatically the likelihood of this failure due to malfunction of or damage to the single sensor involved.

In cases of such breath-taking obtuseness, it is difficult to say which is the greater fault – failure to analyse correctly the true sources of danger, or the recourse to incomprehensibly large numbers to provide a misleading quantitative "fig leaf" of rationality for an unfounded assurance of safety.

2.7. Pressures Generated by the Rush to Market

Although Leveson and Turner do not comment directly on this, the fact that earlier Therac models had been collaboratively developed suggests that AECL may have been anxious to pre-empt possible competition from CGR by expediting the introduction of the Therac-25. This would explain in part the deficiencies in testing, hazard analysis and documentation they uncovered.

In the case of the Boeing 737 Max, these pressures were explicit. Boeing rushed to prevent Airbus from securing market dominance in sale of single-aisle passenger aircraft. Several authors detail the accelerated pace of releasing blueprints and pressures on all aspects of engineering, software development, and testing, all citing internal sources at Boeing. Even an experienced commercial pilot for a major airline was aware of the effects of Boeing's procrustean efforts to preserve the 737 type certificate for the new aircraft. (Campbell, 2019; Gates & Baker, 2019; and Travis, 2019)

2.8. Latest Developments As We Go to Press

With the Therac-25, we have the benefit of closure. In the course of the series of accidents, several software errors that were directly connected to the instances of severe radiation overdose were uncovered. In reaction to the evasive and unsatisfactory responses of AECL, a Therac-25 user's group was constituted with the goal of providing more transparent communication between users and AECL as well as for the timely exchange of information about protective measures adopted by individual treatment centers. Eventually, the combined pressure of the Therac-25 user's group, and the regulatory agencies of the U.S. and Canadian governments forced AECL to address and correct the serious problems affecting the operation of the Therac-25.

In the case of the Boeing 737 Max accidents, the story is still unfolding. The 737 Max has been taken out of service and is undergoing extensive study, modifications, and testing in order to establish its safety as a precondition for recertification. During this process, three new software problems have been uncovered, one of which, though unrelated to MCAS, "could cause the plane to dive in a way that pilots had difficulty recovering from in simulator tests." (Savov, 2019) (Pasztor, 2020) (O'Kane, 2020)

One recent development, however, establishes another noteworthy point of correspondence with the case of the Therac-25. After steadfastly resisting this measure throughout the process of development, marketing, and introduction into commercial service, Boeing has finally admitted that simulator training is recommended and will be necessary for 737 Max pilots once the plane is again declared safe to fly. (Hawkins, 2020)

3. A MYSTERY

Throughout the period of investigation into the background of the Boeing accidents, there was one question that continually and insistently arose in our thoughts. Why was it, how could it be that at no point in the design process for the 737 Max did anyone highlight the danger of relying on a single sensor to indicate the problematic nose-up flight condition that engineers feared might cause the plane to stall and, if not corrected, cause the loss of the aircraft?

In fact, as the result of congressional investigation and tenacious reporting by several teams of journalists, leading to disclosure of internal Boeing documents, we now know that this problem was broached three years before the 737 Max crashes by at least one individual, someone identified only by his or her function as an Aero-Stability & Control advisor. (Helmore, 2019) Although we have inquired of the author of that report as to the specific engineering or computer background of this individual, we have yet to receive a response (and reluctantly conclude that none will be forthcoming.) In the absence of any further information, we consider that it is most likely that he or she is an aeronautical or mechanical engineer.

That this warning, three years in advance, about the eventual proximate cause of two fatal 737 Max crashes appears to have dropped like a stone to the bottom of the ocean seems consistent with reporting on hundreds of additional pages of internal documents recently disclosed to congressional investigators that “reveal chaos and incompetence at 737 Max factory” and indicate that Boeing executives themselves “mocked their [FAA] regulator, joked about safety and said the Max had been ‘designed by clowns.’” (Rushe, 2020) One could be forgiven for regarding with just the most infinitesimal grain of skepticism the company’s latest attempt to salvage the aircraft’s reputation through a series of videos currently being produced maintaining that Boeing is committed to a culture of aviation safety. (Horton, 2019; Boeing, 2020)

The danger with this facile characterization of Boeing as a corporation gone (temporarily) rogue is that it induces forgetfulness of the fact that actions of ethical weight and serious human consequence are not taken by corporations but by individual human beings. Further, in this case, such a perspective obscures an important question relating to shared responsibilities of technical professionals.

4. REFLECTIONS ON CODES OF ETHICS

If we make the most favourable assumptions about the motivation and actions of computing professionals, comparison of the Boeing and Therac-25 cases suggests that at least some of the problems they encounter are rooted in the nature of their interactions with engineers and professionals from other disciplines, the latter possibly acting in a supervisory role. The revised ACM Code of Ethics does speak of the duty to report risks, but in a document of many words, emphasis on this point is lacking and there is no explicit reference to the type of risks that arise in the interactions to which we have alluded.

We believe computing professionals have certain explicit proactive responsibilities in regard to system development tasks they implement where specifications are set by other engineers and managers. **“Are you really asking me to write code to implement a potentially catastrophically dangerous manoeuvre of the aircraft on the basis of input from a single, fragile and easily compromised sensor?”** This, ultimately, would seem to be the critical, unasked question that ought to have been raised by the software development team working on MCAS. Certainly, if members of this team were appropriately informed – either through direct participation in the engineering discussions concerning the potential problem of stalling caused by the aerodynamic characteristics of required upward and forward placement of the new larger and heavier engines, or through documentation in the software specifications indicating the nature of the sensor input that would trigger the dangerous nose-down manœuver – then a large measure of blame for the loss of life in the Boeing accidents accrues to this team for failure to insist that this single point-of-failure condition be modified.

The responsibility is more subtle, and the weight of blame shifted dramatically toward the aeronautical and mechanical engineers responsible for the engineering design of MCAS, if the nature of this critical single point-of-failure design element was concealed or omitted in the specifications provided to the team responsible for writing the code that implemented this feature of MCAS.

Here, the affirmative responsibility would seem to require active questioning on the part of the software development team concerning the engineering considerations and calculations on which the specifications were based. In light of the specious reasoning and the preposterous quantitative claim regarding the likely occurrence of the relevant failure mode (see sub-section 2.6), it is far from certain that questioning by the software engineers would uncover the danger. On the other hand, the requirement of articulating the reasoning underlying the engineering specifications to an audience having an active interest in understanding potential hazards might well lead someone to say, **"Wait! Are you really asking me to write code to implement a potentially catastrophically dangerous manoeuvre of the aircraft on the basis of input from a single, fragile and easily compromised sensor?"** or even, **"Wait! Are we really asking you to write code to implement a potentially catastrophically dangerous manoeuvre of the aircraft on the basis of input from a single, fragile and easily compromised sensor?"**

At any rate, the scenario depicted in the preceding paragraph – in cases where software engineers are developing life- or safety-critical systems based on specifications laid down by engineers and professionals from other disciplines – strikes us as identifying a condition that merits prominent mention and an unequivocal statement to the effect that computing professionals have an affirmative responsibility to insist on appropriate disclosure concerning potential critical points of failure in code they are commissioned to write. Given the nature of the tasks they are likely to be given, we think inclusion of a warning of this nature is essential in order for a code of ethics to have any meaning or force, at least in regard to holding paramount the safety and well-being of the public.

If the objection is raised that such a condition is unrealistic at the contemporary interface of engineering and software engineering practice, the sceptical observer might be forgiven for advancing the opinion that the current diffuse, word-logged, and self-apologetic ACM Code of Ethics has very little value in regard to what one would assume is its most critical purpose: protecting the health, safety, and welfare of the public.

We believe further that including illustrative case studies highlighting these types of interactions, with explicit reference to real world experience, would be useful in alerting computing professionals to the pitfalls they are likely to encounter in such situations. We suggest an approach similar to that of the American Society of Civil Engineers (ASCE) whose code of ethics website includes a sidebar featuring persuasive case studies based on the real experiences of practicing civil engineers. (American Society of Civil Engineers, 2017) In the present context, for example, a compact comparison of common problematic factors in the Therac and Boeing cases might be effectively presented among the case studies accompanying the ACM Code of Ethics.

5. GOVERNMENT IS THE PROBLEM

Ever since Ronald Reagan told us that "The government's not the solution. The government is the problem," and "The most terrifying words in the English language are: I'm from the government and I'm here to help," we have been conditioned to take as the default assumption that, to the extent possible, government regulation should be curtailed. At the present moment in the U.S., we are in the throes of a veritable bacchanalia of disparagement and nullification of long-standing governmental regulations designed to protect the air we breathe, the water we drink, the land we inhabit (including public lands set aside for the enjoyment of multitudes), and the climate in which we and our children and grandchildren will live. The scientific basis for the abolition of these inconvenient regulations is abundantly clear: Greed is good, especially when it concerns the interest of extractive industries and despoilers of the public treasure.

We would like to suggest that these two histories of fatal accidents involving badly conceived technology (specifically, software) tell a different story and provide a badly needed corrective to the reflexive mantra, "Government is the problem." Each, in its own way, reveals something noteworthy about the limitations and virtues of government regulation and the hazards of a regime of laissez faire.

5.1. Government Regulation and the Therac-25: the FDA and the CRPB

As we have already noted, the Therac-25 was approved for use under the FDA's pre-market notification process rather than the more rigorous and time-consuming process of pre-market approval in the early 1980's (Leveson & Turner, 1993). This meant that all AECL had to do was establish or assert that it was substantially equivalent in safety and effectiveness to a product already on the market. In the wake of the Therac-25 accidents, the inadequacy of this regulatory protocol was clearly understood. The idea that one could replace protective measures built around hardware interlocks by software controls and, by means of software alone, seamlessly achieve the same degree of safety, is inconsistent with our current understanding of the limitations of software.

The relevant question about the adequacy of the regulatory process when the Therac-25 was approved for use is, "Where, in the early 1980's, would one have expected to find the expertise necessary to evaluate an application made under pre-market notification in which the safety capabilities of software controls were equated with those of hardware interlocks?" The realistic answer is that individuals with expertise of that nature would be much more likely to be working in industry or private enterprise than for a regulatory agency of the government. From this standpoint, the balance of responsibility would seem to rest more heavily on the manufacturer, AECL. In applying to have the Therac-25 approved under pre-market notification, its engineers asserted a questionable equivalence that they were in a better position to understand than regulatory personnel. If, in fact, they did not appreciate the risk, this was a lapse which they subsequently compounded through their dismissive responses to suggestions that the Therac-25 had caused injuries to several patients. If, however, they were aware of that risk, their actions were not merely irresponsible but criminal as well.

It is noteworthy that the one individual who, according to every account, "got the problem right" from the very beginning was Gordon Symonds, head of the division of advanced X-ray systems of the Canadian Radiation Protective Bureau (CRPB). Perhaps the nature of the CRPB, having a narrower focus than the FDA, whose mandate covers broad and disparate aspects of public health and safety, made it more likely that there would be someone on its staff in a position to identify the source of the problem with the Therac-25.

In his reports on the second Therac accident, which occurred at the (Hamilton) Ontario Cancer Foundation, Symonds expressed prudent skepticism about the advisability of entrusting safe operation of the Therac-25 to software alone and, at that early moment, made several of the crucial recommendations for ensuring safe operation of the device that were eventually incorporated (two years and three patient deaths later) into the final Corrective Action Procedure (CAP) that was imposed on the manufacturer by Canadian and U.S. authorities and the Therac-25 Users Group before its use in radiation therapy was again permitted. (alternatively, "before the recall and suspension of use was lifted.") (Leveson & Turner, 1993)

At any rate, "[o]nce the FDA got involved in the Therac-25, their response was impressive, especially considering how little experience they had with similar problems in computer-controlled medical devices." (Leveson & Turner, 1993)

5.2. Government Regulation and the Boeing 737 Max: the FAA

In the present moment of wholesale abandonment of a commitment to the value of governmental regulation, it is difficult to imagine a story that more dramatically conveys the irresponsibility of this stance. (On second thought, there are probably many equally illuminating stories that carry the same urgent warning but since they mainly involve poor children affected by toxic levels of lead in urban water supplies, or economically disadvantaged people, often predominantly people of color, living in the vicinity of waste fields where cancer-inducing toxic chemicals or coal ash are routinely buried under current, permissive "environmental" regulations, these stories don't rise to the same urgent level of concern as the Boeing accidents for important people like us for whom air travel is a necessity of life.) (Lartey & Laughland, 2019) (Costley, 2020)

Given the unequal balance in expertise between Boeing and the FAA, it has long been customary for some regulatory activity to be delegated back to qualified Boeing personnel. This was an arrangement that worked well as long as the priorities of the regulator and manufacturer were satisfactorily aligned – concern for the safety of the public being of paramount interest to both parties.

However, when reductions in funding for most U.S. regulatory agencies began to affect personnel levels at the FAA and their capacity to adequately oversee the full range of regulatory activity, the balance of responsibility began to shift in significant degree toward *de facto* self-regulation by Boeing. In this process, the reciprocal respect concerning competence and integrity that had existed between Boeing and the FAA based on a shared commitment to the safety of the public must have begun to erode in dramatic fashion. How else explain language like that found in reports describing Boeing employees expressing open contempt for regulatory personnel and process? (Rushe, 2020)

The detrimental effect of the FAA policy of acquiescence in *de facto* Boeing self-regulation is further shockingly demonstrated by its failure to act on an in-house analysis after the first fatal 737 Max accident that anticipated the likelihood of additional tragedies. "US regulators allowed Boeing's 737 Max to keep flying even after their own analysis found the plane could have averaged one fatal crash about every two or three years without intervention." (Rushe, 2019)

5.3. Where Does This Leave Us?

Given the clear evidence of the failure of a scheme that privileges self-regulation over rigorous governmental scrutiny, it is difficult to pronounce even a neutral judgment on the relentless dismantling of government oversight in the interest of public safety. "Nor is oversight likely to get much of a boost from the Trump administration. Donald Trump has used two executive orders to cut regulatory oversight and hand more of that supervision over to businesses. Trump's 2019 budget proposed an 18% cut to the transportation department." (Rushe, 2020)

Let us try to put a fine point on it: This is criminal recklessness on the part of the author of this policy, but also criminal dereliction of responsibility on the part of the U.S. Congress in failing to act to reverse such a dangerous course. In the end, someone has to say, amending in just the subtlest shading the timeless wisdom of Ronald Reagan, "**Government is indeed the problem – when it is a matter of government by corrupt, venal, foolhardy imbeciles**."

REFERENCES

American Society of Civil Engineers (2017). ASCE Code of Ethics. Retrieved from https://www.asce.org/code-of-ethics/

Association for Computing Machinery (2018). ACM Code of Ethics and Professional Conduct. Retrieved from https://www.acm.org/code-of-ethics

Boeing Corporation (2020, January 10). 737 Max Updates: Culture of Safety. Retrieved from https://www.boeing.com/commercial/737max/737-safety.page

Campbell, D. (2019, May 2). Redline: The many human errors that brought down the Boeing 737 Max. *The Verge*. Retrieved from https://www.theverge.com/2019/5/2/18518176/boeing-737-max-crash-problems-human-error-mcas-faa

Costley, D. (2020, January 9). The Guardian. The blackest city in the US is facing an environmental justice nightmare. Retrieved from https://www.theguardian.com/us-news/2020/jan/09/the-blackest-city-in-the-is-us-facing-an-environmental-justice-nightmare

Gates, D. & Baker, M. (2019, June 22). The inside story of MCAS: How Boeing's 737 MAX system gained power and lost safeguards. *Seattle Times*. Retrieved from https://www.seattletimes.com/seattle-news/times-watchdog/the-inside-story-of-mcas-how-boeings-737-max-system-gained-power-and-lost-safeguards/

Hawkins, A. (2020, January 7). Boeing will recommend simulator training for pilots of its troubled 737 Max jets. *The Verge*. Retrieved from https://www.theverge.com/2020/1/7/21055367/boeing-737-max-pilots-simulator-training-recommend

Helmore, E. (2019, October 30). Boeing employee raised concern over Max sensor three years before crashes, email shows. *The Guardian*. Retrieved from https://www.theguardian.com/business/2019/oct/30/boeing-hearings-dennis-muilenburg-737-max-sensor

Horton, W. (2019, December 26). Boeing promotes 737 Max safety to the public, where 40% don't want to fly on a Max. *Forbes Magazine*. Retrieved from https://www.forbes.com/sites/willhorton1/2019/12/26/boeing-promotes-737-max-safety-to-the-public-where-40-dont-want-to-fly-on-a-max/#32e5d2137416

Langewiesche, W. (2019, September 18). What Really Brought Down the Boeing 737 Max? *The New York Times Magazine.* Retrieved from https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html

Laris, M. (2019, June 19). Changes to flawed Boing 737 Max were kept from pilots, DeFazio says. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/trafficandcommuting/changes-to-flawed-boeing-737-max-were-kept-from-pilots-defazio-says/2019/06/19/553522f0-92bc-11e9-aadb-74e6b2b46f6a_story.html

Lartey, J. & Laughland, O. (2019, May 6). ‘Almost every household has someone who has died from cancer.’ *The Guardian*. Retrieved from https://www.theguardian.com/us-news/ng-interactive/2019/may/06/cancertown-louisana-reserve-special-report

Leveson, N. & Turner, C. (1993). An investigation of the Therac-25 accidents. *IEEE Computer*, 26(7), 18-41.

Nicas, J., Kitroeff, N., Gelles, D., & Glanz, J. (2019, June 1). Boeing built deadly assumptions into 737 Max, blind to a late design change. *The New York Times*. Retrieved from https://www.nytimes.com/2019/06/01/business/boeing-737-max-crash.html

O’Kane, S. (2020, February 6). Boeing finds another software problem on the 737 Max. The Verge. Retrieved from https://www.theverge.com/2020/2/6/21126364/boeing-737-max-software-glitch-flaw-problem

Pasztor, A. (2020, January 17). Boeing Finds New Software Problem That Could Complicate 737 MAX’s Return. The Wall Street Journal. Retrieved from https://www.wsj.com/articles/boeing-finds-new-software-problem-that-could-complicate-737-max-return-11579290347

Rushe, D. (2019, December 19). FAA let Boeing 737 Max continue to fly even as review found serious crash risk. *The Guardian*. Retrieved from https://www.theguardian.com/us-news/2019/dec/11/boeing-737-max-plane-faa-regulators-crash-risk

Rushe, D. (2020, January 10) Boeing: internal emails reveal chaos and incompetence at 737 Max factory. *The Guardian*. Retrieved from https://www.theguardian.com/business/2020/jan/10/boeing-shocking-internal-emails-reveal-chaos-incompetence-737-max-factory

Savov, V. (2019, June 27). Newly discovered safety risk will keep Boeing’s 737 Max grounded for longer. *The Verge*. Retrieved from https://www.theverge.com/2019/6/27/18715207/boeing-737-max-faa-risk-flaw-vulnerability-problem-airworthiness

Travis, G. (2019, April 18). How the Boeing 737 Max disaster looks to a software developer. *IEEE Spectrum*. Retrieved from https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer

PERCEIVED RISK AND DESIRED PROTECTION: TOWARD A COMPREHENSIVE UNDERSTANDING OF DATA SENSITIVITY

Yasunori Fukuta, Kiyoshi Murata, Yohko Orito

Meiji University (Japan), Meiji University (Japan), Ehime University (Japan)

yasufkt@meiji.ac.jp; kmurata@meiji.ac.jp; orito.yohko.mm@ehime-u.ac.jp

ABSTRACT

This study clarifies the characteristics of the perceived risk of personal data release and the required protection as a preliminary step toward a comprehensive understanding of data sensitivity, and has long been regarded as the key parameter distinguishing data that should be protected from data that should be utilised. Few studies have empirically considered the cognitive characteristics of data sensitivity. It is essential to consider both perceived risk and desired protection when seeking a comprehensive understanding of data sensitivity. Thus, we quantitatively examined the characteristics of both components using four types of personal data (two of which are considered sensitive under Japanese law). The authors surveyed 420 Japanese subjects and analysed the results using the Friedman test and the Mann-Whitney U-test. The perceived risks and desired protections differed significantly among the four types of data, and legally defined data sensitivities did not always explain the observed differences. The extent of interest in personal data increased the perceived risk and the desire for protection. The effects of various personal factors including gender and the tendency to self-protect were relatively weak, so further analysis is required. We discuss the remaining issues and future research directions.

KEYWORDS: non-parametrical analysis, personal data, protection request, risk perception, sensitive data.

1. INTRODUCTION

We are becoming increasingly dependent on personalised online services. Every individual is part of a vast service ecosystem involving personal data collection, distribution, and storage. The more personal data is disclosed, the greater the benefit to the individual: disclosure improves quality of life as the entire ecosystem responds. However, this enhanced service convenience and quality may be accompanied by negative trade-offs including invasion of privacy, unfair discrimination, and fraud. It is difficult to maximise advantages while minimising disadvantages. The protection of certain types of personal data is the essence of good data management, and personal data sensitivity has long been regarded as key in this context. Scholars have been investigating this issue for at least 40 years. Turn and Ware (1976) used personal data sensitivity as a classification axis in a pioneering discussion of how such data should be categorised. The 2012 EU Data Protection Directive (later replaced by the 2018 General Data Protection Regulation) distinguished some forms of personal data, the disclosure of which would seriously affect fundamental human rights, from other types of personal data (European Commission 2012), and strict conditions are imposed on handling those particularly personal – 'sensitive' – data. Japan's Act on the Protection of Personal Information (APPI; revised in 2015) defines some sensitive data as yo-hairyo (special care is required). Specifically, this includes data related to race, religion, social status, medical history, any criminal record, whether an individual was a victim of crime, and anything else prescribed by cabinet order as requiring special care to preclude discrimination, prejudice, or another disadvantage (paragraph 3, Article 2, Japan APPI 2015). Thus, the concept of data sensitivity is widely used when managing personal data and respecting privacy, but few studies have empirically examined the cognitive characteristics thereof. Here, as a first step toward a comprehensive understanding of data subtleties, data sensitivity is quantitatively analysed in terms of perceived risks and required protections.

2. LITERATURE REVIEW

2.1. Personal data sensitivity

No clear consensus had been reached on a definition of personal data sensitivity. Most studies have conceptualised such sensitivity in terms of the potential negative consequences suffered by an individual if personal data were inappropriately collected, distributed, stored, or used. Turn and Ware (1976) suggested that personal data would become sensitive "when its uncontrolled dissemination may have adverse effects on the individual concerned and on his activities" (p. 303), noting that adverse effects ranged from mild annoyance to serious physical and mental harm. Others have defined personal data sensitivity or 'sensitive data' in terms of various adverse effects such as privacy risks (Sapuppo, 2012), personal identification (Malheiros, Oreibusch and Sasse, 2013), and unfair discrimination and prejudice (Japan APPI, 2015). The higher the estimated probability of negative consequences, the higher the sensitivity. Therefore, personal data sensitivity is the extent to which an individual does not want anyone to use or disclose data because of a perceived risk of negative consequences. Ackerman, Cranor and Reagle (1999) considered someone who is 'comfortable' with disclosure to be the opposite of someone who is 'sensitive' about data disclosure. Sapuppo (2012) defined data sensitivity as an unwillingness to share. Thus, data sensitivity is a cognitive and/or affective feeling. However, the person evaluating data sensitivity is not necessarily the person who 'owns' the data. Two methods have been used to evaluate sensitivity (Fule and Roddick, 2004; Al-Fedaghi, 2012). The first involves having well-trained and highly experienced experts scientifically and comprehensively explore the status quo and define what is socially acceptable; this is the principal approach used for legislation (PPC Japan, 2016). The other method involves assessing the perceptions of data subjects via quantitative or qualitative methods (e.g., surveys and in-depth interviews); data from many subjects can be aggregated to define sensitivity. Although these methods differ (Fukuta et al., 2017), they both apply the concept of data sensitivity, and use of both methods may be essential when seeking a comprehensive understanding of such sensitivity.

2.2. Conceptual structure and research tasks

Our conceptual review of personal data sensitivity revealed how such sensitivity may be structured. Sensitivity reflects the risk perceived by a data subject when his/her personal data are collected, distributed, stored, and used. The extent of perceived risk determines the subject’s attitude about data utilisation and disclosure. If a high risk is perceived, the subject wishes to keep the data secret and/or assigns high priority to data protection. It is thus important to explore how data subjects perceive risk. Subjects may request that sensitive data be protected. As mentioned above, data sensitivity is an unwillingness to allow anyone to access or use the data. Thus, data-handling entities and lawmakers will receive requests to prioritise protection over utilisation. Such requests are very important because they connect data sensitivity to data management and privacy protection. Finally, data sensitivity should be discussed at a collective, not an individual, level. Our conceptual review revealed that data sensitivity is intrinsically subjective, i.e. an attitude about personal data disclosure and use. However, given the roles played by others in data management and privacy protection, it is best to analyse the topic collectively regardless of whether sensitivity is based on expertise or a ‘general feeling.’ In recent years, online services have begun to automatically measure data sensitivity (e.g. Fule and Roddick, 2004). However, most research to date has not focused on customised data disclosure or use, and has instead focused on data sensitivities at the level of a nation, a demographic population, or a cultural group.

Here, we empirically examine the personal data sensitivities of ordinary Japanese people in terms of risk perception and protection requirements. To collect this basic material, which is required for a comprehensive understanding of sensitivity, we explore differences in perceived risk and protection requirements among four types of personal data: political orientation (e.g. party membership and political beliefs); health status (e.g. mental and physical medical histories; diagnosis and treatment records); economic/financial status (e.g. deposits, real estate holdings, and debt); and consumption (shopping history and service records). The former two types of data are defined as ‘sensitive’ by the APPI and the latter two are not. Differences in the perceived risks and protection requirements among these types of data reveal some of the cognitive characteristics of data sensitivity. We also explore the effects of personal factors (gender, interest in personal data, and a self-protective tendency) based on our expectation that data sensitivity is multi-layered.

2.3. Measurements and data collection

Perceived risk was measured using a two-component model; the risk was the product of its subjective probability and the perceived magnitude of damage if the risk eventuated (Mitchell, 1999). We took a multi-dimensional view of perceived risk. Although various risks are assumed during personal data use (Solove, 2008), the perceived risks here included only public surveillance, discrimination/prejudice, commercial use, embarrassment, and exposure to criminals. All respondents were asked to evaluate the probability of risk and the extent of possible harm if the risk eventuated. Each subjective probability was scored from 0 to 100 and converted to a 10-point interval scale (e.g. 0–9% was converted to 1). The extent of harm was measured using a six-point scale (1: “does not harm me at all” to 6: “harms me greatly”). The perceived risk score was the extent of harm multiplied by the subjective probability of risk eventuation.

The required protection level was measured using a single-item method employing a six-point scale. The question was: “To what extent do you require data-handling entities (e.g. businesses and public institutions) to rigorously manage the following personal data?” (scores ranged from 1: “no need at all for rigorous management” to 6: “absolutely must be managed rigorously”). The required protection levels were analysed using a six-point scale: “How much legal regulation do you require when the following personal data are handled?" (1: "no need for any legal regulation at all" to 6: "absolutely must be legally regulated"). We evaluated the sensitivities of political orientation, economic/financial status, health status, and consumption. For example, all respondents were asked to rate the subjective probability of public surveillance (and four other risks), the extents of harm if the risks eventuated, and the required protection levels for the four types of data. We compared the means of the perceived risk scores for the four data types. We employed a related-sample Friedman analysis of variance (ANOVA) to determine whether risk levels differed among the four data types. Use of this non-parametric ANOVA is preferable to use of a parametric test because the distributions of the perceived risk scores were extremely distorted, lying far from a normal distribution. The effects of gender, interest, and self-protective behaviour were analysed by comparing the mean scores by gender (male or female), awareness (high or low interest), and self-protective status (high or low). Interest level was assessed based on responses to two questions using a six-point scale: "In daily life, how much do you care about handling of your personal data?" (1: "I do not care about it at all" to 6: "I care very much") and "To what extent are you interested in news and articles on personal data and privacy?" (from 1: "Not interested at all" to 6: "Very interested"). The extent of self-protection was measured based on responses to two questions using a four-point scale: "Have you ever changed the privacy settings of your PC or smartphone?" and "How often do you read privacy policies when downloading software and apps?" (1: "Never" to 4: "Frequently"). We grouped respondents by quantiles in terms of interest and self-protection levels. For example, a respondent whose mean interest score was below the first quantile point was categorised as "low interest" and a respondent whose mean score was above the third quantile point was categorised as "high interest." The Mann-Whitney U-test was employed to explore whether significant differences in mean ranks were evident between pairs of groups. Effect sizes were calculated with the aid of U statistics. The Mann-Whitney test compares non-parametric means; the test power is usually less than that of the t-test, which compares parametric means. However, as noted above, non-parametric tests were more appropriate because the data were not normally distributed.

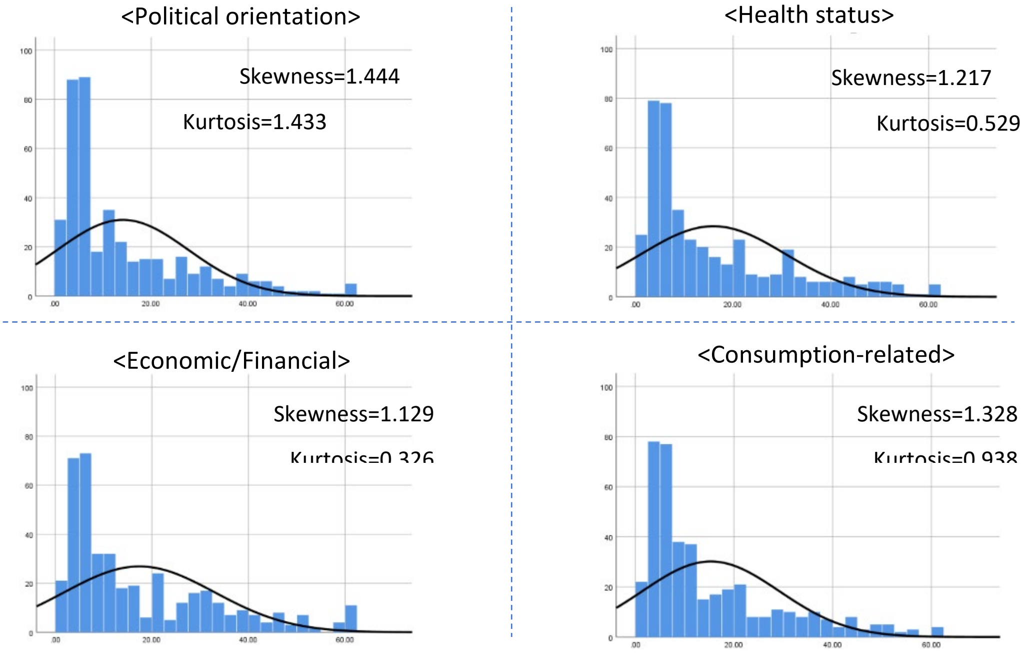
The questionnaire survey was conducted in March 2019. All 420 respondents were Japanese; we enrolled 42 males and 42 females in each of their 20s, 30s, 40s, 50s, and 60s. The questionnaire featured three sections: the first explored respondent attributes and general attitudes about privacy and personal data; the second explored the perceived risks associated with the release of four types of personal data; and the third explored the required protection levels.

3. RESULTS AND DISCUSSION

3.1. Distribution of perceived risk scores

Many previous researches on perceived risk of data release have used a single measure asking respondents about level of risk directly, and the acquired data has been applied to parametric methods such as structural equation modeling (Mitchell, 1999). However, some researchers insisted that a two-component model of risk measurement had several advantages of reliability and validity of over other types of risk measurement models (e.g., Gemünden, 1985; Mitchell, 1999). Therefore, this study adopted the two-component model for risk measurement. The histograms of perceived risk scores, shown in Figure 1, indicated that the distribution of perceived risk scores had a significant positive skew for each data type. Furthermore, the result of the Kolmogorov-Smirnov test showed that the distribution deviated significantly from normal one for each data type [Political orientation: D(420)=0.186, p=0.000; Health status: D(420)=0.179, p=0.000; Economic/Financial: D(420)=0.174, p=0.000; Consumption: D(420)=0.184, p=0.000]. Based on these results, nonparametric methods were used in the following section in order to analyse effects of data type and several personal factors on perceived risk and required protection levels.

Figure 1. Distribution of perceived risk for each data type.



3.2. Differences in risk perceptions and protection requirements among personal data types

Friedman’s test revealed significant differences in the perceived risk levels of release of the four types of data [chi-square (3)=30.364, p=0.000]. We performed pairwise comparisons to locate the differences (Figure 2). The perceived risk of economic/financial data release (mean rank=2.75) was the highest and the perceived risk of political orientation data release (mean rank =2.31) was the lowest (corresponding figures for health and consumption data were 2.51 and 2.43). The p-values for each pair (adjusted using the Bonferroni correction for multiple tests) revealed a significant difference at the 5% level in the economic/financial (ranked 1st)-health status (2nd) pair (p=0.043, z=2.686, p=0.131); the economic/financial (1st)-consumption data (3rd) pair (p=0.002, z=3.595, r=0.175); and the economic/financial (1st)-political orientation (4th) pair (p=0.000, z=4.998, r=0.244), as indicated by the thick solid arrows in Figure 2. Thus, release of personal economic/financial data was perceived as significantly riskier than release of any other data, the perceived release risks of which did not differ. We similarly explored differences in protection requirements, which can be classified into two types: data-handling entities must rigorously manage data; and lawmakers must provide legal protection. In terms of data-handling entities, Friedman’s test revealed significant differences in handling requirements among the four data types [chi-square (3)=228.205, p=0.000] and also in the requirements for legal protection [chi-square (3)= 203.312, p=0.000]. We performed pairwise comparisons to locate the differences [Figure 3a, data-handling entities; Figure 3b, legal protection].

Figure 2. Pairwise comparison of perceived risks.

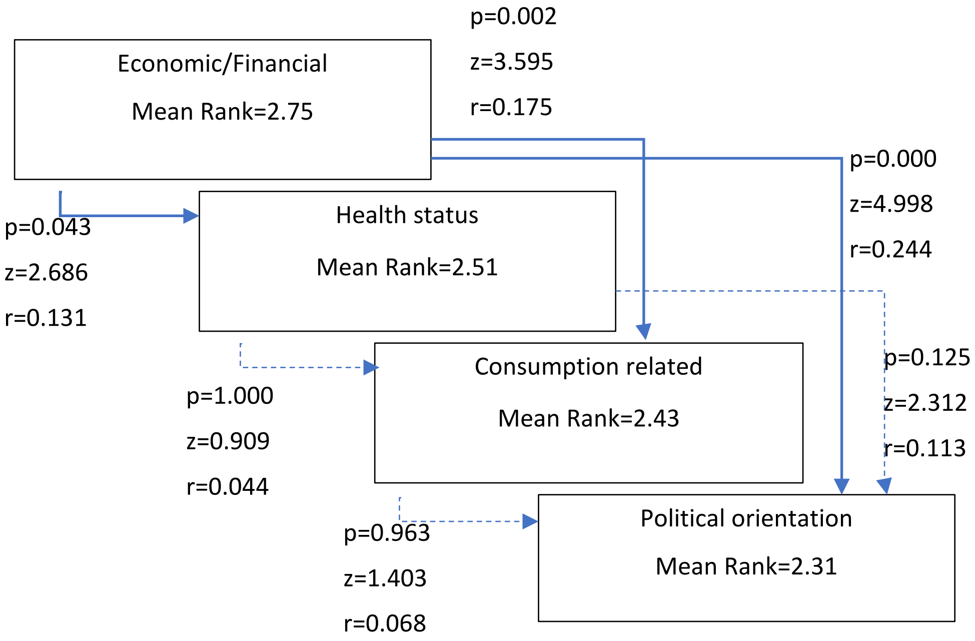
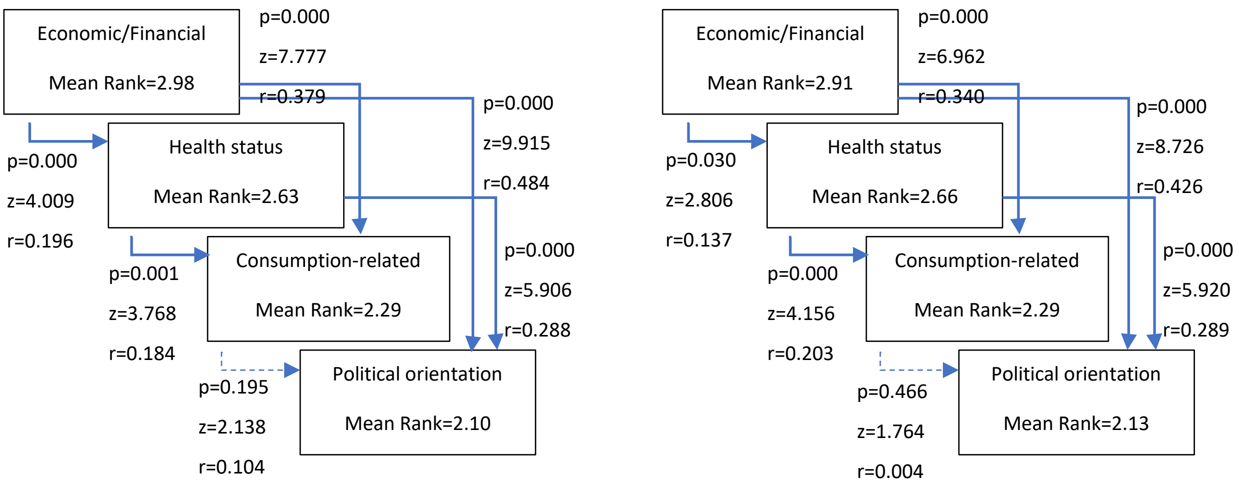


Figure 3. Pairwise comparison of protection requests.


-a. Requests for managerial protection

-b. Requests for legal protection

The results were similar. Economic/financial data required the highest level of protection (mean 2.98 for rigorous management and 2.91 for legal protection), followed by health status (means 2.63 and 2.66), consumption data (means 2.29 and 2.29), and political orientation (means 2.10 and 2.13). The p-values adjusted using the Bonferroni correction revealed that the data pairs differing significantly were identical for both forms of protection. Specifically, the required protection for economic/financial data was significantly higher than that for (second-placed) health data (p=0.000, z=4.009, r=0.196 for managerial protection and p=0.030, z=2.806, r=0.137 for legal protection); and the required protection for health data was significantly higher than that for (third-placed) consumption data (p=0.001 z=3.768, r=0.184 and p=0.000, z=4.156, r=0.203). We found no significant difference in the required protection for consumption data and (fourth-placed) political orientation data (p=0.195, z=2.138, r=0.104 and p=0.466, z=1.764, z=0.004). Thus, for both types of protection, respondents desired stronger protection of economic/financial and health data than consumption and political orientation data; the required protection for economic/financial data was particularly high.

These results suggest that perceived risk and protection requirements vary similarly by data type. Friedman’s test also revealed that these parameters were ordered: respondents were most sensitive to economic/financial personal data release, consistent with the results of a previous study showing that ordinary Japanese people were most sensitive to economic data among 13 personal data categories (Fukuta et al., 2017). The results also suggest that the perceived risks and protection requirements are not unidimensional. In previous studies on privacy and transaction risks, the perceived risks of personal data release encompassed all such data (Glover and Benbasat 2010). However, given the mean ranks and the results of pairwise comparisons, the perceived risk of release of economic/financial data and the required protection clearly differed from those of political orientation release. The effect sizes (r) for this pair were 0.244 (perceived risk), 0.484 (managerial protection), and 0.426 (legal protection). According to Cohen (1992), r=0.1 indicates a small effect, r=0.3 a medium effect, and r=0.5 a large effect. Thus, for both types of protection, the differences verged on large. The levels varied markedly, so unidimensionality was not in play. Finally, the results suggest that the levels of perceived risks and protection requirements were not always consistent with expertise-based analyses. Health and political orientation data are sensitive in the legal sense, but the data indicate that these are considered less important than economic/financial data, which are not legally protected. Moreover, clear differences were evident in terms of protection requirements, with effect sizes of about 0.3 (medium) for both. It remains unclear whether the gaps are caused by differences between evaluations that are expertise-based and those based on ‘feelings’ or by the unexpectedly weak relationship between perceived risk/protection requirements and data sensitivity. It may be necessary to redefine, or develop a new taxonomy of, sensitive personal data to replace the legal definition.

3.3. Effects of personal factors on risk perceptions and protection requirements

Previous research has revealed factors influencing perception of, and behaviours associated with, privacy and personal data disclosure. Barth and Jong (2017) systematically reviewed the privacy paradox and provided a comprehensive list of parameters that affect perceived risk of disclosure. The list includes general privacy concerns, the need for institutional trust, situational characteristics, the affective state, and perceived benefits of disclosure including economic rewards and convenience. Malhotra, Kim, and Agarwal (2004) developed a conceptual model in which several factors served as covariates of risk perception. They assumed that sex, age, and Internet experience confounded the relationships between risk perception, on the one hand, and its antecedents and consequences, on the other. The following discussion explores the effects of gender, interest in personal data, and self-protective tendencies on perceived risks and protection requirements.

1) Effects of gender: It is widely accepted that gender affects risk perception and behaviour associated with personal data disclosure (Gustafson 1998). Most studies have found that females perceive more risks than males (Siegrist 2000). The Knowledgeable Support, Institutional Trust, and Safety Concern Hypotheses have been developed in efforts to explain this difference (Siegrist 2000, Hitchicock 2001). Hypotheses focus on the effects of traditional gender roles (Freudenburg and Davidson 2007). Societal role expectations define 'good' (standard) ways of thinking and behaving, creating gender differences in terms of risk perception. In other words, gender per se may not affect risk perception; gender may interact with sociocultural factors. In general, Japanese society tends to resist changes in social norms, and traditional gender roles remain stronger than in the West. The effects of gender on risk perceptions and the protection requirements for all data types were analysed with the aid of the Mann-Whitney U-test (Table 1). In terms of risk perception, mean female ranks were higher than those of males for all data types, and all differences were significant at the 1% level (political data: U=26818, z=3.834, p=0.000; economic data: U=27424, z=4.321, p=0.000; health data: U=26699.5, z=3.738, p=0.000; consumption data: U=26877, z=3.881, p=0.000). Female respondents perceived greater risks of personal data disclosure than males. Effect sizes were all about 0.2, so gender explained 4% (the effect size squared) of the total dependent variable variance (the rank order of risk); these effects can be considered small to medium. However, the effects of gender on protection requirements were mixed. Table 1 shows that the mean ranks of female groups were all higher than those of male groups. However, no significant gender effect at the 5% level was evident for three of the eight pairs: political-managerial (U=22311.5, z=0.216, p=0.829), political-legal (U=22513.5, z=0.381, p=0.703), and consumption-legal (U=24233.5, z=1.808, p=0.071). Although the remaining five cases exhibited significant effects, the effect sizes were only about 0.1. As mentioned above, this means that gender explains only about 1% (very little) of the total rank order variance in protection requirements. Therefore, the effects of gender on protection requirements varied, but even when significant, they were small.

Table 1. Result of Mann- Whitney's U test: effect of gender.

Perceived risk	Political orientation		Economic/Financial		Health status		Consumption-related	
	Male	Female	Male	Female	Male	Female	Male	Female
Mean Rank	187.80	233.20	184.91	236.09	188.36	232.64	187.51	233.49
Sample size	210	210	210	210	210	210	210	210
Mann-Whitney U	26818		27424		26699.5		26877	
St'd Test Statistic	3.834		4.321		3.738		3.881	
Asymptotic Sig. (2-sided)	0.000		0.000		0.000		0.000	
Effect Size	0.18		0.211		0.182		0.189	
Request for managerial protection	Political orientation		Economic/Financial		Health status		Consumption-related	
	Male	Female	Male	Female	Male	Female	Male	Female
Mean Rank	209.25	211.75	194.11	226.89	199.06	221.94	197.79	223.21
Sample size	210	210	210	210	210	210	210	210
Mann-Whitney U	22311.5		25491		24451.5		24720	
St'd Test Statistic	0.216		3.026		2.017		2.215	
Asymptotic Sig. (2-sided)	n.s. (0.829)		0.002		0.044		0.027	
Effect Size	0.011		0.148		0.098		0.108	

Request for legal protection	Political orientation		Economic/Financial		Health status		Consumption-related	
	Male	Female	Male	Female	Male	Female	Male	Female
Mean Rank	208.29	212.71	194.08	226.92	196.81	224.19	200.10	220.90
Sample size	210	210	210	210	210	210	210	210
Mann-Whitney U	22513.5		25497.5		24924.5		24233.5	
St'd Test Statistic	0.381		2.959		2.409		1.808	
Asymptotic Sig. (2-sided)	n.s. (0.703)		0.003		0.016		n.s. (0.071)	
Effect Size	0.019		0.144		0.118		0.088	

2) Effects of interest in personal data handling: The level of interest in what personal data are collected and how data are handled varies, and differences in interest levels critically affect information processing. In a broad sense, it has been widely accepted that the level of interest in a thing determines the attention paid to, and the intention to learn about, the thing (Klapper 1960). According to one information processing model (the Bettman Model) of consumer behavioural research, a high level of interest incentivises integration and memorisation of relevant internal and external information (Peter and Olson, 2010). The Elaboration Likelihood Model has been used in research about communication and advertising; it suggests that people who have a high interest in, and considerable knowledge of, a certain object, tend to process information principally via a central route that imposes a large cognitive burden (e.g. the need to understand text in an advertisement) (Cacioppo et al., 1986). Together, these models suggest that highly motivated central information processing, triggered by a high degree of interest in personal data, may establish a lifelong belief in the negative outcomes of personal data disclosure. Therefore, the level of interest positively influences perceived risk and protection requirements.

We used the Mann-Whitney U-test to explore the effects of interest on perceived risk of personal data disclosure and protection requirements. Table 2 shows that the mean rank of the high interest group exceeded that of the low interest group in all 12 pairs. Furthermore, the p-values for all pairs indicated that interest significantly affected risk perceptions on disclosure of, and protection requirements for, each type of data at the 1% level. Thus, the level of interest positively affects risk perception and protection requirements, regardless of the type of personal data. Effect sizes ranged from 0.3–0.4. Based on the Cohen estimation, the effects of interest on risk perception and protection requirements were medium or greater; the level of interest thus explained 10–15% of the total variance in perceived risk and protection requirements. The effect sizes of interest clearly exceeded those of gender in all 12 cells, particularly in terms of protection requirements.

Table 2. Result of Mann-Whitney's U test: effect of interest.

Perceived risk	Political orientation		Economic/ Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	152.09	102.50	152.63	101.77	150.01	105.31	151.74	102.97
Sample size	150	111	150	111	150	111	150	111
Mann-Whitney U	11488		11570		11176.5		11436	
St'd Test Statistic	5.247		5.383		4.730		5.160	
Asymptotic Sig. (2-sided)	0.000		0.000		0.000		0.000	
Effect Size	0.325		0.333		0.293		0.319	

Request for managerial protection	Political orientation		Economic/ Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	150.12	105.16	154.36	99.44	153.28	100.89	151.96	102.68
Sample size	150	111	150	111	150	111	150	111
Mann-Whitney U	11193.5		11828.5		11667.5		11468.5	
St'd Test Statistic	4.891		6.379		5.812		5.392	
Asymptotic Sig. (2-sided)	0.000		0.000		0.000		0.000	
Effect Size	0.303		0.395		0.360		0.334	
Request for legal protection	Political orientation		Economic/ Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	150.84	104.19	155.26	98.22	155.38	98.06	154.56	99.16
Sample size	150	111	150	111	150	111	150	111
Mann-Whitney U	11301		11963.5		11981.5		11859.5	
St'd Test Statistic	5.057		6.506		6.316		6.031	
Asymptotic Sig. (2-sided)	0.000		0.000		0.000		0.000	
Effect Size	0.313		0.403		0.391		0.373	

3) Effects of self-protective behaviour: Data subjects can protect their own data independent of data-handling entities and governments. For example, smartphone network settings can be changed to self-protect data. Reading the privacy policies of products and services is another form of self-protection. The need to protect personal data triggers both self-protection and protection requests to other entities, because the perceived risk of personal data use is high. Thus, self-protection might be positively related to both perceived risk and protection requirements. Conversely, self-protection might complement or substitute for protection by other entities, reducing the perceived risk. Under such circumstances, self- protection would exhibit negative relationships with both perceived risks and the need for protection by others. The Mann-Whitney U-test yielded mixed results (Table 3). The mean rank of the high self-protection group was greater than that of the low self-protection group for all 12 pairs. However, for some pairs, no significant effect of self-protection at the 5% level was evident. Specifically, there was no significant effect on the perceived risk of health data release (U=9629.5, z=1.881, p=0.060) or legal protection of data on political orientation (U=9599, z=1.876, p=0.061). The paired effect sizes varied widely from about 0.1–0.3, and many were below 0.2. Compared to the effects of interest, the effects of self-protection on perceived risk and protection requirements were generally low.

Table 3. Result of Mann- Whitney's U test: effect of self-protection.

Perceived risk	Political orientation		Economic/Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	143.84	116.21	140.23	120.68	139.41	121.70	143.22	116.97
Sample size	145	117	145	117	145	117	145	117
Mann-Whitney U	10271.5		9749		9629.5		10182.5	
St'd Test Statistic	2.934		2.077		1.881		2.788	
Asymptotic Sig. (2-sided)	0.003		0.038		n.s. (0.060)		0.005	
Effect Size	0.181		0.128		0.116		0.172	

Request for managerial protection	Political orientation		Economic/Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	143.36	116.80	140.80	119.97	142.99	117.26	150.97	107.37
Sample size	145	117	145	117	145	117	145	117
Mann-Whitney U	10202.5		9831.5		10148		11306	
St'd Test Statistic	2.896		2.43		2.869		4.793	
Asymptotic Sig. (2-sided)	0.004		0.015		0.004		0.000	
Effect Size	0.179		0.150		0.177		0.296	

Request for legal protection	Political orientation		Economic/Financial		Health status		Consumption-related	
	High	Low	High	Low	High	Low	High	Low
Mean Rank	139.20	121.96	144.34	115.59	146.33	113.12	148.72	110.16
Sample size	145	117	145	117	145	117	145	117
Mann-Whitney U	9599		10344.5		10633		10979.5	
St'd Test Statistic	1.876		3.292		3.686		4.218	
Asymptotic Sig. (2-sided)	n.s. (0.061)		0.001		0.000		0.000	
Effect Size	0.116		0.203		0.228		0.261	

4) The effects of personal factors: Of the three personal factors examined, the effects of interest were the strongest and the most stable. Statistically significant (positive) effects were evident for all 12 pairs, and classified as medium (over 0.3) in 11. The higher the extent of interest in personal data, the higher the level of perceived risk and need for managerial and legal protection, regardless of type of data. The effects of gender were more complicated. In terms of effects on perceived risk, statistically significant effects were evident for all data types. Females tended to perceive higher risks than male. However, the effects on protection requirements were mixed: sometimes significant and sometimes not. The effect size of gender was less than that of interest; the reason for this is unclear, so more research is needed. We measured only direct effects of gender, and as noted above, it may be useful to examine the interactions between gender and personal and sociocultural factors. Similar mixed results and relatively low effect sizes were also observed for self-protective tendencies. If self-protection is viewed as a need for personal data protection, self-protection would be expected to co-vary with perceived risks and protection requirements. Conversely, if self-protection has a complementary or other relationship with data protection performed by other entities, self-protection would be negatively related to perceived risks and protection requirements. Given the small positive direct effect, the former hypothesis may be more suitable, but both hypotheses may be correct; the small effect size may reflect offsetting of positive and negative effects.

4. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Data sensitivity is the extent to which data subjects do not want anyone to know or use their personal data because of a perceived risk of negative consequences. Such unwillingness is reflected in data protection requirements. It is essential to understand the characteristics of risk perception and protection requirements; these are the essence of data sensitivity. We quantitatively evaluated these features by focusing on four types of personal data. Comparisons of the perceived risks and protection requirements revealed that these differed by data type. Expertise-based data sensitivity does not explain such variation. Thus, a personal data taxonomy with a definition of sensitive data that differs from the legal definition is required if such sensitivity is to be comprehensively understood. We examined the effects of several personal factors on perceived risk and protection requirements. We found that the level of interest in personal data exerted significant positive effects, but the effects of gender and self-protection were vaguer and weaker. In terms of gender, interaction with sociocultural factors may be in play. More study of complementary or other relationships between protection requirements and self-protective behaviour may be needed to clarify why the effects of self-protective behaviour were so weak.

We sought to clarify the characteristics of the perceived risks of personal data use and the associated protection requirements. This is the first step toward a comprehensive understanding of data sensitivity. Several more steps are required. First, we plan to complement the present work using both qualitative data and data mining. Although our quantitative results demonstrated that perceived risks differed significantly by the extent of a subject's interest in his/her personal data, additional qualitative differences may also be in play between high and low interest groups. Second, the relationships among perceived risk, protection requirements, and other aspects of data sensitivity require more attention. If a triadic relationship is in play among perceived risk, protection requirements, and unwillingness to expose personal data, clarification of this would greatly aid a comprehensive understanding of data sensitivity. Finally, data sensitivity varies socio-culturally, so international comparisons are required. Expertise-based definitions of data sensitivity (such as those employed in privacy laws) have been compared among nations, but data sensitivities among ordinary people must also be compared to appropriately balance globalisation and localisation. Overall, a comprehensive understanding of data sensitivity is required to develop appropriate data management and privacy protection systems.

ACKNOWLEDGEMENTS

This study was supported by the JSPS KAKENHI Grant-in-Aid Scientific Research (C) 17K03879 and 17K04013, and the Kurata Grants subsidised by the Hitachi Global Foundation.

REFERENCES

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*,1–8.

Al-Fedaghi, S. (2007). How sensitive is your personal information? In *Proceedings of the 2007 ACM Aymposium on Applied Computing,* 165–169.

Barth, S., & De Jong, M. D. (2017). The privacy paradox–Investigating discrepancies between expressed privacy concerns and actual online behavior–A systematic literature review. *Telematics and informatics*, *34*(7), 1038-1058.

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology*, 51(5), 1032–1043.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

European Commission (2012). Proposal for a regulation of the European parliament and of the Council- on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). 11 April 2012. Retrieved from https://ec.europa.eu/transparency/regdoc/rep/1/2012/EN/1-2012-11-EN-4-1.PDF

Freudenburg, W.R. and Davidson D.J. (2007). Nuclear families and nuclear risks: The effects of gender, geography, and progeny on attitudes toward a nuclear waste facility. *Rural Sociology*, 72(2), 215-243.

Fukuta, Y., Murata,K., Adams, A.A., Orito, Y., & Palma, A. M. L. (2017). Personal data sensitivity in Japan: An exploratory study. *ORBIT Journal*, 1(2), Retrieved from https://doi.org/10.29297/orbit.v1i2.40

Fule, P., & Roddick, J. F. (2004). Detecting privacy and ethical sensitivity in data mining results. In *Proceedings of the 27th Australasian Conference on Computer Science-Volume 26*, 159–166.

Gemünden, H. G. (1985). Perceived risk and information search. A systematic meta-analysis of the empirical evidence. *International Journal of Research in Marketing*, *2*(2), 79-100.

Glover, S., & Benbasat, I. (2010). A comprehensive model of perceived risk of e-commerce transactions. *International Journal of Electronic Commerce*, 15(2), 47–78.

Gustafsod, P. E. (1998). Gender Differences in risk perception: Theoretical and methodological erspectives. *Risk analysis*, *18*(6), 805-811.

Hitchcock, J. L. (2001). Gender differences in risk perception: broadening the contexts. *Risk*, *12*, 179-204.

Japan APPI (2015). *Act on the Protection of Personal Information* (revised in 2015) Retrieved from http://www.japaneselawtranslation.go.jp/law/detail/?id=2781&vm=&re=

Klapper, J.T. (1960), *Effects of Mass Communication*, Free Press.

Malheiros, M., Preibusch, S., & Sasse, M. A. (2013). “Fairly truthful”: The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In *International Conference on Trust and Trustworthy Computing*, 250–266.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research*, *15*(4), 336-355.

Mitchell, V. W. (1999). Consumer perceived risk: Conceptualisations and models. *European Journal of Marketing*. 33(1/2), 163–195.

Peter, J.P. and J.C. Olson (2009), *Consumer Behavior and Marketing Strategy*, McGraw-Hill Higher Education.

PPC Japan (2016). *The Proceedings of the 19th Personal Information Protection Commission of Japan*. 30 September 2016. Retrieved from https://www.ppc.go.jp/files/pdf/280930_giziroku.pdf

Sapuppo, A. (2012). Privacy analysis in mobile social networks: The influential factors for disclosure of personal data. *IJWMC*, *5*(4), 315–326.

Siegrist, M. (2000). The influence of trust and perceptions of risks and benefits on the acceptance of gene technology. *Risk analysis*, *20*(2), 195-204.

Solove, D.J. (2008). *Understanding Privacy*, Harvard University Press.

Turn, R., & Ware, W. H. (1976). Privacy and security issues in information systems. *IEEE Transactions on Computers*, (12), 1353–1361.

RESPONSIBILITY IN THE AGE OF IRRESPONSIBLE SPEECH

Benjamin Mitchell, William Fleischman

Villanova University (USA)

benjamin.r.mitchell@villanova.edu; william.fleischman@villanova.edu

ABSTRACT

We discuss the impact of language on some ethical problems surrounding the interactions of technology and society. We focus on problems of careless and irresponsible speech in the contexts of artificial intelligence and social media. As these areas are central to modern public discourse, the inappropriate use of language by computer professionals in these contexts has the potential for serious harm.

KEYWORDS: responsibility, language, artificial intelligence, machine learning, social media.

1. INTRODUCTION

The language used to express a concept is important, particularly when the concept is a new one. This is by no means a novel insight: Joseph Weizenbaum commented on this in the context of computing many years ago (Weizenbaum, 1972). In spite of this history, current public discourse suggests that a reminder and an update may be needed. From a linguistic standpoint, when a new concept is encountered, there are essentially two options; either an entirely new term can be created, or an existing term can be re-purposed. The latter is easier and more common, but when we attempt to re-purpose an already existing word into a new context, there is always some conceptual "bleed-through." Connotations of the original usage are ascribed to the new word even when they are not truly warranted.

Creating new words from scratch is difficult, and many attempts to get such terms into circulation fail. It is therefore perhaps unsurprising that the field of computing has a long history of simply borrowing conceptually linked terms and re-purposing them, rather than attempting to define new words. Even the term "computer" originally referred to a human who performed mathematical calculations as a career. But as convenient as repurposing existing terms is, there are clear hazards to doing so, and it must be the responsibility of computing professionals to ensure that the terms we use are not mis-interpreted by those who lack the background to directly understand their intended use.

Sometimes this careless use of language is unintentional, but frequently it appears to be done with malice aforethought. The dominant mode of political rhetoric in our society, for example, seems to revolve around the idea that perception is more important than truth, and that carefully selected terminology can be used to appeal to people's baser instincts while still maintaining some façade of impartiality. To take one example from a story currently dominating the news in the United States, informed reporters, government officials, and casual observers all commonly repeat the mantra "We've seen the transcript [of the phone call between the presidents of the U.S. and Ukraine that may result in articles of impeachment entered against the American president]…" In fact, there are very few individuals who have actually seen the full, accurate transcript of that call, because its potentially incendiary nature led to the "reconstructed transcript" being quickly locked down on a server in the White House's most classified computer system. But by now, everyone in the public "understands" that the transcript of that call is a matter of common knowledge. This imprecision is convenient for those who wish to dismiss the importance of the conversation since at least one national security individual, who listened in on the call as a matter of his official duties, has openly criticized the omission of crucial words and phrases in the publicly disclosed "transcript." (Barnes, Fandos, & Hakim, 2019)

Whether inadvertent, reflexive, or calculated, imprecise or careless speech can have serious consequences and influence the thoughts and actions of individuals and collectives. In this paper, we consider the dangers of such speech in two distinct contexts: First, in public understanding of the capabilities and limitations of machine learning and, more generally, artificial intelligence; and second, in the ways in which careless and irresponsible speech by prominent executives of social media companies can undercut responsible behavior by computing and information professionals and frustrate efforts to find sensible measures to regulate the practices of social media platforms.

2. ANTHROPOMORPHIZATION AND SILICOMORPHIZATION

There is a complex interplay between science, science fiction, and public perception. Artificial Intelligence (AI) has always been deeply entangled in science fictional narratives. This manifests in many ways and affects both researchers and members of the public at large. The result is that many people's reasoning about the world is based on a mythologized version of AI that can lead to dangerous conclusions.

AI researchers have a long history of underestimating the difficulty of the field's core problems. In one memorable anecdote from the founding era of the field, several prominent computer scientists estimated that ***programming a computer system to replicate all the important functionality of a human mind might take several graduate students as much as a few months to accomplish.*** Some seventy years later, we have yet to even come close. This has not stopped popular portrayals of AI from ascribing human-like behaviors and capabilities to these systems. HAL 9000 from *2001: A Space Odyssey*, Skynet from *Terminator* and the eponymous cute robot from WALL-E are just a few of the many iconic examples. Whether implacable foe or compassionate helper, these systems are presented as being "not so different from you and me." Perhaps they have certain affective deficiencies (e.g. Skynet displays a total lack of empathy for the suffering of others), but nothing outside the range of behaviors displayed by actual humans (e.g. psychopaths display a total lack of empathy for the suffering of others).

These narratives paint a highly misleading picture of the capabilities of real-world AI systems. In actuality, all "Artificial Intelligence" systems to date are just purpose-built software tools designed to automate specific and narrowly defined processes. An "AI" has less in common with a human, and more in common with a kitchen knife; both are useful tools for assisting a human to get something done faster and better, but neither has any "agency" of its own. We give human names to these systems (Eliza, Siri, Alexa, Watson, etc.), although they are no more 'human' than a toaster. We use words that imply human-like thought processes (attention, understanding, belief, etc.) as labels for simple mathematical equations and algorithms that have only loose conceptual ties to the conventional meaning of the terms. Once these anthropomorphic characterizations of “AI” are internalized by the public, sweeping extrapolations of these tools’ potential are almost inevitable, resulting in misplaced trust in the capabilities of such systems.

The flip side of this exaggerated conception of the power of “AI” is something that perhaps deserves the name “silicomorphization,” the reductive view of human intelligence and decision making based on the conflation of human intelligence with the operation of a digital computer. Naturally, in any comparison of capabilities based on this view, humanity comes off rather badly; a computer will always do a better job of being a computer. The result is a systematic denigration of the reach and richness of human intelligence and the robustness of human judgment. This devaluation of human capacity seems particularly harmful when taken as received wisdom concerning the relations between humans and machines, and the future of humanity itself. It is also a pillar of the myth of technological inevitability, which in many spheres serves to anaesthetize the conscience of those whose work involves the development and utilization of AI for purposes that are ultimately destructive of human values and human moral agency.

As we have indicated, AI provides useful tools for assisting decision making in cases where the context of the decision process is well understood and is seen to advance human well-being. But obeisance to technological inevitability in the form of unthinking substitution of automated decision-making for human judgment is fraught with danger. Once again, Joseph Weizenbaum understood the bargain: “Technological inevitability can thus be seen to be a mere element of a much larger syndrome. Science promised man power. But, as so often happens when people are seduced by promises of power, the price exacted in advance and all along the path, and the price actually paid, is servitude and [moral] impotence. Power is nothing if it is not the power to choose. Instrumental reason can make decisions, **but there is all the difference between deciding and choosing**.” (emphasis added) (Weizenbaum, 1976)

3. IRRESPONSIBLE SPEECH IN THE CONTEXT OF SOCIAL MEDIA

One important consequence of the anthropomorphization of computer systems is to make it much easier to assign *blame* to these systems when failures occur. The purveyors of such systems often encourage this type of thinking, as it absolves them of culpability when harm is done, though they are quick to take credit when the results are good. Again, however, there is nothing truly “inevitable” about this line of reasoning, and there are plenty of reasons for pushing back against this narrative.

Helen Nissenbaum (1994) urges the adoption of a robust standard of accountability for computing professionals. She rightly observes that it is no more appropriate to assign blame to a computer system than it is to assign blame to any other tool; ultimately, accountability requires moral agency, and we should no more attribute moral agency to an algorithm than we would to any other technological tool. When a Boeing 737 MAX airplane falls from the sky, we might consider a variety of humans as worth investigating for possible responsibility (the pilot, the maintenance crew, the manufacturer, etc.), but we would never allow Boeing to place the blame on the plane itself. Yet when FaceBook’s recommender system is found to be amplifying political disinformation, we are asked to believe that it is the *fault* of that algorithm, and FaceBook is merely a hapless bystander.

How can accountability survive in an atmosphere in which someone like Mark Zuckerberg can deny and distance himself and his company from one scandal after another? The examples are numerous – the misappropriation of user data by Cambridge Analytica, the dissemination of false information in its newsfeed, and the strange policies regarding whether political advertisements may contain verifiably false information, to name just a few. In testimony before the U.S. Congress, when asked directly whether a political actor could run an ad containing politically inflammatory false statements, "Mr. Zuckerberg said the platform would take down posts from anyone, including politicians, that called for violence or tried to suppress voter participation." (BBC News, 2019) This, of course, is a reply to an entirely different question. Additionally, as far as its final clause is concerned, it is demonstrably false.

Among many other examples that can be adduced of evasion, shading the truth, or proclaiming ignorance of contentious action on the part of Facebook, we can cite Zuckerberg's narrow, legalistic denial that the exfiltration of personal information concerning up to 87 million users constituted a data breach, in flagrant contempt of the common understanding of the meaning of the term. And, when asked about Facebook's contract with Definers Public Affairs, a consulting firm it used in an attempt to discredit Facebook's critics, Zuckerberg claimed that "he did not know what Definers' activities were, or who at Facebook authorized that work. Probably 'someone on the communications team,' he offered." If ignorance, or the pretense of ignorance, is an effective defense for the well-placed, why should it not be equally available to the subordinate? (Lee, 2018)

Sadly, Facebook is merely the tip of the technological iceberg. YouTube has shown similar problems with irresponsible use of algorithms promoting falsehood and political bias (Lewis, 2018), and encouraging the radicalization of users (Ribeiro, 2019). In a response on their official blog, YouTube said "Our systems are…getting smarter about what types of videos should get [recommended less], and we'll be able to apply it to even more borderline videos moving forward." (YouTube, 2019) While this is perhaps better than simply ignoring the problem, it still amounts to an instruction to ignore past failures and blindly trust them to do what's best in the future. It is also worth note that this is presented as an improvement to an already great system; at no time do they acknowledge their responsibility for harm, or even that any harm has been done in the first place. Note also that in saying that the "systems are…getting smarter," the company is using anthropomorphic language to imply that the system itself has agency and responsibility here; the company is presented as merely an assistant who is helping the system, but ultimately cannot be held accountable for the system's actions. As noted by Tufecki (Tufecki, 2015), there is every reason to believe that these big tech firms are merely the most public, and therefore most studied, examples of a far more pervasive problem.

4. SOME GUIDANCE FOR THE PERPLEXED

The complexity of the technological systems is often used to justify a disconnect between stated intentions and observed outcomes. In fact, there are many recorded cases in which a reasonable and knowledgeable observer might agree that a certain outcome could be difficult to predict; emergent properties of complex systems are notorious precisely for their unpredictability. To a non-expert, nearly any technological system can be made to seem sufficiently complex that this defense has an air of plausible deniability. Since it is impossible to prove a negative, particularly where intent is concerned, we are left with the rather sticky task of evaluating when a denial is *sufficiently* plausible, and when it stretches credulity beyond our willingness to tolerate. It should come as no surprise that the result will differ from individual to individual, making consensus building on these issues problematic.

It is easy to say that this problem is complex, and that like all social issues it is not amenable to easy quick-fixes. But as with all such problems, this claim is disingenuous. Certainly, an ideal long-term solution would likely involve a combination of legislation, education, and an overall shift in cultural norms. Yet there are straightforward steps that can be taken in the near term to both begin to ameliorate the problem and to help create the conditions necessary for a more sweeping solution further down the line.

In spite of claims to the contrary, the problem of accountability is not one which has remained unsolved in other domains. The primary goal of engineering as a discipline is to take the chaotic world in which we live and create systems which will behave in understandable and predictable ways. A skyscraper is an extremely complex artifact, but while there may be some details that are difficult to predict (e.g. why that one room is always so cold), there are a wide range of behaviors we simply will not tolerate. If a skyscraper collapses, "we didn't expect (or intend for) that to happen" is simply not a sufficient defense. The simple fact that software is a newer technology than suspension bridges, airplanes, and skyscrapers does not mean that we need to tolerate a complete lack of accountability in software systems. It is possible to specify unacceptable behaviors (for which the developer might be held strictly liable) without requiring software that is perfect and 100% free of bugs.

Similarly, we cannot allow the blame to be placed on the systems themselves; a machine learning algorithm trained on a biased data set is no more "at fault" for its poor performance than a bridge constructed on ground too soft to support its weight is to blame for its own collapse. Responsibility requires moral agency, which only humans possess. In spite of their anthropomorphic name, "artificially intelligent agents" are tools, nothing more. Until and unless we develop true artificial general intelligence (an event which has been estimated as "about 20 years away" at every point during the last 70 years), ultimate responsibility for the behavior of any system must fall to the humans who design, create, test, and deploy that system.

In his prophetic paper, "On the Impact of the Computer on Society," Joseph Weizenbaum exhorts us, as computer professionals, to recognize that "[t]he nonprofessional has little choice but to make his attributions to computers on the basis of the propaganda emanating from the computer community and amplified by the press. The computer professional therefore has an enormously important responsibility to be modest in his claims." (Weizenbaum, 1972) In the context of modern AI and ML systems, it is particularly important to be humble not only in our explicit claims, but also in the claims implicit in our choice of language. By choosing humble language over hyperbolic, we can redirect responsibility to humans and improve the public understanding of the systems at the same time. As an example, perhaps YouTube could state that its systems are "being made better at maximizing their scoring function," or perhaps that "we are changing the objective to better reflect our desires". It might require a few extra words of explanation to replace the term "smarter," but it would lead to a much more useful discourse on the matter (particularly if it led the public to question *whose* desires are being reflected by that objective function).

Until the tech giants can be held to this standard, the mid-20th century judgment of Friedrich Dürrenmatt seems uncannily pertinent to our moment in history: "In the Punch-and-Judy show of our century … there are no more guilty and also, no responsible men. It is always, 'We couldn't help it' and 'We didn't really want that to happen.' And, indeed, things happen without anyone in particular being responsible for them. … That is our misfortune, but not our guilt… Comedy alone is suitable for us." (Dürrenmatt, 1964) The comedy, alas, is often of a rather mordant nature (in which we are the bitten.)

REFERENCES

Barnes, J., Fandos, N. & Hakim, D. (2019, October 29). White House Ukraine expert sought to correct transcript of Trump call. *The New York Times*, Retrieved from https://www.nytimes.com/2019/10/29/us/politics/alexander-vindman-trump-ukraine.html

BBC News (2019, October 24), Facebook's Zuckerberg grilled over ad fact-checking policy, *BBC News*, Retrieved from https://www.bbc.com/news/technology-50152062

Dürrenmatt, F. (1964, at 31), *Problems of the Theatre*, translated by Gerhard Nellhaus. Grove Press, New York.

Lee, D. (2018, November 16), Mark Zuckerberg, missing in inaction, *BBC News*, Retrieved from https://www.bbc.com/news/technology-46231284

Lewis, P. (2018, February 2), 'Fiction is outperforming reality': how YouTube's algorithm distorts truth, *Guardian News*, Retrieved from https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

Nissenbaum, H. (1994), Computing and accountability. *Communications of the ACM*, vol. 37, no. 1, pp. 72-80.

Ribeiro, M., Ottoni, R., West, R., Asmeida, V., & Meira, W. (2019), Auditing radicalization pathways on YouTube, *arXive*, Retrieved from https://arxiv.org/abs/1908.08313

Tufecki, Z. (2015), Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency, *Journal on Telecommunications & High Technology Law*, pp. 203-218.

Weizenbaum, J. (1972), On the impact of the computer on society: How does one insult a machine? *Science*, vol. 176, no. 4035, pp. 609-614.

Weizenbaum, J. (1976, at 259), *Computer Power and Human Reason*, W.H. Freeman, New York.

YouTube (2019, June 5), Our ongoing work to tackle hate, *YouTube Official Blog*, Retreived from https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html

THE EMPLOYMENT RELATIONSHIP, AUTOMATIC DECISION AND THEIR LIMITS. THE REGULATION OF NON-UNDERSTANDABLE PHENOMENA

Enrico Gragnoli

Università degli studi di Parma (Italy)

enrico.gragnoli@unipr.it

ABSTRACT

There is currently a debate about the problem of possible limitations to the decision making of workers/employees based on automatic systems that in the end caused choices to not be made by humans. This issue has now become extremely relevant, especially in contemporary labour law, and for quite evident reasons. However, law scholars are the victims, not the protagonist of this debate since they must discuss the regulation of profiles and elements which they do not understand, not even superficially. This is the basic question: how can law interact according to reason towards decision-making methods based on mathematical resources so complex even their fundamental features are impossible to understand? This can very well be called a challenge to regulate the unknown. Can there be a rational regulation of a phenomenon incomprehensible to law and its scholars?

KEYWORDS: Automatic systems, contemporary labour law, decision-making, methods mathematical resources, decision making of workers/employees.

1. THE EMPLOYMENT RELATIONSHIP AND DECISIONS BASED ON SOPHISTICATED ALGORITHMS

There is currently a debate about the problem of possible limitations to the decision making of workers/employees based on automatic systems that in the end caused choices to not be made by humans. This issue has now become extremely relevant, especially in contemporary labour law, and for quite evident reasons. However, law scholars are the victims, not the protagonist of this debate since they must discuss the regulation of profiles and elements which they do not understand, not even superficially. This is the basic question: how can law interact according to reason towards decision-making methods based on mathematical resources so complex even their fundamental features are impossible to understand? This can very well be called a challenge to regulate the unknown. Can there be a rational regulation of a phenomenon incomprehensible to law and its scholars?

The issue does not only concern the work organised by the so called digital platforms, but also (and above all) the industrial and commercial company structures with a more traditional organisation. The continuity of the relationship and the inclusion of performance in a business context makes automatic decision paths a common feature of many employment relationships, and obviously companies exploit this feature with increasing frequency. Which regulatory strategies can be adopted with automatic decision-making systems? Barring the anachronistic prohibition of similar mechanisms, since Luddism has never had much space in Western thought, the traditional idea was to force the company to disclose the basic assumptions and premises employed by these automatic evaluation systems so as to allow the workers to understand how they are judged. The public availability of said data would help workers/employees to react accordingly, but this strategy has two obvious limitations. On the one hand, given the current structure of the civil trial, even if the company revealed the way automatic judgments were made, verifying whether their statements are true would have unrealistic costs and times and, therefore, there would be no realistic penalty for those who lie. It is difficult to calculate whether and to what extent companies are tempted to conceal their evaluation systems behind convenient descriptions. However, due to the complexity of the mathematical models and the nature of a civil law trial, even if the companies presented unreliable theses, said unreliability would never be discovered, in the ordinary proceedings of a judgment, in which a reconstruction of the setting of an algorithm is actually impossible, due to the very nature of the issue. This statement may surprise and the demonstration is not easy, especially for those who have no experience of a civil law trial. However, the conclusion is obvious, because a trial is not the place where such complex scientific questions can be explored. The referring to technical advice and consultancy follows well-established models that are however based on issues that are quite different from the installation of an algorithm and, due to their very nature, these problems cannot be defined with traditional procedural schemes.

On the other hand, it is indeed possible to discuss the protection afforded to the workers from providing information beforehand on the significant elements for the automatic assessment; the idea is trying to recreate organizational situations similar to those where the decisions entrusted only to men are made, while the technological innovation generates a completely different context. When examining an algorithm, the position of those who feel the effects of an automatic decision is not comparable to that of those who are evaluated by a human, as the latter is ruled by emotions and by that complex articulation of opinions and suggestions typical of our mind. The algorithm cannot be considered similar to a subjective decision, due to the profound difference in the way the decision is made.

The law system must accept automatic decisions for what they are, without thinking of dealing with them by forcing them to take features or components of the traditional ones. Modern technologies and the systems built with them cannot be regulated by reusing schemes and duties designed for natural persons and, first of all, any norm must be discussed on the basis of its effectiveness. Since there cannot be a decision made by a judicial body on the structure and working of an algorithm and our civil trial is not ready to face such a challenge in general (let alone at the costs reasonably affordable by an employee), the law system must openly state its acceptance of reality. Which goals can be pursued in the context of the employment relationship and especially of relationships based on continuous collaboration , so that the inevitable recourse to automatic decisions is compatible with the protection of the workers' rights and demands? Very often, when discussing this topic, many ask for a sort of "humanization" of technologies and their methods, and while this is a fascinating perspective from an ethical point of view, it is altogether not possible on the practical side.

Regardless of the fact that, no matter how one can imagine it, an algorithm is different from human reasoning, if only because the latter is never neutral in considering the elements submitted to its analysis as it is influenced by internal conflicts and by the impact of emotions, as well as of whatever element may be beyond the reach of a precise calculation, one must ask what can we actually learn about the way automatic decisions work during a civil trial - and the answer is an utterly disheartening one. Even if it were possible to set up an analysis on the algorithm and even if we overcame the momentous complexity of the issue, using all available experts and having an acceptable amount of time for the process, the costs would be unsustainable in any labour dispute, thus going against the interests protected therein. For this reason the automatic decision processes are impossible to understand especially considering the actual dynamics of the disputes, because it is not possible to clarify how these processes actually work and, in fact, the only information one has is the version provided by the companies, whose credibility cannot be verified. Everything that lies beyond the empirical analysis of the judge is presented as a mechanism that cannot be dominated, a kind of unknown terrain which, on the one hand, conditions the development of the employment relationship and on the other hand, escapes full understanding because the way it works is described by the company without anyone else knowing whether this description is actually true.

2. THE APPROACHES AND TRENDS OF EUROPEAN LAW AND THEIR STRUCTURAL LIMITS

Article 22 of the European Union regulation n. 679 of 2016 only apparently prohibits decisions based on automatic processing, since it allows them in the execution of contracts. In fact, on the one hand, art. 22, paragraph 1, has a generalist approach because it states the principle that the data subject has "the right not to be subject to a decision based solely on automated processing", a principle which is reinforced by a clarification that is difficult to interpret, whereby the prohibition operates if the decision "produces legal effects" which concern the interested party or which "similarly significantly affects him or her", with a significant space left to subsequent judicial interpretation. However, for the employment relationship, the main problem is caused by art. 22, paragraph 2, lett. a), which states that the principle of paragraph 1 is not applied if the "automated decision [...] is necessary for entering into, or performance of, a contract between the data subject and a data controller;" and this is the case of an employee, which means we can reasonably expect this provision to be applied extensively, both for the decisions taken in a pre-hiring selection phase and to those concerning the subsequent employment.

Nor can we imagine a restrictive reading of the concept of "necessity" of art. 22, paragraph 2, lett. a), because the behaviour the sentence refers to is not "indispensable" objectively but rather it is so only within the limit of a company assessment which is therefore driven by the interest of the company, albeit inherent in the stipulation or implementation of the agreement, and of this reading there are significant traces in motivation (point n. 71: "decision making based on such processing, including profiling, should be allowed," albeit regulated pursuant to art. 22). If, then, in an indeed unpredictable way, art. 22, paragraph 2, lett. a), had a selective exegesis by the judges, the condition of the workers would not improve much, since the determination based on the algorithm would still be allowed in the same way as art. 22, paragraph 2, lett. c), for which "the data subject's explicit consent" is enough.

The worker, who find themselves at the apex of social weakness at the time they enter into an employment contract (as they are finally getting a job which often they have sought for a long time) cannot exercise any negotiating power and cannot offer resistance towards the company, which requires consent to automatic decision-making processes and, according to a well known scheme in labour law, with the minimum emphasis on forms of protection based on consensual instruments. The company's interest in introducing binding evaluations based on algorithms is fully satisfied, as far as the employment relationship is concerned, by art. 22, paragraph 2, lett.

c), since its full, straight implementation makes any rigorous and restrictive interpretation of art. 22, paragraph 2, lett. a) irrelevant.

Nor can we see a significant rebalancing of the positions of the employer and the employee carried out by art. 22, paragraph 3, whereby, in the event of implementation of art. 22, paragraph 2, lett. a) and lett. c), "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" of the person involved, first of all due to the generic nature of the provision, which lacks any limiting specification (especially so when complex mathematical tools, whose functioning has no realistic possibility of being revealed in a judgment and, even less, of being understood in advance are used). The rule does not specify any constraint and, if any, refers to the judicial knowledge, so that forms of reconciliation between opposing expectations can be identified, with a deferred effectiveness over time (that is, at that time where there will be a significant jurisprudence) and limited by the impossibility of dominating the action of the algorithms.

Not surprisingly, art. 22, paragraph 3, introduces a minimum protection, indicative of the low value of the initial statement of principle, since it is expected "at least" (and the word is important, as it reveals the creation of… a sort of rear guard!) "the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision." As can be deduced from point n. 71 of the reasons, the three limitations were not designed for the employment relationship, but for commercial activities and for large-scale sales strategies, in particular electronic marketing and sales, with the identification of possible buyers of general consumer goods. In fact art. 22, paragraph 3, has no chance of affecting the company organisation in any way since, in the first place, requesting the so-called "human intervention" does not change the impact of the "automated decision", not only because the personal evaluation can be subsequent to that produced by the mathematic mechanisms, but also because human evaluation can very well entail a simple confirmation of the judgment and a subsequent endorsement of the results achieved, without causing an amendment or erasure of the results nor a change in either their genesis, or in their consequences.

Secondly, the same remarks apply to the employee's right to "express his or her point of view", with a debate of little scope and depth, as it is based on the assumption of the determination, and destined to limitedly condition the "human intervention", so that both are weak and of limited scope, in the frenetic organizational dynamics of most of the productive contexts. Nor is of any help the principle that the worker can "challenge the decision", since this is taken for granted in every system, which, however, admits judicial knowledge of the company's unilateral acts. As far as out-of-court oppositions are concerned, there is still no immediate effect on the exercise of corporate power, which would make the condition of the employee, considering the expensiveness of the process far better off due to the poor possibility of an actual, valid court decision on the functioning of the algorithms.

Indeed, if the company wished (and it is not always plausible), it could easily use the "special categories of data" of art. 9, paragraph 1, since, pursuant to art. 22, paragraph 2, the prohibition is apparent, as the exception of art. 9, paragraph 2, lett. a), and the prohibition is removed in case of consent of the so-called data subject. For the very reason that the worker is in serious contractual weakness at the time of hiring, he/she is led to consent and can rarely exercise resolute opposition. The situation is not actually much rebalanced by the last part of art. 22, fourth paragraph, for which, for the purposes of the evaluation and of the "special categories of data", there must be "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" once again with general expressions of limited limitation and with the reference to a dubious, slow-to-form jurisprudential initiative conditioned by the scarce ability to understand the phenomenon.

Nor is there much to expect from the national regulations which so far, as it happened to Italy, have confirmed the European forecasts, also due to the fear of long-range impacts on international competition, and damage to the companies' ability to compete in each country, as the companies are not eager to see restrictions or compromises on the adoption of innovative organizational techniques. Article 22 as a whole is designed to regulate phenomena different from those of the corporate structure, especially for the protection of the consumer, subject to the interference of others, but, in any case, as the depositary of the economic resources that feed the market, in a position of greater power than that of the employee. Even if art. 22 can give some protection to whoever is subject to capillary commercial actions based on the collection and cross-referencing of information, the same does not apply to the workers, who are rarely against expressing ... any consent, for finding the long-awaited professional position.

3. THE USE OF ALGORITHMS IN EMPLOYMENT RELATIONSHIPS, THE CONSENSUS PRINCIPLE, TRADE UNION ACTION AND ADMINISTRATIVE ACTION

Consistently with more general considerations relating to the meeting between labour law and technological innovation, art. 22 of the European Union regulation n. 679 of 2016 fails to provide effective protection to the workers, due to its trust in strategies based on the consensual principle, despite the evident lack of contractual power of the employee. As labour law has known very well, in any system and for almost a century, the consent of the worker can be obtained easily as he/she has few resources to oppose the request of the company, so that protection objectives of standards such as the aforementioned art. 22 are useless in a context of strong psychological pressure, which is the one in place in any company, regardless of its mode of action and of its economic resources.

The European Union regulation n. 679 of 2016 does not significantly strengthen the weak safeguards of art. 22 for "automated decisions". In particular, the adoption of the codes of conduct of art. 40 is voluntary and, if they can identify (art. 40, paragraph 2, lett. b) "the legitimate interests pursued by controllers in specific contexts;" the benefit is possible and limited, if only because the unilateral acts are conceived by the person carrying out the treatments or by representative organizations, with control (art. 41) and certification procedures (art. 42) which in any case come after the original status of free choice of setting codes and without direct relevance to the theme of "automated decisions".

In this regard, at least in Italy, there were no significant interventions by the supervisory authorities (art. 51 and following of the European Union Regulation n. 679 of 2016) and this can be easily explained, for several reasons. First of all, with regard to deliberative power, it is difficult to devise and develop an overall regulation, at least at a national level, due to the heavy sector-specific character of each IT solution, with a high sophistication and strong personalization. Although in the workplace, the possible decisions pertain to infinite profiles and their setting is related to organizational models of various companies who, often, are in competition, and have an obvious interest in extreme confidentiality. In addition to that so far, although the topic has been known and relevant for over twenty years and has led to numerous scientific reflections, there have not been many reports of employees, with regard to their fate, both for the difficulty of understanding the functioning of the algorithms, both for the psychological pressure inherent in the desire to acquire or to maintain a coveted job.

Not surprisingly, in December 2019, on the basis of news published in the press, there has been a strong challenge to the use of decisions obtained with the use of algorithms by a company from northern Italy, but the initiative was collective, proposed by a trade union association and, based on what has been understood, without immediate complaints addressed to the control body, but rather with threats of legal action. In particular, the company has been accused of exasperating the work rhythms, with methods of detecting individual production linked to rewarding systems. If these reports are true, they confirm the low confidence of the employees in the Italian Authority, a low confidence which to be honest has an implicit endorsement in the regulation, whose articles 57 and 58 summarize the functions of the control body, but contain no indication about the execution of art. 22.

Article. 88 of the regulation envisages a more stringent protection "in the context of employment", but does not expressly dwell on the implementation of art. 22 (instead mentioning the "transfer of personal data within a group of enterprises engaged in a joint economic activity" and the "monitoring systems at the work place") (art. 88, paragraph 2). The regulation authorizes implementation of stricter national laws and regulation, even with the intervention of collective agreements, but, at least in Italy, there is no regulation about "automated decisions". In art. 88, paragraph 2, the most significant prescriptive fragment is given by the reference to the incentive of the "transparency" of the "processing", in line with the traditional idea for which the setting of the processes leading to the "automated decision" should be known, but, for all the reasons set out and for its indirect reference to the consensus principle, this strategy is weak and leaves companies without effective restrictions.

There is not much confidence in union action, both because, in the current macroeconomic situation, it is forced to direct its initiatives towards the protection of the minimum wage and, if ever, towards its partial improvement, because the cultural complexity of the problem makes it hostile and resistant to collective examination, since any measure targeting national contracts is blocked by the personalized character of each corporate organizational project and, in this context, there are rarely the conditions for balanced negotiations. Unsurprisingly, in the hypothesis mentioned in December 2019 by the press, the complete failure of the dialogue with the company was stated clearly and, thanks to the wide freedom granted by art. 22, the latter can be authorised to exercise its power in technological innovation and to introduce extreme "automated decisions" solutions.

Article 22 approaches the phenomenon with a transversal logic, as if every aspect of it, especially on the technological side, could be regulated basing on its intrinsic features. This cannot be done. Since labour law is connected to the principle of non-negotiability of the regulations of the law and of collective agreements, it requires a selective strategy, which takes into account its issues and imposes protection models which are based on the selection of realistic and limited objectives with targeted actions. For these reasons, dissatisfaction with article 22 should coincide with the search for new paths, especially if it affects the growing importance of the issue.

4. AN ALTERNATIVE PROPOSAL ON THE REGULATION OF AUTOMATIC DECISIONS IN THE EMPLOYMENT RELATIONSHIP

The company must be free to use mathematical tools of its choosing for decision-making, but the laws must set down and codify rights with binding prerequisites for fundamental measures. An oppositional strategy, as far as technical innovations are concerned, cannot lead to lasting success, even more so if the phenomenon to be regulated cannot be credibly and effectively regulated by the judges. The laws and regulations should divide the possible automatic choices into two veins, one left free for the unconditional application of the mathematical tools, the other strictly regulated, because it pertains to the necessary protection of fundamental interests of the employee, in which the automatic interventions can create disturbances in the implementation of protected values.

In these areas, regardless of the results of the automatic analyses, the decisions must correspond to selected reasons, and the employer must provide a demonstration. To comply with the law the employer should set their computer aids (however sophisticated) using a conditioning evaluation grid as a reference and they should be able to prove compliance to said grid in court. In other areas, the employer would be free to use their preferred decision-making strategies. The protection would be selective and would not concern any decision, but only those on crucial assets, of extreme importance for the worker. This stronger but limited protection would force companies to compare their merits with mandatory regulatory indications. In other words we should move from regulating all automatic decision-making procedures to a much stronger protection only for some, for example those regarding pay, assignments, transfers, training opportunities and dismissal.

The protection based on the explanation of decision-making methods is illusory; the worker cannot possibly dominate the algorithms that surround him/her and that are used for decisions against him. We should rather, with a more modest but also more effective approach, be content with substantial constraints on the decisions made in some matters, in which, free to process information, the company should motivate according to predetermined selection lines, while remaining free from constraints in other areas. Two alternative routes should be created, one regulated by laws and norms that pertain to the object of the evaluation and the other left in an empty space by law, which makes application of the preferred automatic solutions possible.

Let us give some examples; if a system has to select the employee worthy of a prize, it is foolish to think that the decision-making processes may be made clearer by disclosing in advance the parameters used for setting the algorithm, it goes without saying that if it is complex, it uses many information and applies to a large number of employees. In the event of a dispute and in the context of a civil trial, it will never be possible to clarify the actual functioning of the algorithm and, therefore, to establish whether the statements made by the company are true or whether, on the contrary, the information given on the setting of the selection are incorrect. Nor can one think of banning such an initiative, since the strategy would be anti-historical. A different case would be whenever, in a litigation started by a dissatisfied worker, the company is forced to demonstrate positively (with the ordinary procedural instruments) the fact that the choice corresponds to the criteria established by law. This solution could be excessive in the cases pertaining to rewards, since these would be about a benefit exceeding the minimum remuneration and connected to individual merit.

But what about the cases of transfers to workplaces located very far from one's hometown or promotions? In such cases, little it matters to establish which parameters are used by the algorithm, since, in fact, it is impossible to verify in court a similar statement. Having established the legal limits to the decision, the company must be free to resort to the preferred technological solutions but, in case of a trial, it must prove to have based its decision on a motivation consistent with the laws in force. For example, with regards to the transfer, the law can prescribe that only management profiles shall be considered for transfers and that, when evaluating profiles meeting the professional requirements, family requirements shall be taken into account.

Consider the case (far too “popular”) of the so-called digital platforms and of the mechanisms for selecting workers in charge of a service. Even if the company owning the so-called platform declared which facts are relevant for the purposes of the decisions taken with complex mathematical tools, one would never be able to establish in judgment the veracity of such statements and any attempt would have appalling costs, in no way proportionate to the economic interest of the individual worker. The law must take an opposite stance; the company decides as it wishes, with its instruments, but, in the event of a reliable objection, based on data that reveals an objective problem, it must demonstrate that no discrimination has been made and that the functioning of the algorithm does not affect subjects basing on of sex, religion, race, and so on.

In case the collaboration is an articulated and continuous, prolonged one, such as that typical of subordinate employment, opportunities for decisions based on automatic choices multiply. Precisely for this reason, the legal response must be selective, since an overall regulation of the court decisions entrusted to algorithms would presuppose their precise understanding and demonstration in court of their functioning, but such results cannot be achieved, nor are there hopes or actual prospects of overcoming these obstacles, since, if well understood, the complexity of the IT resources is destined to increase making them increasingly difficult to regulate. If one wanted to mitigate the difficulties proposed by the technological context, the selection of targeted and limited objectives is the first step, as it allows focusing one's attention on issues in which the goal of protecting the employee is more evident, without having to engage in an all-compassing, and therefore only generic, law.

In fact, if one wanted a law to hope to be fruitfully applied, one should take into account not the modern technological resources, but the specific aspects of the employment relationship. On these issues, the company must be forced to justify its decisions, in the form of a predefined and circumscribed group of elements, considered consistent with the objectives of the system. The passage from regulating algorithms to regulating certain areas of the employment relationship shifts the focus on safeguarding the person from decisions that have a harmful potential and on the basis of traditional legal reasoning and analysis.

5. THE REASONS OF LABOUR LAW AND THE REASONS OF TECHNOLOGY

The problem of automated decisions is a sign of a more complex comparison between technological transformations (and their impact on company organizational models) and labour law. The latter, left without a direction, because it is used to regulating a consolidated structure, in the face of a range of solutions very different from each other, with that originality allowed precisely by the extension of computer mechanisms that leave wide berth to the initiative to the employer. The solution cannot be a "transversal" approach to regulation of the so-called "data processing" and the insistence of the Community system demonstrates the little awareness of the ineffectiveness of this approach, with regards to the protection of workers. The European Union regulation n. 679 of 2016 does not grasp the specific profiles of their condition and looks at the company's computer and knowledge resources without distinguishing the different contexts in which they are applied, in particular without seeing the typical protection needs of employees.

On the one hand, the authority of the employer creates the premises for the repeated appearance of decisions destined to become "automated" to an increasing extent and without any reasonable chance of actual judicial control. On the other hand, the continuity of the relationship exposes the interests of the employee to stronger threats just because of its length and its being characterised by constant technical evolution, with modifications of the tools that becoming even more significant the longer the relationship lasts. Furthermore, the social weakness of the worker does not allow him to effectively exercise the contractual powers recognized to him, in particular by the regulation of the European Union n. 679 of 2016, and the collective agreements are weakened by the lack of information and action capacity of the trade unions, unable to condition the organizational change, of which they are mere spectators.

The error of European Union regulation no. 679 of 2016 lies in its wanting to regulate information in all areas, without reflecting on the needs of the employment relationship, in which, especially regarding the "automated decisions", what needs regulation (with an accurately selective approach) is the powers of the employer, rather than their analytical background, which escapes the judicial understanding. Rather than trying to change organizational paths, if labour law wants to have hopes of success it must look at the individual provisions influencing the employee's sphere. The problem is not the application of the algorithms, but the justness of the decisions stemming from them and, for this purpose, it is necessary to identify each decision's binding assumptions, with the company having the burden to prove the consistency of its decisions with these assumptions.

As much as this may displease some, there is no possible resistance to technological change in the company and the laws and regulation must be careful not to set such an unachievable goal. There is another goal to pursue, that is to condition the exercise of the powers, which can exploit the knowledge or indications governed by complex mathematical models, but must also, in court proceedings, be able to generate an argument consistent with the regulatory parameters and submitted to the judge for the final decision. Although their background is "automated", decisions fall within the province of man when they become the possible subject of a judgment and this happens if the law creates a grid of significant factors, which the company must confront itself with and with respect to which there is the possibility of a coherent appreciation. In essence, corporate decision are in any case human decision not if they are forced to renounce mathematical elaborations, but if, in order to be legitimate, they must respect axiological parameters that can be dominated by the judge and mandatory for the setting and conclusion of the decision-making process itself.

In this, labour law rediscovers its original nature of containing the authority of the company, rather than changing organizational strategies, despite the fact that the illusion of being able to do the latter is still a popular one. The core and focus should not be the corporate structural criteria, but the prerequisites for unilateral powers affecting the position of the worker. Modesty in the selection of ends must be combined with realism in setting the rules, so that they can be effective and contribute to the rebalancing of the reasons of companies and employees.

REFERENCES

Appelbuam, E., Batt, R. (2011), *Private equity at work. When Wall Street manages main street,* New York;

Bayern, S. (2014), Of bitcoins, independently wealthy software, and the Zero – member LLC*,* in *Northwestern University law review,* n. 108, 1485 ss.;

Bostrom, N. (2014), *Superintelligence. Paths, dangers, strategies,* Oxford University press, Oxford;

Cherry, M. A. (2015), Beyond misclassification: the digital transformation of work*,* in *Comparative labor law and policy journal,* n. 37, 577 ss.;

Fagnino, E. (2019), *Dalla fisica all'algoritmo: una prospettiva di analisi giuslavoristica,* Bergamo;

Dan, M., Cohen (2016), *Rights, persons and organizations: a legal theory for bureaucratic society,* New Orleans;

Domings, P. (2015), *The master algorythm: how the quest for the ultimate learning machine will remake our world,* New York;

Faioli, M. (2018), *Mansioni e macchina intelligente,* Torino;

Lopucki, L. M. (2018), Algorithmic entities*,* in *Washington University law review,* n. 95, 887 ss.;

Peyronnet, M. (2019), *Take easy controlle e sanctionne des salariès,* in *Revue de droit du travail,* n. 1, 36 ss.;

Prassl, J. (2017), *Humans a service. The promise and the perils of work in the gig economy,* Oxford University press, Oxford;

Rosenblat, A. (2018), *Uberland: how algorithms are rewriting the rules of work,* Los Angeles;

Salazar, C. (2019), Diritti e algoritmi: la gig economy e il "caso Foodora", tra giudici e legislatore*,* in Liber amicorum *per Pasquale Costanzo,* 2019;

Thomaz, A. L., Hoffman, G., Cakmak, M. (2016), Computational human – robot interaction. Foundations and trends, in *Robotics,* n. 4, 105 ss.;

Todoli Signes, A. (2018), La gobernanza colectiva de la protecciòn de datos en las relaciones laborales: big data, creaciòn de perfiles, decisiones empresariales automatizadas y los derechos colectivos*,* in *Revista de derecho social,* n. 84, 69 ss.;

Trijsi, A. (2013), *Il diritto del lavoratore alla protezione dei dati personali,* Torino;

Weaver, J. F. (2013), *Robots are people too: how Siri, Google car and artificial intelligence will force us to change our laws,* Preager, Santa Barbara.

THE ROLE OF DATA GOVERNANCE IN THE DEVELOPMENT OF INCLUSIVE SMART CITIES

Damian Okaibedi Eke, Obas John Ebohon

De Montfort University (United Kingdom), London South Bank University (United Kingdom)

damian.eke@dmu.ac.uk; ebohono@lsbu.ac.uk

ABSTRACT

As more cities turn to information and communications technology (ICT) for the efficient delivery of public services, there are huge potentials to improve access to better infrastructure and services, including water supply and waste disposal facilities, urban transport networks, safer public spaces and improved public engagement or interaction. At the heart of this smart city initiatives is a big and robust data ecosystem that generates insights, stimulates innovation and efficiency, improves productivity and delivers wider social, economic and cultural benefits. The nature of the available data for providing these solutions, predictions and decisions can advance or impede inclusion in cities. Harnessing the benefits of smart cities for all communities is therefore mainly dependent on a functional data economy with good quality data and responsible governance approaches. The paper identifies the roles data governance can play in developing inclusive smart cities and then suggests a sustainable smart city data governance framework that is aimed at fostering inclusion by aligning with diverse objectives for and by residents.

KEYWORDS: Data governance, smart cities, inclusion, sustainability, social exclusion.

1. INTRODUCTION

In the face of the current big data economy, the idea of ‘smart cities’ have gained more traction in Europe and beyond. As more cities turn to information and communications technology (ICT) for the efficient delivery of public services, there are huge potentials to improve access to better infrastructure and services, including water supply and waste disposal facilities, urban transport networks, safer public spaces and improved public engagement or interaction (Pla-Castells, et al, 2014). However, smart city initiatives have been criticised for overemphasizing technological solutions and business interests over social inclusion (Paskaleva et al., 2017) - an integral part of sustainable urban development (Chan and Lee, 2008). In a 2018 discussion panel on ‘The Invisible Smart city’, urban designer Gil Peñalosa stated that ‘‘we currently design our cities as though everyone is 30 and active’’, indicating the exclusions of a sizeable proportion of the population outside of this energetic and active segment of the population. In a similar vein, findings of the Microsoft-backed initiative- Smart Cities for All, ‘‘most of today’s smart cities, in both the global north and the global south, are not fully accessible’’. These indicate that smart city designs reflect traditional urban design biases that exclude parts of resident communities such as children, women, older population, the disabled, low income households and the mentally ill (O'Dell et al., 2019). With about 15% of the world population living with some sort of disability (WHO, 2011), about 12.3% of the global population over the age of 60 (ONS, 2018) and with nearly half of the world's population living under the poverty line (World Bank, 2020), there is great need for prioritising inclusion in 'smart city' urban development initiatives . This is particularly important because making "cities and human settlements inclusive, safe, resilient and sustainable" where equality of access and outcome of urban opportunities is a key goal of the UN's 2030 Agenda for sustainable development. Sustainable Smart cities therefore, should advance or reinforce inclusion; appreciating the diversity of different communities, and eliminating identifiable digital, economic, physical and cultural barriers to social inclusion in urban development are critical to successful implementation of the UN SDG 2030 Goals.

At the heart of a successful smart city is a big and robust data ecosystem that generates insights, stimulates innovation and efficiency, improves productivity and delivers wider social benefits (Bibri, 2018; Hashem et al., 2016). ''Big data'' refers to the datasets that represent relevant activities that are characteristically big in volume, velocity, variety, veracity and value (Chen, et al., 2012) (Fothergill et al., 2019). Data plays a central role in the services provided in smart cities. Digital data platforms and cloud-based systems enable smart cities to collect multimodal, cross-functional, big, complex but mostly unstructured data (Chen, et al., 2014) of residents activities with associated individual and collective risks like; data protection, privacy, data sharing, environmental neglect, economic discrimination, social bias and data subject rights. Data is extracted from sources like healthcare systems, transportation, power grids, crime records, irrigation systems and other public service networks which are then used to recognize patterns and needs of the residents. While these different types of data can fuel innovation in smart cities, they can also facilitate exclusion. For instance, many of the smart city data are collected using facial recognition software but a recent study has revealed that commercial facial-recognition software show error rate of 0.8 percent for white male and 34.7 percent for black females (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019). The findings of the study demonstrate inherent racial and gender bias and further evaluation into the cause of this evident bias in the technology shows that the algorithms are informed by datasets that were lacking in diversity (Buolamwini and Gebru, 2018). This is further evidence that datasets have large influences on how technology discriminates certain groups of people in today's society.

The nature of the smart city data, its method of collection and usage have great impact on issues such as; respect for human rights, equitable distribution, respect for diversity and inclusion. While available regulations particularly the EU's General Data Protection Regulation (GDPR) focus on data subject rights, little attention is paid to the impacts of the inferential decisions made with the data which constitute exclusion of some sections of the population. The nature of the available data for such decisions can advance or impede inclusion in cities. Therefore, harnessing the benefits of smart cities for all communities is dependent on a functional data economy with good quality data and responsible governance approaches. The paper identifies the roles data governance can play in smart cities with regard to fostering inclusion. With the understanding of a smart city as ''a blend of institutions, processes, people, and technology'' (Paskaleva et al., 2017), this paper argues that an inclusive and sustainable smart city requires a sustainable data governance characterized by diverse datasets and approaches that address community, environmental, social and economic risks and concerns. The argument here is that the UN's SDG goal 11 of sustainable cities and communities cannot be achieved without a collaborative, dialogical approach to data governance where only datasets that reflect the diverse nature of the population should be used to make inferential decisions affecting the people, the environment and the economy.

In this paper, two major questions are addressed. First, what is the relationship between data processing in smart cities and inclusion/exclusion? And second, how can data governance foster inclusion in smart cities? Answers to these questions are provided through a critical literature review. This is a non-empirical paper supported by critical review of literature on urban inclusion, smart cities and data governance. Academic literature helped us to construct an emerging narrative of smart city exclusion and the relationship with data governance. The paper is deeply rooted in analysis of the conceptual relationships between inclusion/exclusion, smart city technologies and data governance. What emerges is the overview of the roles data governance can play in a sustainable smart city that is then used to provide recommendations for an inclusive framework for data governance. We start with a detailed clarification of urban exclusion, dove-tailing into the issues of smart city exclusion exacerbated by pervasive and ubiquitous technologies. The identified roles of data governance in inclusive smart cities are then used to provide recommendations on a framework that can prioritise inclusion in smart cities. This paper offers a unique contribution to the general discourse of sustainable and inclusive smart cities. The focus on data governance illustrates the interrelatedness of data and wider social issues, particularly the inferences drawn from such data analysis. The conclusions contribute to the ever-growing discussion on the responsible data governance for AI applications. These will not only be of interest to developers of smart cities but also other experts working on AI systems and inclusion.

2. THE CONCEPT OF SOCIAL EXCLUSION AND INCLUSION

Writing about exclusion, Murard (2002 p.41) described it as an empty box given by the French state to which has since been ‘‘filled with a huge number of pages, treaties and pictures, in varying degrees academic, popular, original and valuable’’. Even though the historical roots of this concept can be traced back to ancient Greece, Murard was referring to the contemporary emergence of exclusion in France which is linked to the documented civil unrest in the late 1960s on the heels of growing unemployment and socio-economic inequalities (Ibid). It was not until the concept became prominent in national, regional and international policy agenda that attention began to shift to defining and specifying its meaning. In the late 1990s, the UK government, the European Union and the International Labour Organization helped to popularize this concept (Mathieson et al., 2008). However, a uniform definition remains difficult because the concept is largely described by different people according to its constituent elements. Wolfe, (1995) explained this by offering examples of what people can be excluded from which include; livelihood, social services, consumer culture, political choice, community solidarity and knowledge of the society and oneself. The UK Social Exclusion Unit defined it as what can happen when people or communities suffer from unemployment, poor skills, low incomes, poor housing, high crime environment, bad health and family breakdown (SEU, 1997). This definition portrays exclusion as a consequence of a number of risk factors. In the same vein, Estivill (2003) described it as a result of a combinational process that puts persons or communities at a disadvantage in relation to power, resources and prevailing values. These definitions do not provide a direct description of this concept but provide potential causes or predictors of social exclusion In 2001, the first round of the EU social inclusion process was launched and produced a Joint Inclusion Report that agreed that this concept should be defined on the basis of a number of risk factors including but not limited to low income, unskilled labour, poor health, immigration, low education, gender inequality, discrimination and racism, age, marital status and health (Council of the European Union, 2001). These risk factors were subsequently adopted and documented as the Laeken indicators from the 2001 EU Summit in Leaken-Brussels, Belgium.

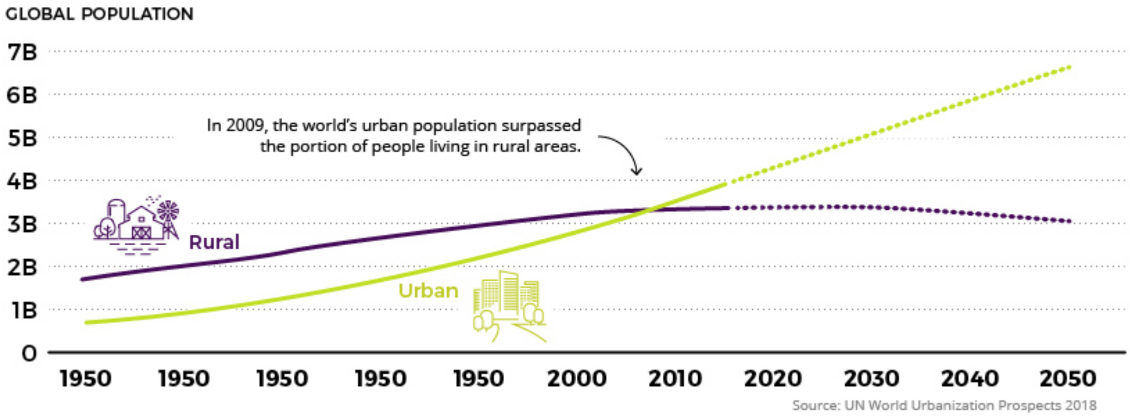
Furthermore, (Jehoel-Gijsbers and Vrooman, 2007) argued that the concept of exclusion should not only be about the process of being socially excluded but should also be about the condition of being socially excluded because exclusion can have a relational dimension (economic and structural exclusion) and a distributional dimension (socio-cultural exclusion). Both dimensions imply a lack of access to benefits owing to some socio-cultural and economic variables. From the individual level perspective, it is about any factor including race, nationality, ethnicity, age, sex, sexuality, poverty and disability that causes a person's incapacity to participate or access social activities and to build meaningful social relations (Silver, 2007). In Urban areas, this can involve insufficient integration into the social and cultural life of the society, or the insufficient access to basic needs including housing, foods, and health services amongst others. Exclusion and its opposite concept of inclusion has become a central concept in today's urban discourse.

Urban cities tend to reflect the intricate social differentiations of the society, exclusion is a critical concept in urban studies and as Nowosielski (2012) argued, it is the most studied subject in the fields of urban development studies. For the purposes of this paper, we will define exclusion in urban areas as a multidimensional phenomenon where persons or communities are explicitly or implicitly denied full access or participation in the social, economic, cultural and political life of the city. It occurs when people are denied equal access to the labour market, education, healthcare, transport, judicial system, participation and other benefits in relation to others in the same city. This revolves around the concepts of poverty, marginalization and deprivation that hinders sustainable urban growth and development. Thus, myriads of initiatives are being deployed by city authorities and policy makers to tackle urban exclusions and foster inclusion to ensure that benefits of urban economies, policies and programmes benefit all sections of the society. This supports Murie and Musterd (2004) who asserted that social exclusions is a normative concept that highlights the need to address apparent inequalities or lack of participation.

Exclusion is a concept that is getting more attention considering the significant changes in contemporary cities caused by globalization and advancements in technology. Urbanization is increasing rapidly owing to increase in urban birth rates, rural-urban migration and movements of people across international borders (Serageldin, et. al., 2004).

According to Figure 1, it is observable that in 2008, the global population living in urban areas surpassed the rural population. According to United Nations Statistics which correlates with figure one above, about 55% percent of the world's population currently live in urban areas and by 2050, this is expected to reach about 68%. The dynamics of urban demographics are rapidly changing; new communities, and new industries and business demanding new set of skills and competences are emerging, resulting in new social formation and structures. It should be noted that accelerating urbanisation in many cities of the world, predominantly in the developed countries, is equally matched by cities experiencing depopulation mainly in the developed countries. Accompanying this new urban reality across global cities is the diversity and polarization, which is hardly surprising as the old economies defined by heavy industries are displaced by the new economies that are knowledge and technologically driven, requiring new skills and competences. Such transitions in the urban economy have intensified urban inequalities that continue to manifest economically, socially, and spatially.

Figure 1. Global population growth.



The challenge for all urban development stakeholders has been how to generate inclusive urban growth and prosperity accessible to all residence irrespective of income, race, gender, sexual orientation and age; (Kearns and Paddison, 2000; Wrigley, 2002). In other words, how can the benefits of urbanization - infrastructures, services and social networks be equitably accessed to ensure inclusion. Thus, inclusion has become a popular aspirational concept in recent developmental discourse (Kasper, 2003). It is a term that has been defined by the World bank (2013: 3) as ‘‘the process of improving the terms of individuals and groups to take part in the society… improving the ability, opportunity, and dignity of people, disadvantaged on the basis of their identity, to take part in society’’. In the United Nations 2030 Sustainable Development Agenda, the concept had about 45 references as ‘inclusion’ or ‘inclusive’ and about six times in the list of sustainable development goals. It was aptly captured by the term “leaving no one behind” in SDG2. The New Urban Agenda, agreed at the Habitat III Conference also has about 45 references to the concept of inclusion (Atkinson, 2000). The aspiration reflects the desire for inclusive urban development underpinning to sustainable development. As McGranahan et al (2016) asserted, inclusion transcends more than elimination of exclusion, it involves an active process of creating equitable services and policies including proactive pursuit and guarantee of human rights. But how can this be contextualized in critical discourse of smart cities?

3. SMART CITIES

The Urban theorist and historian, Lewis Mumford, defined a city as ‘‘geographical plexus, an economic organization, an institutional process, a theater of social action, and an aesthetic symbol of creative unity’’ (Donald and Williams, 2011; Hudson, 2011). This definition portrays a city as a dynamic populous urban settlement that is essentially characterized by geographical, economic and social activities. Despite the accumulation of a significant body of literature on urban development, there is no consensus on what a city is or should be at the national and international level (Kasper et al., 2017). Kasper et al (2017) explored different concepts of this term that have emerged in academic and policy literature including; world or global cities, charter cities, prosperous cities, inclusive cities and smart cities.

Smart cities are fundamentally characterized by the use of information and communications technology (ICT), to develop, deploy, promote and improve practices and policies to address social, economic, political, cultural and environmental challenges (Caragliu et al., 2011; Leydesdorff and Deakin, 2011). It is a concept that does not enjoy a universally accepted definition. Smart city means different things to different people; with different variations in different cities depending on a lot of factors including resources and levels of aspirations, willingness to change and capacities available. However, smart city's underlying framework is that it essentially uses an intelligent network of connected objects and machines (Internet of Things - IoTs) and data to drive digital transformations to improve the lives of citizens and visitors. The World Bank offered further insights to the attributes of smart city[1] by indicating the technological intensity of smart cities, particularly the prevalence of ubiquitous sensors to access real time data. Furthermore, a smart city is projected as a city that uses technology to foster better relationships between citizens and governments. These attributes capture key elements of a smart city: technology and creating efficient relationships between key stakeholders of the city. It involves the use of data-driven technologies to connect different components of the city, optimizing public service operations and infrastructures. It is about the use of data-driven technological approaches to address challenges in waste management, public transport, policing, public health services, welfare systems, emergency response systems, infrastructure challenges and other aspects of the city life. Smart cities are driven by big, multimodal and multidimensional data from connected devices, public agencies, private companies and residents.

Data is a common element of smart cities around the world. Every smart city requires data which forms the base of the smart city model. IoT devices, sensors, networks and applications are all used to gather relevant data to enable efficiency in the technology solutions. Big data generated by IoTs and other applications are then processed and analysed to improve services and infrastructures. Big data analysis therefore plays an important role in smart city operations. From the monitoring of environmental pollution levels, wildlife counts, health monitoring of buildings and dams, traffic lights, CCTV's, connected vehicles, to smart home applications, smart cities use sensors to gather data. Smart applications are then deployed to process and analyse the data, deriving insights which inform actions such as notifications (such as parking space availability, highway incident alerts). Cisco estimates that cities that run on information/data can improve their energy efficiency by 30% within 20 years.[2] Such is the power of data in the context of smart cities. These cities are no longer aspirations but realities in Europe, the US, China and many parts of the world. Cisco, IBM, Intel, Silver Spring Networks, Build.io and Siemens are among the many tech companies providing smart city solutions covering a range of areas: hospitals, traffic/transportation, power plants, water supply and waste management. So far, global smart city spending has reached 34.35billion USD and counting.[3]

Citizens engage with smart city technologies and policies in a variety of ways involving IT skills/knowledge. For instance, some smart city infrastructures require smartphone skills such as the ability to pair mobile phones to city services. This raises the possibility of excluding some sections of the society that lack such skills or lack the financial ability to own smartphones. It does not mean that these disparities in required skills are necessarily caused by smart cities but

[1] https://www.worldbank.org/en/topic/digitaldevelopment/brief/smart-cities

[2] https://newsroom.cisco.com/press-release-content?articleId=4766225

[3] https://mobility.here.com/smart-city-technologies-role-and-applications-big-data-and-iot

the solutions they provide can exacerbate the inequalities that already exist in the society (Gilbert, 2010). Citizens are also affected by these technologies differently depending on the width and breadth of ICT applications used. An example of this is the use of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in the criminal justice system in some US states. COMPAS is a risk-assessment algorithm being used by counties in the US to predict crime rates, determine jail time and provide information used for sentencing. In 2016, reporters working for ProPublica analyzed about 7,000 COMPAS assessments in Broward County. The conclusion of this research was that the algorithm was biased against blacks because ‘‘blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend’’. This is a typical example of how deep learning analytics can contribute to the exclusion of residents in a smart city. These demonstrate that the data that offers powerful value to smart city initiatives can also contribute to social discrimination and exclusion. Inclusion in a smart city is therefore about the elimination/reduction of social exclusion but also about creating processes that actively protect the rights of all residents. The argument in this paper is that what and how data is used in smart cities can provide solutions to exclusion challenges in the smart city discourse. So, the question is how can data that drive smart cities be governed?

3.1. Data governance in smart cities

As it was described in the last section, data is at the core of the decision-making process of all smart cities. Multidimensional and multimodal data that include human, animal and technical data, inform smart city solutions. Like in many data-intensive operations, the multifaceted nature, volume and value of data in smart cities call for an effective data governance structure. But key questions here are: what is data governance and how can it be understood in the context of smart cities? Data governance and information governance are often used interchangeably in literature (Godinez et al., 2010) but scholars such as Kooper et al, (2011) and Nielsen (2017) stress that there are fundamental differences between the two. Kooper et al (2011) described data governance as a process that focuses on data assets while information governance is related to interactions. The data governance literature paints a picture of a term that is evolving both as a concept and as a discipline (Zhang and Yuan, 2016). Its different definitions are informed by the goal of the institution or discipline where the concept is contextualized.

Many of these definitions describe it as a compliance process to fulfil internal organizational (Donaldson and Walker, 2004) or external legal or ethical requirements (Chalcraft, 2018). This perspective focuses on responsive processes to policies, principles and regulations such as the GDPR. Others see data governance as an organizational decision-making process about data related issues (Putro et al., 2015; Weber et al., 2009). These scholars conceptualize data governance as a framework for accountability that encourages better use of data to achieve the organizations’ objectives. Data governance is also defined according to how its goals are perceived. In this vein, Nielsen (2017) stated that the goal of data governance is to enhance ‘business goals’. For business organizations that conceive data as key business assets, data governance becomes a way of managing assets (Aiken, 2016). This is done through critical alignment and organization of data management strategies with the business strategies (Brous et al., 2016). These definitions can all be situated in organizational literature where ‘business goals’ are determined by the pursuit of the bottom-line.

In a bid to offer a robust definition of data governance that offers full consideration of ethics and implementation challenges, Fothergil et al (2019) defined data governance as the overall management of the availability, usability, integrity, quality and security of data in a given organization with the intention of ensuring maximum creation of value from the data while adhering to ethical and legal requirements. In this paper, we extend this definition to also involve the establishment of processes, policies, roles and responsibilities that foster the effective management of data for the benefit of relevant stakeholders that will be affected by the decisions derived from the data.

The EU DECODE project recently called for a different understanding of data as a common good rather than an asset in the context of smart cities (Bass, 2017). They observed that the current digital economy fosters an ecosystem where data is collected and used in ways that, 'create stark new imbalances of power' which means that 'cities will need to play a more active role in leveraging more responsible innovation with data in the local economy'. This is a call for a different kind of data governance in smart cities where data can be seen as a common good that can deliver significant personal and public benefits. We believe that one of these benefits is the building of inclusive communities where improved participation of residents can be achieved. Through a sustainable data governance framework, smart cities can achieve improved inclusion of different communities.

3.2. Sustainable data governance and Inclusion in Smart Cities

According to the European Commission, the concept of smart city means striving for sustainability through smart solutions.[4] The EC report on the Impacts of Information and Communication Technologies on Energy Efficiency[5] identified areas in a city where ICT can positively influence. Smart city is also a place where efficient ICT networks and services create benefits for all communities and for businesses through the use of cross-functional data. The underlying logic is to be more responsible, sustainable and inclusive. Achieving these sustainable, inclusive goals requires setting up a sustainable data governance structure that can enable cities to turn data into benefits for businesses, improve the quality of life of citizens (including increased access and participation), and at the same time, ensuring effective response to environmental challenges. A smart city data governance framework should focus on the people, businesses and the environment. The neglect of any of these elements can implicitly or explicitly constitute exclusion.

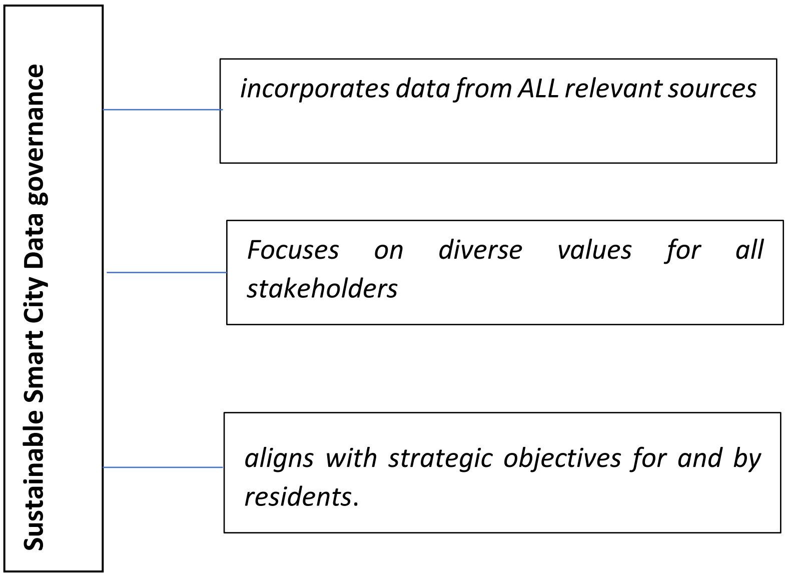
The data at the core of smart city solutions goes beyond human data to animal, technical and environmental data. Human data alone cannot provide a robust understanding of urban challenges. To harness the benefits of smart technologies (IoT, artificial Intelligence- AI) in cities, rich and diverse data from more sources provides high values. That means data governance in smart cities should not be limited to human related issues of privacy, confidentiality and data protection. It should also focus on efficiency gains and economic optimization and also on improving the environment (such as air quality, lower emissions etc). It requires a data governance framework that can help to balance competing interests of all stakeholders in the city, especially concerns of the most vulnerable residents of the city, generate economic value to support needed services and leave less environmental footprints. This is our idea of sustainable data governance which incorporates data from all relevant sources, focuses on

[4] https://ec.europa.eu/digital-single-market/en/smart-cities

[5] https://ec.europa.eu/information_society/activities/sustainable_growth/docs/studies/2008/2008_impact-of-ict_on_ee.pdf

diverse values for all stakeholders and aligns with strategic objectives for and by residents and is demonstrated in the figure below.

Figure 2. Sustainable smart city data governance.


Sustainable Smart City Data governance
incorporates data from ALL relevant sources
Focuses on diverse values for all stakeholders
aligns with strategic objectives for and by residents.

Cities have differing needs but a city seeking sustainability and inclusiveness must focus on three key elements: socio-cultural sustainability, economic sustainability and environmental sustainability (Basiago, 1999). Socio-cultural sustainability is about how the city respects citizens' intrinsic value by ensuring that their rights to social justice, health, education, culture, religion, peace, privacy and confidentiality are protected. Economic sustainability concerns how the city provides the capacity for citizens and businesses to develop economic potentials while environmental sustainability refers to the ability of cities to protect and maintain environmental resources for future generations. Smart city data governance must focus on these three elements in order to achieve inclusivity. A narrow focus on only one or two of these areas of sustainability can exclude some communities from benefiting from smart city initiatives. A sustainable city is essentially an inclusive city. The quality, availability and integrity of the data and how they are governed have impacts on how human rights (including privacy and confidentiality) are protected, how attractive the city is to businesses, the level of government's productivity and inclusiveness, environmental sustainability and ultimately how livable the city is.

Paskaleva et al (2017) observed that this idea of harnessing smart city data to improve urban sustainability and inclusion is yet to be explored. It is something we believe is important considering the significant value data can provide and the increasing level of connectivity and collaborations between data, technology, citizens, the environment and private enterprises in smart cities. Smart city data governance goes beyond compliance and encompasses how data can be managed to create values for citizens, businesses and the environment. To derive maximum value of smart city data, sustainable data governance promotes collaboration of key partners/stakeholders in the collection of sufficient data for smart city decisions. Sustainable Development Goal is about the establishment of a resilient framework that will manage the exploding quantity of data and disparate data sources to deliver enduring value to relevant stakeholders in the city. It is a framework that should shape the quality, availability, obtainability, usability, security and effectiveness of the data used for smart city decisions; striking the balance between data protection and acceleration of ICT potentials.

3.3. Inclusion as a priority of Sustainable data governance

Smart city data can come in many different formats, collected by multiple devices and agents in different sectors including: energy, health, government, economy, environment, community life and directly from citizens (Batty, 2013). Smart city systems will only be as good as the quality and diverse nature of the data that inform them. That is why it is necessary for smart city data to reflect the intersectional differences of the demography of the city. Non-representative datasets can lead to biased decisions exacerbating the issues of inequality and discrimination. Lack of focus on the diversity of the data that informs smart city decisions will lead to overlooking the diverse needs and preferences of the people. Aura Vasquez, a former commissioner of the Los Angeles Department of water and power observed that, ‘‘without understanding the people that are going to live in this smart city… what their priorities and problems are - we’re not going to get to them’’. It is the nature and quality of the data that will provide an improved understanding of the diverse needs of all stakeholders in the city. To be inclusive, smart cities need to use representative stakeholder data to provide critical understanding of the people and their needs.

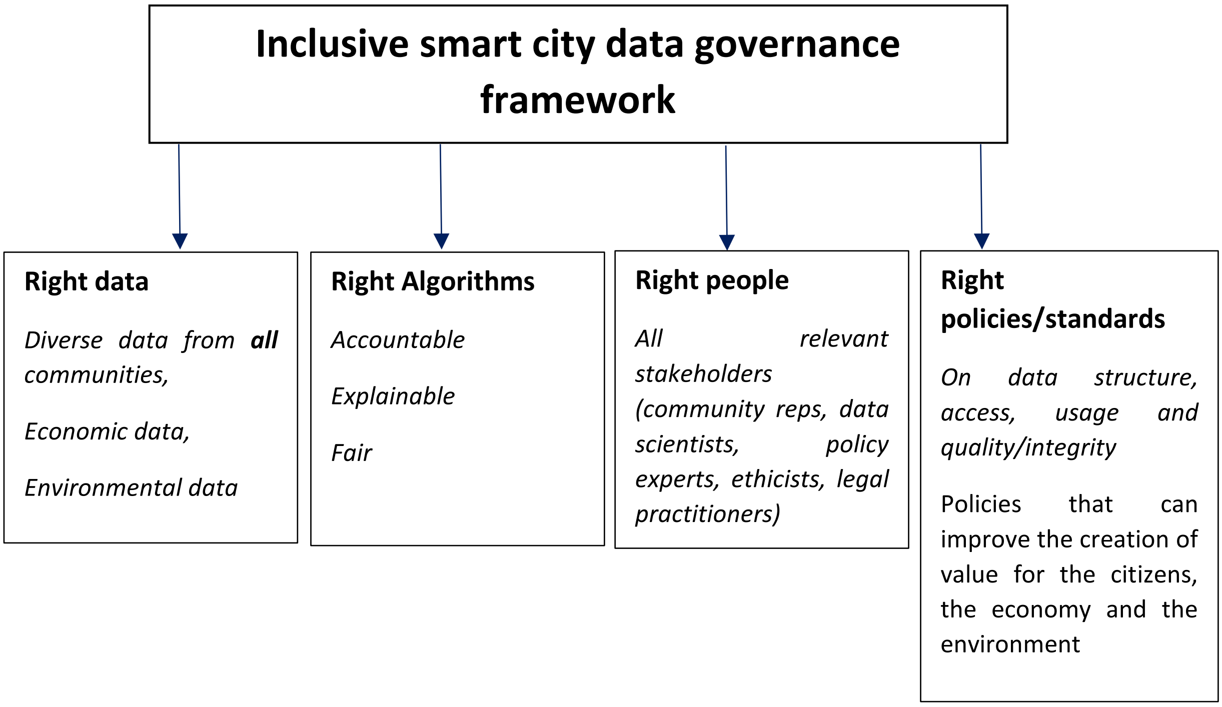
A sustainable and inclusive smart city should also make use of diverse economic and environmental data which are traditionally collected, analysed, aggregated and utilized by different actors working towards their specific goals without appreciating the value of collaborating with others beyond their operational ‘silos’. For instance, fuel consumption records from factory machineries are required for financial accounting but can also be used for emission reporting (Gerrard, 2014). Geological information collected during mineral resource drilling for technical drilling purposes can also be used for determining waste management options and potential for acid mine drainage. These and many more environmental data have been recognized as for decision making in recognising, minimizing and environmental mitigating risk (Gerrard, 2014).

Managing the use of smart data is also the domain of data governance. Setting up a governance framework that ensures the protection of stakeholders’ human rights (including rights to privacy, to equal socio-economic and cultural participation) and builds trust should be the goal of smart cities. Sustainable data governance should proactively encourage the use of smart city data to improve the quality of life of the residents but also for economic growth and environmental sustainability. The last two are particularly important in the context of sustainable data governance. Poverty and employment status are two causal factors of social exclusion. When smart city data is used for public/private economic growth, there are possibilities generating jobs and economic opportunities that can alleviate issues of poverty and employment. Cities must develop responsible approaches to sharing relevant smart city data for business uses in ways that it will benefit the citizens. For instance, Cia et al (2018) believe that a symbiotic approach to data in a smart city can increase Cellular network operators’ efficiency and ensure lower operating costs which implicitly benefits the citizens. The physical environment can also determine how some people participate in city life. According to Williams et al., (2008), buildings design and structure, inaccessibility of public spaces and other environmental factors restrict disabled and elderly people from participating in community and outdoor activities. Sustainable data governance of smart city data should not only focus on data protection issues but also on how to use data to ensure ‘‘equal access to resources, poverty alleviation, disaster and hazard management, land use to reduce biodiversity loss… creating a low-carbon and energy efficient society’’ (Dwevedi et al., 2018). But the practical challenge remains how to manage the diverse and big smart city data to ensure inclusive distribution of resources, fair access to services, improve mobility of all residents (including disabled people) and the livability of the city while being prepared to respond to environmental changes.

4. A FRAMEWORK FOR INCLUSION

Using the above definitions and discussion, we propose a framework for social inclusion in smart cities through data governance. This framework has no methodological or theoretical underpinnings but serves as a prescriptive framework of how the application of data can improve smart city inclusion. It is not intended to be a tool that provides solutions to all concerns of social exclusion in smart cities but rather it is presented as a structure that can help to address social, political, legal, cultural and ethical concerns that have impacts on exclusion in smart cities. Figure 3 shows should be considered in developing a data governance structure for smart cities; the right data, algorithms, people and policies/standards.

Figure 3. Framework for inclusive smart city data governance.



Inclusive smart city data governance framework

Right data

*Diverse data from **all** communities,*

Economic data,

Environmental data

Right Algorithms

Accountable

Explainable

Fair

Right people

All relevant stakeholders (community reps, data scientists, policy experts, ethicists, legal practitioners)

Right policies/standards

On data structure, access, usage and quality/integrity

Policies that can improve the creation of value for the citizens, the economy and the environment

4.1. Right data

Data that inform smart city decisions must be right to produce inclusive outcomes. According to a recent survey of bias and fairness in machine learning, data and the algorithm are two key sources of bias in AI and machine learning (Mehrabi et al., 2019). Bias in data can be in the form of representation bias, measurement bias, evaluation bias, aggregation bias, population bias, sampling bias or data linking bias (Ibid). It is important for smart city decisions to be based on data that reflects the diverse stakeholders in the city. When the collected data does not accurately represent the environment, the system is meant to serve, this can lead to unfair and non-inclusive decisions. The right data for inclusive smart city decisions should data that represents even the ‘hard-to-reach' or ‘digitally invisible’ residents (O’Dell et al, 2010). Smart city data can also contain prejudice (Kallus and Zhou, 2018) or group attribution bias which can lead to non-inclusive decisions. Data that contains prejudicial views about individuals based on race, social class, nationality, gender, sexuality, educational status is not the right data for smart city decisions. Data that contains a lopsided view of a certain community is also not the right data for smart city decisions. The reliability of the insights derived from smart city data depends on the quality, nature and integrity of the data used. Smart city data governance framework must therefore include processes, partnerships and programs that can overcome trust, resource and language barriers to data collection (Stonewell et al., 2017) and ensures the mitigation of bias associated with data.

4.2. Right algorithms

In addition to using the right data, data-driven algorithms used for smart decisions should be right. According to Mehrabi et al., (2019) algorithmic bias is when the bias is not associated with the data but is added by the algorithm. Examples of biased algorithms can be found in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism algorithm used in judicial decisions in the US. This software is more likely to assign higher risk scores to black offers than Caucasians that committed similar or more grievous offences (Rudin, Wang and Coker, 2018). In another study evaluating the fairness of algorithms for predicting juvenile recidivism, Tolan et al., (2019) discovered biases in machine learning algorithms in data-driven risk assessment. Sustainable data governance in smart cities needs to ensure that algorithms applied to smart city data are evaluated for biases. One city that has effectively done this is Johnson county Kansas where the county partnered with a private enterprise to develop a recidivist software that predicts the likelihood of re-entry into the criminal justice system (Sullivan, 2018). But rather than using the predictions derived from data from police departments and public health centres to police the residents, they use it to allocate preventive and proactive mental health resources to reduce re-offending.

4.3.Right people

Smart city data governance structure should involve the right mix of people and expertise to adequately address the issues of inclusion. Smart city governance requires a lot of companies monitoring, policy development, data quality/integrity assessment, data security and access evaluation and data management. While data scientists can bring valuable expertise in smart city data governance, social scientists, policy makers, legal practitioners and civil society groups should all be included in the data governance decision making process. The multimodal data involved in smart city decisions requires effective engagement of citizens. Therefore, the input of people who understand the issues and values associated with the associated data is highly valuable. Representatives from different communities should be represented in smart city data governance to represent the interests and goals of their communities. The shortcomings of the smart city data can only be effectively uncovered when the interests of the many different communities are equally considered. When communities are underrepresented in governance decisions, their needs and interests may likely be neglected.

4.4. Right policies/standards

To achieve inclusiveness, data governance framework in smart cities needs to adopt policies and standards that can build trust, transparency, accountability and responsibility. The depth, size and types of data being generated by smart cities are growing exponentially by the day and many of these are strongly guarded by government or proprietary databases. The idea of limiting access to smart city data not only prevents the useful application of the data to solve common problems, it also exacerbates the lack of trust between the government and the people. Sustainable data governance policies should promote an open data platform that optimizes interoperability and ultimately contributes to the common good. One advantage of this is that it opens the door for more innovations that can restructure the economy and improve environmental sustainability. An example is using open smart city data to develop GPS-apps for the blind and visually impaired (Ryu, Kim and Li, 2014). The Barcelona City council's CityOS is based on an open-code big data platform enabling private tech company innovations boasting the economy and improving the quality of life.

Open data platforms in smart cities also contribute to improved public access to data. The citizens should be able to know what data the city holds and how they are used. This helps to build trust in government initiatives by increasing citizens participation. The citizens collectively share the ownership of the smart city data and as such should be granted a platform to know, understand and appreciate how their data is being applied. Amsterdam's Tada-data disclosed is an example of an open data initiative giving citizens control over data. The city is using active campaign and operational tools to provide clarity about data to all participating parties in the city, citizens and businesses alike. Smart city data governance policies should indeed make inclusion a strategic imperative.

5. CRITICAL DISCUSSION AND CONCLUSION

Observably, cities play critical roles in economic development of nations, this is evidenced by the disproportionate share cities account in total gross domestic product of nation states. Nonetheless, cities also account for significant proportion of factors responsible for environmental degradation hence the desire to develop cities into sustainable and livable communities. This translates to creating employment and income generating opportunities, safe and affordable housing, and building resilient economies and societies, as we have seen emerging in many countries, particularly the developed countries. Other manifestations of sustainability in cities include investment in sustainable public transport and public spaces to afford recreational facilities for healthy living and wellbeing.

However, the challenge has been how to ensure that the opportunities sustainable cities have to offer is equitably accessed by all citizens, and many cities have mainstreamed citizens' participation into urban planning and management to bridge the widening inequality in cities. For effective participation and delivery of adequate and right amount of services in real time, the use of information technology has become inevitable. Evidence abound showing that huge amount of data is required to determine the kinds of urban services demanded by citizens and also to determine the efficiency of such deliveries in terms of indications of who are able to access services and those who might have been left behind. The governance of the huge data input to smart cities has become just as important as the inequality the data set serves to bridge. For inclusiveness, data governance framework is a necessary prerequisite in smart cities to effect sustainable and livable urban environment. Wider access to smart city data is the only way equality of access and outcome of the opportunities cities have to offer its citizens, allowing efficient and effective delivery of urban services. Thus, data governance is critical to smart cities and the delivery of 21st century sustainable and livable cities.

REFERENCES

Aiken, P. (2016) Experience: Succeeding at data management—BigCo attempts to leverage data. *Journal of Data and Information Quality (JDIQ)*, 7(1–2), pp. 1–35.

Atkinson, R. (2000) Combating Social Exclusion in Europe: The New Urban Policy Challenge. Urban Studies, Vol. 37, No. 5-6, 1037-1055

Basiago, A. D. (1999). Economic, social, and environmental sustainability in development theory and urban planning practice

Bass, T. (2017) *Reclaiming personal data for the common good*. [Online] https://decodeproject.eu/blog/reclaiming-personal-data-common-good.

Batty, M. (2013) Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), pp. 274–279.

Bibri, S.E. (2018) The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustainable Cities and Society*, 38, pp. 230–253.

Brous, P., Janssen, M. and Vilminko-Heikkinen, R. (2016) Coordinating decision-making in data management activities: a systematic review of data governance principles. In: *International Conference on Electronic Government*. Springer, pp. 115–125.

Buolamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proceedings of Machine Learning Research*. Conference on Fairness, Accountability, and Transparency. pp. 1–15.

Caragliu, A., Del Bo, C. and Nijkamp, P. (2011) Smart cities in Europe. *Journal of urban technology*, 18(2), pp. 65–82.

Chalcraft, J. (2018) Drawing ethical boundaries for data analytics. *Information Management*, 52(1), pp. 18–25.

Chan, E., Lee, G.K.L. Critical factors for improving social sustainability of urban renewal projects. Soc Indic Res 85, 243–256 (2008). https://doi.org/10.1007/s11205-007-9089-3

Chen, H., Chiang, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36, pp. 1165–1188.

Chen, M., Mao, S. and Liu, Y. (2014) Big Data: A Survey. *Mobile Networks and Applications*, 19(2), pp. 171–209.

Council of the European Union (2001) *Joint Report on Social Inclusion*. European Commission.

Dalla Cia, M. et al. (2018) Using Smart City Data in 5G Self-Organizing Networks. *IEEE Internet of Things Journal*, 5(2), pp. 645–654.

Donaldson, A. and Walker, P. (2004) Information governance—a view from the NHS. *International journal of Medical informatics*, 73(3), pp. 281–284.

Donald, A., & Williams, A. (2011). Introduction: The Paradoxical City. In The Lure of the City: From Slums to Suburbs (pp. 1-11). London: Pluto Press. doi:10.2307/j.ctt183pc6x.3

Dwevedi, R., Krisha, V. and Aniket Kumar (2018) Environment and Big Data: Role in smart cities of India. *Resources*, 7(4), p. 64.

Estivill, J. (2003) *Concepts and strategies for combating social exclusion: an overview*. International Labour Organization.

Fothergill, B.T. et al. (2019) Responsible Data Governance of Neuroscience Big Data. *Frontiers in Neuroinformatics*, 13, p. 28.

Friedmann, J. and Wolff, G. (2017) World city formation: an agenda for research and action. In: *The Globalizing Cities Reader*. Routledge, pp. 46–54.

Gilbert, M (2010) THEORIZING DIGITAL AND URBAN INEQUALITIES, Information, Communication & Society, 13:7, 1000-1018, DOI: 10.1080/1369118X.2010.499954

Hudson, A. (2011). The Dynamic City: Citizens Make Cities. In Williams A. & Donald A. (Authors), The Lure of the City: From Slums to Suburbs (pp. 12-31). London: Pluto Press. doi:10.2307/j.ctt183pc6x.4

Godinez, M. et al. (2010) *The art of enterprise information architecture: a systems-based approach for unlocking business insight*. Pearson Education.

Hashem, I.A.T. et al. (2016) The role of big data in smart city. *International Journal of Information Management*, 36(5), pp. 748–758.

Jehoel-Gijsbers, G. and Vrooman, C. (2007) Explaining social exclusion.

Kallus, N. and Zhou, A. (2018) Residual Unfairness in Fair Machine Learning from Prejudiced Data. *arXiv:1806.02887 [cs, stat]*, [Online] Available from: http://arxiv.org/abs/1806.02887 [Accessed 12/03/2020].

Kasper, E. et al. (2017) *Inclusive Urbanisation and Cities in the Twenty-First Century*. IDS.

Kearns, A. and Paddison, R. (2000). New Challenges for Urban Governance. Urban Studies, Vol. 37, No. 5-6, 845-850.

Kooper, M.N., Maes, R. and Lindgreen, E.R. (2011) On the governance of information: Introducing a new concept of governance to support the management of information. *International journal of information management*, 31(3), pp. 195–200.

Leydesdorff, L. and Deakin, M. (2011) The triple-helix model of smart cities: A neo-evolutionary perspective. *Journal of urban technology*, 18(2), pp. 53–63.

Mathieson, J. et al. (2008) Social Exclusion Meaning, measurement and experience and links to health inequalities. *A review of literature. WHO Social Exclusion Knowledge Network Background Paper*, 1, p. 91.

McGranahan, G., Schensul, D. and Singh, G. (2016) Inclusive urbanization: Can the 2030 Agenda be delivered without it? *Environment and Urbanization*, 28(1), pp. 13–34.

Mehrabi, N. et al. (2019) A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Murard, N. (2002) Guilty victims: social exclusion in contemporary France. *Biography and social exclusion in Europe. Experiences and life journeys*, pp. 41–61.

Murie, A. and Musterd, S. (2004) Social exclusion and opportunity structures in European cities and neighbourhoods. *Urban studies*, 41(8), pp. 1441–1459.

Nielsen, O.B. (2017) A comprehensive review of data governance literature. *Selected Papers IRIS*, 8, pp. 120–133.

Nowosielski, M. (2012) Challenging Urban Exclusion? Theory and Practice. *Polish Sociological Review*, 179(3), pp. 369–383.

O'Dell, K. et al. (2019) *Inclusive Smart Cities: Discovering Digital solutions for all*. Deloitte Center for Government Insights.

O'Dell, L. et al. (2010) Constructing 'normal childhoods': Young people talk about young carers. *Disability & society*, 25(6), pp. 643–655.

ONS (2018) *Living longer: how our population is changing and why it matters*. [Online] https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/livinglongerhowourpopulationischangingandwhyitmatters/2018-08-13.

Paskaleva, K. et al. (2017) Data Governance in the Sustainable Smart City. *Informatics*, 4(4), p. 41.

Pla-Castells, Marta & Martinez-Durá, Juan & Samper Zapater, J. Javier & Cirilo Gimeno, Ramon. (2015). Use of ICT in Smart Cities. A practical case applied to traffic management in the city of Valencia. Conference: Smart Cities Symposium Prague 2015, Volume: 978-1-4673-6727-1 EEE10.1109/SCSP.2015.7181559.

Putro, B.L., Surendro, K. and Herbert (2016) Leadership and culture of data governance for the achievement of higher education goals (Case study: Indonesia University of Education). In: *AIP Conference Proceedings*. AIP Publishing LLC, p. 050002.

Raji, I.D. and Buolamwini, J. (2019) Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*. the 2019 AAAI/ACM Conference. Honolulu, HI, USA: ACM Press, pp. 429–435.

Romer, P. (2010) *Technologies, rules, and progress: The case for charter cities*.

Rudin, C., Wang, C. and Coker, B. (2018) The age of secrecy and unfairness in recidivism prediction. *arXiv preprint arXiv:1811.00731*.

Ryu, H.-G., Kim, T. and Li, K.-J. (2014) Indoor navigation map for visually impaired people. In: *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*. pp. 32–35.

Sassen, S. (2013) *Deciphering the global: its scales, spaces and subjects*. Routledge.

Serageldin, M., Cabannes, Y., Solloso, E,. Valenzuela, L. (2004). Migratory Flows, Poverty and Social Inclusion in Latin America. World Bank Urban Research Symposium. https://www.researchgate.net/profile/Yves_Cabannes/publication/32893681_Migrator y_Flows_Poverty_and_Social_Inclusion_in_Latin_America_By/links/55a3e9db08ae5e82ab1f25c6/Migratory-Flows-Poverty-and-Social-Inclusion-in-Latin-America-By.pdf

SEU (1997) Social Exclusion Unit: Purpose, work priorities and working methods. *London: The Stationery Office*.

Silver, H. (2007) The process of social exclusion: the dynamics of an evolving concept. *Chronic Poverty Research Centre Working Paper*, (95).

Stonewall, J. et al. (2017) Best practices for engaging underserved populations. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, pp. 130–134.

Sullivan, R. (2018) *Innovations in Identifying people who frequently use criminal justice and healthcare systems*. [Online] https://www.prainc.com/innovations-identification-cj-healthcare/.

Tolan, S. et al. (2019) Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. pp. 83–92.

Weber, K., Otto, B. and Österle, H. (2009) One size does not fit all—a contingency approach to data governance. *Journal of Data and Information Quality (JDIQ)*, 1(1), pp. 1–27.

WHO (2011) *World report on disability 2011*. World Health Organization.

Williams, B. et al. (2008) Experiences and expectations of disabled people. *London: Office for Disability Issues*.

Wolfe, M. (1995) Globalization and social exclusion: Some paradoxes. *Social Exclusion: Rhetoric Reality Responses/International Institute for labour studies. United Nations development program*, pp. 81–102.

World Bank (2020) *Poverty*. [Online] https://www.worldbank.org/en/topic/poverty/overview. Available from : https://www.worldbank.org/en/topic/poverty/overview.

WorldBank (2013) *Social Inclusion*. [Online] https://www.worldbank.org/en/topic/social-inclusion#4.

Wrigley, N., Guy, C., Lowe, M. (2002). Urban Regeneration, Social Inclusion and Large Store Development: The Seacroft Development in Context. Urban Studies, Vol. 39, No. 11, 2101–2114.

USE OF GUILT APPEALS IN NPO CAMPAIGNS AND ITS ETHICAL CONSIDERATIONS

Pamela Simón, Jesús García-Madariaga, Ingrit Moya, María-Francisca Blasco

Complutense University of Madrid (Spain)

fblasco@ucm.es; jesgarci@ucm.es; ingritvm@ucm.es; pamsimon@ucm.es

ABSTRACT

Negative appeals and specially guilt appeals are quite common in non-profit advertisement as a persuasion technique to increase donations for their causes. However, these messages might not always be effective and even though it is widely known that guilt might lead people to donate more or engage into prosocial behaviour, the effects of guilt might be damaging for the public and for the brand (Graton & Mailliez, 2019). In this article we review the use of guilt in non-profit organizations advertising and how evoking these emotions and the over exposure to these emotions might be consider unethical, since they might increase distress and general discomfort in the public. For which, we advise marketers and advertisers to use these appeals with caution and consideration towards the donors.

KEYWORDS: Ethics, Non-Profit Organizations, Advertising, Negative valence image, Guilt.

1. INTRODUCTION

It is well known that Non-Profit Organizations (NPOs) use emotional images to persuade people to donate to their cause. One of the most common ways for NPO´s to promote their cause is by employing emotional appeals in their advertising (Septianto & Tjiptono, 2019). Since NPO advertising tries to touch the desire of people of helping those less fortunate, this type of sector tend to use affective effect in their advertisings. It is widely recognized that emotional appeals are very effective tools for persuasion (Burt & Strongman, 2005; Bagozzi & Moore, 2006; Poels & Dewitte, 2009). Emotional appeals are instrumental in providing the creative punch to enhance persuasion (Bebko, Sciulli, & Bhagat, 2014), also emotions have the ability to capture attention, influence attitudes, and affect consumer behaviour.

But the problem is that some NPO have been collecting a huge amount of donations but maybe doing a lot of harm at the same time. Using images of people in developing countries that manifest suffering by starving people, begging eyes, and distended bellies, to gather donations. These types of biased images are known as the “Starving Baby Appeal” (Fine, 1990, p. 154). This kind of campaigns involves inherent ethical concerns and dilemmas.

Advertising of NPO that uses a guilt appeal are supposed to make the audience feel guilty and advertisers expect that the audience have some feeling of failing at their own ideals or ethical principles, but not always make them feel guilty (Cotte, Coulter, & Moore, 2005).

Several ethical dilemmas emerge in how NPOs request donations, because of a pressure between what the NPO needs to do for their beneficiaries (requesting in the most efficient and effective ways to guarantee enough money to help them) and what the donor wants (been asked less, asked in different ways, or simply not asked at all) (MacQuillin & Sargeant, 2018).

Thus, it is possible that advertising may have harmful effects for individuals and society that deserve ethical analysis. Furthermore, social marketers should notice that target viewers can have reservations about the ethicality of social advertising, even when they know their intentions are good (Hastings, Stead, & Webb, 2004).

So, the purpose of this paper is to examine how negative-valence images used by NPOs to communicate their causes may affect the public by making them feel guilty and how the saturation of this type advertising might create aversion.

2. LITERATURE REVIEW

2.1. Advertising

Advertising can be defined as a mechanism that companies use to persuade, remind, and inform people about their brands. It is considered that advertising is a mean of communication that comprises all the actions in which companies present a visual or oral message regarding a service, product or idea (Stanton, 2000). Nowadays, due to digital technologies the way companies interact and communicate with consumers have change via digital media, this growth in digital advertising correlates with this increase in digital media consumption (Truong, McColl, & Kitchen, 2010).

There are many kinds of advertising: commercial consumer advertising, small ads as in the classified section, prestige, business and financial advertising, trade and technical advertisements and government and charity advertising. This last one is usually non-profit but often uses the persuasive techniques of commercial advertising (Dyer, 2008). The use of persuasion in advertising is adopted to motivate people into action by influencing desires and beliefs. So, advertisements should be able to connect with what concerns the public, and anything that concerns the public and impacts values or core concerns has the ability to arouse emotions (O´Shaughnessy & O´Shaughnessy, 2003).

In advertising, the two most important benefits consumers perceive are utilitarian and emotional benefits. Utilitarian benefits are related to consumer´s basic motivation levels, and information is one of the major utilitarian benefits consumers look for. While emotional benefits are linked to consumer´s basic needs for personal expression, self-esteem, social approval and stimulation. Considering these benefits, advertising has developed two important message appeal strategies: rational and emotional appeals (Zhang, 2014). This last one is the most used by NPOs. According to a meta-analysis by Hornik, Ofir, & Rachamim (2017), approximately 76 percent of all advertisements use these type appeals.

On the other hand, the use of technology in advertising can be a powerful facilitator of development goals, because it can improve communication, can create new social and economic networks and can boost the exchange of information (Sheombar, Urquhart, & Kayas, 2018). That’s why more and more, different formats of advertising are now propagating online. But is important to note that according to different studies, attitudes toward online advertising are more negative than to offline advertising (Drèze & Hussherr, 2003; Cho & Cheon, 2004; Ha &

Mccann, 2008; Howe & Teufel, 2014). For advertisers, this quick acceleration of usage of online media will provide both opportunities and challenges (Truong et al., 2010).

There is a saturation of advertisement exposure, everywhere we look at, we will find an advertisement, for which, institutions have created conduct codes to avoid unethical advertising. This concerns for unethical advertising begun since 1942, when the Advertising Federation of America created a 39-point code of ethics for advertising (Pratt & James, 1994). According to the British Code of Advertising, Sales Promotion and Direct Marketing, 2010 they claim that to have responsible advertising it is necessary that marketing communications should be decent, legal, truthful and honest. And should be responsible for the consumers and for society (Hyman, 2009; CAP Code, 2010). Cunningham (1999) defined advertising ethics as: “what is right or good in the conduct of advertising function. It is concerned with questions of what ought to be done, not just with what legally must be done” (p.500).

2.2. Non-profit Sector Advertising

The non-profit sector has an important role in many countries, since they provide social services and community services that the government of those countries can´t afford. These organizations help coordinate humanitarian activities for the less fortunate and attempt to build a humane society (Chang & Lee, 2009). Therefore, non-profit sector uses advertising to promote their cause. For which, NPOs advertisements aim to motivate people to donate either money or time (Reed, Aquino, & Levy, 2007).

The theory states that for messages of social causes to be effective, they must motivate some form of behavioural change (Sciulli et al., 2005) and it seems that emotion is better for encouraging behavioural change (Andreasen, 1994; Bagozzi, Baumgartner; Pieters & Zeelenberg, 2000). Different researches in advertising and marketing had conclude that emotions are important for human behaviour and decision making (Bagozzi, Gopinath, & Nyer, 1999; Poels & Dewitte, 2006). Based on that premise, there are two types of emotional framings that are most commonly used in non-profit advertisement: to increase the intention to donate or to enhance the attitude towards the ad (positive and negative).

According to the Prospect Theory people respond differently to the same information, and this may depend if the information is framed on positive terms, in which it emphasizes in the potential gain, or if it is framed in terms of loss, which emphasizes on potential losses (Tversky & Kahneman, 1992). This is known as the “message framing effect” (Randle, Miller, Stirling, & Dolnicar, 2016).

According to various researches the advertisements used by NPOs that elicit negative emotions contribute to more donations than the advertisements that elicit positive emotions. In the same way, if the negative emotion is stronger, the intention to donate will be greater (Burt & Strongman, 2005). Whereby, according to Dekker (2011) empathy and greater levels of guilt actually result into a greater amount of funds donated. Ford & Merchant (2010) analyzed the effect on donation intention when charity advertisement recall personal nostalgia, they conducted three experiments, and the results showed that using nostalgic charity appeals can provoke greater levels of emotions and also increase the donation intentions than using non-nostalgic appeals; also charity appeals that recall nostalgia has a better result if it recalls important memories for the consumer.Many NPOs have created a presence via websites, to generate awareness of their causes, for fundraising and for managing their brands. But now they are also employing the potential of online social network sites (Quinton & Fennemore, 2013) to achieve their goals. There is no doubt that online social networks are now integrated into the daily lives of millions of people worldwide. In Spain 85% of the internet users from 16 to 65 years old are using social networks (ELOGIA, 2019). So, this is a very essential tool for NPOs to communicate their advertisement campaigns.

The transformation of advertising to the digital world began in the early ´90s. NPOs also followed this transformation and in 1992 the first NPO started using email marketing, this NPO was Amnesty International. Nowadays, 92% of NPOs use digital advertising and have a website (Tonetti, 2019). The channels most used by NPOs to communicate with public, promote their campaigns, inform news regarding the organization and recruit volunteers are email marketing and social media. Among them, social media has the greatest potential for NPOs since it can help connect with others, create and share content, collaborate with other people and find, use, organize and reuse content (Sheombar et al., 2018).

2.3. Guilt

As mentioned before, the emotional appeal is the most used by NPO´s. There are different types of emotion-based appeals; the most common are fear, shock and guilt appeals. The feeling of guilt is one of the most common negative emotions across cultures (Chen & Moosmayer, 2020; Huhmann & Brotherton, 1997), since guilt-induced messages not only arouse or emotionally activate consumers, but also positively influence their attitudes and behaviour (Chang 2011).

Consumer guilt has been defined as an emotional state in which people may experience self-blame, remorse or self-punishment after doing something that infringe or a future infraction of his own standards of acceptable behaviour (Lascu, 1991). When people feel guilty, they are concern of a real or potential negative situation and tries to reduce their level of guilt by making retribution (Chang, 2012).

Huhmann & Brotherton (1997) suggested three types of guilt: (1) anticipatory, (2) reactive and (3) existential. The anticipatory guilt is the guilt expirienced when there is a potential violetion against one´s own standard of acceptable behavior. The reactive guilt refers to the guilt felt when violeting one´s internalized standards of acceptable behavior. Finally, existencial guilt is when one becomes aware of the discrepancy between one´s fortune or contentment and others, and one may feel more fortunate than others which results in feeling empathy.

The existencial guilt, also known as social-responsibility guilt, is experienced as a consequence of a discrepancy between one’s well-being and of others (Cotte et al., 2005). It is used in persuasive communications of NPOs since in the context of donation, the desire to reduce feelings of guilt are associated to the egoistic motives for helping, which NPOs benefit from (Hibbert, Smith, Davies, & Ireland, 2007).

On the other hand, Chang (2014) identified three forms of guilt based on the analysis of where the guilt comes from: (1) existing guilt, an enduring affective state that exists prior to the ad exposure; (2) integral guilt, aroused by ad content; and (3) incidental guilt, which is elicited by contextual messages (i.e. preceding magazine articles).

Despite the different forms to classify guilt, it is important to differentiate guilt from other negative emotions such as shame, since both include the feeling that the person is responsible for a negative outcome, but guilt is focused on the personal feeling for failing others or violating his own standards. In contrast, shame is centred on how other people may evaluate the failure. Another feeling that is related to guilt is regret, but the difference is that someone may feel regret when is not satisfied with a choice, and a better choice could be made.

Fear is also an emotion usually related to guilt. However as Pinto and Priest (1991) point out there are differences between these two constructs because while guilt is defined as a posteriori emotional reaction, fear represent a priori response. Thus, fear is an anticipatory feeling related to danger while guilty inform us that we have transgressed a moral, social or ethical principle (Ghingold, 1980).

According to Izard (1977), the use of guilt appeal can lead to change in consumer behaviour, since in a mature conscience, guilt is the primary motivational factor. Because as aforementioned when someone feels guilty, he will want to reduce the level of this feeling by making a compensation. So, advertisers use this primary motivation factor by showing that the compensation can be made and persuade them to adopt a behaviour to reduce the guilt feeling.

Symbolic self-completion theory (Wicklund & Gollwitzer, 1981) states that if one´s self-identity or self- defining is perceives as incomplete, it will trigger a restorative action to remedy the situation (Lalot, Quiamzade, Falomir-pichastor, & Gollwitzer, 2019). This restorative actions leads to pro-social behaviour to compensate a transgression (Chen & Moosmayer, 2020). As a result, many advertisements are designed to evoke the feeling of guilt and promote a restorative action to reduce that feeling. This relationship between guilt and pro-social behaviour demonstrate why NPOs advertisements use this emotion.

In the same line, Hibbert et al. (2007) pointed out that the more guilt a charity ad arouses, the greater people’s donation intentions. However, other studies have shown that guilt not always lead to restorative actions. Coulter & Pinto (1995) suggested that the influence of guilt on persuasive message and consumer behaviour was described by an “inverted U” curve. Thus, high levels of guilt may create reactance because people will feel forced to experiment the guilt through the ad so, they will reject the message and will feel anger and irritation (Lwin, 2011).

2.4. Ethical Issues

As mentioned before there are some ethical issues in advertising of NPOs. Mimi Drumwright (2012) examined ethical issues in advertising in three levels, micro-meso-macro level. The “macro” perspective focuses on advertising´s effects on society, the “meso” perspective evaluates actions of advertising agencies, while the “micro” perspective focuses on the individual effect- individual advertising practitioners, individual ads, individual consumers (Tellis & Ambler, 2007). The evaluation of ethics of advertising in these three levels suggests that the influence of advertising is extensive, as anyone is likely to be exposed and, as a result, have a reaction to advertising (Sheehan, 2013). This article focuses on the “micro” perspective effect of NPO advertising on the individual consumers.

According to MacQuillin & Sargeant (2018) fundraising ethics have received little scholarly attention. There is sparse work of ethical theorizing by scholars in philanthropy and fundraising which haven´t proposed a coherent normative theory that might inform the ethics applied in this profession. But the problem is not new, there have been critics that while images of suffering and desperate people may capture the attention, move emotions and promote donations, they also illustrate people from developing countries as hopeless and helpless, without the support of these organizations (Nathanson, 2013). The term 'degree zero images' refers to those images that want to illustrate a given portion of reality in an precise way (Grancea, 2015).

Another of the ethical dilemmas of NPO advertising is that the images used in their campaigns makes that people have a view of the Third World as a place of misery. The extensive sense of hopelessness, which is strengthen by the news and NPO campaigns, understandably incite responses that range from indifference to aversion (Nathanson, 2013).

So, negatively-valence images that present victims of different social problems may suggest ethical problems. The accusations most frequently appealed include, increased anxiety among vulnerable viewers, increased satisfaction from those not affected by the problem, lack of respect for the dignity of the people presented in the advertisement (Grancea, 2015).

As stated in the Association of Fundraising Professionals' International Statement on Ethical Principles in Fundraising (2018) "Funds will be collected carefully and with respect of donor's free choice, without the use of pressure, harassment, intimidation or coercion." This statement presents another ethical problem since a person might consider felt pressured because they saw an advertisement where they made him feel guilty by the use of explicit images and threatening messages and felt they must do something but couldn´t afford it, making him feel guilty about it. So, the general ethical question of whether it is appropriate for donors to feel guilty if they decide not to donate to a cause, might be considered a form of pressure (MacQuillin & Sargeant, 2018). This is the aspect we are going to focus in this paper.

As mentioned before NPOs advertisements use negative-valence images to elicit negative emotions such as guilt to increase the willingness to donate, it is a question that is important to aboard if it is appropriate for donors to feel guilty when they are requested for a donation through advertising. MacQuillin & Sargeant (2018) aboard this question through different ethical theories (trustism, donorcentrism and service of philanthropy). According to the trustism it is not acceptable for a donor to feel guilty for not donating, since it could suggest that in the long-term making donors feel guilty will weaken public trust, which will affect lon-term incomes. From the donorcentrism (consequentialist) point of view it is not adequate making donors feel guilty, since this might generate short-term benefit, in the long-term people less likely donate. Also from the donocentrism (deontological) point of view it is not adequate since this might evoke negative emotions in donors and it is wrong to make people feel guilty. And according to the service philanthrophy it is not ethical, considering that donors won´t experience significance in their charity if they are been pressured to make a donation.

So, even though the use of the feeling of guilt may be effective in some cases or in short-term to promote pro-social behaviours and increase the donations for NPOs, it is important to consider that the use of guilt may make people experience distress and discomfort (Graton & Mailliez, 2019).

Another aspect to consider is that as mentioned before with the increase of digital advertising, NPOs have also increased their presence in social media, websites and email marketing. This makes that people is always exposed to negative advertisements of NPOs without permission. Exposing a person to against their will to harmful images may create unnecessary consumer anxiety (LaTour & Zahara, 1989; Hastings et al., 2004). Also, according to Reactance Theory, forceful messages are rejected by audiences due to perceived loss of freedom to choose their own course of action (Lwin, 2011) so, an over exposure to guilt advertising may create aversion or rejection of the message (Coulter & Pinto, 1995).

3. CONCLUSIONS

Even though there is evidence that guilt appeals are persuasive and negative emotions could contribute to more donations than the advertisements that elicit positive emotions, marketers and advertisers should use these appeals with caution. Since is still not clear the relationship between guilt and pro-social behaviour and taking into account that the results might be more negative than positive, levels of guilt must be controlled because as aforementioned if guilt surpasses the tolerable limits it might have reactance effects.

There are genuine concerns about the ethical issues these guilt-inducing advertising may have, considering ethical theory and ethical codes that suggest there is potential danger and discomfort for individuals in the use of guilt appeals. It is important for marketers and advertisers to examine the negative effects and the long-term effects these appeals may have on donations and pro-social behaviour.

Social marketers must re-examine their preference for negative appeals in advertising and consider also the welfare of the donors and not only look at the immediate results of donations, since after all, these donations may be affected in the future because guilt may encourage avoidance and reactance effects. And the continuous exposure to these types of advertisements may reduce its attention and thus its effectiveness. It is a fact that NPO provides well-being and that they work to improve the lives of many people but in the end, the question is does the end justify the means?

4. FUTURE INVESTIGATIONS

Following up with the fact that there is no systematic link between guilt and donations or pro-social behaviours, it is necessary to go deeper into the analysis of the guilt. In order to clarify if it is possible to make ethical and responsible management of this resource, it is necessary to study the cognitive processes involved in guilt appeal and how factors like the implication with the cause, its proximity to people or even the ethical position of the individual could be mediators in the relationship between guilt and consumer behaviour. Besides, although the present study is focused on guilt, it would be interesting to examine how unethical could be other advertising resources. Our goal as marketers must be making advertising more effective but also more ethical.

ACKNOWLEDGMENTS

This work is supported by the Spanish Ministry of Economy under grant RTC2016-4718-7.

REFERENCES

Association of Fundraising Professionals. (2017). International statement on ethical principles infundraising. http://www.afpnet.org/Ethic s/IntlA rticl eDeta il.cfm?ItemN umber =3681.

Andreasen, A. R. (1994). *Social Marketing: Its Definition and Domain*. *13*(I), 108–114.

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*(2), 184–206. https://doi.org/10.1177/0092070399272005

Bagozzi, R. P., & Moore, D. J. (2006). Public Service Advertisements: Emotions and Empathy Guide Prosocial Behavior. *Journal of Marketing*, *58*(1), 56. https://doi.org/10.2307/1252251

Bebko, C., Sciulli, L. M., & Bhagat, P. (2014). Using Eye Tracking to Assess the Impact of Advertising Appeals on Donor Behavior. *Journal of Nonprofit and Public Sector Marketing*, *26*(4), 354–371. https://doi.org/10.1080/10495142.2014.965073

Burt, C., & Strongman, K. (2005). Use of images in charity advertising: Improving donations and compliance rates. *International Journal of Organisational Behaviour*, *8*(8), 571–580. Retrieved from http://www.usq.edu.au/extrafiles/business/journals/HRMJournal/InternationalArticles/Volume 8/Burt Vol 8 no 8.pdf?origin=publication_detail

CAP Code. (2010). *The CAP Code Editon 12*. Retrieved from https://www.asa.org.uk/uploads/assets/uploaded/7c856612-4a2b-4c89-bd294ab8ff4de0a7.pdf

Chang, C.-T., & Lee, Y.-K. (2009). *Framing Charity Advertising: Influences of Message Framing, Image Valence, and Temporal Framing on a Charitable Appeal 1*.

Chen, Y., & Moosmayer, D. C. (2020). When Guilt is Not Enough : Interdependent Self - Construal as Moderator of the Relationship Between Guilt and Ethical Consumption in a Confucian Context. *Journal of Business Ethics*, *161*(3), 551–572. https://doi.org/10.1007/s10551-018-3831-4

Cho, C.-H., & Cheon, H. J. (2004). Why do people avoid advertising on the internet.pdf. *Journal of Advertising*, *33*(4), 89–97.

Cotte, J., Coulter, R. A., & Moore, M. (2005). Enhancing or disrupting guilt: The role of ad credibility and perceived manipulative intent. *Journal of Business Research*, *58*(3 SPEC. ISS.), 361–368. https://doi.org/10.1016/S0148-2963(03)00102-4

Coulter, R. H., & Pinto, M. B. (1995). Guilt Appeals in Advertising: What Are Their Effects? *Journal of Applied Psychology*, *80*(6), 697–705. https://doi.org/10.1037/0021-9010.80.6.697

Dekker, E. (2011). *A sad little child , a charity ' s cash cow? An experiment into the effects of photographs on donation behavior in case of a natural disaster*.

Drèze, X., & Hussherr, F.-X. (2003). INTERNET ADVERTISING : IS ANYBODY WATCHING ? *Journal of Interactive Marketing*, *17*(4), 8–23. https://doi.org/10.1002/dir.10063

ELOGIA. (2019). Estudio anual de Redes Sociales IAB 2019. *IAB Spain*, *2019*, 52. Retrieved from https://iabspain.es/wp-content/uploads/estudio-redes-sociales-2018_vreducida.pdf

Ford, J. B., & Merchant, A. (2010). Nostalgia drives donations: The power of charitable appeals based on emotions and intentions. *Journal of Advertising Research*, *50*(4). https://doi.org/10.2501/S0021849910091592

Grancea, I. (2015). Visual Arguments and Moral Causes in Charity Advertising: Ethical Considerations. *Nursing Ethics*, *2*(2), 167–185. https://doi.org/10.1191/0969733005ne828oa

Graton, A., & Mailliez, M. (2019). A theory of guilt appeals: A review showing the importance of investigating cognitive processes as mediators between emotion and behavior. *Behavioral Sciences*, *9*(12), 1–10. https://doi.org/10.3390/bs9120117

Ha, L., & Mccann, K. (2008). *An integrated model of advertising clutter in offline and online media*. *2007*(4), 569–592. https://doi.org/10.2501/S0265048708080153

Hastings, G., Stead, M., & Webb, J. (2004). Fear appeals in social marketing: Strategic and ethical reasons for concern. *Psychology and Marketing*, *21*(11), 961–986. https://doi.org/10.1002/mar.20043

Hibbert, S., Smith, A., Davies, A., & Ireland, F. (2007). Guilt Appeals: Persuasion Knowledge and Charitable Giving. *Psychology & Marketing*, *24*(August 2007), 723–742. https://doi.org/10.1002/mar

Hornik, J., Ofir, C., & Rachamim, M. (2017). *Advertising Appeals , Moderators , and Impact on Persuasion : A Quantitative Assessment Creates a Hierarchy of Appeals*. *57*(3).

Howe, P., & Teufel, B. (2014). Native advertising and digital natives : The effects of age and advertisement format on news website credibility judgments Native Advertising and Digital Natives : The Effects of Age and Advertisement Format on News Website Credibility Judgments. *Spring*, *4*(1).

Huhmann, B., & Brotherton, T. (1997). A content analysis of guilt appeals in popular magazine advertisements. *Journal of Advertising*, *26*(2), 35–45.

Hyman, M. (2009). *Responsible Ads : A Workable Ideal*. 199–210. https://doi.org/10.1007/s10551-008-9879-9

Lalot, F., Quiamzade, A., Falomir-pichastor, J. M., & Gollwitzer, P. M. (2019). When does self-identity predict intention to act green ? A self-completion account relying on past behaviour and majority-minority support for pro- environmental values. *Journal of Environmental Psychology*, *61*(April 2018), 79–92. https://doi.org/10.1016/j.jenvp.2019.01.002

LaTour, M. S., & Zahara, S. A. (1989). Fear Appeals As Advertising Strategy : Should They Be Used ? *Journal of Consumer Marketing*, *6*(2), 61–70.

MacQuillin, I., & Sargeant, A. (2018). Fundraising Ethics: A Rights-Balancing Approach. *Journal of Business Ethics*, (0123456789), 1–12. https://doi.org/10.1007/s10551-018-3872-8

Nathanson, J. (2013). The Pornography of Poverty: Reframing the Discourse of International Aid's Representations of Starving Children. *Canadian Journal of Communication*, *38*(1), 103–120. https://doi.org/10.22230/cjc.2013v38n1a2587

O´Shaughnessy, J., & O´Shaughnessy, N. J. (2003). *Persuasion in Advertising*.

Poels, K., & Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*(1), 18–37. https://doi.org/10.2501/S0021849906060041

Poels, K., & Dewitte, S. (2009). Getting a Line on Print Ads: Pleasure and Arousal Reactions Reveal an Implicit Advertising Mechanism. *Journal of Advertising*, *37*(4), 63–74. https://doi.org/10.2753/joa0091-3367370405

Pratt, C., & James, E. L. (1994). Advertising Ethic: A contextual response based on classical ethical theory. *Journal of Busines*, *13*(6), 455–468.

Quinton, S., & Fennemore, P. (2013). Missing a strategic marketing trick? The use of online social networks by UK charities. *International Journal of Nonprofit and Voluntary Sector Marketing*, *18*, 36–51. https://doi.org/10.1002/nvsm

Randle, M., Miller, L., Stirling, J., & Dolnicar, S. (2016). Framing advertisements to elicit positive emotions and attract foster carers: An investigation into the effects of advertising on high-cognitive-elaboration donations. *Journal of Advertising Research*, *56*(4), 456–469. https://doi.org/10.2501/JAR-2016-049

Reed, A., Aquino, K., & Levy, E. (2007). Moral identity and judgments of charitable behaviors. *Journal of Marketing*, *71*(1), 178–193. https://doi.org/10.1509/jmkg.71.1.178

Septianto, F., & Tjiptono, F. (2019). The interactive effect of emotional appeals and past performance of a charity on the effectiveness of charitable advertising. *Journal of Retailing and Consumer Services*, *50*(April), 189–198. https://doi.org/10.1016/j.jretconser.2019.05.013

Sheombar, A., Urquhart, C., & Kayas, O. (2018). *Social Media and Development : Understanding NGO practices and perceptions Social Media and Development : Understanding NGO*. (December).

Tonetti, A. (2019). *Fundraising and Online Marketing. How social media have an impact on the growth of a charity organization*. Instituto Universitario de Lisboa.

Truong, Y., McColl, R., & Kitchen, P. (2010). Practitioners' perceptions of advertising strategies for digital media. *International Journal of Advertising*, *29*(5), 709–725. https://doi.org/10.2501/s0265048710201439

Zhang, H. (2014). *Be rational or be emotional : advertising appeals , service types and consumer responses*. https://doi.org/10.1108/EJM-10-2012-0613

The ETHICOMP Book series fosters an international community of scholars and technologists, including computer professionals and business professionals from industry who share their research, ideas and trends in the emerging technological society with regard to ethics. Information technologies are transforming our lives, becoming a key resource that makes our day to day activities inconceivable without their use. The degree of dependence on ICT is growing every day, making it necessary to reshape the ethical role of technology in order to balance society's 'techno-welfare' with the ethical use of technologies. Ethical paradigms should be adapted to societal needs, shifting from traditional non-technological ethical principles to ethical paradigms aligned with current challenges in the smart society.

UNIVERSIDAD
DE LA RIOJA

UNIVERSITAT
ROVIRA i VIRGILI