ETHICOMP 2024

The leading role of smart ethics in the digital world

ETHICOMP BOOK SERIES

Edited by
MARIO ARIAS-OLIVA
JORGE PELEGRÍN-BORONDO
KIYOSHI MURATA
MAR SOUTO ROMERO

UNIVERSIDAD DE LA RIOJA
Cátedras Telefónica
UNIVERSIDAD COMPLUTENSE MADRID
UNIVERSITAT DE BARCELONA
UNIVERSITAT ROVIRA i VIRGILI
Logroño

Edited by

Jorge Pelegrín-Borondo

Mario Arias-Oliva

Kiyoshi Murata

Mar Souto Romero

ETHICOMP 2024

The Leading Role of Smart Ethics in Society

ETHICOMP Book Series

UNIVERSIDAD DE LA RIOJA

Cátedras Telefónica

UNIVERSIDAD COMPLUTENSE MADRID

UNIVERSITAT DE BARCELONA

UNIVERSITAT ROVIRA i VIRGILI

Logroño

ETHICOMP BOOK SERIES

Title	The Leading Role of Smart Ethics in the Digital World
Edited by	Mario Arias-Oliva ((Complutense University of Madrid), Jorge Pelegrín-Borondo (University of La Rioja), Kiyoshi Murata (Meiji University), Mar Souto Romero (Rey Juan Carlos University)
ISBN	978-84-09-58161-0
Local	Logroño, Spain
Date	2024
Publisher	Universidad de La Rioja

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

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

ISBN 978-84-09-58161-0

To those who care about the human side beyond technology

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

THE COUNTERVAILING POWER OF AI DAOS INFLUENCES VALUE TRANSFORMATION; BITCOIN (POW) VS. ETHEREUM (POS)

Kazuyuki Shimizu

School of Business Administration, Meiji University (Japan)

shimizuk@meiji.ac.jp

ABSTRACT

In this paper, we investigate the process of value transformation influenced by the countervailing power of decentralised autonomous organisations (DAO) controlled by artificial intelligence (AI). There are various values in society. The Internet was believed to lead humanity better by further decentralizing various values. However, the uneven distribution of information causes many problems, such as "cyber cascades", "filter bubbles", and "echo chambers" etc. To solve these problems, DAO using blockchain is expected to become a method of solving those problems in the Internet space.

This paper intends to capture the value change in three steps. In the first step, we take two philosophical approaches. First, using Hegel's dialectic, we attempt to compare A. current social values (thesis), B. the value of, for example, AI DAOs (antithesis), and C. new values (synthesis) in an Internet world where these contradictions exist (A + B = C). The second philosophical approach is to consider the mutuality of Bitcoin (POW) and Ethereum (POS) by adapting Popper's World 1, 2, and 3 models.

In the second step, we examine the three core capabilities that amplify "credit" in decentralised finance (DeFi): exchange swap rates, staking, and indices (portfolio). Finally, we consider the value dispersion within the Bitcoin and Ethereum networks as the main two poles. The countervailing power between these two poles and the upcoming expanding AI DAOs, especially Bitcoin-halving, coordinate the interests of stakeholders due to the uneven distribution of values and encourage the interaction between diverse values in cyberspace.

KEYWORDS: The countervailing power, AI, DAO, Web3, DeFi, artificial society.

1. METHODOLOGY

Humanity has various values: legitimacy, human rights, freedom, security, dignity and human life. First, consider two methodologies. Dialectics considers matter to be the root of spirit. Hegel's dialectics: For example, consider value relationships using this dialectic. Also, we applied Popper's World 1/2/3 theory to the current value relationships. Idealism emphasizes the compositional ability inherent in the mind rather than matter.

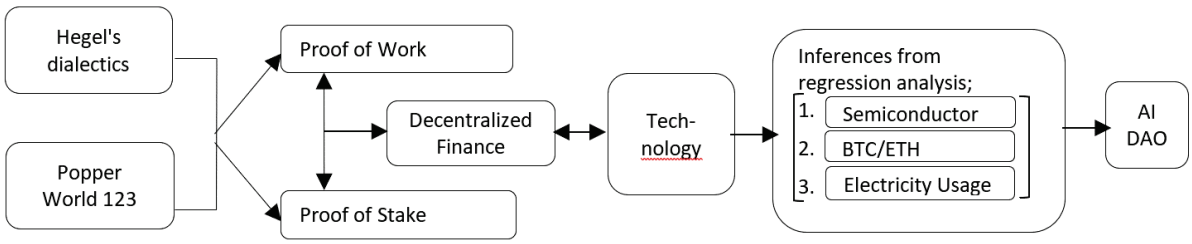
- Based on **Hegel's dialectics**, A. Thesis: ex. "openness of blockchain" + B. Antithesis: "private information" = C. Synthesis: "using hybrid technology of public and private chain".
- **Popper's Worlds I, II. III**; Interaction between World I: the "openness of blockchain" ~ the dilemma between World I and World II: "private information",+ World III: the new Web 3 interaction with the captured values. (Karl, 1978)

The global crypto market cap is $1.16T on June 2023. As mentioned above, Bitcoin's dominance is around half of the total crypto market (46.53% in June 2023). The share of Ethereum is around 19% (Omkar, 2023). The total crypto market trading volume is $32.26B. The total volume in DeFi is about 8% ($2.37B), and the volume of all stablecoins is about 92% ($29.69B) of the total crypto market. We will examine the concept here with POW, POS and the rest of the consensus algorithms for simplicity and understanding.

- Based on Hegel's dialectics, A. Bitcoin's POW (thesis) + B. Ethereum's POS (antithesis) = C. co-evolution of both consensus algorithms (synthesis). The simple formula is A + B = C.
- Popper's Falsifiability and Worlds I, II. III; Bitcoin's POW (world 1), Ethereum's POS (world 2), and co-evolution of both consensus algorithms (world 3). Bitcoin's POW (world 1), Ethereum's POS (world 2), and co-evolution of both consensus algorithms, such as NPoS (Nominated Proof of Stake), PoH (Proof of Histrory), PoA (Proof of Authority), PoC (Proof of Consensus) etc (world 3).
  - ✧ Diagram of knowledge growth by Popper's idea:

P□ (Problem 1) – TT (Tentative Theory) – EE (Error Elimination) – P (Problem 2)

Figure 1. The logic tree of this paper.



2. DEFI (DECENTRALIZED FINANCE) AS A VALUE ACCELERATOR

2.1. Brief history of Defi

What we look for in a financial institution is trust. For this reason, financial institutions are also called credit institutions. Financial institutions are trusted because they can settle transactions smoothly. The source of trust in financial institutions is that they do not lie. A similar trust is formed if this source of trust is unbreakable Cryptography. The POW blockchain project, Bitcoin, secures this trust.

Personal overseas money transfers and on-demand transactions became possible, enabling several online value transformations. As a result, cryptocurrencies have significantly impacted traditional centralised financial (Cefi). However, decentralised finance (Defi) in the early 2000s was still in the prototype technologically, and its business application had just begun (Nakamoto, 2008). Ethereum was developed by Vitalik Buterin in 2013 as the research and development of blockchain technology progressed (Buterin, 2014). By incorporating smart contracts based on cryptocurrencies, blockchain technology can be used for various purposes.

Furthermore, by changing Bitcoin's Proof of Work (PoW) problem to Proof of Stake (PoS), the problem of energy consumption can now be addressed. By decentralising such diverse values, the Internet was believed to lead society in a more pluralistic and better direction. To solve such issues, DAO on Web3, powered by blockchain, is expected to promote the democratisation of the "human" internet space (Simon, 2023).

2.2. Orderbook vs Liquidity Pool at Leyer 1

Uniswap is a decentralised exchange (DEX) and part of the decentralised finance (Defi) product ecosystem launched on Ethereum mainnet in November 2018. It replaces the traditional order book type trading on centralised exchanges (CEX). At a very high level, an AMM (Automated Market Maker) replaces the buy and sell orders in an order book market with a liquidity pool of two assets valued relative to each other. As one asset is traded for the other, the relative prices of the two assets shift and a new market rate for both is determined. In this dynamic, a buyer or seller trades directly with

the pool rather than with specific orders left by other parties (Uniswap, 2023). Liquidity providers can be regarded as investors in the decentralised exchange and earn fixed commissions per trade. They lock up funds in liquidity pools for distinct pairs of currencies allowing market participants to swap them using the improved price function**.** Liquidity providers take on market risk as liquidity providers in exchange for earning commissions on each trade. In short, Investors as a customer accept market risks usually taken by traders. This new pool trading has a risk profile of a liquidity provider and the so-called impermanent (unrealised) loss in particular (Andreas & Gurvinder, 2021).

Table 1 below explains the following; **Uniswap** introduced the constant product market maker formula to ensure continuous liquidity in exchanging tokens on Ethereum. The formula follows: x is token 1, y is token 2, and k is a constant.

Curve's primary distinction from other decentralised exchanges, such as Uniswap, is its low slippage and low fee algorithm specifically designed for trading between assets of the same value. Curve makes it very useful for stablecoin swaps, as they are expected to hold roughly the same value [Perpetual Protocol, 2022].

Balancer's pools are like index funds that construct a portfolio of assets with fixed weights. The balancer protocol allows each pool to have two or more assets and to supply them in any ratio. Each asset reserve is given a weight when the pool is created and the weights sum to one.

Table 1. A pattern of AMM.

AMM	Connection of individual tokens
Uniswap	X * Y = K
Balancer	(X * Y * Z) ^ (1/3) = K
Curve	(X * Y) + (X + Y) = K

Source: JBA Defi study group.

3. BITCOIN (POW) VS. ETHEREUM (POS)

The competitive principle of getting Coinbase through mining functions as a solution to the 51% problem (double spending problem). After mining, the BTC transaction ledger is recorded and held as a single block on c.a. 16,000 nodes worldwide, and these nodes confirm this block once every 10 minutes. This is the final approval of transaction information on the final distributed ledger. This approval method is based on a block number (nonce: a hash value with approximately 16 leading zeros) that is encrypted using a consensus-building process (perfectly competitive authentication process) by an unspecified number of participants called POW (Proof of Work). It is a competition for discovery. This discovery competition is held worldwide, and because it is a winner-take-all system, the problem of wasted energy use has been pointed out. Considering the community debate surrounding the future trajectory of Bitcoin versus Ethereum consensus protocols, there's ongoing discussion about the feasibility of transitioning Bitcoin to a Proof of Stake (PoS) mechanism (ethereum.org, 2024). While Ethereum has successfully implemented PoS, Bitcoin continues to rely on the Proof of Work (PoW) mechanism. [Alex , 2023]

3.1. Consensus mechanisms

Fundamental differences: PoW requires miners to solve complex mathematical problems to validate transactions, while PoS relies on validators who hold a certain amount of cryptocurrency. Here we believe that Community consensus applies to the idea of the Popper 123 world, The Bitcoin community

is divided on whether transitioning to PoS is necessary or beneficial, with some advocating for maintaining the current PoW model.

Concerning energy efficiency, PoS is considered more energy-efficient than PoW, which consumes significant electricity. the gas fee and time (transaction cost) using POW are much higher than POS. Security concerns are Some argue that PoS may be more vulnerable to attacks or centralization, while others believe it can provide equal or even higher security compared to PoW. The consensus is different for both algorithms, which are mentioned above. Ultimately, whether Bitcoin can switch to PoS will depend on the consensus reached within the community and the perceived benefits and drawbacks of such a transition.

Table 2. Differences between BTC and ETH.

Feature	**Bitcoin (BTC)**	**Ethereum (ETH)**
Primary Purpose	Digital Gold, Store of Value	Platform for Smart Contracts and Decentralized Applications
Transaction Speed	Average 10 minutes (block generation time)	About 15 seconds (block generation time)
Transaction Cost	Can be high in times of network congestion	Gas fees (variable transaction fees)
Consensus Mechanism	Proof of Work (PoW)	Transitioning to Proof of Stake (PoS)
Supply Limit	Capped at 21 million BTC	No fixed supply limit (though issuance rate varies)

Source: [Alex , 2023] (Alistair & Eleftherios, 2020)

"The Merge" of Ethereum on 15 September 2022, fundamentally altered the network's architecture. The upgrade has considerable implications for the network's electricity usage, which we estimate to have decreased by a significant 99.99%. This reduction is caused by PoS not requiring the same level of computational power as PoW, as it relies on validators who hold a stake in the network, rather than miners, who utilise powerful hardware to solve cryptographic puzzles.

POS consensus mechanism is used to ensure trustless, fundamentally different from proof-of-work. Unlike POW, POS substitutes the resource cost associated with the computationally intensive process of guessing random numbers to solve a cryptographic puzzle with a requirement to pledge financial resources in the form of the blockchain's native tokens as collateral, so-called "staking". To participate in attesting or proposing new blocks, those who are validators must lock or "stake" a set number of native tokens, for instance, ether (ETH) in the case of Ethereum.

How do people usually give value? We adapt Weber's concept of authority to consensus algorithms like Bitcoin. Barnard describes the acceptance of authority (the realm of indifference) in relation to standards. This criterion is a "statistically significant difference." Interests and motivations become a way to unify the boundaries between general and individual values.

3.2. Inferences from regression analysis

The multiple regression analysis here will explain the dominance situation of BTC and ETH (CoinMarketCap, 2023), using by R. Therefore, in this study, the dependent variable is semiconductor sales (WSTS, 2023), and the independent variable is the dominance status of BTC and ETH. Semiconductor sales and the BTC and ETH dominance status (each amount of market capitalization) are based on 10-year time series data from 2013 Jan to 2023 Jan weekly basis.

The dependent variable is global semiconductor sales, and the independent variables are the market capitalization of BTC and ETH. As a result, BTC and ETH's market capitalisation had 1% significance in

predicting global semiconductor sales. Different letters indicate significant differences by *** test (P < 0.01). All data were Z-transformed, but the same results were obtained by inverse Z-transformation.

Adjusted R-squared: 0.7015 , **BTC (b=0.43, SE=0.07, t(552)=5.97, P=4.09e-09 ***)**,

ETH (b=0.42, SE=0.07, t(552)= 5.72, P=1.80e-08 *)**.

It was confirmed that BTC (POW) and ETH (POS) have a 1% significance in terms of semiconductor sales using both consensus algorithms. It seems to confirm the relationship between POW and POS consensus algorithms and technological innovations in the semiconductor (silicon) cycle that occur every four years.

The dependent variable is Asian semiconductor sales, and the independent variables are the market capitalization of BTC and ETH. As a result, the market capitalization of BTC and ETH significantly predicted semiconductor sales in Asia.

Adjusted R-squared: 0.6734, **BTC (b=0.68, SE=0.024, t(552)=8.78, P=<2e-16 ***)**,

ETH (b=0.16, SE=0.024, t(552)= 2.09, P=0.0372 *)

The dependent variable is U.S. semiconductor sales, and the independent variables are the market capitalization of BTC and ETH. As a result, ETH's market capitalization was a strong predictor of U.S. semiconductor sales. However, BTC did not show any significance for U.S. semiconductor sales.

Adjusted R-squared: 0.6394, **ETH (b=0.695, SE=0.08, t(552)=8.68, P=<2e-16 ***)**,

BTC (b=0.11, SE=0.08, t(552)= 1.375, P=0.17).

We conducted a multiple regression analysis of the relationship between two types of semiconductor sales by region (America and Asia) and BTC and ETH. As a result, the 1% significance in semiconductor sales in the Asian region over BTC was calculated to be higher than the 10% advantage over ETH. Additionally, the 1% significance of ETH in semiconductor sales in the Americas region was calculated to be more significant than BTC.

As pointed out in Dániel Kondor's paper (Dániel , Márton , István , & Gábor , 2014), “The Bitcoin system has 6.28% of addresses owning 93.72% of the total assets”. BTC mining is done by one specific Chinese mining site, where electricity is cheap. This Chinese consensus is communism and a POW-like decision-making method, and in the United States it is famous that the richest 1% holds a 90% of total American financial assets.

The WSTS 2024 Semiconductor Market Outlook, forecast in the fall of 2023, predicts that the global semiconductor market will increase by 13.1% and reach a valuation of US$588 billion. This growth is primarily driven by the memory sector, which is expected to soar to approximately US$130 billion in 2024, an increase of more than 40% year-on-year. IDC says that from 2024 to 2026, accelerating demand for AI servers and AI-enabled endpoint devices will drive semiconductor content growth and drive new upgrade cycles across the enterprise [Shirer, 2023]. Current cloud-based AI processing is inefficient because it requires energy to transfer data because it is processed by a distant system. AI-enabled endpoint devices separate simple and complex information processing and perform most of their processing on general-purpose memory, CPUs, GPUs, and ASICs (application-specific integrated circuits) where data is generated.

3.3. Supply and Demand of Semiconductor and Electricity

What does it mean that the Bitcoin system is in a Pareto situation, with 93.72% of the total assets held in 6.28% of addresses? Does BTC mining reflect the geographical consensus of China? BTC's consensus

algorithm is POW. POW mining involves an unspecified number of participants creating cryptography in perfect competition. This competition of free will makes a system that does not require a third party (Adam, 2007). Prateek Goorha explains gold holding and its circulation from a historical perspective. Building a concentrated inventory of scarce goods is pointless if that inventory cannot make any circulation between social, economic, and political values. He also points out that a rare item has little meaning as a subjective meditated value if the concentration does not change. After all, there are far rarer elements (or technology) on the periodic table than gold (Prateek, 2019). In other words, the act of people exchanging ``trust'' that establishes accumulated truths has extremely important meaning for economics.

Table 3. BTC Halving Schedule.

Event	**Date Span**	**Duration (Month)**
1st BTC halving	From 03/01/2009 to 28/11/2012	47
2nd BTC halving	From 28/11/2012 to 09/07/2016	40
3rd BTC halving	From 09/07/2016 to 18/05/2020	44
4th BTC halving	From 18/05/2020 to around ??/05/2024	Around 48
5th BTC halving		Latest block No.825,278
	Continue to 2140 every 210,000 blocks	

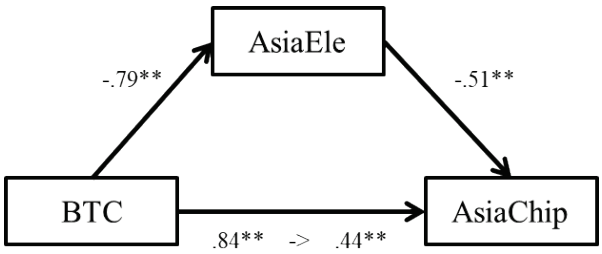
Source: (Hasan, Spring 2021)

Bitcoin halving, also known as "Coinbase halving," is a significant event in the Bitcoin network. It refers to the reduction by half of the rewards granted to miners for mining a new block on the blockchain. Initially, miners received 50 BTC per block. This reward halves approximately every four years, or after every 210,000 blocks mined. This process will continue until the total supply of Bitcoin reaches 21 million BTC. Halving reduces the rate of new Bitcoin entering circulation, acting as a deflationary mechanism to potentially increase the value of Bitcoin over time due to its reduced supply growth. As of 2024, people are focusing on the aspect of POW's BTC that the special feature has the issuing cap, being determined in the context of central banks of various countries implementing monetary easing and accelerating global inflation after the COVID-19 pandemic that hit the world.

BTC mining should not be used in an unsustainable model that ignores environmental issues. Therefore, The Cambridge Bitcoin Electricity Consumption Index (CBECI) analyses the relationship between BTC and ETH (using POW) electricity consumption (Cambridge Center for Alternative Finance, 2022). Consequently, we conducted a regression analysis using electricity usage data as a mediating variable on the above-mentioned regional semiconductor sales data and market capitalisation data of BTC.

We conducted a mediation analysis for the period from September 2019 to January 2022 to see if electricity demand in Asia mediates the impact of BTC on semiconductor sales in Asia. First, we conducted a regression analysis using Asian semiconductor sales as the dependent variable and BTC as the explanatory variable. As a result, BTC had an advantage in predicting semiconductor sales in Asia (b=0.84, SE=0.84, t(125)=17.39, P=0.00). Furthermore, by adding power consumption in Asia as an explanatory variable, Asia's electricity consumption significantly predicted semiconductor sales in Asia (b=-0.51, SE=-0.51, t(125)=-7.92, P=0.00). And the effect of BTC also had an advantage. (b=0.44, SE=0.44, t(125)=6.92, P=0.00).Since this data has been normalized, the standardization coefficient and standard error are the same value.

Figure 2. DAG for Asian power supply and demand as mediating effects.



Source: HAD

What is very interesting here is that the coefficient in Asia is negative. This reflects the impact of China's mining regulations since mid-2021. If such regulations are excluded, in order to confirm whether the US power demand will lead to the effects of BTC in the United States in the United States, the same period from September 2019 to January 2022 can be analyzed. Was implemented. First, a regression analysis was implemented with subordinate variables and BTC as explanatory variables with US semiconductor sales. As a result, the BTC was advantageous to prediction of semiconductor sales in Asia (b = 0.73, se = 0.73, t (125) = 11.98, p = 0.00). Furthermore, if the US power consumption (AmeEle) was added as a description variable, the US power consumption was greatly predicted to the US semiconductor sales, AmeChips (b = 0.64, SE = 0.64, T (125) = 9.2, p = 0.00). The effect of BTC was also advantageous. (B = 0.52, SE = 0.52, T (125) = 8.18, p = 0.00). Since this data is normalized, the standardization coefficient and the standard error are the same value.

The case in the United States reflects the trade relationship between the United States and China, which was also pointed out in The Cambridge Bitcoin Electricality Consumption Index (CBECI). The CBECI indicators indicate the relationship between semiconductors and power used as hardware. It is ETH's merge that changed this relationship.

4. AI DAO, ARTIFICIAL SOCIETIES AND POST-TRUTH

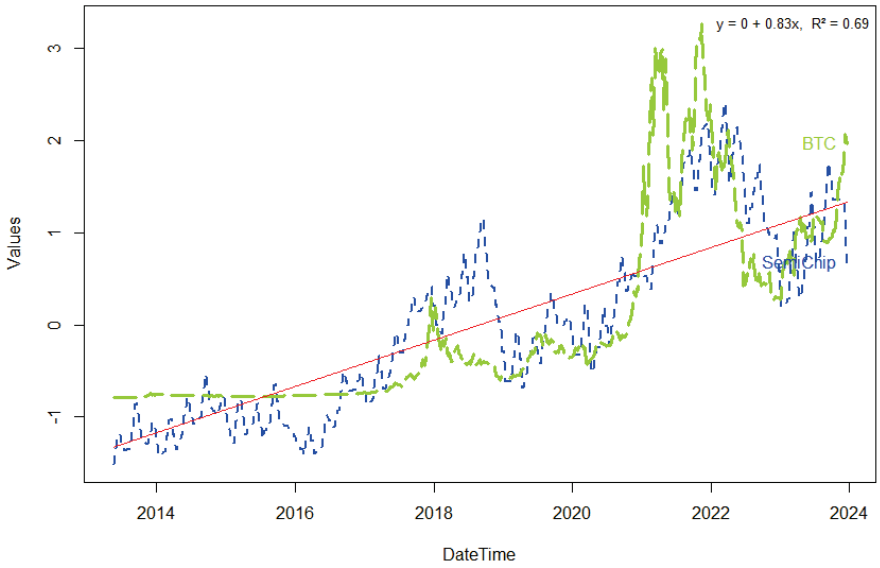
Following Hegel's dialectics, we consider the current values (thesis), the AI DAOs values (antithesis), and the new values (synthesis) in which the two are in a countervailing relationship in the Internet world (Stanford, 2020). In this research, we seek a procedure for sustainability that transcends human interests and seeks technocracy ("rule by mechanism") advocated by Thorstein Veblen, John K. Galbraith and others (Galbraith, 1952). Specifically, governance by AI and DAO. The issues of privacy, monopoly/oligopoly, human rights, freedom, security, dignity and human life thus require changes in our value orientation. There are two value distribution models: the centralised client-server and decentralised P2P models. In the former, AI is at the centre and distributes (unevenly distributed) value (Freeman, Harrison, Wick, Parmar, & Colle, 2001). On the other hand, in the P2P model, DAO by various AIs and values by humans are distributed.

In natural ecosystems, many animal populations are self-organised. For example, the behaviour of a single ant is simple and limited, but ant colonies automatically control complex behaviours such as foraging, feeding, nest building, and defence. Significantly, individuals forming ant colonies can adjust according to the division of labour and their behaviour based on environmental changes. In addition to ants, fish, bees, and swarms also exhibit similar self-organising behaviours. All these animal populations are characterised by centralised control and a lack of hierarchy. In other words, it shows that group behaviour on the Internet can be controlled autonomously. (Shuai, et al., 2019)

Also, in this network, traditional values, AI programs and his DAO democratic compete and interrelate (Computer Politics; Algocracy). This is governance based on mutual relationships (value chains) with

DAOs and AIs and countervailing power (Jacques, 1999). Schumpeter believes that innovation is fostered by entrepreneurship. Furthermore, considering economic cycles, we can assume a 4- to 50-year cycle of Chicken, Juggler, and Kondratyev (Joseph, 1939). The short-term chicken cycle is consistent with the silicon cycle. While Hegel's materialist view of history and Marx's historical materialist view of history have an influence on people's way of thinking, the ideas of idealism, such as Popper's, have also had an influence.

Figure 3. SemiChips and BTC Correlation and Regression graph.



Source: R (Public, n.d.).

Table 3 above shows the relationship between semiconductor sales and BTC market capitalization in chronological order. Then, when explaining the explained variable (SemiChips) using the explanatory variable BTC, the multiple regression coefficient is 0.69, which has moderate explanatory power, and the regression equation is Y = 0 + 0.38x, and BTC and SemiChips are 2e-16 ***, which are highly significant.

The problem of blockchain's openness and "privacy" can influence the recognition of the technology itself; anonymity destroys all possibilities for post-truth understanding (Kiyoshi & Yohko, 2021). However, value transformations are occurring through Web3 via Blockchain in privacy, monopoly/oligopoly, human rights, freedom, security, dignity, and human life. Our collective intelligence and the countervailing power of AI DAOs using blockchain technology will form this value transformation (Humans.ai, 2022). Its value transformation process is carried out in three stages according to the following Hegelian dialectics and Popper's world1/2/3 model;

1. the current values (thesis), such as the dilemma with "privacy" and POW vs POS.
2. The AI DAOs' values (antithesis) and Post-truth, such as appropriate information, are provided to professionals and lead to appropriate decisions.
3. the new values (synthesis), in which the two are in a countervailing power in the Internet world. In this network, traditional values, AI programs and his DAO democratic compete and interrelate (Computer Politics; Algocracy).

The most important characteristic of blockchain is "creditability" due to its "openness", which creates a dilemma with "privacy". Therefore, the value transformation in the " privacy " issue in blockchain varies according to the concepts of Hegel and Popper. In general, privacy is divided into a wide range

of contents. Two of the most important issues are "credit" and "falsification". This paper focuses on the characteristics (limitations) of blockchains that expose privacy. The countervailing power of values and the elimination of value boundaries coincide with the fading of the border between real- and cyberspace. The logic is that Web3 solves the problems of the traditional concept of privacy and also creates a new way of thinking (Post-truth). The "openness and trust" of blockchain prevent "lack of trust" and contribute to better decisions (Bernd , Doris , & Rowena , 2022). The two ideas of post-truth interact, and it would be more natural to think of them as an interaction rather than as polarities between POW and POS. As a future research topic, we will continue to monitor the progress of POW and POS with great interest. If AI-like capabilities were to amplify value automatically, we would have to encourage economic circulation to take place and be used in human interaction.

REFERENCES

Adam, S. (2007). *An Inquiry into the Nature and Causes of the Wealth of Nations* (Vols. 1,2,3,4 and 5). MetaLibri Digital Library, 2007. Retrieved from https://www.ibiblio.org/ml/libri/s/SmithA_WealthNations_p.pdf

Alex , S. (2023, 11 4). *Bitcoin Vs Ethereum Consensus: PoW Vs PoS Mechanisms Explained*. Retrieved from https://www.doubloin.com/learn/bitcoin-vs-ethereum-consensus

Alistair, S., & Eleftherios, K.-K. (2020, June). *GRANDPA: a Byzantine Finality Gadget.* Retrieved from grandpa.pdf: https://github.com/w3f/consensus/blob/master/pdf/grandpa.pdf

Andreas, A. A., & Gurvinder, D. (2021). UNISWAP: Impermanent Loss and Risk Profile of a Liquidity Provider. *Trading and Market Microstructure*, 16. Retrieved from https://arxiv.org/abs/2106.14404

Bernd , S. C., Doris , S., & Rowena , R. (2022). *Ethics of Artificial IntelligenceCase Studies and Options for Addressing Ethical Challenges.* Springer. https://doi.org/10.1007/978-3-031-17040-9

Buterin, V. (2014). *Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform.* ethereum.org. https://ethereum.org/669c9e2e2027310b6b3cdce6e1c52962/Ethereum_Whitepaper_-_Buterin_2014.pdf

Cambridge Center for Alternative Finance. (2022, Jan). *Evolution of network hashrate.* Retrieved from https://ccaf.io/cbeci/api/v1.2.0/download/mining_countries

CoinMarketCap. (2023, Dec). *Bitcoin (BTC) dominance*. Retrieved from https://coinmarketcap.com/charts/#What%20is%20Bitcoin%20(BTC)%20dominance?

Dániel , K., Márton , P., István , C., & Gábor , V. (2014). Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. https://doi.org/10.1371/journal.pone.0086197

ethereum.org. (2024, Jan). *The Merge*. Retrieved from https://ethereum.org/en/roadmap/merge/

Freeman, E. R., Harrison, J., Wick, A., Parmar, B., & Colle, S. (2001). *A Stakeholder Approch to Strategic Management.* Masachusetts: Blockwell.

Galbraith, J. K. (1952). *American Capitalism: The Concept of Countervailing Power.* (K. Nikawa, Trans.) Hakusuisya.

Hasan, K. (Spring 2021). Market Efficiency for Bitcoin. *Master Thesis NEKN02*, p29. Retrieved from https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9051578

Jacques, F. (1999). *Multi-Agent System: An Introduction to Distributed Artificial Intelligence.* JASSS. Retrieved from https://jasss.soc.surrey.ac.uk/4/2/reviews/rouchier.html

JIN. (2022, Feb). *The Introduction to Web 3.0 (NFT, DeFi, DAO, DApp, Cryptocurrency, GameFi, etc)*. Retrieved from https://medium.com/experience-stack/the-introduction-to-web-3-0-nft-defi-dao-dapp-cryptocurrency-gamefi-etc-8285d9525a7e

Joseph, S. A. (1939). *BUSINESS CYCLES.* McGraw-Hill Book Company. Retrieved from https://discoversocialsciences.com/wp-content/uploads/2018/03/schumpeter_businesscycles_fels.pdf

Karl, P. (1978). Three Worlds. *The Tanner Lecture on Human Values*, 1-27. Retrieved from https://tannerlectures.utah.edu/_resources/documents/a-to-z/p/popper80.pdf

Kiyoshi, M., & Yohko, O. (2021). The Privacy Paradox: Invading Privacy While Protecting Privacy. *ETHICOMP 2021* (pp. 199-201). La Rioja: Universidad de La Rioja, Universidad Complutense Madrid, CCSR De Montfort University, CBIE Meiji University.

medium.com. (2019, Jan). *Comparison of PoW, PoS And DPoS Governance Models.* Retrieved from https://medium.com/@salmanmiah/comparison-of-pow-pos-and-dpos-governance-models-dcea481140f8

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Bitcoin.org. Retrieved from https://bitcoin.org/ja/bitcoin-paper

Omkar, G. (2023, Apr 12). *CoinDesk*. Retrieved from Bitcoin, Not Ether, Builds Crypto Market Dominance Ahead of Ethereum's Shanghai Upgrade: https://www.coindesk.com/markets/2023/04/11/bitcoin-not-ether-is-becoming-more-dominant-in-crypto-market-ahead-of-ethereums-shanghai-upgrade/

Perpetual Protocol. (2022, Sep). *medium.* Retrieved from What is an Automated Market Maker (AMM)?: https://medium.com/perpetual-protocol/what-is-an-automated-market-maker-amm-a71ea1d80ea9

Prateek, G. (2019). *Bitcoinomics 101: Principles of Bitcoin's Supply, Demand & Price.* http://doi.org/10.2139/ssrn.3473279

Public. (n.d.). *Unchained Capital*. Retrieved from Bitcoin Age Distribution: https://chart-studio.plotly.com/~unchained/37/bitcoin-utxo-age-distribution/#plot

Shimizu, K. (2021). *Blockchain and Biometrics Authorization: What We Actually Count Truly Counts?* Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=8037017

Shirer, M. (2023, Dec). *Worldwide Semiconductor Market Outlook Upgraded to GROWTH from TROUGH: Semiconductor Market to Grow 20.2% in 2024 to $633 Billion, According to IDC*. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS51383823

Shuai, W., Wenwen , D., Juanjuan , L., Yong , Y., Liwei , O., & Fei-Yue , W. (2019). *Decentralized Autonomous Organizations: Concept, Model, and Applications.* IEEE Transactions on Computational Social Systems . Retrieved from https://ieeexplore.ieee.org/abstract/document/8836488

Simon, R. (2023). *Technology Ethics: The Ethical Digital Technology Trilogy.* London: Routledge.

Stanford. (2020, Oct). *Stanford Encyclopedia of Philosophy.* Retrieved Dec 2022, from Hegel's Dialectics: https://plato.stanford.edu/entries/hegel-dialectics/

Stephen, S. (2015). BITCOIN: THE NAPSTER OF CURRENCY. *J.D., University of Houston Law Center*, 581-641. Retrieved from http://www.hjil.org/articles/hjil-37-2-small.pdf

Uniswap. (2023, June). *Welcome to Uniswap Docs.* Retrieved from Order Book VS AMM: https://docs.uniswap.org/concepts/uniswap-protocol

WSTS. (2023, Dec). *Historical Billings Report*. Retrieved from https://www.wsts.org/esraCMS/extension/media/f/WST/6280/WSTS-Historical-Billings-Report-Oct2023.xlsx

Yamazaki, T., Murata, K., Orito, Y., & Shimizu, K. (2020). *Post-Truth Society: The AI-driven Society Where No One Is Responsible.* Universidad de La Rioja: ETHICOMP 2020. Retrieved from https://www.researchgate.net/publication/343099318_Post-Truth_Society_The_AI-driven_Society_Where_No_One_Is_Responsible

IS A BRAIN–MACHINE INTERFACE USEFUL FOR PEOPLE WITH DISABILITIES? CASES OF SPINAL MUSCULAR ATROPHY

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), 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

Today, a Brain Machine Interface (BMI) or Brain Computer Interface (BCI) has widespread application in various fields. In particular, the use of BMI systems in social welfare fields benefits people with disabilities who cannot control limb movement. BMI systems enable users to control external devices using brain signals. However, to date, the opportunities for people with disabilities to access and use such cyborg devices have been limited because of the high cost and technological difficulties associated with device use; hence, the potential benefits and risks of BMI device use for people with disabilities have not been sufficiently recognised or discussed in a practical sense. Accordingly, this study involves experiments in which a non-invasive wearable BMI device is used to operate a robotic arm using participants' brain signals. Further, semi-structured interviews are conducted with the participants before, during, and after the experiments to investigate the possible benefits and social risks of BMI use. Two individuals with congenital disabilities were invited to participate in the experimental surveys. Survey results provide impressive insights into the benefits and risks of BMI use in the context of human dignity, social isolation versus independence/autonomy, and self-determination of people with disabilities.

KEYWORDS: brain–machine interface, support for people with disabilities, cyborgisation, spinal muscular atrophy.

1. INTRODUCTION

Brain machine interface (BMI) or Brain Computer Interface (BCI) systems have been developed and used for various purposes, such as gaming and marketing. In the field of social welfare, BMI systems are expected to be used as an assistive cyborg technology by people with disabilities, particularly those who cannot move their limbs at will. BMI systems enable individuals without limb movement to control external devices through brain signalling (Orito et al., 2020a). Recently, BMI was successfully used to help people with speech/vocal disabilities regain their lost voice (Nikkei Shimbun, 2023). In this case, brain electrodes were implanted in the brain of a patient suffering from stroke through a hole in the skull to read brain signals that would otherwise be sent to the lip, teeth, and jaw muscles and generate a text of approximately 80 words per minute. Dedicated software was developed to move the mouth and mouth parts of the patient's avatar in accordance with a conversation and reproduce facial expressions.

The developments in and wide availability of BMI systems have generated concerns on the systems' possibilities, utilities, and potential social risks (e.g., Bernal et al., 2023, Wahlstrom, 2018, Grübler and Hildt, 2014). The potential risks and ethical issues associated with BMI device use by people with disabilities should be analysed before deploying these devices extensively in society. However, to date, people with disabilities have relatively limited access to such devices; further, even when they are aware of BMI devices, they require the aid of specialised engineers to operate the devices. Therefore,

the potential benefits and risks of BMI device use by people with disabilities have not been comprehensively discussed.

Accordingly, this study examines the ethical and social issues associated with BMI use by performing experiments using BMI systems and conducting semi-structured interview with people with disabilities before, during, and after the experiments (Orito et al., 2022a, 2022b, 2021a, 2021b). In the experimental survey, participants wore a headset-type, non-invasive wearable BMI device (an electroencephalogram input device) to remotely operate a robotic arm; subsequently, relevant semi-structured interviews were conducted. Question items were developed to investigate participants' attitudes towards BMI devices' utilities and potential risks based on the findings of authors' earlier studies (Orito et al., 2022a, 2022b, 2021a ,2021b, 2020a, 2020b; Murata et al., 2018, Murata et al., 2017, Isobe, 2013).

Earlier survey results (Orito et al., 2022a, 2022b) have identified several ethical and social issues pertaining to BMI use by people with disabilities, for example BMI devices' influence on the self-identity and relationship development of people with disabilities, gaps between the expectations and reality of using cyborg technology, and economic disparities in accessing cyborg devices. However, in a previous survey conducted by the authors, two participants with acquired disabilities indicated that this type of experimental survey should target people with congenital disabilities (Orito et al., 2022a, 2022b). According to their opinions, the attitudes, feelings, and recognition of assistive cyborg technologies, such as BMI devices, may differ between people with congenital and acquired disabilities since these groups differ in their backgrounds and acceptance of their disabilities. Therefore, this study examines the benefits and social risks of BMI devices for people with disabilities by analysing the experimental results and interview responses of two people with congenital disabilities having different demographics than the participants of the earlier study (Orito et al., 2022a, 2022b).

2. OVERVIEW OF THE EXPERIMENTAL SURVEY

2.1. Experimental survey

In this study, semi-structured interview surveys were conducted before, during, and after the experiment using BMI devices, since participants could easily evaluate the usefulness, benefits, and risks of BMI technology while actually using the devices rather than answering interview questions. In this experimental survey, the participants were asked to place a non-invasive wearable BMI/EEG (where EEG refers to electroencephalography) device on their head and control a robotic arm using their brain signals. Before the experiment, as a training operation to control the robotic arm, we recorded the participants' brain signals at the 'relaxed' and 'in-operation' states (for this purpose, the participants were asked to imagine that they were pushing a small box displayed on a computer screen). These brain signal data were transmitted and utilised through a specific application software that operated the robotic arm. After this training, the participants were required to be in the in-operation state by imaging that they were pushing the robotic arm backward. In this experiment, the brain signal patterns were appreciated in the in-operation state and registered in the software to move the robotic arm turn towards the back (see Orito et al., 2022a, 2022b, 2021a, 2021b, 2020a, 2020b).

The participants were asked to manipulate the robotic arm several times during the experiment, and they were required to respond to interview questions before, during, and after using it. The interview question items were categorised as follows: (a) privacy and personal data protection, (b) human autonomy and dignity, (c) identity development and personal transformation, (d) the acceptance of a body extension in an individual and organisational context, (e) workplace cyborgisation, (f) social responsibility and informed consent, and (g) the technological and social situation of people with disabilities (Orito et al., 2022a, 2022b, 2021a, 2021b, 2020a, 2020b).

2.2. Survey participants

In this study, experiments and interview surveys were conducted in February and March 2023 at Ehime University, Matsuyama, Japan. The study adhered to relevant reporting guidelines, all procedures were performed in accordance with the ethical standards of the Research Ethics Committee of the Faculty of Collaborative Regional Innovation, Ehime University, and the two participants provided informed consent to participate in the study.

Table 1 depicts participants' attributes. The two participants were suffering from spinal muscular atrophy (SMA). Although SMA involves muscle weakness and atrophy, the exact symptoms differ among individuals and vary widely. Both the participants used wheelchairs for mobility and independence and received 24-hour care; however, they had different symptoms. In addition, they belonged to and worked at the Center for Independent Living (CIL) to support people with disabilities who wish to have independent life. Before the survey, the participants' health conditions were confirmed. Interviews with Participant 1 were conducted once online after the experiment; all other interviews were conducted face-to-face.

Table 1. Details of the participants of the experimental survey (*n*= 2).

ID	Age	Gender	Types of disabilities, conditions	Expectation/anxiety about the experiment (Weak 0–Strong 7)
1	40s	Male	Spinal Muscular Atrophy. He could move only the thumb of the right hand. Usually, he operated a PC with his jaw and breath, and a smartphone with his thumb. He also had a tracheostomy and could not speak when equipped with a ventilator during rest or sleep.	6/2
2	30s	Female	Spinal Muscular Atrophy. Her body was inclined, and she did not use a respirator but had a respiratory illness; one of her lungs was partially collapsed. She could move her hands and neck freely.	5/2

3. SURVEY RESULTS

In the experimental survey, both the participants successfully operated the robotic arm using their brain signals. Subsequently, they were asked to explain the possible purposes of using BMI devices and the ethical issues associated with device use which they thought of during the experiments. The following section summarises the participants' responses.

3.1. Potential purposes and benefits of BMI device use

Survey results revealed that BMI devices were expected to help patients call caregivers during emergencies, control digital devices (e.g., personal computers and smartphones), digitalise identification cards, and support communication with others in daily life. These findings are similar to the results of the 2021 survey conducted with people with acquired disabilities (Orito et al, 2022a, 2022b). According to Participant 1,

> P1: 'When I come back to my home and get into bed, I cannot speak anything because of my ventilator. So, I can't even call a caregiver. Then, I think it would be very useful if there was a

switch that would allow me to call the caregivers using brain signals. It would be nice if there was something that could be done without using Wi-Fi.'

Moreover, he stated that the BMI system could support caregivers having some degree of physical burden and that it should be used to improve the caregivers' motivation and working conditions.

P1: 'I think it would be good that this system makes the caregivers' burdens easier when their assisting me in, say, taking a wheelchair. This does not mean that I consider a robot should take the place of a human caregiver. It would be somewhat wrong that a person being assisted thought "I don't need human caregivers anymore" (through using a BMI or other cyborg devices). I don't consider it is desirable that the use of machines will reduce the number of human caregivers or make them unnecessary. It would be not nice if the caregivers' jobs disappear gradually owing to cyborg devices. Rather, I think it would be good if the human caregivers' burden could be reduced. I believe many people with disabilities supported by human caregivers think in a similar way.'

On the other hand, Participant 2 revealed her desire to express ideas using such devices:

P2: 'I'm not good at talking in appropriate words what I imagine or intend. On the other hand, I am quite capable of imagining something. So, I wonder if cyborg devices could support me to make my imaginations, for example, written texts.'

3.2. Security risks and reliability

Participants were concerned about the operational risks and issues of implantable BMI. For example, they were worried about the risk of the occurrence of malfunctions or errors within the BMI device and the issues associated with maintaining adequate electronic power to operate the devices. The possible occurrence of these incidents made the participants anxious and highlighted the necessity of developing a multi-layered backup system, as follows:

P1: 'In fact, if the machines are installed in my body, and a disaster occurs and the electronic power stops, I wonder if the machines will really work, and I also think malfunctions could occur, because the machines are machines. I have a portable power supply for disasters; otherwise, if the respirator stops, I won't be able to breathe'.

P1: 'For example, even if an IC chip is implanted into my head and operated completely, it does not mean that the human caregivers are no longer needed at all. In case that the device goes out of order, I do need the help of them. But, once they are replaced by the device, I would not be able to expect the immediate support from them in case of emergency; it would take time to secure the manpower to take care of me. This means that I would not be able to do what I can do now thanks to the help from human caregivers. So, I think the implantation of such a brain chip would increase our anxieties. To ease the anxieties, the development of the environment to provide immediate backup in the event of a breakdown or an accident, in terms of both personnel and technical solutions, should be necessary'.

P2: 'I think any computer-based systems cannot be completely reliable; they would break down in the worst case. Even if we use such a BMI or cyborg machine, we should not totally depend on it, but have a plan B. For example, a communication board is a good backup device for those who have ALS, which is a slow communication tool, but enables them to ensure clear and precise communication with others. The most terrible case may be that they become totally accustomed

> to using the cyborg devices and cannot use any other communication tools like a communication board'.

Participant 1 was worried about using implantable BMI. It could cause serious damage to his postsurgical body, especially because his muscular abilities were already compromised.

> P1: 'The more severe the disability, the greater the anxiety about having surgery for an implantable device. If people with disabilities have a muscle disease like me, there is a lot of worry that, for example, they will lose spontaneous breathing after the operation, or that there will be more parts of the body that don't move due to anaesthesia. It is necessary to reduce the time and effort required for the implantation procedure as much as possible. If the implantation is as simple and convenient as receiving an injection, I would not feel anxiety about it so much'.

3.3. Sense of human dignity

None of the participants preferred to use BMI devices or cyborg machines to maintain their lives or be cared for totally; instead, they preferred to be supported by human caregivers. Participant 1 commented that when he used or was cared for by an emotionless machine, his human emotions disappeared; he mentioned that he too became an 'emotionless machine'.

> P1: 'For me, human emotion is important, and I don't like to be moved mechanically; so, I think it is important to be assisted by human caregivers. It is more enjoyable to be assisted by human caregivers who can chat with me. If a machine cared for me, it would just help me, say, take a wheelchair without saying anything; this would not evoke any emotions in me, and there wouldn't be no warm-heartedness. If that were to be the case, I probably wouldn't be able to live as cheerfully as I do now. I am afraid that people with disabilities would lose their emotions if they were cared for by machines alone, which have no emotions'.

According to Participant 2, although it was easy to control her body and communicate her intentions using BMI devices and brain signals, she was uncomfortable with her intentions being considered just a 'code' in the manner of an object. She considered it sad to be supported like a physical object.

> P2: 'I think it would be convenient for me if I could use such a device to communicate with others when I can no longer say anything or communicate my intentions. However, I also think that if my intentions are read from my brain signals and "coded", they might still be my own intentions or feelings, but I would feel as if I'm being regarded as an object. I don't like this.
>
> Indeed, this may be better than that I cannot communicate anything with others. I have no idea what happens when one's brain is dead and thus one cannot respond to anything at all. Anyway, I think it's sad that we are treated as physical objects.'

3.4. Personal data protection and social isolation

Participant 1 had no serious concerns regarding the privacy issues associated with BMI use and expected the personal information or brain signal data collected using BMI devices to be used in the future research and development of assistive cyborg devices for people with disabilities. However, he suggested that different cases should be assumed.

> P1: 'I have no sense such that "my brain signal is my personal information". I have no problem with researchers using my brain signal information; rather, I prefer it to be used more and more.

> Because my disease is a rare one, I think not so much information is available about it. So, I would be happy if my personal data related to the disease are utilised for academic research on it. I'll welcome presenting or publishing papers based on the outcomes of such research.
> However, some people with this disease do not like this kind of information being disclosed in public or even utilised for academic research. They are embarrassed that their data including ones on their disabilities and diseases go public. I think there are a significant number of people who do not like to be pitied by others as unfortunate guys who suffer from such a rare disease'.

On the other hand, Participant 2 believed that although brain signals are useful in controlling BMI devices, making this information available to the public or third parties and using the brain signals to analyse and predict people's intentions without agreement are not permissible.

> P2: 'Actually, I have thought that it would be easy and convenient that my intentions or feelings could be understood or identified by others using my brain signals even when I could not say my intention well using words.
> But, I do not prefer to disclose my information to people unnecessarily or irrelevantly. I don't want people I meet for the first time to know that I am feeling or considering something like this based on my brain signals'.

Moreover, Participant 2 noted the benefits of full-computer-mediated support for people with physical disabilities: since only electronic devices were used to assist people with disabilities, no risk was caused by human caregivers' unintentional leakage of privacy-related information, under the assumption that the personal data of users were appropriately protected within the system. Although Participant 2 herself was not worried about such risks and trusted her caregivers, she expected such a cyborg-supported scenario to substantially benefit the people with disabilities who prefer to live as independently as possible and are not willing to develop close relationships with human caregivers.

> Interviewer: 'If all the assistance is provided entirely by machines, is there no risk that the caregiver will not maintain the confidentiality of the information acquired when you are assisted? Will you prefer that?'

> P2: 'It may depend on how people with disabilities think about their private life. My private life is composed of me and my caregivers. But, if an assisted person with disabilities really prefers to be alone and has difficulty in establishing a relationship with a human caregiver, I consider a cyborg system or the like is extremely beneficial for such a person.
> The community of people with disabilities is relatively small, compared to one of able-bodied people for whom social communities play an important role as a basis for their life; if an assisted person or people with disabilities who need support are willing to just stay within a world where only cyborg or robotic supportive technologies are working, or less than that, they may think that they don't need to get involved with others. Conversely, if they wish to make society more inclusive and more collaborative or they would like to communicate with many other people, they may feel that their small world would become even smaller if they use and depend on this kind of technology'.

As suggested by Participant 2, automated/unmanned supportive cyborg technologies, like BMI, enable people with disabilities to move and communicate on their own, which is beneficial in cases where these people want their private information to remain unknown even to human caregivers. However,

it raises the question of whether such technology use causes social isolation, given the relatively limited social communication opportunities for people with disabilities.

3.5. Self-determination

Finally, the participants' overall responses to the semi-structured interview survey suggested that people with congenital disabilities tended to be left in environments where they were isolated from educational opportunities or general social communication, which made it difficult for them to make appropriate autonomous decisions regarding the use of information and communication technologies, including cyborg technology. They explained the backgrounds and social environments of people with disabilities, as follows:

> P1: 'As is often the case, people with disabilities are dependent on their parents, and the parents are very protective about them. They ask their parents about everything including cyborg device usage and data protection, regardless of whether they are minor or not. There are the cases such that, even if the people with disabilities themselves would like to use a BMI, they pretend not to like to use following the opinions of their parents who want to keep the secret that children have a disability. The communications between people with disabilities and attending physicians are often intermediated by their parents. So long as they live with their parents, they tend to feel they must do what their parents say.
> Actually, when I was living at home, it was the same for me. Everything was said by my parents and, to be honest, it was more comfortable for me that way. In fact, I didn't know about my disability until I was in primary school and, until then, my parents hid my disability from me, as well. But, when my parents told me "You should do your best" in the hospital, I asked them to explain to me about my physical conditions with a bit of anger. After that, they respect my opinions'.

> P2: 'When I was in the hospital, I felt a kind of resignation, as if I knew that I had to continue to stay there. Of course, some people with disabilities are willing to try out something actively. However, many people with disabilities seem to end up there while trying to live their lives, as if they just play games and end up there. I think they feel hopeless.
> In hospitals and care institutions, the feeling "out of here is unsafe" is commonly shared among people with disabilities. Owing to this, they tend to be inactive and become reluctant to go out on being constantly told that it's not safe.'

In addition, they described the difficulties of making any type of self-determination in such an environment.

> Interviewer: 'In case that parents hide their child's disability, when people with disability (child) are asked whether they would like to use cyborg or BMI devices, can they actively express his or her desire to use it actually?'

> P1: 'I think that the people with the disability by themselves will not indicate their intentions. Instead, I think, many people with disabilities show their preferences or decide to use them only if their parents say they should try it'

> Interviewer: 'Is it difficult for them to insist that they would like to use those devices, being at odds with their parents?'

> P1: 'I think it is difficult as long as they live with their parents. When people talk without their parents, sometimes, they can express their own opinions quite well. But, when parents are nearby, they tend to talk while looking at their parents' faces.'

Participant 2 mentioned parents' influence over decisions made by people with disabilities and clarified that such parental influence might be stronger for people with congenital than acquired disabilities.

> P2: 'One of the most prominent features of people with congenital disabilities is that they have experienced their parents or others around them have often decided many things for a long time saying "this is what is best or desirable for you". This is the case for me; my own decisions have been influenced by my parents. So, I think people with congenital disabilities are less able to make their own decisions in many cases as if they are becoming more and more dependent on others. They feel that someone else will decide for them, or that what someone else says is right and they don't need to have a different opinion, or that if someone says it's okay, then it's okay.
> In the case of people with congenital disabilities, in particular, they may have no criteria for making decisions by themselves. Such criteria are, in general, developed through experience; so, they may not know what a criterion they have. It's like 'if someone said this is the normal or average, but I don't have any way of verifying it'.

Regarding the use of embedded or implantable BMI devices, Participant 2 commented that in the absence of a sense of self-determination, there would be 'no sense of being invaded'.

> P2: 'I don't have an idea that an implantable BMI invades my body and mind, from the beginning. In case that I use the device being said "this is the way to make your parents easier", I would not have a feeling that I'm controlled by those around me. If we can see the world or society where a lot of things are going on, and if we can perceive the world or if we can feel something about it by ourselves, we can recognise the world or society as it is; but, because our decisions are always being made by others, we have no such sense of being controlled. So, we cannot have a sense of being invaded by embedded cyborg devices or the like'.

4. DISCUSSIONS

The results of the experiment and semi-structured interview survey suggest that BMI use potentially benefits people with disabilities in many ways such as operating personal computer devices and supporting communication with others. Participants were also concerned about the risks associated with BMI device malfunction and uncontrollability and insisted on maintaining multiple backup systems in the BMI device to avoid the case where the device stops functioning. These aspects are basically similar to the results of our previous surveys. Although practical BMI use generates such security and safety issues or concerns, these operational risks are expected to be technically addressed, to some extent, by technological development as soon as possible.

However, the two participants expressed common and strong concerns on continuing to be cared for by human caregivers, rather than cyborgs only, and their unwillingness to be treated as physical objects following the application of cyborg technologies, such as BMI. In other words, they strongly expressed resistance to being treated like physical objects and desired respect for human dignity. While there is some ambiguity as to what is meant by human dignity (Rosen, 2018), it is considered

that human dignity is violated when human beings are treated as objects (Matsui, 2003). This point marks the novelty of this survey's results.

On the other hand, the machine operations involved in caring for people with disabilities may encourage such people to engage in risk aversive behaviour to avoid human caregivers' personal data leakage and to become independent. Cyborg devices enable people with disabilities to do various personal activities and move around. This can reduce the sense of their privacy being exposed and help them become independent and socially active. Some people with disabilities are expected to use cyborg technologies for these reasons; however, this use may reduce their opportunities to communicate with others or human caregivers. Further, these issues may raise ethical questions on clarifying how the boundary between social isolation and an independent life is determined appropriately, what are the relevant policies that determine this line, and what do independent life, isolation, and 'human dignity' mean for people with disabilities. Of course, the technical development of the BMI system itself is not necessarily free from ethical problems, which should appropriately be addressed by relevant parties like developers both before and after the fact. For example, in the case of BMI's helping regain lost voices mentioned in the introduction, it is believed that such machine usage may raise a privacy issue: the brain signal data can be constantly monitored, and more complicated situation where a BMI user's voice which can otherwise be kept a secret is revealed.

Moreover, an important aspect that should be considered with respect to the aforementioned issues is self-determination. Particularly, if people with congenital disabilities who have only limited experience in making autonomous personal decisions over a long time and limited opportunities to develop their social skills that are important for decision-making in society, then they may find self-determination difficult. However, in the future, they may be required to make autonomous decisions and express their opinions on the extent to which they want to use computing or cyborg technologies to perform their personal activities and enhance their mobility. In other words, they may have to make decisions at a meta-level, for instance, to specify the parts of cyborg devices that should be operated by themselves and those that should be supported by human caregivers. Although this is an expected situation, if the individuals' ability to make such autonomous decisions is not sufficiently developed or mature, use of cyborg technologies such as BMI systems will be in accordance with governmental policies, other people's opinions, or financial factors. As pointed out by a participant, while cyborg technology potentially enables autonomous living, and the ability to live autonomously or independently is believed to help develop socialisation, there is a risk that this could result in social isolation. This implies the importance of considering their social background, practical skills, and ability to make autonomous decisions of people with disabilities when implementing cyborg technology.

This definitely includes a cost-effectiveness consideration, and if we want to be financially efficient in this matter, we may eventually be required to use computers or cyborg devices in many cases, similar to how unmanned technology was introduced and is currently progressing in our society. However, since not all processes need to be automated or cyborgised for people with disabilities, we must consider the meaning of autonomy and isolation and consider the findings of research on the social acceptance of unmanned technology. It is noted that this study was exploratory since it involved only a small number of participants with disabilities, and the examinations were not sufficiently comprehensive; thus, these aspects should be examined in more detail, and it is necessary to consider the possibility that many aspects of unmanned and mechanised operations that are considered rational and efficient for people without disabilities may not necessarily be so for people with disabilities, particularly those with congenital disabilities.

Furthermore, if the use of implantable BMI is realised and spread in the future, the so-called 'disappeared body' (Lyon, 2003) phenomenon would progress, and a digital-human fusion state that

uses brain signal data—a critical part of human consciousness—would emerge as data double. The data double in the form of brain signals may enable human beings to become one with cyborgs and help them become completely active even if they cannot move their bodies. However, the social influences of such situations have not been examined sufficiently. If it is possible to do everything using brain signals and other bioinformation, the human being could transform into trans-human beings. This would require us to consider whether the concept of people with disabilities would disappear or change into something entirely different.

5. CONCLUDING REMARKS

In this study, experimental and semi-structured interview surveys were conducted with two individuals suffering from congenital disabilities to examine the potential benefits, usefulness, and ethical concerns of BMI devices for patients with SMA. Since the study had only two participants, the responses cannot, by any means, be interpreted as representative of all the people with disabilities or the unique perceptions of people with SMA. On the other hand, it is a fact that finding such people who willingly participate in the experiment is itself difficult. The two study participants shared some important views on the use, as well as social risks or concerns, of BMI devices for people with disabilities who did not have sufficient opportunities to access latest technologies until now.

Based on their survey responses, we concluded that, although cutting-edge cyborg technologies, such as BMI devices, are useful for and appreciated by people with disabilities, it is important to discuss how BMI use by people with disabilities should be considered in terms of not only technological development but also the enhancement of the human dignity with respect to them and the individuals who care for them. Certainly, people with disabilities may not start using BMI devices immediately, since this technology's practical implementation involves considerations such as technological developments, cost-effectiveness, and social welfare policies. Undoubtedly, technological, financial, and political factors are paramount considerations in the application of the BMI system. However, the social use of cyborg technology and BMI systems should not be discussed without considering how society respects the dignity of people with and without disabilities who use cyborg technology.

Everyone probably finds it difficult to evaluate a technology that they have limited knowledge of. In this case, it is necessary to provide opportunities to discuss any relevant ethical and social issues and consider a wide range of multifaceted issues after establishing an environment in which such a technology can be accessed and used simply, rather than conducting research that is merely focused on technological development potential. To deepen the discussion of this study and enhance its generalisability, more experimental surveys should be conducted on a large and diverse population with disabilities. Simultaneously, quantitative research methods, such as questionnaire surveys, should be used to examine the BMI utilisation by individuals with disabilities from various perspectives.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 23K01545, 22K02063, 20K01920, and the Meiji University Grant-in-Aid for the international collaborative research project "Cyborg Ethics." We also appreciate Professor Shizuka Suzuki of Ehime University, Mr. Yukio Takeda, Dr. Yoshitaka Moritsugu and all the participants in the experiments and researchers for their support of our study.

We certify that all procedures of the experiments performed in this study were in accordance with the ethical standards of the research ethics committee established at the Faculty of Collaborative Regional Innovations, Ehime University (issued June 2021, No. 2021-01).

REFERENCES

Bernal, S. L., Celdrán, A. H., & Pérez, G. M. (2023). Eight reasons to prioritize Brain-Computer Interface cybersecurity. *Communications of the ACM*, 66(4), 68-78.

Grübler, G. & Hildt, E. eds. (2014). *Brain-Computer Interfaces in their ethical, social and cultural context*. Dordrecht: Springer.

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).

Lyon, D. (2003). *Surveillance as social sorting: Privacy, risk, and digital discrimination*. London: Routledge.

Matsui, F. (2003). What is human dignity What is human dignity?: A double function of differentiation and leveling, *Bioethics*, 13(1), 58-62 (in Japanese).

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.

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. Retrieved from https://www.researchgate.net/publication/327904976_Cyborg_Athletes_or_Technodoping_How_Far_Can_People_Become_Cyborgs_to_Play_Sports

Nikkei Shimbun (2023, September 23). Into the realm of "God" (1) Age of "super-humans" with expanded human body, Manipulation of avatars by thoughts in the brain. Retrieved from https://www.nikkei.com/article/DGKKZO74706440V20C23A9MM8000/

Orito, Y., Murata, K. & Suzuki, S. (2020b). Possibilities and ethical issues surrounding brain-machine interfaces in the realm of social welfare: Potential for use by people with disabilities based on results from psychokinesis experiments. *E-poster at the 68th Academic Conference of Japanese Society for Social Welfare* (in Japanese).

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2020a). The ethical aspects of a "psychokinesis machine": An experimental survey on the use of a brain-machine interface. In Arias-Oliva, M. et al. (Eds.), *Societal challenges in the Smart Society Ethicomp book series* (pp. 81-91). Logroño, Spain: Universidad de La Rioja.

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021a). How a brain-machine interface can be helpful for people with disabilities?: Views from social welfare professionals. In Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata, and Ana María Lara Palma (eds.), *Moving Technology Ethics at the Forefront of Society, Organisations and Governments* (pp. 103-115).

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021b). The ethical issues of the use of BMI in social welfare: An experimental and semi-structured interview study with professionals. *National Conference of Japanese Society for Management Information* (in Japanese).

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2022a). The social implications of brain machine interfaces for people with disabilities: Experimental and semistructured interview surveys. *Proceedings of the ETHICOMP 2022: Effectiveness of ICT ethics – How do we help solve ethical problems in the field of ICT?,* 487-501.

Orito, Y., Yamamoto, T., Sai, H., Suzuki, S., Murata, K., & Fukuta, Y. (2022b). Brain machine interface ethics: The ethical issues on the usage of a psychokinesis machine for people with disabilities. *Proceedings of Conference of Japan Society for Information Management* (in Japanese), 115-118.

Rosen, M. (2018). *Dignity: Its History and Meaning*. Cambridge, MA: Harvard University Press.

Wahlstrom, K. (2018). *Privacy and brain-computer interfaces* (Publication No. 10169573) [Doctoral dissertation, De Monfort University, UK].

SECURING HEALTHCARE DATABASES: A COMPREHENSIVE POLICY-BASED FRAMEWORK INTEGRATING RELATIONAL AND BLOCKCHAIN TECHNOLOGIES

Olga Siedlecka-Lamch, Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology (Poland)

olga.siedlecka@icis.pcz.pl; sabina.szymoniak@icis.pcz.pl

ABSTRACT

In the evolving healthcare landscape, increasing data accessibility poses security and privacy concerns. This article proposes a novel hybrid model for enhancing medical database security by integrating conventional relational databases with blockchain technology. The hybrid model aims to fortify data integrity, ensure privacy, and enforce access control.

The discourse within this article will explore diverse security facets inherent in a medical database founded on the hybrid model. We will introduce the implementation of user certificates and the associated permission assignments, streamlining the management of data access. The adoption of proposed solutions holds the potential to establish resilient medical databases and elevate the security of patient data. The explicit organization and immutability of data stored in blockchains can instill heightened trust among patients. Anchored in the hybrid model, the solutions dissected in this article stand as a foundational platform for future research endeavors and the realization of systems for medical data management. By proposing a nuanced approach to data categorization and storage, our model aims to bridge the gap between legacy infrastructure and the advanced data management solutions, ensuring a smooth and efficient transition without compromising on the vital aspects of security and accessibility.

KEYWORDS: Blockchain data model, Healthcare data, Hybrid relational-blockchain model, Privacy Security.

1. INTRODUCTION

In recent years, the burgeoning application of blockchain technology has emerged as a transformative force across diverse industries, promising unprecedented advancements in security, transparency, and efficiency. Originally conceptualized as the underlying technology for cryptocurrencies (Nakamoto 2008), blockchain has transcended its initial confines and found novel applications in realms beyond finance (Bodkhe 2020, Peck 2017, Raikvar et al. 2020, Rajasekaran et al. 2022, Zheng et al. 2022). This paradigm shift is underscored by its decentralized nature, cryptographic security, and the ability to create tamper-resistant, transparent, and immutable ledgers.

In sectors ranging from finance to healthcare (Mougayar 2016, Patel et al. 2022, Hou 2017, Cagigas 2021, Rivera et al. 2017, Kar et al. 2021, De Aguiar 2020, Roman-Belmonte et al. 2018, Yaqoob et al. 2022), logistics to supply chain management (Jabbar et al. 2021, Queiroz et al. 2019), the multifaceted utility of blockchain technology is reshaping traditional paradigms. Its promise lies not only in fortifying the security of digital transactions but also in providing a decentralized framework that fosters trust and collaboration.

In the midst of the process of converting healthcare into digital format, the extensive collection of medical data presents a complicated obstacle - the need to guarantee its security, accuracy, and confidentiality. With the increasing amount of sensitive health information, there is a corresponding rise in concerns over unauthorised access, data breaches, and the smooth interchange of patient records. Blockchain technology presents itself as a promising option to address these urgent concerns, providing innovative methods to strengthen the fundamentals of medical data management.

Blockchain technology has become pervasive across various domains ((Tan et al., 2022), (Shi et al., 2022), (Khanna et al., 2022)), with the healthcare industry emerging as a prominent beneficiary. In the medical field, the integration of blockchain addresses critical aspects such as information management, drug tracking, data security, and privacy. A comprehensive exploration of blockchain and IoT in healthcare systems was undertaken by Farouk et al. (Farouk et al., 2020), shedding light on the synergies between these technologies. Demonstrating the efficacy of blockchain in health data storage and exchange, Lin et al. showcased the benefits of this technology (Lin et al., 2016). Further emphasizing its viability, Yaqoob et al. illustrated the potential applications of blockchain in healthcare (Yaqoob et al., 2021).

Notably, Jabbar et al.'s study delved into the challenges and prospective trajectories in the intervention of pharmaceutical supply chains (Jabbar et al., 2021), underscoring the multifaceted applications of blockchain in enhancing various facets of the healthcare sector. Building upon these foundational works, our article introduces an innovative hybrid model that amalgamates traditional relational databases with blockchain technology. This novel approach aims to fortify the security of medical databases, addressing concerns related to data integrity, privacy protection, and access control. Through an in-depth exploration of the proposed hybrid model, we aim to contribute to the ongoing discourse on leveraging blockchain solutions for advancing security measures in healthcare data management.

As we navigate the intricate landscape of healthcare data challenges, the subsequent sections of this article delve into how blockchain's integration with medical databases, specifically through a hybrid model, can foster unparalleled advancements in security, privacy, and access control. By harnessing the potential of blockchain, the healthcare industry can embark on a transformative journey towards a more resilient, transparent, and secure future in medical data management.

2. PROPOSED MODEL

In the pursuit of technological innovation, practical and financial constraints often stand as formidable obstacles to the seamless integration of entirely new systems. Recognizing this reality, there arises a compelling need to navigate a middle ground—a harmonious coexistence of existing, often relational databases, with novel and advanced solutions. This imperative is the driving force behind our proposed model, which advocates for a pragmatic approach to data management.

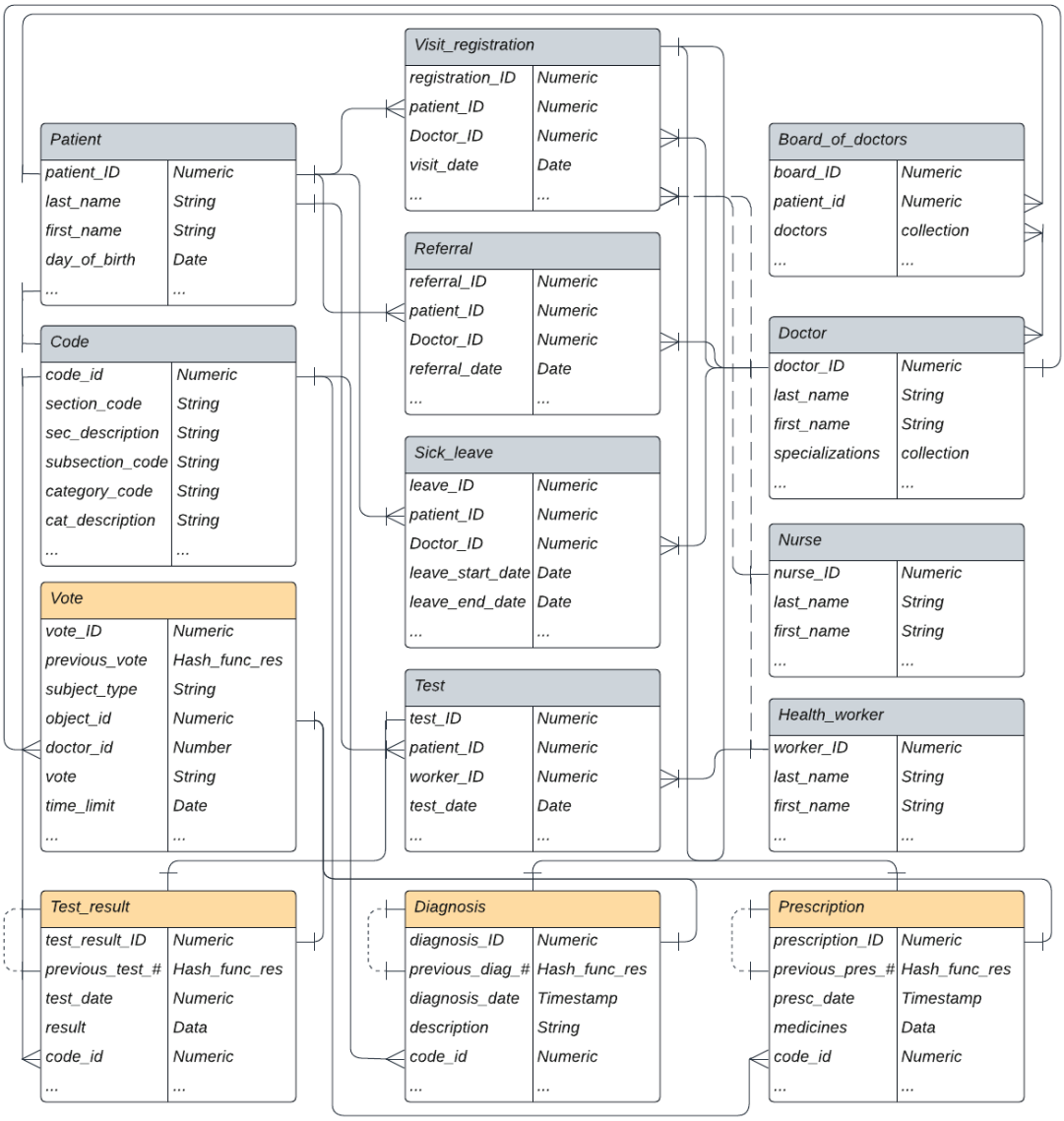
The inherent challenge lies in finding a balance between the established infrastructure of relational databases, which may house extensive historical data, and the imperative to embrace new, more secure, and technologically advanced solutions. Our model seeks to reconcile this dichotomy by proposing a strategic division of data into categories. Some data can continue to reside in conventional databases, leveraging the reliability and familiarity of existing systems. Simultaneously, particularly sensitive data, where security, accessibility, and even the chronology of entries are paramount, can find a home in cutting-edge solutions, notably the hybrid model of relational and blockchain databases.

2.1. Database Model

Our framework includes a wide range of healthcare data, such as patient information, visit records, medical leave records, test results, diagnoses, referrals, prescribed medications, medical staff details, disease codes, and their corresponding categories. This data has been thoroughly examined and integrated into our system. Significantly, data specifically associated with the treatment procedure, such as test outcomes, diagnoses, and recommended drugs, is securely stored within a blockchain framework, as indicated in orange in Figure 1.

This implementation ensures the creation of an immutable, transparent, and readily analyzable chronology of events originating from each medical device. Noteworthy is the accessibility granted to patients, who can peruse this information. However, modifications are exclusively within the purview of medical professionals possessing the requisite certificates meticulously assigned to their profiles. This intricate system establishes a secure and controlled environment, fostering data integrity and privacy while facilitating transparent information dissemination within the healthcare ecosystem.

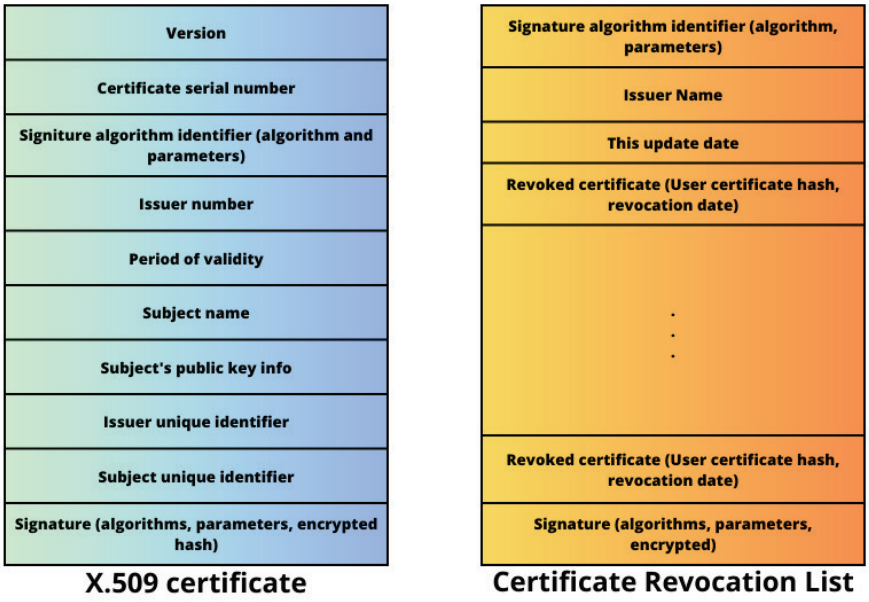
Figure 1. Simplified conceptual model.



2.2. Certifcates System

The X.509 standard, a widely adopted format for public key certificates in various applications (Cooper et al., 2008), seamlessly integrates cryptographic key pairs with identities representing individuals, organizations, or websites. This amalgamation empowers organizations to verify their identities and engage in secure, digitally signed communication. The Platform swiftly addresses compromised X.509 certificates by revoking them, subsequently disseminating a Certificate Revocation List across the network and issuing new certificates. Moreover, the Oracle Blockchain Platform employs X.509 certificates to ensure the security of the blockchain network and uphold data integrity.

Figure 2. X.509 certificate and Certificate Revocation List structures.



In Figure 2, the most recent rendition of X.509 certificate and Certificate Revocation List structures is illustrated. The certificate encapsulates crucial details about the certificate subject (including the subject's name, public key information, and unique identifier) and the certificate issuer (comprising the issuer's number and unique identification). Additionally, the certificate encompasses information about the employed algorithm, signature, version, serial number, and validity period. The certificate revocation list enumerates revoked certificates, their updated dates, signatures, and the issuer's names. Notably, only users equipped with valid certificates possess the authorization to input data into the key tables of the database. This stringent control mechanism ensures the integrity of the certificate system and reinforces the security measures within the database infrastructure.

2.3. Security Aspects

A medical data security policy is crucial in protecting the confidentiality, integrity and availability of patient-related information. It includes several measures to minimise the risk of violating patient privacy and securing medical information against unauthorized access, loss or modification.

Medical data security involves various essential components within the scope of a security strategy. The security policy is a comprehensive and official document created for a particular organization. It includes detailed assessments and precise directives related to managing risks and safeguarding assets within that unique environment. The process involves identifying and analyzing several hazards, including terrorism, cyberattacks, criminal activity, and natural disasters. This analysis allows for the identification of priorities and the implementation of suitable protective measures. The objectives in the security sector will be determined by these assessments, which include protecting life and property, guaranteeing personal stability, securing information, and considering specific dangers, tactics, and actions to achieve these goals.

Furthermore, the security policy should clearly establish the organizational hierarchy and accountability for overseeing security while outlining comprehensive protocols and directives for addressing any threats or incidents. These may encompass protocols for addressing cyberattacks, evacuating during disasters, and securely storing data. In addition, the security policy will include a framework for overseeing, assessing, and consistently improving security-related operations.

Moreover, the issue of medical data security is closely linked to ethics because of its susceptibility to unauthorized access. Security significantly impacts the choices and behaviours related to protecting data, resources, and the interests of both organizations and individuals. Especially regarding medical data, the security policy must incorporate principles of privacy, data confidentiality, and operational protocols. The document must comprehensively describe the methods used to store, process, and access the obtained information, along with the limitations and procedures associated with this data. Furthermore, the policy must ensure sufficient safeguarding of personal data and adherence to privacy legislation and standards.

From a technical perspective, we will address concerns regarding system uptime, system authentication, data access privileges, recovery procedures following a failure or event, and user entitlements and available services. When utilizing medical data, it is necessary to consider three distinct user roles: administrators, medical professionals, and patients.

The administrators oversee the entire system, monitor data access, configure and maintain the infrastructure, and restore the system following any faults. We can designate individuals as server administrators, schema administrators, and administrators responsible for managing blockchain and certificates. Medical personnel, such as physicians or nurses, can retrieve their patients' medical records, including their medical histories, test results, diagnoses, and prescription drugs. Typically, their permits are chosen based on their specialization, field of expertise, or other relevant characteristics. Medical professionals and diagnosticians will possess certifications to contribute data to blockchains. Patients utilize the system to acquire information about their health, medical records, diagnostic outcomes, medication prescriptions, and other healthcare-related data. Patients exclusively possess access to their data. They cannot modify their data.

In addition, utilising blockchain technology necessitates carefully considering measures to safeguard the confidentiality, integrity, and accessibility of information held within the blockchain network. The security policy for blockchain data will be customized according to unique requirements. However, the following guidelines can be highlighted for this matter. Implementing robust encryption techniques securely protects the data recorded on the blockchain. The security policy outlines the procedures for verifying the identities and granting permissions to blockchain users. The policy should establish explicit guidelines for accessing data stored on the blockchain, encompassing periodic upgrades of blockchain software and the implementation of security fixes to rectify identified flaws.

3. IMPLEMENTATION

In this chapter, we delve into the intricate details of our proposed system's implementation, shedding light on the intricacies of the relational database management system (RDBMS) and the meticulously designed objects within it. The focal point of our exploration lies in the meticulous integration of the blockchain component, the imposition of integrity constraints, the implementation of key procedures, and the nuanced allocation of permissions.

Our chosen RDBMS serves as the backbone of the system, providing a robust foundation for storing and managing diverse datasets critical to healthcare operations. Within this database, special attention is directed towards the implementation of key objects, notably the blockchain-enabled tables. These tables serve as secure repositories for critical data related to the treatment process—test results, diagnoses, and prescribed medications. This strategic integration ensures an immutable and transparent ledger, fostering trust and accountability in the medical data ecosystem.

Integral to the success of our implementation are the carefully imposed integrity constraints, safeguarding data accuracy and reliability. These constraints serve as guardians against inadvertent

data corruption or manipulation, upholding the sanctity of medical records and contributing to the overall security of the system.

As we navigate the intricacies of implementation, detailed procedures are unveiled, providing a comprehensive understanding of the system's inner workings. From data access management to the execution of secure transactions, each procedure is meticulously crafted to align with our overarching objectives of data integrity, privacy protection, and access control.

Furthermore, our exploration extends to the nuanced allocation of permissions, delineating the boundaries of data manipulation within the system. This facet is crucial in ensuring that only authorized entities, equipped with the necessary certificates, can modify, access, or contribute to the database, fostering a secure and controlled environment.

In essence, this chapter serves as a detailed roadmap through the implementation intricacies, providing a comprehensive view of the technological underpinnings that define our proposed model.

3.1. Implemented Model

We researched using a server with Ubuntu 22.04.2 LTS operating system, equipped with Intel(R) Xeon(R) CPU E5-2609 v4 @ 1.70GHz processor and 8 GB RAM. We used Oracle Database 21c database management system.

Thanks to DBMS_USER_CERTS package, we can manage certificates with row signing. The certificate infrastructure contains administrator certificate, which is basis for other certificates, and users certificates, that are stored in one common oracle directory on the server. The DBA_CERTIFICATES or USER_CERTIFICATES views show the certificates.

The administrator certificate should be generated using openssl command, according to following exemplary code (Code 1).

Code 1. Example of the openssl command for administrator certificate creation.

```
openssl req \
 -newkey rsa:2048 -nodes -sha512 \
 -x509 -days 3650 \
 -outform der \
 -keyout /home/oracle_adm/my_wallet/ cer-adm-key.der \
 -out /home/oracle_adm /my_wallet/ cer-adm-cert.der \
 -subj "/C=GB/ST=state/L=city/O=company/OU=Devs/CN=time_hall/emailAddress=admin@email.com"
```

To create a user certificate, we can use following procedure (Code 2).

Code 2. Example of the code for the procedure to create a new user certificate.

```
CREATE PROCEDURE create_certificate

DECLARE
 l_dir          VARCHAR(20) := 'CERT_DIR';
 l_file_name    VARCHAR (20) := 'user_1_signature.dat';
 l_cert         BLOB;
 l_bfile        BFILE;
 l_destoffset   INTEGER:= 1;
 l_srcoffset    INTEGER:= 1;
 l_cert_id      RAW(16);
BEGIN
```

```
 dbms_lob.createtemporary(l_cert, false);
 l_bfile := bfilename(l_dir, l_file_name);
IF(dbms_lob.fileexists( l_bfile ) = 1) THEN
  dbms_lob.fileopen( l_bfile );
  dbms_lob.loadblobfromfile(
    dest_lob    => l_cert,
    src_bfile   => l_bfile,
    amount      => dbms_lob.getlength(l_bfile),
    dest_offset => l_destoffset,
    src_offset  => l_srcoffset
  );
  dbms_lob.fileclose( l_bfile );

  dbms_user_certs.add_certificate(l_cert, l_cert_id);
  dbms_output.put_line('certificate ID: ' || l_cert_id);
 ELSE
  raise_application_error(-20001, 'Prioritize the creation of user certificates');
 ENDIF;
END;
```

After loading a certificate into the database, we can utilize it to digitally sign rows. We get certain concealed column values for the desired row and provide them as input to the DBMS_BLOCKCHAIN_TABLE.GET_BYTES_FOR_ROW_SIGNATURE operation in order to determine the specific data that has to be signed for the row. Subsequently, we proceed to save this data to a file. The file is signed using the private key associated with our certificate, resulting in the creation of the "user_1_signature.dat.sha512" file in previously created directory for certificates. Our signature is the file that is produced as a result. After that, we can now use created signature to sign the row.

The previously presented model has been instantiated in the form of tables, as illustrated in Figure 2. Tables delineated in grayscale denote traditional relational tables, while those highlighted in orange represent blockchain tables.

Each of the blockchain tables is characterized by unique parameters, including the specification of a time window during which the structure or data must not be deleted and the selection of a hashing algorithm. For testing purposes, specific parameters were configured:

Code 3. Example of the code for a blockchain table.

```
CREATE BLOCKCHAIN TABLE diagnosis (
 diagnosis_id NUMBER
  GENERATED BY DEFAULT ON NULL AS IDENTITY (START WITH 46463),
 diagnosis_date TIMESTAMP NOT NULL,
 doctor_id NUMBER NOT NULL,
 board_id NUMBER NOT NULL,
 code_id NUMBER NOT NULL,
 description VARCHAR2(500) NOT NULL,
 CONSTRAINT diagnosis_pk PRIMARY KEY(diagnosis_id),
```

```
 CONSTRAINT diag_doctor_fk FOREIGN KEY (doctor_id)
  REFERENCES doctors(doctor_id),
 CONSTRAINT diag_board_fk FOREIGN KEY (board_id)
  REFERENCES Board_of_doctors(board_id),
 CONSTRAINT diag_code_fk FOREIGN KEY (code_id)
  REFERENCES codes(code_id),
) NO DROP UNTIL 100 DAYS IDLE
 NO DELETE LOCKED
 HASHING USING "SHA2_512" VERSION "v1";
```

The final three lines of code establish conditions for structure removal: if no data has been inserted for 100 days, the structure can be deleted. Additionally, data deletion is disallowed, and the chosen hashing algorithm is specified as SHA2_512. The relational data within the table is conventionally visible to authorized users. However, a significant portion of the blockchain tables comprises metadata accessible solely to authorized administrators. These metadata include, among other details, the result of the hash function for data from the previous entry and the certificate data authorizing data insertion.

Exemplary data has been introduced into the database, including the blockchain tables, and queries have been tested. From the user's perspective, both of these operations resemble the handling of conventional databases. However, attempts to modify or delete data from blockchain tables are entirely restricted. The greater workload is borne by developers and administrators, as the system accumulates additional metadata associated with blockchains and certificates.

4. CONCLUSION

This work has introduced a new method to enhance the security and reliability of medical databases by combining traditional relational databases with blockchain technology in a hybrid paradigm. The incorporation of this hybrid model effectively tackles the growing apprehensions regarding the security and confidentiality of medical data, offering a holistic solution that harmonises the advantages of both systems.

The paper provides a clear explanation of the intricate design of the relational database management system (RDBMS), the integration of blockchain tables, the enforcement of integrity constraints, and the assignment of permissions. This connection guarantees a robust and protected basis for handling a wide range of healthcare datasets.

Our investigation has demonstrated that, from the user's point of view, interactions with the system, such as inserting data and answering queries, closely resemble typical database processes. However, the blockchain tables provide an additional level of immutability and transparency, hence improving data security and integrity.

Importantly, the system enforces limitations on altering and removing data in the blockchain tables, providing enhanced security for confidential medical data. The increase in burden for developers and administrators is recognised as the system gathers a more extensive collection of metadata associated with blockchains and certificates.

Despite the presence of constraints, such as complex implementation and a higher administrative burden, the suggested hybrid model has great potential as a solution for the healthcare sector. It not

only deals with the urgent issues with data security and privacy, but also establishes itself as a fundamental platform for future research and innovation in the administration of medical data.

Future research endeavours will prioritise the enhancement of the system, tackling any arising obstacles, and investigating possibilities for scalability. The incorporation of blockchain technology into current relational databases signifies a crucial advancement towards a more secure and robust future for managing healthcare data.

REFERENCES

Bodkhe, U., Tanwar, S., Parekh, K., Khanpara, P., Tyagi, S., Kumar, N., Alazab, M., (2020). Blockchain for industry 4.0: A comprehensive review. IEEE Access 8, 79764-79800.

Cagigas, D., Clifton, J., Diaz-Fuentes, D., Fern´andez-Guti´errez, M., (2021). Blockchain for public services: A systematic literature review. IEEE Access 9, 13904-13921.

Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., & Polk, W. (2008). RFC 5280: Internet X. 509 public key infrastructure certificate and certificate revocation list (CRL) profile.

De Aguiar, E.J., Faic¸al, B.S., Krishnamachari, B., Ueyama, J., (2020). A survey of blockchain-based strategies for healthcare. ACM Computing Surveys (CSUR) 53, 1-27.

Farouk, A., Alahmadi, A., Ghose, S., & Mashatan, A. (2020). Blockchain platform for industrial healthcare: Vision and future opportunities. Computer Communications, 154, 223-235.

Hou, H., (2017). The application of blockchain technology in e-government in china, in: 2017 26th International Conference on Computer Communication and Networks (ICCCN), IEEE. pp. 1-4.

Jabbar, S., Lloyd, H., Hammoudeh, M., Adebisi, B., Raza, U., (2021). Blockchain-enabled supply chain: analysis, challenges, and future directions. Multimedia systems 27, 787-806.

Kar, A.K., Navin, L., (2021). Diffusion of blockchain in insurance industry: An analysis through the review of academic and trade literature. Telematics and Informatics 58, 101532.

Khanna, A., Sah, A., Bolshev, V., Burgio, A., Panchenko, V., & Jasiński, M. (2022). Blockchain–Cloud Integration: A Survey. Sensors, 22(14), 5238.

Linn, L. A., & Koo, M. B. (2016, September). Blockchain for health data and its potential use in health it and health care related research. In ONC/NIST Use of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States: ONC/NIST (pp. 1-10).

Mougayar, W., (2016). The business blockchain: promise, practice, and application of the next Internet technology. John Wiley & Sons.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.

Patel, R., Migliavacca, M., Oriani, M., (2022). Blockchain in banking and finance: is the best yet to come? a bibliometric review. Research in International Business and Finance, 101718.

Peck, M.E., (2017). Blockchain world-do you need a blockchain? this chart will tell you if the technology can solve your problem. IEEE Spectrum 54, 38-60.

Queiroz, M., Telles, R., Bonilla, S., (2019). Blockchain and supply chain management integration: a systematic review of the literature. Supply Chain Management: An International Journal, 25.

Raikwar, M., Gligoroski, D., Velinov, G., (2020). Trends in development of databases and blockchain, in: 2020 Seventh International Conference on Software Defined Systems (SDS), IEEE. pp. 177-182.

Rajasekaran, A.S., Azees, M., Al-Turjman, F., (2022). A comprehensive survey on blockchain technology. Sustainable Energy Technologies and Assessments 52, 102039.

Rivera, R., Robledo, J.G., Larios, V.M., Avalos, J.M., (2017). How digital identity on blockchain can contribute in a smart city environment, in: 2017 International Smart Cities Conference (ISC2), pp. 1-4.

Roman-Belmonte, J.M., de la Corte-Rodriguez, H., Rodriguez-Merchan, E.C., (2018). How blockchain technology can change medicine. Postgraduate Medicine 130, 420-427.

Shi, Z., Zhou, H., de Laat, C., & Zhao, Z. (2022). A bayesian game-enhanced auction model for federated cloud services using blockchain. Future Generation Computer Systems, 136, 49-66.

Tan, W., Zhu, H., Tan, J., Zhao, Y., Xu, L. D., & Guo, K. (2022). A novel service level agreement model using blockchain and smart contract for cloud manufacturing in industry 4.0. Enterprise Information Systems, 16(12), 1939426.

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2021). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. Neural Computing and Applications, 1-16.

Zheng, X.R., Lu, Y., (2022). Blockchain technology–recent research and future trend. Enterprise Information Systems 16, 1939895.

PRIVACY-RELATED CONSUMER DECISION-MAKING: RISK ASSESSMENTS BY COGNITIVELY FRUGAL CONSUMERS

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 explores privacy-related cognition and assessments within the context of consumer choices. Existing researches have often treated privacy decision-making as independent of other decision processes, with limited exploration into the actual incidence of privacy decision-making. In this study, under the fundamental assumption that assessments of privacy risk compete with other perceived risks for consumers' cognitive resources, we tested hypotheses regarding the occurrence of privacy risk assessments and their relationship with privacy concerns through two studies utilising verbal protocols and recall set data. The findings revealed the following insights: privacy risk assessments may not frequently occur in the consumer decision-making process; within the context of consumer choices, there tends to be a greater emphasis on performance risk than on privacy risk; and the presence or absence of cues, which prompted the recall of privacy risks, leads to a transformation in the relationship between privacy concerns and the occurrence of privacy risk assessments. We also examined the significance of these results in the context of research on privacy decision-making, the privacy paradox, and the "notice-and-consent regime".

KEYWORDS: privacy risk, privacy concern, consumer perceived risk, protocol, recall set.

1. INTRODUCTION

The opportunities for individuals to disclose personal information are increasingly prevalent. One prominent instance of such information disclosure is in consumer decision-making. When utilising various services, service providers often require consumers to disclose their personal information as a condition for service provision. The evolution of online services and the increasing trend toward personalisation have significantly broadened the opportunities through which consumers divulge personal information (Karwatzki et al., 2017). Conversely, several consumer surveys suggest that personal information disclosure is a factor where risks are notably high in online transactions, highlighting it as a significant concern among many consumers (Lieberman & Stashevsky, 2009). The Privacy Calculus Model (PCM) explains the decision-making process related to consumers' disclosure of personal information. In this model, the intention to disclose personal information is the dependent variable. Factors such as the benefits of services made available through information disclosure and the trust of the data recipients are the promoting factors; concerns related to privacy and the perceived level of risks are inhibitory factors (Dinev & Hart, 2006; Kehr et al., 2015). Based on this framework, researchers recognise the current phenomenon of individuals with high-risk concerns readily engaging in information disclosure as a paradox.

Numerous studies have sought to explain the "Privacy Paradox" and modify existing models (Norberg et al., 2007; Barth & De Jong, 2017). Among these, two main directions have emerged. One is refining the PCM to enhance its explanatory power regarding the dependent variable of disclosure intention. The other is attempting to elucidate the unstable relationship between disclosure intention and disclosure behaviour. Many of these endeavours have focused on understanding the connection between privacy-related cognition and actual disclosure behaviour from the perspective of privacy decision-making. However, there has been a lack of attempts to approach the cognitive and evaluative

aspects of privacy risks from the perspective of the consumer choice process. This study, starting from the premise that privacy risk is one of the diverse consumer-perceived risks, seeks to reassess privacy concerns and risks by shedding light on the intrinsic nature of privacy decision-making embedded within consumer choices. Our endeavours offer new insights into the connection between privacy and consumer decision-making while providing implications for the conditions under which privacy decision-making occurs and the aspects of the notice-and-consent regime.

2. CONCEPTUAL REVIEW

Researchers have discussed the consumer's decision-making process in selecting products or services as an integrated response to various perceived risks (e.g. Bettman, 1973). Cox and Rich (1964), a pioneering study on consumer-perceived risks, have highlighted that the root of consumer risk perception lies in the uncertainty associated with achieving consumption goals. Consumers perceive the diverse uncertainties during the selection process and the resulting negative outcomes as risks. Thus, scholars have portrayed consumers as entities holistically responding to such risks. The Consumer Choice Model (CCM) does not have a comprehensive list of relevant perceived risks, but traditionally, scholars have addressed several kinds of risks. For instance, performance risk refers to the likelihood that the selected service may not fully satisfy the consumer's needs due to factors such as functionality, design, and usability (Jacoby & Kaplan, 1972; Featherman & Pavlou, 2003). Financial risk indicates the potential monetary loss associated with selecting a product or service, while physical risk represents the likelihood of bodily harm or health damage resulting from that choice (Jacoby & Kaplan, 1972; Mitchell, 1999). Time taken for the selection process has been considered a risk factor, as well. Time risk is associated with learning to use and maintain the product and the time spent waiting during usage (Featherman & Pavlou, 2003).

With the widespread adoption of online transactions, new risks arise in consumer perception. Harridge-March (2006) emphasises that the characteristics of online transactions, such as a lack of control over highly anonymous transaction partners and their opportunistic behaviours, result in a perception of privacy-related risk compared to conventional transactions. Featherman and Pavlou (2003) explicitly delineate privacy risk as a risk dimension perceived by consumers in the online purchase process. They define privacy risk as the potential loss of control over personal information, such as when entities utilise the consumer's personal information without the consumer's understanding or permission.

The conceptual definition of privacy risk is nearly identical in the PCM and CCM. However, based on the differing objectives of both models, we observe a significant distinction in the assumptions underlying privacy risk. In the PCM, where the disclosure intention of personal information is the dependent variable, researchers position privacy risk as a predictor to explain this dependent variable. Scholars evaluate privacy risk in a context independently from other risks which are unrelated to information disclosure. Furthermore, since the researcher formulates disclosure intention as a function of perceived benefits and risks, they do not consider any scenarios that completely exclude privacy risk assessment.

On the other hand, the CCM, with its dependent variable "service choice" positions privacy risk as one of perceived risks that consumers consider simultaneously. Given the combination of the assumption of simultaneous consideration and another premise that consumers are cognitively frugal (Fiske & Taylor, 1991), it appears that aspects and arguments related to privacy risks, which are overlooked in the PCM, come to the forefront. The subsequent sections will guide the development of three hypotheses to explore these aspects.

3. HYPOTHESES DEVELOPMENT

In light of the findings of studies on consideration sets or consideration studies, we infer that consumers do not necessarily conduct privacy risk assessments in all their choices consistently because they perceive privacy risk as one of the various risks within the decision-making process. Consideration studies focus on the research domain of forming choices in consumer decision-making. These studies underscore that, due to the limited cognitive capacity dedicated to their purchasing behaviours, consumers do not consider all available options; instead, they concentrate on a subset known as the consideration set (Andrews & Srinivasan, 1995). Numerous studies suggest that incorporating the consideration set into the consumer choice model enhances the explanatory power of it (Roberts & Lattin, 1997), establishing it as a key tool to elucidate the consumer choice process.

Conchar et al. (2004) present an exemplary study that explicitly applies the concept of the consideration set to consumer-perceived risks. They present a process model regarding consumers' risk perception, emphasising a crucial phase termed "risk framing". This initial phase is essential for determining which risks warrant attention and provides the groundwork for a more comprehensive scenario of perceived risk assessment. As a result, a risk consideration set (RCS) forms, comprising the risk factors chosen as the actual subjects for evaluation. The extent of importance associated with avoiding negative outcomes determines the selection of risks, assigning higher importance to risks prioritized for consideration (Mitchell, 1999; Conchar et al., 2004). If one considers the avoidance of privacy risk a critical factor, one would infer that this risk should be incorporated into the RCS, prompting a comprehensive assessment for each service for individuals. However, if individuals fail to recognise the importance of avoiding privacy risks, they may exclude these risks from the set of considerations for risk assessment. Therefore, it seems reasonable to posit the following hypothesis:

H1. The privacy risk assessment may not necessarily occur in consumer decision-making.

The fundamental assumption of viewing consumers as cognitive misers closely ties to this hypothesis (Fiske & Taylor, 1991). Consumers tend to allocate their limited cognitive resources efficiently, expending only the effort necessary to make satisfactory choices. Researchers suggest that in certain situations, consumers may willingly assume some risks to save cognitive effort. Most consumer behaviour studies consistently endorse this perspective (Garbarino & Edell, 1997). Incorporating the premise that consumers are cognitive misers into the preceding discussion on the consideration set of risks, it becomes apparent that privacy risk perceptually coexists with other risks and gives rise to the inference that it competes with them for the cognitive effort exerted by consumers. Given this perspective, the perceived importance of avoiding privacy risk and the perceived significance assigned to avoiding other risks shape the consideration of privacy risk in consumer service choices. The presence of risk elements which a consumer strongly desires to avoid during decision-making, other than privacy risk, may diminish the likelihood of his/her cognitive effort devoted to privacy risk.

In the context of consumer choice, one could identify performance risk as likely to influence one's perceived relative importance of privacy risk. The purpose-driven nature of the consumer's choice, aimed at satisfying needs, leads to an increased concern about performance risk, indicating the possibility that the chosen service may not fully meet needs due to factors such as functionality, design, and usability. Empirical evidence consistently demonstrates the heightened relative importance of performance risk across diverse product categories (Jacoby & Kaplan, 1972) and various stages of the purchase process (Cunningham et al., 2005).

One can glean insights into the relative importance of privacy risk and performance risk from concepts such as hyperbolic discounting (Waldman, 2020) and likelihood of positive reinforcement rewards (Meyer & Kunreuther, 2017), which represent human cognitive biases and thought patterns associated

with risks. The former concept indicates a tendency for individuals to overestimate the value of outcomes in the immediate moments of decision-making while undervaluing distant and ambiguous future outcomes. Building upon the framework, one could argue that consumers tend to overestimate the significance of performance risk, given its ease of forming concrete mental images and immediate perceptibility related to usage. In contrast, one might underestimate the content of privacy risk, which encompasses abstract and unpredictable disadvantages. With regard to the latter concept, it is believed that the perception of risk diminishes as the likelihood of positive reinforcement occurring decreases. Successfully addressing privacy risks involves maintaining the current state of "nothing happening", and explicit reward-based motivation for such outcomes is relatively rare. On the other hand, responding to performance risk is more likely to result in tangible rewards, such as satisfaction and joy, if things go well, making reinforcement learning more probable. This approach may render performance risk more prominently considered within the consumer choice process. Taking these various pieces of evidence and factors into account, we propose the following hypothesis:

H2. In the actual consumer choice process, consumers give a higher weight to evaluating performance risk than assessing privacy risk

In line with the studies on consideration sets discussed in Hypothesis 1, we posit that consumers with heightened privacy concerns are more strongly motivated to avoid privacy risks, resulting in an expected increase in the likelihood of privacy risk assessment under such circumstances. However, as indicated by Hypothesis 2, we anticipate the perceived importance of avoiding other risks to influence this fundamental relationship. In other words, in situations characterised by intense competition for consumers' cognitive efforts, there may be instances where consumers do not conduct privacy risk assessments despite elevated levels of concern. In contrast, in situations with low competition and minimal interference from the importance of other risk avoidance, we expect the previously discussed positive fundamental relationship to manifest. Considering these aspects, we posit the following hypothesis:

H3. The relationship between privacy concerns and the occurrence of privacy risk assessments depends on the degree of competition among risk factors: a positive relationship will occur when the competition is low, but in situations with high competition, such a relationship becomes less likely to occur.

4. STUDY 1: EXPLORATION USING VERBAL PROTOCOLS

4.1. Method

In the first empirical test, we conducted a qualitative examination using verbal protocols to gain insights into the assessment of privacy risks within consumer choices. This methodology uncovers the sequence of psychological thoughts that unfold during the execution of a specific task through verbalization (Ericsson & Simon, 1993). We instructed participants to articulate their actions and thoughts while navigating the task so that the analysis of the cognitive processes involved in task execution based on the acquired voice data could be conducted. If a participant utters something relevant to privacy risk in their protocol, we can infer that the participant is attentive to this risk, suggesting its inclusion in their RCS. Being an outcome of instructing participants to engage in "think-aloud" (Ericsson & Simon, 1993), the verbal protocol is suitable for acquiring data that represents spontaneously elicited privacy-related awareness rather than one forcibly awakened.

We collected verbal protocols in May 2022. The survey respondents comprised 21 Japanese students from Meiji University (13 females, eight males) aged 19 to 21. Before the data collection, participants

engaged in a practice session where they vocalised their thoughts into a voice recorder while solving simple area calculation problems and playing Sudoku, a logic puzzle on a 9×9 grid. After a briefing session that explained the purpose and details of the survey and confirmed their will to take part in the survey, we instructed participants to verbalize their actions and thoughts while navigating the task of selecting a diary app from the available downloadable options. Following the voice data submission, participants responded to a questionnaire survey. In this survey, we assessed privacy concerns using a direct questioning method. To establish a work environment to encourage participants to perform tasks without apprehension regarding monitoring, and to maintain anonymity, we systematically conducted all activities through Zoom with audio and video functions suspended; communication and data exchange used pseudonyms and email addresses exclusively created for this survey.

The coding of the collected protocols followed the procedures outlined by Bettman & Park (1980). Two coders conducted the coding task: one of the authors (Fukuta) was the lead coder, and a consumer behaviour research expert, who was not the co-author of this paper, was a sub-coder. We extracted events related to information search from the protocols and, based on predefined coding rules, categorised these events into information searching for privacy, performance, and other perceived risks. Information-seeking related to privacy policy, online security, and trust in service providers was categorised as a privacy risk. We classified the search for information related to the diary app's functionality, design, and ease of use as performance risks, while other search events involving information exploration were "other risks". Each coder independently categorised information-seeking events, and we assessed the inter-rater reliability of the coding results.

4.2. Results

In the data cleansing stage, we excluded four protocols from the analysis due to difficulty in discriminating voice and interruptions in the audio stream. As a result, we coded the verbal data of 17 participants. We used intra-class correlation coefficients to evaluate the inter-rater reliability of the coding results, assessing the number of occurrences for each of the three risk categories (privacy risk, performance risk, and other risks) within each respondent's protocol. The results of the evaluation are shown in Table 1. Generally accepted guidelines for interpreting intra-class correlation coefficients are as follows: 0.00–0.20 indicates none to slight agreement, 0.21–0.40 is fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–1.00 almost perfect agreement (Landis & Koch, 1977). According to these criteria, the coding of three categories by two coders exhibits a considerably high reliability.

Table 1. Intraclass Correlation Coefficient as an Inter-Rater Reliability Metric.

	Privacy risk (Pr)	Performance risk (Pe)	Other risks (Ot)
ICC (2,1)	0.818 (95%CI [0.559, 0.931])	0.902 (95%CI [0.756, 0.963])	0.715 (95%CI [0.372, 0.887])

Table 2. Overview of 17 Verbal Protocols.

pseudonym	Information search flow	Time	pseudonym	Information search flow	Time
516uihas	○→Pe→□→Pe→☆→Pe→✓	3:50	Miutan23	○→Pe→？→Pe→□→✓	4:15
10100907	○→Ot→Pe→□→✓	2:45	Msaler11	○→Pe→Ot→✓	2:50
Bunbabab	○→Pe→？→✓	1:05	Mutsuki	o→Pe→Ot→Pe→Ot→Pe→Pe→Ot→？→Pe→Pe→✓	5:15
Celaptnn	○→Pe→□→Pe→Pe→□→Ot→Pe→✓	4:20			
Chongkan	○→Ot→Pe→？→✓	1:10	Saku89ut	○→Pe→**Pr**→□→Ot→**Pr**→Pe→✓	3:00
Cotton73	○→☆→Ot→☆→Pe→Pe→☆→✓	3:10	Syousin3	○→Pe→Ot→Pe→Ot→✓	3:05
Dorachan	○→Pe→Pe→✓	2:50	Uniikwow	○→Pe→□→Pe→Pe→□→Ot→Ot→✓	5:00
Dtco4869	○→☆→□→Ot→Pe→Pe→✓	6:10	Wanpiz37	○→？→Pe→✓	1:10
Jist0909	○→Pe→Ot→✓	1:50	Y6u11a2r	○→Pe→Pe→**Pr**→Ot→Pe→Pe→Ot→**Pr**→□→✓	14:50

○: Start of Protocol ✓: Finish of Protocol □: Consideration on search result list

☆: Consideration on web pages (outside the official app store) ?: Unintelligible/Unclear

Pr: Info-seeking on privacy risk Pe: Info-seeking on performance risk Ot: Info-seeking on other risks

Table 2 shows the overview of the verbal protocols collected from 17 participants. This table includes the pseudonym, the flow of information-seeking events, and the time taken for decision-making for each of the 17 verbal protocols. In the protocols, we identified 59 information-seeking events which were for privacy risk ('Pr' in Table 2), performance risk ('Pe') and other risks ('Ot'). Information-seeking related to privacy risks is notably low, constituting 6.8% of the total number of information-seeking instances (4 of 59) and 11.8% of respondents (2 of 17). In contrast, information-seeking related to performance risks, where all 17 participants engaged in at least once, accounts for a substantial 64.4% of the total number of information-seeking instances (38 of 59). This finding reveals a significant disparity in prevalence between information-seeking related to privacy and performance risks. The results are consistent with Hypotheses 1 and 2.

We categorised those respondents who answered the question, "When using online services, I am concerned about how my information is being collected and used", with "7. Strongly Agree" or "6. Agree", into the high-concern group, resulting that nine participants were classified into this group. We also classified the two individuals who mentioned privacy risk (Saku89ut and Y6u11a2r) into the high-concern group. Although seven respondents expressing notable concerns about privacy, they did not actively pursue information searches related to privacy risks, and privacy risk assessment did not occur. This trend mirrors the latter part of Hypothesis 3.

It is noteworthy that the support for these hypotheses is highly limited in meaning, as the current protocol analysis faced challenges due to a small dataset, making parameter estimation difficult. This limitation emphasises the need for cautious interpretation within the constrained scope of sample statistics. Verbal protocols provide valuable insights into the state of privacy risk assessment during actual decision-making without forcibly triggering respondents to consider privacy risks explicitly. However, collecting reliable verbal data necessitates that respondents possess high cognitive processing capabilities and undergo significant training, as the highly sophisticated task of articulating one's thoughts aloud while making decisions is necessary during data collection (Eckersley, 1988). These characteristics give rise to issues in the sample, such as small sample size and demographic biases, and exert negative influences on the reliability of the data and the generalizability of the findings. Considering these aspects, Study 2 undertakes a parallel examination using an alternative methodology for a more extensive pool of research participants.

5. STUDY 2: EXAMINATION USING UNAIDED RECALL AND AIDED RECALL

5.1. Method

In Study 2, we employed the concepts of recall and recall set to attempt to understand the actual occurrence of privacy risk assessments, advancing discussions related to our hypotheses. In general, recall can be used to understand the level of awareness of information within memory. It denotes the state or ability to access particular memory information in a given situation (Cianfrone et al., 2006; Leigh et al., 2006). Scholars broadly categorise recall into two types: unaided and aided (Leigh et al., 2006). Unaided recall refers to the ability to retrieve information from memory without cues in a specific situation or task. The set of information recalled without any aid is the unaided recall set. In contrast, aided recall involves retrieving information from memory using cues.

Higgins et al. (1996) posit that the difference in recall between these two types is rooted in the accessibility of the recalled information. If the unaided set encompasses privacy risks, researchers presume consumers attribute high accessibility to this risk, making it more salient during consumer decision-making. Conversely, including privacy risks in the aided set refers to a state where one conducts risk assessments in situations with provided cues. In cases where the aided set does not

include privacy risks, we can interpret it as a state with a very low likelihood of consideration about privacy risks during the decision-making process. Moreover, in cases of aided recall, the support offered for cognitive efforts in recall creates a relatively mild competition among perceived risks. On the other hand, for unaided recall, a situation characterised by intense competition among risks arises due to the limited retrieval potential. Considering these aspects, we tested the hypotheses using empirical data from the survey.

We utilised Macromill's internet survey services to survey in June 2023. The total number of participants was 420, with 42 individuals assigned to each of the ten demographic groups: five age categories, ranging from individuals in their 20s to 60s+, and the two gender categories of females and males. Participants were randomly divided into two groups, Groups A and B. We implemented this division to mitigate any potential learning effects between the measurements of the unaided and aided sets. We excluded 62 cases identified as inappropriate from the collected data due to incorrect responses to trap questions and straight-lining and analysed the remaining 358 cases. The key characteristics of the sample are shown in Table 3. We used a between-group comparison to assess the homogeneity of the two groups. As indicated by the results of the chi-square tests within the table, we observed no significant differences between the two groups in all four coverage areas (gender, age range, residence, operating system (OS) type of a participant's smartphone) and four basic statistical measures (duration of smartphone use, app download count, perceived app selection difficulty, private concerns assessment). Consequently, both groups were homogeneous, allowing for further analysis.

Table 3. Sample Characteristics and Test Results for Homogeneity of Two Groups.

Gender and Age Range (Gender: chi-square (1) =0.306, p=0.597; Age Range: chi-square (4) =2.456, p=0.652)										
	Female					Male				
	20-29	30-39	40-49	50-59	60-	20-29	30-39	40-49	50-59	60-
Group A	16	21	20	20	20	10	16	16	16	21
Group B	20	17	20	20	18	18	16	15	17	21

Residence (chi-square (7) =2.213, p=0.947)								
	Hokkaido	Tohoku	Kanto	Chubu	Kinki	Chugoku	Shikoku	Kyushu
Group A	8	9	67	33	26	17	3	13
Group B	9	12	75	26	28	14	4	14

OS_Type (chi-square (3) =1.222, p=0.748)				
	iOS	Android OS	Others	Unspecified
Group A	77	96	1	2
Group B	81	98	0	3

Duration of Smartphone Use						Private Concerns Assessment				
	n	Mean (SD)	t-value	df	p-value	n	Mean (SD)	t-value	df	p-value
Group A	176	9.943(5.069)	-0.024	356	0.981	176	5.256(1.115)	0.436	356	0.663
Group B	182	9.956(4.93)				182	5.203(1.155)			
App Download Count						Perceived App Selection Difficulty				
Group A	176	58.483(120.633)	-1.066	356	0.287	176	2.733(1.345)	-0.102	356	0.919
Group B	182	76.571(191.389)				182	2.747(1.305)			

We created both types of recall sets based on Cianfrone et al. (2006). The unaided recall set in this study involved responses to the open-ended question "What points do you want to check when selecting an app? Please write freely". To form the aided recall set, we used responses to the question "Among the listed items, which ones would you like to check when selecting an app? Please mark all that apply". Following the terminology "risk consideration set" introduced by Conchar et al. (2004), we called these two types of recall sets the unaided risk consideration set (RCS) and the aided RCS. In Group A, we conducted unaided RCS measurements of the diary app, followed by aided RCS measurements of the health management app. In Group B, we initially performed unaided RCS measurements of the health management app, succeeded by aided RCS measurements of the diary app.

The list of consideration items as a cue used for measuring the aided RCS includes 11 items, including app functionality, privacy policy and financial cost. We formulated this list based on the existing body of research on consumer perception risk and the results of preliminary investigations. To identify the privacy and performance risks mentioned in the hypotheses, similarly to Study 1, we categorised these risk items into three distinct groups: privacy risk, performance risk, and other risks. We deemed the aided RCS for each category "mentioned" if the respondent selected at least one item within that category. To ensure comparability between unaided RCS and aided RCS, we categorised the free-text responses used for unaided RCS measurement into the same three groups as aided RCS. We deemed unaided RCS for each category "mentioned" if the respondent referenced any content related to that category in their free answer. Two coders, both of whom are the authors of this study (Fukuta and Orito), carried out the coding process and subsequently assessed inter-rater reliability.

5.2. Results

The frequencies of mentioned risks for each risk factor category, segmented into unaided RCS and aided RCS for two respective apps, are described in Table 4. The right side of Table 4 illustrates the inter-rater reliability of unaided RCS coding. We used Cohen's Kappa coefficient, recommended by Grant et al. (2017), as an inter-rater reliability measure for a binary task conducted by two evaluators. Following the conventional criteria for interpreting Kappa coefficients, where values of 0.8 or higher indicate "almost perfect agreement" (Landis & Koch, 1977), the inter-rater reliability of the coding in this study is exceptionally high. A chi-square test for independence (Chi-square (1) = 0.423, p = 0.534; Phi = −0.019, p = 0.515) confirmed the absence of significant differences in the distribution of the two RCS across apps (Table 5).

Table 4. Frequency Distribution of Each RCS and Interrater Reliability.

		Diary app		Health management app		Cohen's Kappa coefficient	
		Not mentioned	Mentioned	Not mentioned	Mentioned	Diary app	HM app
Unaided RCS	Privacy risk	159	17 (9.7%)	175	7 (3.8%)	0.856	0.815
	Performance risk	31	145	41	141	0.851	0.921
	Other risks	139	37	133	49	0.817	0.916
Aided RCS	Privacy risk	93	89 (48.9%)	106	70 (39.8%)		
	Performance risk	52	130	53	123		
	Other risks	10	172	12	164		

Table 5. Cross-Tabulation for Testing Homogeneity of Retrieval States Across Apps.

	Unaided RCS	Aided RCS
Diary app	199	391
Health management app	197	357

To test Hypothesis 1, we examined the occurrence of privacy risks in RCS. In the unaided RCS, the occurrence rates were notably low, with 9.7% for diary apps and 3.8% for health management apps. While variations may arise based on factors such as the content of the apps, we overall anticipated that a relatively small number of consumers place a high priority on avoiding privacy risks in their choices. We expected many cases to involve an absence of evaluations regarding privacy risks. The likelihood of considering privacy risks in aided RCS is 48.9% for diary apps and 39.8% for health management apps. Approximately half of the respondents considered privacy risks in situations with available cues. However, this percentage also suggests that more than half of the respondents did not consider privacy risks, even with aided RCS. Although within the range of sample statistics, the results

offer limited support for Hypothesis 1, suggesting that there are many instances where consumer decision-making regarding privacy does not integrate with consumer choices.

As mentioned in Hypothesis 2, we examined the relationship between privacy and performance risk through chi-square tests (Table 4). The number of individuals mentioning performance risks in the unaided RCS was significantly higher than those mentioning privacy risks in the diary app (Chi-square (1) = 187.367, $p < 0.001$; Phi = −0.730, $p < 0.001$) and health management app (Chi-square (1) = 204.454, $p<0.001$; Phi = −0.749, $p < 0.001$). In addition, in the aided RCS, the cases where participants selected performance risks statistically significantly surpassed those choosing privacy risks, as evidenced in the diary app (Chi-square (1) = 19.269, $p < 0.001$; Phi = −0.230, $p < 0.001$) and health management app (Chi-square (1) = 32.221, $p < 0.001$; Phi = −0.303, $p < 0.001$). We applied the Bonferroni correction to the obtained p-values to address the issue of multiple testing arising from the repetition of the chi-square tests. The adjusted significance level for all tests became 0.0025. Despite correcting for type I error inflation, all four test results demonstrated significance probabilities below 0.001, maintaining the statistical significance of the findings. These findings confirm that in the context of consumer choices, consumers expend more cognitive effort on performance risks than on privacy risks. We observed this tendency widely, regardless of recall styles or types of apps, supporting Hypothesis 2.

We used logistic regression analysis to examine Hypothesis 3, focusing on the correlation between privacy concerns and the occurrence of privacy assessments. We measured the independent variable, privacy concerns, on a seven-point Likert scale in response to the question “I am concerned about whether my privacy is protected” (Kehr et al., 2015). We determined the binary dependent variable based on whether mentions were present or absent in each app’s unaided RCS and aided RCS. When analysing the impact of privacy concerns on unaided recall of privacy risk, the odds ratio was 1.519 ($p = 0.083$, 95% CI [0.946, 2.438]) for diary app choice and 1.346 ($p = 0.392$, 95% CI [0.682, 2.655]) for health management app choice. Both sets of results suggest the absence of a statistically significant correlation between privacy concerns and the likelihood of unaided recall of privacy risks. However, we observed a significant positive association with privacy concerns when analysing the aided recall of privacy risk as the dependent variable.

In the case of choosing a diary app, the odds ratio was 1.572 ($p < 0.001$, 95% CI [1.197, 2.064]), indicating a significant association between privacy concern and the likelihood of aided recall of privacy risks, with each unit increase in privacy concern corresponding to a 57.2% increase in the odds of aided recall of privacy risks. When opting for a health management app, the odds ratio stood at 1.594 ($p = 0.002$, 95% CI [1.190, 2.134]), signifying a noteworthy connection between privacy concerns and the likelihood of aided recall of privacy risks. Additionally, each incremental unit of privacy concern corresponds to a 59.4% elevation in the odds of aided recall of privacy risks. These results align with the content of Hypothesis 3, indicating that the relationship between the level of privacy concerns and the occurrence of risk assessments is not constant but rather situation-dependent. In situations with mild competition for cognitive effort in aided RCS, we observed a positive relationship between privacy concerns and risk assessments, which inherently exists. However, in situations with intense competition in unaided RCS, this relationship diminishes, and we identified cases where, despite high levels of privacy concerns, there are no risk assessments. Thus, Hypothesis 3 is supported.

6. DISCUSSION AND CONCLUDING REMARKS

This study has approached privacy risk not as a variable to explain individuals’ intention to disclose personal information but rather framed it as one of the perceived risks in consumer choice. In a more specific sense, this study derived three hypotheses incorporating the premise that privacy risk is

perceived concurrently with other consumer risks into concepts such as the formation of consideration sets and competition over consumers' cognitive efforts. We examined these hypotheses through empirical research employing methods other than direct questioning, such as verbal protocols and recall set data. The findings are as follows: in the actual consumer choice process, there are cases where the evaluation of privacy risk itself does not occur; there is a tendency for the assessment of performance risk to take precedence over the evaluation of privacy risk; in situations with intense competition for cognitive effort among other risks, individuals with high concerns may not engage in the evaluation of privacy risk.

These findings provide insights for refining the theoretical understanding of the connection between privacy decision-making and consumer decision-making, a dimension that studies haven't extensively nuanced before. To further advance our understanding of these insights, it is imperative to conduct a comprehensive analysis that delves into the content of privacy risk assessment, specifically the perceived level of risk, and a meticulous examination of the underlying processes involved in privacy risk assessments. This detailed exploration is essential in comprehending the conditions that give rise to the Privacy Calculus Model in consumer decision-making, establishing a significant nexus with the existing knowledge framework.

In addition, these findings suggest implications for the validity and efficacy of privacy consent during consumer decision-making, specifically privacy policy consent. Even individuals with high privacy concerns may forgo the assessment of privacy risks when confronted with exceptionally high levels of other consumer risks. This situation challenges the assumed simplistic decision-making flow that the "notice-and-consent regime" (Waldman, 2020, p.105) anticipates, suggesting that individuals would decline consent if concerned. Furthermore, the revealed results may give an alternative perspective on the privacy paradox. Acknowledging this paradox relies on a causal chain in which privacy concerns mediate the impact on information disclosure behaviour through privacy risk assessment and disclosure intent pathways. Despite heightened concerns, if judgments regarding risk assessment and disclosure do not occur, the causal chain is disrupted, and the paradox fails to manifest. Put differently, a heightened awareness of the context of consumer choice necessitates scrutiny of the prerequisites for the privacy paradox. In this context, research on the occurrence of privacy risk assessment suggests a fresh perspective for studies on the privacy paradox, indicating the potential for a more sophisticated approach.

Lastly, although not directly related to the content of hypothesis testing, using verbal protocols and recall set data in measuring perceived privacy risks is crucial in privacy risk research. Researchers have so far tended to predominantly employ the method of directly questioning individuals about their privacy risk perceptions. However, this approach entails coercively inducing the occurrence of risk assessments. Instead of evaluating the occurrence of risk assessments, employing this measurement method, which assesses the perceived level of risk under the assumption that risk assessments have occurred, is superior in yielding more detailed data. However, applying such methods proves challenging when attempting to ascertain the occurrence of risk assessments. It is crucial to include risk assessments as a variable in the analysis to foster discussions on privacy risks while staying connected to the context of consumer choice. As such, it is imperative to prioritise alternative approaches to assess risk perception without depending on direct questioning. This study contributes to these efforts.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 20K02000 and 23K01545.

REFERENCES

Andrews, R. L., & Srinivasan, T. C. (1995). Studying consideration effects in empirical choice models using scanner panel data. *Journal of Marketing Research*, 32(1), 30-41.

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.

Bettman, J. R. (1973). Perceived risk and its components: a model and empirical test. *Journal of Marketing Research*, 10(2), 184-190.

Bettman, J. R., & Park, C. W. (1980). Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: A protocol analysis. *Journal of consumer research*, 7(3), 234-248.

Cianfrone, B., Bennett, G., Siders, R., & Tsuji, Y. (2006). Virtual advertising and brand awareness. *International Journal of Sport Management and Marketing,* 1(4), 289-310.

Conchar, M. P., Zinkhan, G. M., Peters, C., & Olavarrieta, S. (2004). An integrated framework for the conceptualization of consumers' perceived-risk processing. *Journal of the Academy of Marketing Science*, 32, 418-436.

Cox, D. F. & Rich, S. U. (1964). Perceived risk and consumer decision-making—the case of telephone shopping. *Journal of Marketing Research*, 1(4), 32-39.

Cunningham, L. F., Gerlach, J. H., Harper, M. D., & Young, C. E. (2005). Perceived risk and the consumer buying process: internet airline reservations. *International Journal of Service Industry Management*, 16(4), 357-372.

Dinev, T. & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, 17(1), 61-80.

Eckersley, M. (1988). The form of design processes: a protocol analysis study. *Design Studies*, 9(2), 86-94.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). The MIT Press.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. *International Journal of Human-Computer Studies*, 59(4), 451-474.

Fiske, S. T., and Taylor, S. E. (1991). *Social Cognition 2nd edn*. McGraw-Hill, New York.

Garbarino, E. C., & Edell, J. A. (1997). Cognitive effort, affect, and choice. *Journal of Consumer Research*, 24(2), 147-158.

Grant, M. J., Button, C. M., & Snook, B. (2017). An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement*, 41(4), 264-276.

Harridge-March, S. (2006). Can the building of trust overcome consumer perceived risk online? *Marketing Intelligence & Planning*, 24(7), 746-761.

Higgins, E. T. (1996). Knowledge activation: accessibility, applicability, and salience. *Social psychology: Handbook of Basic Principles*, 133-168.

Jacoby, J., & Kaplan, L. B. (1972). The components of perceived risk. Advances in consumer research. Ed. M. Venkatesan. Chicago: *Association for Consumer Research*, 382-393.

Karwatzki, S., Dytynko, O., Trenz, M., & Veit, D. (2017). Beyond the personalization–privacy paradox: privacy valuation, transparency features, and service personalization, *Journal of Management Information Systems*, 34(2), pp.369-400.

Kehr, F., Kowatsch, T., Wentzel, D., & Fleisch, E. (2015). Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, 25(6), 607-635.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Leigh, J. H., Zinkhan, G. M., & Swaminathan, V. (2006). Dimensional relationships of recall and recognition measures with selected cognitive and affective aspects of print ads. *Journal of Advertising*, 35(1), 105-122.

Liebermann, Y., & Stashevsky, S. (2009). Determinants of online shopping: examination of an early-stage online market. *Canadian Journal of Administrative Sciences*, 26(4), 316-331.

Meyer, R., & Kunreuther, H. (2017) *The Ostrich Paradox: Why We Underprepare for Disasters*, Wharton Digital Press.

Mitchell, V. W. (1999). Consumer perceived risk: conceptualisations and models. *European Journal of Marketing*, 33(1/2), 163-195.

Norberg, P.A., Horne, D.R., & Horne, D.A. (2007). The privacy paradox: personal information disclosure intentions versus behaviors." *Journal of Consumer Affairs*, 41(1), 100-126.

Roberts, J. H., & Lattin, J. M. (1997). Consideration: review of research and prospects for future insights. *Journal of Marketing Research*, 34(3), 406-410.

Waldman, A. E. (2020). Cognitive biases, dark patterns, and the 'privacy paradox. *Current Opinion in Psychology*, 31, 105-109.

IMPACT OF ETHICAL JUDGMENT ON UNIVERSITY PROFESSORS ENCOURAGING STUDENTS TO USE AI IN ACADEMIC TASKS

Jorge Pelegrín-Borondo, Cristina Olarte-Pascual, Luis Blanco-Pascual, Alba García-Milon

Universidad de La Rioja (Spain)

jorge.pelegrin@unirioja.es; cristina.olarte@unirioja.es; luis.blanco@unirioja.es; alba.garciam@unirioja.es

ABSTRACT

This research examines how ethical judgment influences the intention of professors to encourage their students to use artificial intelligence for their tasks at the university. Professors were presented with a potential scenario where the ChatGPT artificial intelligence functioned correctly and another scenario where it malfunctioned. Subsequently, they responded to a questionnaire. The scale used was the Composite MES (Shawver and Sennetti, 2009). A sample of 398 university professors in Business Administration studies was obtained. The results demonstrated that ethical judgment has a high capacity to explain the intention of professors to promote the use of AI among their students. The influencing dimensions were moral equity and egoism, with moral equity exhibiting a greater explanatory capacity.

KEYWORDS: Artificial intelligence, ethical concerns, higher education, intention to use.

1. INTRODUCTION

In higher education, Artificial Intelligence (AI) emerges as a phenomenon that both it introduces new possibilities and also poses significant challenges (Silander & Stigmar, 2019). This technology promises to transform the educational experience, offering substantial opportunities to enhance both the effectiveness and efficiency of teaching and academic governance, benefiting not only students but also teachers, administrative staff, and researchers (Nasrallah, 2014). However, this technological advancement is not free of ethical questions and considerable challenges.

The need to integrate AI in higher education is presented as a significant challenge (Stefan & Sharon, 2017). The adoption of AI in this field has the potential to optimize administrative processes, personalize teaching to meet individual student needs, and enhance research with advanced data analysis tools. The promise of efficiency and improvement in educational quality makes the integration of AI a crucial step towards the future of higher education. Notwithstanding, this is not free from ethical dilemmas, especially in the context of using AI-based technologies for teaching and learning (Celik, 2023).

Currently, educators find themselves at an ethical crossroads, torn between encouraging or discouraging students from using AI in their studies. In this decision, the ethical considerations of educators play a crucial role in determining their performance as promoters or opponents of AI in the academic field. Ethics, in this context, acts as a mechanism to address the controversy between the potential benefits of technological progress and not to compromise fundamental values of equity and justice (Olarte-Pascual, Pelegrín-Borondo, Reinares-Lara, Arias-Oliva, 2021).

AI in higher education presents a landscape of exciting possibilities and complex ethical challenges. The integration of AI requires a careful and thoughtful approach, where ethics not only guides decision-making but also drives deeper investigations to effectively understand and address the ethical complexities inherent in this technological advancement. Higher education is at a critical juncture of adaptation, and ethical consideration plays a fundamental role in determining the direction it will take in this era of technological innovation.

The impact of different dimensions of ethical judgment in this decision remains unexplored. This research aims to address this question, focusing on the widely recognized AI platform ChatGPT, which has captured the global attention and public interest. The choice of this focus is supported by recent news in Spain showing an extensive use of ChatGPT by university students (Planas Bou, 2023). By exploring how educators confront the ethical challenges associated with the use of this platform, this paper would try to contribute to the understanding and ethical management of AI integration in higher education.

2. LITERATURE REVIEW

One of the primary concerns regarding AI revolves around the potential for AI tools to exhibit systematic errors, leading to discrimination against students from diverse backgrounds. This situation poses a threat to the pursuit of inclusivity in education, a goal highly sought after (De Cremer & De Schutter, 2021; Dietvorst et al., 2018). Addressing these ethical challenges is paramount, and it demands meticulous attention to guarantee that the incorporation of AI in higher education is both equitable and respectful of diversity.

Additionally, other ethical concerns arise in relation to the use of AI, such as content moderation, environmental impact, and the risk of copyright infringement (Cooper, 2023). Responsibility in the use of AI involves not only pedagogical considerations but also weighing its ethical repercussions in the broader context of society and the environment. Ethical reflection thus becomes an essential component to ensure that technological progress is not only beneficial but also ethically responsible.

In the realm of ethical judgment, evaluations of actions are recognized as individual cognitive processes (Nguyen & Biderman, 2008). Moreover, within the framework of the psychological contract theory, decision-making is inherently subjective (Thompson & Hart, 2006). The amalgamation of these perspectives implies that ethical judgment is fundamentally a subjective assessment. This theoretical foundation finds relevance in situations where there are no absolute rules dictating what is permissible or prohibited (Goel et al., 2016). Decisions and actions often stem more from applied ethical perceptions than from a comprehensive understanding of what can or should be done (Cohen & Wellman, 2005; LaFollette, 2002). According to circular evolutionary ethics, ethical beliefs influence behavior in one phase, and over time, the behavior performed shapes ethical beliefs (Goel et al., 2016). In the context of this study, we find it apt to explore the impact of ethics on the intention of university professors to endorse the use of AI in their students' tasks and academic activities. This exploration is grounded in individual perceptions of what behaviors align with applied ethical considerations (Thompson & Hart, 2006).

Focusing on ethical judgment, Reidenbach and Robin (1990) propose that people rely on more than one reason when making ethical decisions, and therefore, they devised their Multidimensional Ethics Scale (MES). Reidenbach and Robin (1990) considered the existence of five main normative ethical theories that established a person's ethical judgment:

a) "Moral equity" dimension refers to the individual's perception of fairness, justice, and morality in a broad sense (Nguyen & Biderman, 2008, p. 628). According to Leonard and Jones (2017), it encompasses concepts such as fairness, justice, rectitude, and goodness. In an analysis of 155 academic articles, Hofmann et al. (2017) pointed out ethical dilemmas in the use of technology, especially in relation to justice. They concluded that equity is crucial in the development, evaluation, decision-making, implementation, use, and formation of norms related to technology. Additional studies highlight that the use of smart devices can increase the digital divide and create unfair advantages (Weber & Zink, 2014; Bozyer, 2015). "Moral equity" positively influences the intention to use disruptive technologies according to Pelegrín-Borondo et al. (2020).

b) "Relativism" involves the perception that morality is based on social and cultural norms, rather than in individual considerations (Nguyen & Biderman, 2008; Reidenbach & Robin, 1990). Ferenbok et al. (2016) emphasize that novel technologies can challenge social norms. Pelegrín-Borondo et al. (2020) observed that "relativism" is capable of explaining a significant part of the intention to use disruptive technologies.

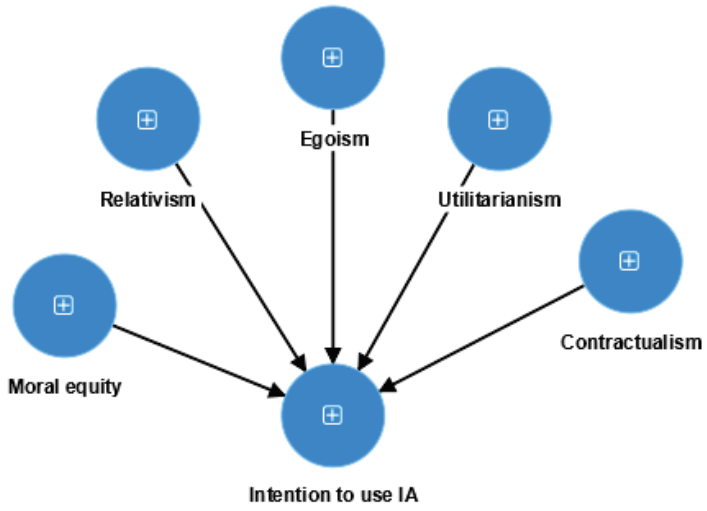
c) "Utilitarianism" involves decision-making based on the analysis of costs and benefits to achieve the greatest good for society as a whole (Nguyen & Biderman, 2008, p. 628). According to Berger et al. (2008), enhanced technologies are considered ethical because they have the potential to contribute to the advancement of society. In this regard, the balance between costs and social benefits is viewed from a utilitarian perspective, and this social utilitarianism can explain the intention to use disruptive technology (Pelegrín-Borondo et al., 2020).

d) "Egoism" is defined as the action that seeks only one's own long-term interest (Nguyen & Biderman, 2008). Leonard et al. (2017) argue that individual ethics are driven by personal benefits. In the field of technological implants to enhance one's own capabilities, Pelegrín-Borondo et al. (2020) find that this dimension of ethical judgment related to "egoism" is the most influential dimension in the intention to use these implants.

e) "Contractualism" reflects an individual's perception of right or wrong based on an implicit contract between the person and society (Nguyen & Biderman, 2008, p. 633). In the realm of technology, Shipp et al. (2014), Mok et al. (2015), and Thierer (2015) examine this perspective, considering ethical aspects such as autonomy, common welfare, privacy, and security.

Based on this theoretical framework, the model proposed is presented in Figure 1.

Figure 1. Proposed model.



This theoretical framework serves as the foundation for the proposed research, which explores how dimensions of ethical judgment affect the intention of university professors to promote the use of AI in the tasks and academic activities of their students. A conceptual model is proposed to address this issue, outlining the complex interaction between ethical dimensions and the promotion of AI in higher education. This approach aims to provide a deeper understanding of the relationship between ethics and the adoption of AI in the academic environment.

3. METHODOLOGY

Commencing with an explanation of the scale employed for assessing ethical judgment, initially, this scale consisted of eight items distributed across three subscales. It was developed through a prior distillation and validation process based on an original inventory of 33 items (Reidenbach & Robin, 1990, p. 639), which, in turn, had its foundation in earlier work by Reidenbach & Robin in 1988. The MES scale (1990) and its subsequent modified versions (e.g., Fleischman et al., 2017; Kadi´c-Maglajli´c et al., 2017; Mudrack & Mason, 2013; Pelegrín-Borondo et al., 2020; Secchi & Bui, 2018) are widely utilized in the literature to elucidate the influence of ethical judgment on human behavior.

Loo (2004, p. 290) argued that the MES scale provides users with a brief and psychometrically sound multidimensional ethical measure, suitable when administration time is limited. However, Reidenbach and Robin (1990) noted that this scale does not address items related to utilitarianism and egoism, a significant deficiency in situations involving utilitarian or egoistic considerations (Loo, 2004, p. 293). These observations are relevant to our study on AI, as its application may be linked to ego and social utility (utilitarianism).

In this vein, Shawver and Sennetti (2009) identified theoretical issues in the Reidenbach and Robin (1990) scale and proposed an alternative called the Composite MES. This new scale incorporates the five major ethical theories into five dimensions: "moral equity," "relativism," "utilitarianism," "egoism," and "contractualism" (deontology). The Composite MES has been widely used to explain the impact of ethical judgments on behavior (e.g., Kara et al., 2016; Manly et al., 2015; Mudrack & Mason, 2013). To a lesser extent, the MES has also been applied in the realm of consumer behavior (e.g., Jones & Leonard, 2016; LaTour & Henthorne, 1994; Leonard & Jones, 2017; Nguyen et al., 2008; Nguyen & Biderman, 2008). However, in the context of disruptive technology acceptance, the influence of ethical judgments and the dimensions of the Composite MES have only been discussed by Reinares-Lara et al. (2018), Pelegrín-Borondo et al. (2020), and Olarte-Pascual et al. (2021) in relation to body implant technologies for enhancing capabilities.

To measure usage intention, the scale developed by Venkatesh & Davis (2000) was adapted. Regarding the data collection for hypothesis analysis, contact was made with all faculties offering Business Management programs, and all professors were requested to respond to the self-administered questionnaire online. The decision was made to focus exclusively on professors of a specific type of university program to control for the potential influence of the type of program on the analyses.

Before responding to the questionnaire, participants were provided with the following information:

> Artificial Intelligence (AI) can be used for academic tasks such as information retrieval, summarization, translation, and improving writing, but there is a risk of incorrect responses. ChatGPT is an example of AI.
>
> Here is an example of an incorrect response from ChatGPT: "Tell me in 20 words for what specific tasks AI can be used in teaching with a reference."
>
> ChatGPT's response: "AI can be used for intelligent tutoring, educational data analysis, and administrative task automation in teaching (Chen, 2020). Reference: Chen, M. & Zhao, J. (2020). Applications of artificial intelligence in educational technology. Journal of Educational Technology & Society, 23(1), 133-146."
>
> UPON CHECKING THE REFERENCE, IT DOES NOT EXIST.
>
> Here is an example of a correct response from ChatGPT: "Summarize your previous answer without citing."

ChatGPT's response: "AI aids in teaching through tutoring, data analysis, and administrative automation."

In this way, a sample was obtained that captures the opinions of 398 university professors who teach in the Bachelor's program in Business Administration.

Regarding the results, the reliability and validity of the scales were examined. One item from the relativism dimension was removed due to convergent validity issues. The final scales demonstrated satisfactory reliability, convergent validity, and discriminant validity, as shown in Table 1.

Table 1. Composite reliability, Cronbach's alpha, AVE (convergent validity) and discriminant validity.

Construct	**Composite reliability > 0.7**	**Cronbach's *Alpha* > 0.7**	**AVE > 0.5**	**HTMT**				
				ME	**R**	**E**	**U**	**C**
Moral Equity (ME)	0.963	0.963	0.897					
Relativism (R)	0.834	0.832	0.713	0.864				
Egoism (E)	0.919	0.919	0.850	0.829	0.848			
Utilitarianism (U)	n.a.	n.a	n.a.	0.715	0.727	0.789		
Contractualism (C)	0.965	0.964	0.932	0.800	0.797	0.746	0.705	
Intention to use (IU)	0.963	0.963	0.929	0.746	0.677	0.723	0.612	0.653

Note: n.a = does not apply

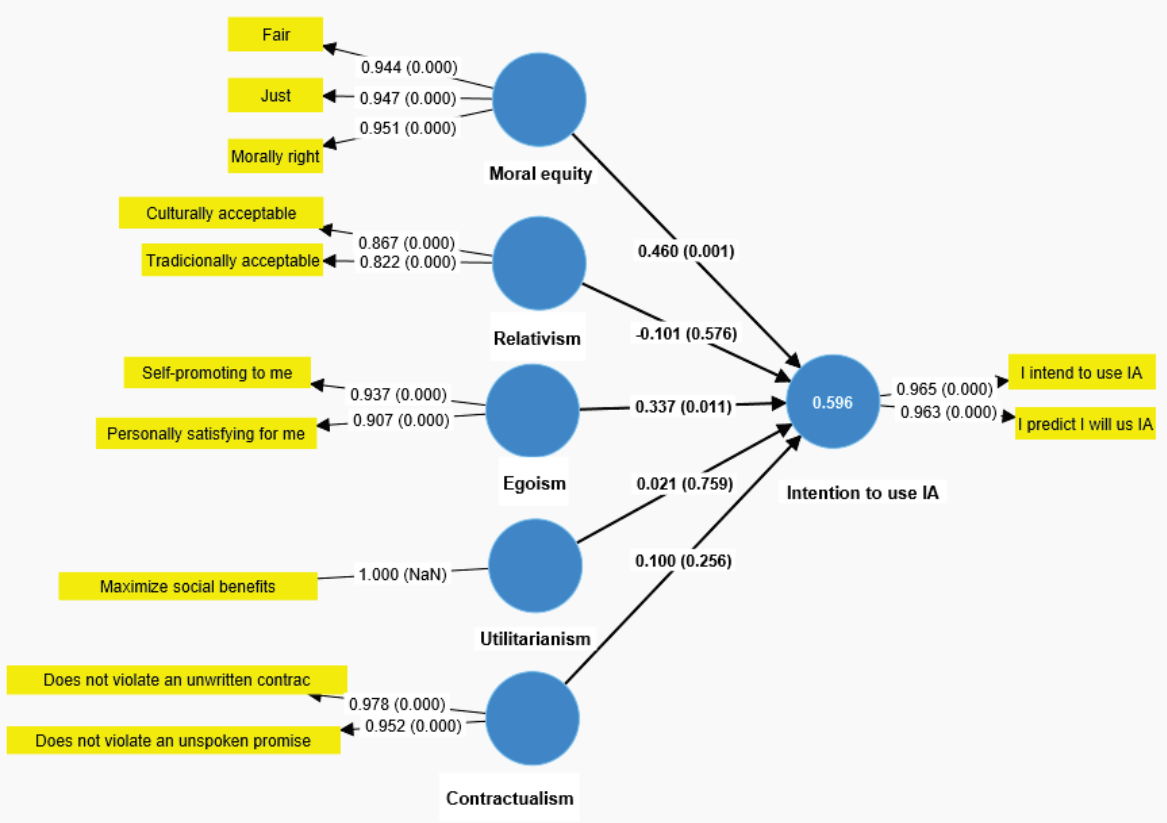
4. RESULTS

In Figure 2, the result of the model is displayed. Table 2 displays the values of R^2 and Q^2, the path coefficients (direct effects), and p-values for each antecedent variable of professors' intention for their students to use AI. The R^2 for the model of AI use intention was high ($R^2 = 0.596$), and the Q^2 provided by PLS Predict was greater than 0.5 ($Q^2 = 0.543$). This indicates that the dimensions of ethical judgment have explanatory and predictive power over professors' intention for their students to use AI. In Table 2, it is shown that the dimensions of moral equity and egoism positively influence the intention to use AI.

Table 2. Effect on the endogenous variables.

	R^2	**Q^2predict**	**Path coefficient**	***p*-value**
INTENTION TO USE AI	0.596	0.543		
Moral Equity =>(+) Intention to use AI			0.460	0.001
Relativism =>(+) Intention to use AI			-0.101	0.576
Egoism =>(+) Intention to use AI			0.337	0.011
Utilitarianism =>(+) Intention to use AI			0.021	0.759
Contractualism =>(+) Intention to use AI			0.100	0.256

Figure 2. Model outcome. Path coefficients (p-values).



5. DISCUSSION AND CONCLUSION

The present research examines how different dimensions of ethical judgment influence professors' intention to encourage their students to use AI for academic tasks. Professors were presented with a scenario in which ChatGPT AI worked correctly and one in which it operated incorrectly.

The results indicate that overall, professors' ethical judgment regarding students using AI for their tasks has a significant impact on their attempts to encourage students to use AI and propose the use of AI for academic tasks at the university. Previous findings had already demonstrated that ethical judgment influenced behavioral intention (Kara et al., 2016; Manly et al., 2015; Mudrack & Mason, 2013), specifically in the intention to use novel technology (Reinares-Lara et al., 2018; Pelegrín-Borondo et al., 2020; Olarte-Pascual et al., 2021). However, this had not been tested in the context of technology in teaching, nor specifically for the use of AI for tasks. In this regard, the first contribution of the research is demonstrating that both in the educational context and in the use of AI, ethical judgment is crucial for technology acceptance. The next key question is to determine which dimensions of ethical judgment affect this behavioral intention.

In this sense, our results show that two dimensions significantly impact professors' intention to promote the use of AI by students in teaching tasks and activities: moral equity and egoism.

Among them, moral equity has the highest explanatory power, indicating that perceiving the use of AI as fair motivates professors to encourage it. Egoism is the second influential dimension, suggesting that personal benefits derived from students' use of AI increase professors' inclination to promote it. Previous studies observed the influence of moral equity (Pelegrín-Borondo et al., 2020; Olarte-Pascual et al., 2021) and egoism (Pelegrín-Borondo et al., 2020; Olarte-Pascual et al., 2021) on the intention to use disruptive technology. However, unlike them, we observe that moral judgment has the greatest explanatory power for professors' intention to promote AI use in their students. In the research by

Pelegrín-Borondo et al. (2020) and Olarte-Pascual et al. (2021), egoism was the dimension with the greatest explanatory power regarding the intention to use disruptive technology, specifically body implants to enhance human capacity. This highlights that the most important dimensions of ethical judgment depend on the type of disruptive technology being analyzed.

Pelegrín-Borondo et al. (2020) and Olarte-Pascual et al. (2021) had determined that relativism, utilitarianism, and contractualism influenced the intention to use other disruptive technologies different from AI, but we have not observed this influence. These authors analyzed the same type of technology: body implants to enhance human capacity, and in one of the articles, the comparison with wearables. The conclusion that can be drawn is that the influence of different dimensions of ethical judgment on the intention to use technology depends on the type of disruptive technology being analyzed.

These conclusions emphasize the importance of considering the ethical perceptions of professors when integrating AI into education and provide valuable insights for developing effective strategies for the integration of AI in teaching.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support from the University of La Rioja for funding Teaching Innovation Projects during the academic year 2023-24, as well as the grant given to the COBEMADE Research Group at the University of La Rioja.

REFERENCES

Berger, F., Gevers, S., Siep, L., & Weltring, K. M. (2008). Ethical, legal and social aspects of brain-implants using nano-scale materials and techniques. NanoEthics, 2(3), 241–249. https://doi.org/10.1007/s11569-008-0044-9

Bozyer, Z. (2015). Augmented reality in sports: Today and tomorrow. International Journal of Science Culture and Sport, 3(4), 314–325. https://doi.org/10.14486/IJSCS392

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468.

Cohen, A. I., & Wellman, C. H. (Eds.). (2005). Contemporary debates in applied ethics. Wiley-Blackwell.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444-452.

De Cremer, D., & De Schutter, L. (2021). How to use algorithmic decision-making to promote inclusiveness in organizations. AI and Ethics, 1(4), 563–567. https://doi.org/10.1007/s43681-021-00073-0

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Ferenbok, J., Mann, S., & Michael, K. (2016). The changing ethics of mediated looking: Wearables, veillances, and power. IEEE Consumer Electronics Magazine, 5(2), 94–102.

Fleischman, G. M., Johnson, E. N., Walker, K. B., & Valentine, S. R. (2017). Ethics versus outcomes: Managerial responses to incentive-driven and goal-induced employee behavior. Journal of Business Ethics. https://doi.org/10.1007/s1055.1-017-3695-z

Goel, L., Hart, D., Junglas, I., & Ives, B. (2016). Acceptable IS use: Conceptualization and measurement. Computers in Human Behavior, 55, 322–328. https://doi.org/10.1016/j.chb.2015.09.029

Hofmann, B., Haustein, D., & Landeweerd, L. (2017). Intelligent-glasses: Exposing and elucidating the ethical issues. Science and Engineering Ethics, 23(3), 701–721.

Jones, K., & Leonard, L. N. (2016). Applying the multidimensional ethics scale in C2C commerce. Issues in Information Systems, 17(1), 26–36.

Kadi´c-Maglajli´c, S., Arslanagi´c-Kalajd˘zi´c, M., Micevski, M., Michaelidou, N., & Nemkova, E. (2017). Controversial advert perceptions in SNS advertising: The role of ethical judgement and religious commitment. Journal of Business Ethics, 141(2), 249–265. https://doi.org/10.1007/s1055.1-015-2755-5

Kara, A., Rojas-M´endez, J. I., & Turan, M. (2016). Ethical evaluations of business students in an emerging market: Effects of ethical sensitivity, cultural values, personality, and religiosity. Journal of Academic Ethics, 14(4), 297–325. https://doi. org/10.1007/s10805-016-9263-9

LaFollette, H. (2002). Ethics in practice (2nd ed.). Oxford: Blackwell Publishing.

LaTour, M. S., & Henthorne, T. L. (1994). Ethical judgments of sexual appeals in print advertising. Journal of Advertising, 23(3), 81–90. https://doi.org/10.1080/00913367.1994.10673453

Leonard, L. N., & Jones, K. (2017). Ethical awareness of seller's behavior in consumer-toconsumer electronic commerce: Applying the multidimensional ethics scale. Journal of Internet Commerce, 16(2), 202–218. https://doi.org/10.1080/15332861.2017.1305813

Leonard, L. N., Riemenschneider, C. K., & Manly, T. S. (2017). Ethical behavioral intention in an academic setting: Models and predictors. Journal of Academic Ethics,15(2), 141-166. https://doi.org/10.1007/s10805-017-9273-2.

Loo, R. (2004). Support for Reidenbach and Robin's (1990) eight-item multidimensional ethics scale. The Social Science Journal, 41(2), 289–294. https://doi.org/10.1016/j.sosci.j.2004.01.020

Manly, T. S., Leonard, L. N., & Riemenschneider, C. K. (2015). Academic integrity in the information age: Virtues of respect and responsibility. Journal of Business Ethics, 127 (3), 579–590.

Mok, T. M., Cornish, F., & Tarr, J. (2015). Too much information: Visual research ethics in the age of wearable cameras. Integrative Psychological and Behavioral Science, 49(2), 309–322. https://doi.org/10.1007/s12124-014-9289-8

Mudrack, P. E., & Mason, E. S. (2013). Ethical judgments: What do we know, where do we go? Journal of Business Ethics, 115(3), 575–597. https://doi.org/10.1007/s10551-012-1426-z

Nasrallah, R. (2014). Learning outcomes role in higher education teaching. Education, Business and Society, 7(4), 257–276. https://doi.org/10.1108/EBS-03-2014-0016

Nguyen, N. T., & Biderman, M. D. (2008). Studying ethical judgments and behavioral intentions using structural equations: Evidence from the multidimensional ethics scale. Journal of Business Ethics, 83(4), 627–640. https://doi.org/10.1007/s10551-007-9644-5

Nguyen, N. T., Basuray, M. T., Smith, W. P., Kopka, D., & McCulloh, D. (2008). Moral issues and gender differences in ethical judgment using Reidenbach and Robin's (1990) multidimensional ethics scale: Implications in teaching of business ethics. Journal of Business Ethics, 77(4), 417–430. https://doi.org/10.1007/s10551-007-9357-9

Olarte-Pascual, C., Pelegrín-Borondo, J., Reinares-Lara, E. Arias-Oliva, M. (2021). From wearable to insideable: Is ethical judgment key to the acceptance of human capacity-enhancing intelligent technologies?. Computers in Human Behavior, 114, 106559.

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2020). Does ethical judgment determine the decision to become a cyborg? Journal of Business Ethics, 161(1), 5–17. https://doi.org/10.1007/s10551-018-3970-7

Planas Bou, C (2023). Universitarios y adolescentes se pasan en masa a ChatGPT para hacer trabajos (y exámenes). El Periódico (9-05-2023). https://www.elperiodico.com/es/sociedad/20230508/chatgpt-universidad-escuelas-inteligencia-artificial-estudiantes-deberes-examenes-86837251

Reidenbach, R. E., & Robin, D. P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. Journal of Business Ethics, 7, 871–879. https://doi.org/10.1007/BF00383050

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. Journal of Business Ethics, 9(8), 639–653. https://doi.org/10.1007/BF00383391

Reinares-Lara, E., Olarte-Pascual, C., Pelegrín-Borondo, J., & Pino, G. (2016). Nanoimplants that enhance human capabilities: A cognitive-affective approach to assess individuals' acceptance of this controversial technology. Psychology and Marketing, 33(9), 704–712. https://doi.org/10.1002/mar.20911

Secchi, D., & Bui, H. T. (2018). Group effects on individual attitudes toward social responsibility. Journal of Business Ethics, 149(3), 725–746. https://doi.org/10.1007/s1055.1-016-3106-x

Shawver, T. J., & Sennetti, J. T. (2009). Measuring ethical sensitivity and evaluation. Journal of Business Ethics, 88(4), 663–678. https://doi.org/10.1007/s10551-008-9973-z

Shipp, V., Skatova, A., Blum, J., & Brown, M. (2014). May). The ethics of wearable cameras in the wild. Proceedings of the IEEE 2014 international symposium on ethics in engineering, science, and technology. Chicago, USA: IEEE Press. https://doi.org/10.1109/ETHICS.2014.6893382

Silander, C., & Stigmar, M. (2019). Individual growth or institutional development? Ideological perspectives on motives behind Swedish higher education teacher training. Higher Education: The International Journal of Higher Education Research, 77, 265–281. https://doi.org/10.1007/s10734-018-0272-z

Stefan, A. D. P., & Sharon, K. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 1, 3–13. https://doi.org/10.1186/s41039-017-0062-8

Thierer, A. D. (2015). The internet of things and wearable technology: Addressing privacy and security concerns without derailing innovation. Richmond Journal of Law and Technology, 21(2). 1–118. https://doi.org/10.2139/ssrn.2494382

Thompson, J., & Hart, D. (2006). Psychological contracts: A nano-level perspective onsocial contract theory. Journal of Business Ethics, 68(3), 229–241. https://doi.org/10.1007/s10551-006-9012-x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. Management Science, 46, 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Weber, H., & Zink, K. J. (2014). Boon and bane of ICT acceleration for vulnerable populations. In C. Korunka, & P. Hoonakker (Eds.), The impact of ICT on quality of working life (pp. 177–190). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-8854-0_11

EXAMINING THE MEDIATING EFFECT OF FINANCIAL FRAUD RISK ON FINANCIAL EDUCATION AND CORPORATE ETHICS AND INTENTION TO USE FINANCIAL SERVICES

Juan Carlos Yáñez-Luna, Pedro I. González-Ramírez

Universidad Autónoma de San Luis Potosí (México)

jcyl@uaslp.mx; pedro.gonzalez@uaslp.mx

ABSTRACT

Based on path analysis, the study evaluates the direct effect of financial fraud risk, corporate ethics, financial education, and the intention to use financial services. Furthermore, the study proposes to assess the mediating impact of financial fraud risk on the direct relationship between financial education and the discretion to use financial services, as well as the mediating effect of financial fraud services on the direct relationship between financial education and the intention to use financial services. The results show that all direct paths are significant. The direct effect between financial fraud risk and intention to use shows a negative value. Corporate Ethics and financial education are positively related to the intention to use financial services. Finally, results indicate that the relationship between financial education and the intention to use financial services is partially mediated by the perception of the risk of financial fraud. A higher level of financial literacy may be associated with a lower perceived risk of financial fraud, which in turn may increase the willingness to use financial services.

KEYWORDS: Financial fraud risk, corporate ethics, financial education, intention to use, PLS-SEM.

1. INTRODUCTION

Fraud poses a significant challenge to the global economy. Its adverse effects extend from the stability of financial markets to businesses, consumers, and governments. Confronting this large-scale issue requires a comprehensive approach that involves multiple stakeholders. In this regard, one of the primary reasons international financial institutions are concerned with combating financial fraud is to preserve the integrity of the global financial system. Fraud can erode investor and market participants' trust, leading to capital flight, financial volatility, and decreased investment. Moreover, fraud can hinder access to credit and essential financial services, thus impeding economic and social development. These consequences can undermine economic growth and financial stability, impacting entire countries and regions.

The term "*fraud*" can be simply defined as a crime involving deceptive activities in a specific sector. These activities are typically carried out by individuals or groups with malicious intent to obtain illicit benefits. With the increased use of technology in recent decades and the interconnectedness of devices (Internet of Things), such illegal activities have become more prevalent. For instance, Cross et al. (2014) note that online fraud arises from "an individual's experience of responding over the internet to a dishonest invitation, solicitation, notification, or offer, by providing personal information or money that results in a loss, with or without financial impact." Likewise, Juhandi et al. (2020) state that fraud refers to "a broad legal concept, describing any intentional fraudulent attempt aimed at taking someone's property or rights, or those of other parties." Marabad (2021) suggests that "*fraud* is defined as an unlawful deception intended to secure financial or personal gain. It is a premeditated behaviour that goes against the law or policy to achieve unfair financial gain."

The abovementioned concepts propose cooperation among local and international financial institutions, government authorities, and regulatory bodies to create public policies that effectively

combat financial fraud in a region. For instance, the (OCDE/CAF, 2020) highlights that implementing effective programs to promote financial inclusion and education can enhance consumer awareness, which, in turn, are crucial elements for protecting against and preventing fraud while fostering a culture of integrity and transparency in the financial sector. In this regard, the economic impact of financial fraud in emerging countries can have significant consequences. That means financial institutions and consumers bear direct financial losses while the overall economy suffers from a lack of trust and decreased investment. Moreover, financial fraud can erode the reputation of companies and the country, leading to long-term effects on economic development.

Given the above, the government in Mexico has implemented several measures to combat financial fraud, including the establishment of specialized units for financial crimes and the enactment of stricter laws and regulations (CONDUSEF, 2021). However, it is worth noting that the country still faces significant challenges related to infrastructure, such as telecommunications and cybersecurity. In this regard, the lack of robust security infrastructure can impact financial institutions' reputation and perceived quality (García Witron, 2021), ultimately affecting customer loyalty. Building trust remains a recurring challenge for institutions, particularly those operating in the financial sector. Institutions must develop a corporate ethics framework as a social normative framework, embracing ethical practices to convey commitment and social responsibility as a loyalty-building strategy (Gómez Pescador & Arzadun, 2019).

This study assesses the relationship between financial fraud risk, corporate ethics, financial education, and the intention to use financial services. Notably, the study will evaluate the direct effect between financial fraud risk, corporate ethics, financial education, and the intention to use financial services. Furthermore, the study proposes to assess the mediating impact of financial fraud risk on the direct relationship between financial education and the discretion to use financial services, as well as the mediating effect of financial fraud services on the direct relationship between financial education and the intention to use financial services.

2. BACKGROUND

Financial fraud risk has been studied concerning the intention to use financial services. Trust, social influence, cyber-security (risks), and privacy risks influence customers' behavioural intention to use financial services. Al-Gasawneh et al. (2022) discovered that perceived risk has a negative impact on the use of financial artificial intelligence services. Kim et al. (2016) developed multi-class financial misstatement detection models to detect fraud intention, indicating the importance of detecting and classifying misstatements according to the presence of fraud intention. These findings suggest that understanding and addressing fraud and perceived risks are crucial in promoting the adoption and use of financial services and applications.

Financial education and corporate ethics significantly impact the relationship between financial fraud risk and intention to use financial services. Studies have shown that financial education programs positively affect financial knowledge and downstream financial behaviours (Liao et al., 2019). That suggests that individuals who receive financial education are more likely to make informed decisions and engage in ethical financial practices. Furthermore, financial education positively impacts the intention to use financial services. Improved financial literacy and capability can motivate and enable the safe and beneficial use of financial services (Ansar et al., 2023). The evidence shows that financial education improves personal finance and influences public choices, voting behaviour, economic reforms, and policy outcomes (Kaiser et al., 2022). Therefore, financial education plays a crucial role in empowering individuals to make informed decisions and reducing the probability of financial fragility, particularly in old age.

Additionally, research has found that corporate social responsibility (CSR) is negatively associated with fraudulent financial activities, indicating that CSR firms are less likely to engage in financial fraud (Jamieson et al., 2019). That implies that companies prioritising ethical behaviour and social responsibility are more trustworthy and less likely to engage in fraudulent activities. Therefore, financial education and corporate ethics play a crucial role in reducing financial fraud risk and promoting responsible use of financial services.

3. METHODS

This study applies a quantitative methodology based on path analysis and PLS-SEM (Hair Jr. et al., 2019). The methodological justification is based on the need to comprehensively address the following predictor variables: Financial Education (Mungaray et al., 2021; Świecka, 2019), interpreted in this study as the knowledge, skills, and understanding individuals possess regarding various financial mechanisms enhance decision-making. Corporate Ethics (Aliyu, 2022; Suresh & Rakesh, 2019), which for this study refers to the standards of conduct, values, and principles guiding the actions of financial businesses concerning social responsibility, transparency, and behaviour in their operations and relationships with stakeholders. Financial Fraud Risk (Abdulrahman & Alshammari, 2022; Murrar, 2022), which for this research refers to the possibility of fraudulent or deceptive activities in the financial realm, ranging from dishonest practices in commercial transactions to the manipulation of financial data or information to gain illegitimate profits.

The dependent variable defined is the Intention to Use Financial Services, based on the original concept (Davis, 1985), referring to an individual's predisposition to use technology. This research aligns with studies such as (Nan et al., 2020; Phonthanukitithaworn et al., 2016), where the technology explicitly used relates to payment methods as financial services available in the market, through other types of services such as bank accounts, loans, insurance, investments, among others, will also be considered.

A survey-based instrument was designed to collect data. This instrument allowed gathering information from 1077 respondents through 23 scales exclusively developed for this research. The survey was conducted electronically using Google Docs and distributed via email and social media.

This study adopts a methodology based on the Partial Least Squares – Structural Equation Model (PLS-SEM) and the Path Analysis Technique, both widely used in academic literature for this type of research on the intention to use electronic financial services (Dewi et al., 2019; Sharma & Aggarwal, 2019). Thus, a model was developed to comprehensively understand the complexities and interrelationships between predictor and dependent variables. The following causal assumptions and hypotheses will be analysed according to the proposed methodology.

Direct relationship between Financial Fraud Risk (FFR) and Intention to Use Financial Services (IU): The perceived risk of financial fraud may decrease people's confidence in the financial system. Suppose there are concerns about the safety of financial services due to fraud risk. In that case, an individual's willingness to use these services is likely to be negatively affected, so this study hypothesizes that:

H1. There is a significant negative relationship between FFR and IU financial services. It is posited that an increased perception of financial fraud risk correlates with a decreased interest in utilising financial services.

Direct relationship between Corporate Ethics (CE) and Intention to Use Financial Services (IU): This association concerns companies with ethical practices tend to generate more consumer trust. A solid corporate ethics can create a positive perception among customers about the company's

transparency, reliability, and responsibility, which could influence individuals' willingness to use the financial services.

H2. A significant positive association exists between CE and IU financial services. This hypothesis suggests that higher ethical standards within a financial institution are related to an increased willingness to engage with their financial services.

A direct relationship between Financial Education (FE) and Intention to Use Financial Services (IU): This relationship expects that users with higher financial education will have a more robust understanding of available financial products and services, which could generate greater confidence in their use. Additionally, financial education can assist in making more informed and prudent financial decisions, increasing the willingness to use financial services. Based on the above, this study postulate that:

H3. A significative positive relationship is anticipated between FE and IU financial services. This hypothesis posits that a heightened level of financial education correlates with an increased disposition to utilise financial services.

Direct relationship between Corporate Ethics (CE) and Financial Fraud Risk (FFR): This relationship examines the perception of financial fraud risk or incidents within an institution, which may raise doubts about organisation's integrity and ethical values. Also, it can lead users to believe that company lacks strong internal controls or an ethical organizational culture. Hence, through this assumption, it can be hypothesised that:

H4: A significant negative relationship is expected between CE and FFR. This hypothesis suggests that higher ethical standards within a financial institution correlate to a reduced perception of financial fraud risk.

A direct relationship between Financial Education (FE) and Financial Fraud Risk (FFR): This relationship analyses how an individual with higher financial education relates to better capability in identifying and evaluating financial risk, including fraud. Furthermore, a robust financial education could prepare individuals with strategies to protect their assets against possible financial frauds. Based on the above, it can be hypothesised in this study that:

H5. There is a significant negative relationship between FE and FFR. This hypothesis posits that a heightened level of financial education is associated with a decreased perception of financial fraud risk.

4. RESULTS AND DISCUSSION

4.1. Convergent validity and reliability

The reliability and validity of measurement scales are fundamental in any research study to estimate internal consistency. These indicators provide the foundation to ensure that the collected data is accurate and consistent, strengthening the validity of the conclusions and generalisations drawn from the study. In this context, this section presents a detailed analysis of the reliability of the scales used to measure the four critical variables in the studied model. Oviedo and Arias (2005) point out some metrics to evaluate Cronbach's Alpha. The proposed metrics range between 0 and 1. Hence, the minimum acceptable threshold for the coefficient is set at 0.70. Elements scoring below this value should be deemed insufficient, while those surpassing 0.90 are considered redundant. According to Hair Jr. et al. (2019), the Average Variance Extracted (AVE) measures the proportion of latent variables captured by measurement error. It suggests that this value exceeds the threshold of 0.5; the Composite Reliability values rhoC and rhoA should exceed 0.7. A higher value suggests a better ability to represent latent variables and their measurements reliably and accurately.

The Table 1 shows the Financial Fraud Risk (FFR), Corporate Ethics (CE), Financial Education (FE), and Intention to Use (IU) reliability values for all variables in the model.

Table 2. Reliability of the constructs.

	∝	rhoC	AVE	rhoA	R2
Financial Education	0.923	0.924	0.802	0.925	.
Corporate Ethics	0.919	0.919	0.792	0.921	.
Financial Fraud Risk	0.944	0.944	0.773	0.947	0.585
Intention to Use	0.926	0.927	0.809	0.930	0.672
Alpha, rhoC, and rhoA should exceed 0.7 while AVE should exceed 0.5					

Source: Self-elaboration based on SmartPLS 4 results.

The reliability analysis in the studied model shows that the scales are highly encouraging; all the values exceed the recommended quality criteria in academic literature, strengthening the validity and reliability of the studied model.

3.3. Discriminant validity

According to Chua (2022, p. 18), discriminant validity refers to the ability of a measurement instrument or test to differentiate between different constructs or concepts. This validation demonstrates not only the capability of the variables to measure the same information as others but also distinguish between different characteristics or dimensions. Discriminant validity ensures the uniqueness and specificity of measurements, avoiding redundancy when assessing similar aspects and allowing for a more precise and detailed understanding of the studied phenomena. In studies based on structural equation modelling (Sarstedt et al., 2017), a specific test is commonly conducted, such as the Fornell-Larcker Criterion, Corss-loadings, and the Heterotrait-monotrait Ratio (HTMT).

Table 2 illustrates the discriminant validity analysis using the Fornell-Larcker criterion (FL) to assess the distinction between constructs within the proposed model. The FL criterion points out that the square root of the AVE for each construct must exceed the correlations between a specific construct and the other constructs (Fornell & Larcker, 1981). The outcomes from the test make it apparent that the values along the diagonal, representing the square root of the AVE for each construct, exceed the corresponding correlations located beneath the diagonal. The findings suggest that the analysed constructs demonstrate adequate discriminant validity, as the shared variance among constructs is lower than the individual variance of each. This evidence supports that the measurements captured distinct conceptual aspects, implying a suitable differentiation among the evaluated constructs within the research model.

Table 3. Fornell-Larcker Criteria and HTMT Ratio.

	Financial Education	Corporate Ethics	Financial Fraud Risk	Intention to Use
Financial Education	**0.895**	0.880	0.732	*0.787*
Corporate Ethics	0.881	**0.890**	0.748	*0.785*
Financial Fraud Risk	-0.733	-0.749	**0.879**	*0.697*
Intention to Use	0.787	0.785	-0.699	**0.900**
Square root of AVE on the bold diagonal and the construct correlations on the lower triangle. On the upper triangle (above diagonal) HTMT values				

Source: Self-elaboration based on SmartPLS 4 results.

The Cross-loadings represent the loadings of each indicator (latent variable's observed measures) on all latent variables in the model, assessing the degree of association or contribution of each indicator to its intended latent variable and other variables in the model. Additionally, in the relationships between the observed variables (indicators) and the latent variables in the model, all indicators align well with their respective latent variables, demonstrating strong associations that might mean the chosen indicators effectively represent the constructs they are intended to measure. That supports the reliability and validity of the measurement model proposed in this study.

According to Henseler et al. (2015), the HTMT assesses discriminant validity in PLS-SEM. This ratio compares the correlations between constructs (measured as Heterotrait correlations) against the average correlations within constructs (Monotrait correlations). Lower HTMT values indicate better discriminant validity, suggesting that constructs are distinct. In HTMT, generally, the values accepted for the discriminant validity should be less than 1; values below 0.85 or 0.90 are considered adequate, indicating that discriminant validity is present among the constructs. Table 2 suggests potential issues with discriminant validity between CE and FE (0.88), which means the constructs might be more closely related or overlapping to some extent, which could impact the distinctiveness of the concepts in the model. However, it is essential to note that the result for this pair falls within the acceptable range for HTMT; therefore, further analysis is not required for this study.

3.4. Collinearity analysis

Multicollinearity is the phenomenon where the predictor variables in a regression model are highly correlated and can pose significant challenges in statistical analysis, leading to unreliable regression coefficients and potentially misleading interpretations of relationships. The academic literature employs the Variance Inflation Factor (VIF) as a diagnostic tool (Chua, 2022; Kyriazos & Poga, 2023) to address this concern in terms of the Structural Equation Model, so this measure quantifies the degree of multicollinearity, adding in the assessment of the validity of regression models. The VIF values of the studied model are as follows: for the Financial Fraud Risk construct, the VIF values for individual items (FFR 1 to 5) range from 3.350 to 4.593; these values, although not exceeding the conventional threshold of 10, suggest a correlation among the different dimensions of FFR.

Regarding variables associated with Corporate Ethics (CE1, CE2 and CE5), the VIF values are between 2.882 and 4.048; these results indicate a moderate correlation among the dimensions of the Latent Variable but do not suggest significant multicollinearity.

The variables corresponding to Financial Education (EF2, EF4. EF5) exhibit VIF values rating from 2.985 to 4.266; the results are similar to the previous case, and a moderate correlation among the dimensions of FE is observed, without problematic multicollinearity being detected. Finally, the variable related to Intention to Use (IU1, IU3, IU4) shows VIF values between 3.132 and 4.241. These Results suggest a moderate correlation among the dimensions of IU, without significant multicollinearity being evident.

3.5. Analysis of the Structural Model and Mediating Effects

Measuring the mediating effects (direct and indirect) between the latent variables outlined in the model is crucial to accomplish the study's objective and making a meaningful contribution to the academic literature on financial inclusion.

Table 3 shows that all direct paths are significant at $p < 0.05$. The direct effect between FFR and IU is negative ($\beta = -0.192$, $p < 0.05$). That indicates that every unit increase in the FFR is associated with a decrease of 0.192 units in the intention to use Financial Services. In other words, this finding suggests that when people perceive a higher risk of fraud in the financial environment, they tend to be more

cautious or less likely to use these services. Corporate Ethics is positively related to the intention to use financial services (β = 0.320, p < 0.05); this outcome suggests an increment in the intention to use financial services for every increase in the corporate ethics perception. That implies that when companies adopt high ethical standards, individuals are more willing to use the financial services offered by these companies.

Table 3. The direct effects of the hypothesized model.

	OE	T Stat.	2.50% CI	97.5% CI	*p*-Value	f^2
FFR → IU	-0.192	4.189	-0.284	-0.101	0.000	0.046
CE → IU	0.320	4.281	0.169	0.469	0.000	0.063
FE → IU	0.365	5.128	0.230	0.509	0.000	0.086
CE → FFR	-0.462	5.798	-0.613	-0.304	0.000	0.116
FE → FFR	-0.326	4.092	-0.487	-0.171	0.000	0.058

Source: Self-elaboration based on SmartPLS 4 results.

On the other hand, there is a positive relation between financial education and the intention to use financial services (β = 0.365, p < 0.05). A higher financial education is associated with a greater intention to use these services. This result indicates that people with a higher level of financial education are more inclined to use financial services. The outcomes in this research also show that corporate ethics is negatively related to the risk of financial fraud (β = -0.462, p < 0.05). An increase in corporate ethics is associated with a decrease in the perceived risk of fraud. That means that when companies maintain high ethical standards, individuals perceive a lower risk of fraud in the financial environment. Likewise, financial education is negatively related to the perceived risk of financial fraud (β = -0.326, p < 0.05). A higher level of financial literacy is associated with decreased perceived risk of fraud. This result suggests that as people with more excellent financial knowledge tend to perceive a lower risk of fraud in the financial field.

In PLS-SEM analysis, the f^2 index is used to assess the relative relevance of each endogenous variable in the structural model. This index indicates the proportion of variance in each endogenous variable explained by predictor variables. The academic literature suggests that values in f^2 close to 1 indicate that the predictor variables explain a significant proportion of the variance in the endogenous variable. On the other hand, values close to 0 suggest that the predictor variables have minimal contribution in explaining the variance; following (Cohen, 1988, pp. 410–414; Sarstedt et al., 2017) suggest a guideline for the f^2 values, for example, values of 0.02 represent a small effect, 0.15 represent moderate effect and values of 0.35 or higher represent large effects.

Table 3 also shows the f^2 sizes for this study. According to the results, the relationship between FFR and IU (f^2 = 0.046) indicates that the perceived risk of financial fraud has a small effect on the intention to use financial services. However, this is a relatively low effect; it is still significant in the context model. Corporate ethics has a small effect on the intention to use financial services (f^2 = 0.063); this value suggests that corporate ethics has a more substantial influence on people's willingness to use financial services than the perceived risk of financial fraud. Financial education moderately affects the intention to use financial services (f^2 = 0.086). This value indicates that the impact of financial education is significant and relatively stronger than the effect of the perceived risk of financial fraud on people's willingness to use financial services. Corporate ethics has a moderate to substantial effect on the perception of financial fraud risk (f^2 = 0.116). That suggests that ethical practices in companies may have a stronger impact on fraud risk perception than other variables in the model. Finally, financial education has a small effect on the perception of financial fraud risk (f^2 = 0.058). Although the effect is not as strong as that of corporate ethics in this regard, it is still significant relative to the model.

Table 4. The indirect effects of the hypothesized model.

	OE	BS-M	BS-SD	T Stat.	2.5% CI	97.5% CI	p-Value
CE → FFR → IU	0.088	0.088	0.027	3.264	0.041	0.148	0.001
FE → FFR → IU	0.062	0.062	0.021	3.012	0.026	0.107	0.003

Source: Self-elaboration based on SmartPLS 4 results.

Table 4 shows the indirect effects of the paths in the proposed model. A positive indirect effect of 0.088 implies that through the mediating effect of FFR, the relationship between CE and IU is positive; this indirect effect is statistically significant with a *p*-value = 0.001 and a confidence interval ranging from 0.041 to 0.148. The result suggests that the relationship between corporate ethics and the intention to use financial services is, in part, mediated by the perception of the risk of financial fraud; that is, when a company adopts high ethical standards (CE), this can positively influence the perception of a lower risk of financial fraud, which in turn can increase people's willingness to use financial services. This indirect relationship is essential in financial decision-making since trust in corporate ethics can determine consumers' acceptance and adoption of financial services.

Like the previous finding, a positive indirect effect of 0.062 is estimated. This suggests that the relationship between financial education and the intention to use financial services is positive through the mediating effect of the perceived risk of financial fraud. This indirect effect is also statistically significant, with a *p*-value = 0.003 and a confidence interval ranging from 0.026 to 0.107. Results also indicate that the perception of the risk of financial fraud partially mediates the relationship between financial education and the intention to use financial services. A higher level of financial literacy may be associated with a lower perceived risk of financial fraud, which may increase the willingness to use financial services. This relationship highlights the importance of financial education in forming perceptions about financial risk and how this can influence people's financial behaviour.

3.6. Analysis of predictive power

Carrying out a predictive analysis within the framework of research based on PLS-SEM becomes a crucial aspect of evaluating the model's effectiveness in terms of its ability to anticipate and explain relationships between variables and conduct subsequent confirmatory analysis. According to Chua (2022), predictive measures such as Q^2predict and MAE are suggested for this type of analysis. These techniques enable a comprehensive evaluation of the predictive suitability of the PLS-SEM model, providing an essential perspective on its competence to predict key variables of phenomena in real-world scenarios.

Table 5 illustrates positive Q^2 prediction values for the Financial Fraud Risk (Q^2 predict = 0.527) and Intention to Use Financial Services (Q^2 predict = 0.583). Moreover, the mean values, equaling zero, indicate that the final mediation model possesses adequate predictive capability. Table 6 displays indicators' values for the two dependent variables in the PLS-SEM model, which are positive, ranging from 0.355 to 0.557. The outcome of the Q^2 prediction analysis further confirms the mediation model's sufficient predictive capacity.

Table 5. Q^2predict values of the dependent variables.

Latent Variable	Q^2predict	RMSE	MAE	Mean
FFR	0.527	0.689	0.500	0.000
IU	0.583	0.647	0.469	0.000

Source: Self-elaboration based on SmartPLS 4 results.

Table 6. Q^2predict values of the indicators of the two dependent variables.

	Q^2predict	PLS-SEM_RMSE	PLS-SEM_MAE	LM_RMSE	LM_MAE
FFR1R	0.492	1.383	1.002	1.387	0.972
FFR2R	0.434	1.43	1.043	1.434	1.044
FFR3R	0.355	1.536	1.167	1.534	1.18
FFR4R	0.410	1.469	1.09	1.477	1.102
FFR5R	0.455	1.341	0.995	1.347	0.997
IU1	0.557	1.218	0.901	1.215	0.88
IU3	0.528	1.228	0.91	1.224	0.904
IU4	0.433	1.384	1.025	1.376	1.035

Source: Self-elaboration based on SmartPLS 4 results.

CONCLUSION

Based on path analysis and PLS-SEM, this study assessed the relationship between financial fraud risk, corporate ethics, financial education, and the intention to use financial services. The results show that all direct paths are significant. The direct effect between financial fraud risk and intention to use shows a negative value. Corporate Ethics and financial education are positively related to the intention to use financial services. Finally, results indicate that the perception of the risk of financial fraud partially mediates the relationship between financial education and the intention to use financial services. A higher level of financial literacy may be associated with a lower perceived risk of financial fraud, which may increase the willingness to use financial services.

The study aimed to enable decision-makers to formulate strategies and policies based on corporate ethics, creating a context of trust in using financial services even in the presence of financial fraud risk. The proposed model shows the importance of incorporating financial education in university curricula, intending to promote the use of financial services. Additionally, by establishing a reliable and ethical environment, financial institutions can strengthen their relationship with clients and encourage greater engagement in financial services, thereby contributing to economic and social development.

ACKNOWLEDGEMENTS

We sincerely thank the Sistema de Bibliotecas at the Universidad Autónoma de San Luis Potosí for their invaluable support and provision of access to a diverse array of academic databases. The accessibility they provide to a wealth of information has significantly contributed to the depth and breadth of this research.

REFERENCES

Abdulrahman, A. T., & Alshammari, A. O. (2022). Factor Analysis to Determine the Actual Causes That Led to the Spread of Financial Fraud. *Advances and Applications in Statistics*, *79*, 11–23. https://doi.org/10.17654/0972361722057

Al-Gasawneh, J. A., Alfityani, A., Al-Okdeh, S., Almasri, B., Mansur, H., Nusairat, N. M., & Siam, Y. A. (2022). Avoiding uncertainty by measuring the impact of perceived risk on the intention to use financial artificial intelligence services. *Uncertain Supply Chain Management*, *10*(4), 1427–1436. https://doi.org/10.5267/j.uscm.2022.6.013

Aliyu, A. N. (2022). Business Ethics and Corporate Social Responsibility for Successful Modern Business Operations. *KIU Journal of Humanities*, *7*(3), 33–41.

Ansar, S., Klapper, L., & Singer, D. (2023). *The Importance of Financial Education for the Effective use of Formal Financial Services*. The World Bank. https://doi.org/10.1596/1813-9450-10345

Chua, Y. P. (2022). *A Step By Step Guide PLS-SEM Data Analysis Using SmartPLS 4*. Kuala Lumpur: Researchtree Education.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

CONDUSEF. (2021). *Fraudes cibernéticos y tradicionales*. Secretaría de Hacienda y Crédito Público. https://www.condusef.gob.mx/documentos/comercio/FraudesCiber-3erTrim2019.pdf

Cross, C., Smith, R. G., & Richards, K. (2014). Challenges of responding to online fraud victimisation in Australia. *Trends & Issues in Crime and Criminal Justice*, *May*(474), 1–5. http://www.bbc.co.uk/news/

Davis, F. D. (1985). *A Technology Acceptance Model for Epirically Testing New End-User Information Systems: Theory and Results*. Massachusetts Institute of Technology.

Dewi, C. K., Mohaidin, Z., & Murshid, M. A. (2019). Determinants of online purchase intention: a PLS-SEM approach: evidence from Indonesia. *Journal of asia business studies*, *14*(3), 281–306. https://doi.org/10.1108/JABS-03-2019-0086

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39. https://doi.org/10.2307/3151312

García Witron, C. (2021). Ciberseguridad en el Sector Financiero. ¿Cómo transformar una amenaza en una oportunidad? [Comillas Universidad Pontificia]. https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/46570/TFG-Garcia%20Wirton%2C%20Carlota.pdf

Gómez Pescador, I., & Arzadun, P. (2019). Responsabilidad social cooperativa en el sector de ahorro y crédito de costa rica. Mediación de la reputación, credibilidad y percepción en la lealtad de los asociados. *Boletín de Estudios Económicos*, *74*(228), 553–578.

Hair Jr., J. F., M. Hult, G. T., M. Ringle, C., Sarstedt, M., Castillo Apraiz, J., Cepeda Carrión, G. A., & Roldán, J. L. (2019). *Manual de Partial Least Squares Structural Equation Modeling (PLS-SEM) (Segunda Edición)* (2nd ed.). OmniaScience. https://doi.org/10.3926/oss.37

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Jamieson, D., Awolowo, I. F., Garrow, N., Winfield, J., & Bhaiyat, F. (2019). Financial shenanigans: The importance of anti-fraud education. *Journal of Governance and Regulation*, *8*(3), 58–63. https://doi.org/10.22495/jgr_v8_i3_p5

Jawad, A. I., Parvin, T., & Hosain, M. S. (2022). Intention to adopt mobile-based online payment platforms in three Asian countries: an application of the extended Technology Acceptance Model. *Journal of Contemporary Marketing Science*, *5*(1), 92–113. https://doi.org/10.1108/jcmars-08-2021-0030

Juhandi, N., Zuhri, S., Fahlevi, M., Noviantoro, R., Nur abdi, M., & Setiadi. (2020). Information Technology and Corporate Governance in Fraud Prevention. *E3S Web of Conferences*, *202*, 16003. https://doi.org/10.1051/e3sconf/202020216003

Kaiser, T., Lusardi, A., Menkhoff, L., & Urban, C. (2022). Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics*, *145*(2), 255–272. https://doi.org/10.1016/j.jfineco.2021.09.022

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, *62*, 32–43. https://doi.org/10.1016/j.eswa.2016.06.016

Kyriazos, T., & Poga, M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, *13*(03), 404–424. https://doi.org/10.4236/ojs.2023.133020

Liao, L., Chen, G., & Zheng, D. (2019). Corporate social responsibility and financial fraud: evidence from China. *Accounting & Finance*, *59*(5), 3133–3169. https://doi.org/10.1111/acfi.12572

Marabad, S. (2021). Credit Card Fraud Detection using Machine Learning. *ASIAN JOURNAL OF CONVERGENCE IN TECHNOLOGY*, *7*(2), 121–127. https://doi.org/10.33130/AJCT.2021v07i02.023

Mungaray, A., Gonzalez, N., & Osorio, G. (2021). Financial education and its effect on income in Mexico. *Problemas Del Desarrollo*, *52*(205), 55–78. https://doi.org/10.22201/iiec.20078951e.2021.205.69709

Murrar, F. (2022). Fraud schemes during COVID-19: a comparison from FATF countries. *Journal of Financial Crime*, *29*(2), 533–540. https://doi.org/10.1108/JFC-09-2021-0203

Nan, D., Kim, Y., Park, M. H., & Kim, J. H. (2020). What motivates users to keep using social mobile payments? *Sustainability (Switzerland)*, *12*(17). https://doi.org/10.3390/SU12176878

OCDE/CAF. (2020). *Estrategias nacionales de inclusión y educación financiera en América Latina y el Caribe: retos de implementación*. https://www.oecd.org/finance/financial-education/Estrategias-nacionales-de-inclusion-y-educacion-financiera-en-America-Latina-y-el-Caribe.pdf

Oviedo, H. C., & Arias, A. C. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista Colombiana de Psiquiatría*, *34*(4), 572–580. http://www.redalyc.org/articulo.oa?id=80634409

Phonthanukitithaworn, C., Sellitto, C., & Fong, M. W. L. (2016). An investigation of mobile payment (m-payment) services in Thailand. *Asia-pacific journal of business administration*, *8*(1), 37–54. https://doi.org/10.1108/APJBA-10-2014-0119

Sarstedt, M., Ringle, C. M., & Hair, J. F. (2017). Partial Least Squares Structural Equation Modeling. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research* (2nd ed., Issue September, pp. 1–40). Springer International Publishing. https://doi.org/10.1007/978-3-319-05542-8_15-1

Sharma, H., & Aggarwal, A. G. (2019). Finding determinants of e-commerce success: a PLS-SEM approach. *Journal of Advances in Management Research*, *16*(4), 453–471. https://doi.org/10.1108/JAMR-08-2018-0074

Suresh, A., & Rakesh, R. (2019). Relationship between Consumerism and Business Ethics in India. In *Journal of Marketing and Management* (Vol. 10, Issue 2).

Świecka, B. (2019). A theoretical framework for financial literacy and financial education. In *Financial Literacy and Financial Education: Theory and Survey* (pp. 1–12). De Gruyter. https://doi.org/10.1515/9783110636956-001

ETHICAL CHALLENGES IN AI INTEGRATION: A COMPREHENSIVE REVIEW OF BIAS, PRIVACY, AND ACCOUNTABILITY ISSUES

Mariusz Kubanek, Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology, Poland,

mariusz.kubanek@icis.pcz.pl; sabina.szymoniak@icis.pcz.pl

ABSTRACT

This article addresses the complex ethical challenges posed by the integration of Artificial Intelligence (AI) across various sectors. It emphasizes critical issues such as biases and discrimination inherent in AI algorithms, particularly in areas like criminal justice and healthcare. The review underscores the urgent need for equitable and unbiased AI applications, highlighting the role of transparency and ethical frameworks in ensuring fairness. Additionally, it explores the paramount importance of privacy and data protection in AI, advocating for robust frameworks and the ethical responsibilities of developers to safeguard individual rights. The review also delves into the need for accountability and transparency in AI decision-making processes, stressing the significance of ethical guidelines and user comprehension. The article examines the societal impact of AI, especially its transformative effects on the workforce, emphasizing the necessity for human-centric approaches and policy interventions in adapting to AI-driven changes. It also critically evaluates the ethical complexities of autonomous systems, particularly in military applications, advocating for responsible governance and human oversight. Conclusively, the review calls for continuous ethical reassessment and collaborative efforts among academia, industry, and policymakers to develop responsible AI practices.

KEYWORDS: Artificial Intelligence, ethics, autonomous systems, biomedicine, decision-making processes.

1. INTRODUCTION

The advent of Artificial Intelligence (AI) heralds an era of unparalleled innovation and potential. However, this rapid advancement brings with it a host of ethical challenges, particularly in the realms of bias, discrimination, privacy, accountability, and societal impact. "Ethical Threats Associated with the Application of Artificial Intelligence: A Comprehensive Review" delves into these challenges, offering a thorough exploration of recent scholarly work and critical insights into the ethical landscape shaping AI's integration across various sectors.

This review critically examines the inherent biases and discrimination in AI, particularly in sectors such as criminal justice and healthcare. It sheds light on research that exposes systemic inequalities embedded within AI algorithms, emphasizing the urgent need for equitable and unbiased applications. The review discusses the vital role of transparency and the implementation of ethical frameworks to address these biases, ensuring fairness and integrity in AI's deployment. Furthermore, the review explores the crucial issue of privacy and data protection in AI applications. It highlights the ethical responsibility of developers to safeguard sensitive data and emphasizes the integration of privacy-by-design principles and robust data protection strategies. This is essential in maintaining ethical standards in the collection, storage, and processing of personal data across various domains, including education, banking, and healthcare. The need for accountability and transparency in AI decision-making forms another cornerstone of this review. It underscores the importance of understandable decision-making mechanisms and ethical guidelines in AI systems to foster user trust and mitigate the opacity of AI algorithms. The review advocates for continual efforts to enhance user comprehension and societal transparency, holding AI systems accountable for their decisions. In addressing the societal impact of AI, particularly on the workforce, the review acknowledges both the opportunities

and risks brought about by AI-driven automation. It emphasizes the necessity for a human-centric approach in AI development, advocating for systems that complement rather than replace human capabilities. The review calls for comprehensive policy interventions and education programs to navigate the transformative impact of AI on employment and socio-economic structures.

Finally, the ethical dilemmas of autonomous systems, especially in military contexts, are scrutinized. The review emphasizes the need for ethical frameworks and proactive governance to ensure responsible deployment and to mitigate potential risks. It highlights the importance of human oversight and adherence to international laws in AI-driven military operations. In conclusion, "Ethical Threats Associated with the Application of Artificial Intelligence: A Comprehensive Review" offers a critical and comprehensive examination of the ethical challenges in AI. It underscores the need for collaborative efforts across academia, industry, and policymakers to prioritize ethical deployment, address biases, enhance privacy, foster accountability, support workforce adaptation, and govern autonomous systems. This review serves as a clarion call for continuous exploration and adaptation of ethical measures to ensure the responsible and ethical integration of AI into societal frameworks.

2. BIAS AND DISCRIMINATION IN AI

Artificial Intelligence (AI) holds immense promise but carries underlying biases perpetuating systemic inequalities, notably evident in crucial sectors like criminal justice and healthcare. Scholarly investigations delve into these biases, unveiling their implications and advocating for ethical rectification. Angwin et al. (2016) laid bare pervasive biases within predictive criminal justice models, exposing disparities across ethnic and racial groups. This research illuminates systemic biases undermining fairness in the justice system, stressing the need for equitable outcomes. Mittelstadt et al. (2016) contributed to the ethical debate surrounding algorithms, emphasizing fairness and accountability. Their work advocates for transparent processes and thorough audits to detect and rectify AI-based biases, ensuring ethical integrity. Wachter et al. (2017) emphasized transparency's critical role in AI, particularly in robotics, advocating for understandable decision-making mechanisms. Jobin et al. (2020) explored AI ethics guidelines in biomedicine, underscoring the need for comprehensive frameworks aligned with ethical principles.

Moreover, studies by Angwin et al. (2022), Osoba et al., Mehrabi et al. (2021), Castro (2019), and others unveiled biases against black individuals within AI algorithms used in criminal justice. These studies expose systemic biases, false identifications of future criminals, and inequitable outcomes within machine learning analyses, emphasizing the urgent necessity to address biases within AI systems, particularly in criminal justice settings. The research collective underscores the imperative of ensuring fair and unbiased AI applications to guarantee equitable outcomes, irrespective of race or ethnicity. Silva et al. (2018) highlighted challenges in discerning biased decisions within machine learning, particularly noting weighted expectations of higher recidivism rates among black individuals. Dressel et al. (2018) identified software biased against black individuals in predicting recidivism, significantly impacting bail and sentencing decisions. Howard et al. (2018) underscored biases ingrained in machine systems resulting in unfair discrimination against African Americans. Jain et al. (2019) addressed bias in models predicting prisoner recidivism, specifically highlighting bias against black individuals. Furthermore, Malek (2022) discussed the risk of biased decisions in AI, noting the preference for machine bias over human bias in certain scenarios. Završnik (2021) questioned societal preferences concerning human versus machine bias, especially within the context of criminal justice. Cowgill et al. (2019) shed light on the opaque nature of machine learning and its potential economic biases. Milaninia (2020) discussed biases in law implications, notably when judges rule on defendants of the same race.

These studies collectively stress the urgency of addressing biases deeply embedded in AI systems. They advocate for transparency, accountability, and implementation of comprehensive ethical frameworks to rectify systemic biases in AI, particularly within critical domains like criminal justice. By rectifying these biases, the aim is to ensure fairness, equity, and ethical integrity in the application of AI technologies, especially in decision-making processes significantly impacting individuals' lives.

3. PRIVACY AND DATA PROTECTION IN AI APPLICATIONS

Privacy and data protection stand as paramount ethical concerns in the integration of Artificial Intelligence (AI) across diverse domains. A multitude of scholarly works has emphasized the critical necessity for robust frameworks and ethical practices to preserve individual privacy rights within the realm of AI. Scholars like Jobin et al. (2019) stress the urgency of implementing privacy-preserving techniques and data anonymization methods in AI applications. Their work underscores the ethical responsibility of AI developers to ensure privacy safeguards in the collection, storage, and processing of sensitive data. Floridi et al. (2018) advocate embedding privacy as a foundational value throughout the lifecycle of AI technologies. Their ethical framework promotes integrating privacy by design principles and encryption techniques, prioritizing preemptive measures for the ethical handling of personal data. Addressing the complexities of privacy in AI, Dignum et al. (2021) explore the trade-offs between privacy and AI advancement. They highlight the need for adaptive regulations that balance the benefits of AI with individual privacy rights. Additionally, Floridi et al. (2018) propose an ethical framework that integrates privacy as a pivotal pillar, advocating for robust data protection strategies and ethical norms in technological design.

Further research by Mazurek et al. (2019), Kuner et al. (2018), Lai et al. (2019), and Mühlhoff (2023) emphasizes the intersection of AI applications with data privacy, urging the analysis of predictive AI models' impact on data protection and societal inequalities. Thapa et al. (2021) delve into complexities concerning data security in precision health AI applications. Efforts to preserve privacy in AI systems include federated learning (Cheng et al., 2020) and Blockchain integration (Duy et al., 2020), enabling cross-enterprise AI applications while adhering to data protection laws.

The literature extends its exploration to AI's practical implications in education, banking, femtech applications, and medical imaging, emphasizing the importance of safeguarding user data across diverse domains. Ethical considerations and the implementation of comprehensive frameworks are advocated by Stahl et al. (2018) and Timan et al. (2021) to navigate AI's ethical challenges and data protection. Studies consistently underscore the pressing need to reconcile data protection laws with the evolving AI landscape. They emphasize the necessity for ethical frameworks, comprehensive regulations, and innovative approaches to ensure privacy and data security in the realm of artificial intelligence. Moreover, studies like Van den Hoven van Genderen (2017) and Sebastian (2023) delve into specific aspects of data protection, examining implications on AI processes and proposing strategies for securing user information, reinforcing the importance of transparency and accountability in AI systems.

The literature collectively stresses the continuous reassessment of privacy-preserving technologies and regulatory measures to navigate the intricate landscape of AI-related privacy, ensuring the ethical and secure use of data across diverse AI applications. Expanding on privacy considerations, various studies examine the intricate challenges and evolving solutions within AI. Federated learning's role in fintech applications (Dash et al., 2022) emphasizes the need for balancing AI advancements with data privacy and security for diverse user groups. In healthcare, particularly medical imaging, Kaissis et al. (2020) delve into privacy-preserving machine learning techniques, bridging the gap between data

protection and advanced diagnostics. Gerke et al. (2022) scrutinize direct-to-consumer AI/ML health apps, evaluating their implications on consumer privacy under potential US federal privacy laws, drawing parallels with stringent EU regulations.

These studies collectively illuminate multifaceted challenges and evolving solutions regarding data privacy in AI across diverse domains. They emphasize the necessity for adaptable regulatory frameworks, innovative technological solutions, and an ethical compass to navigate the intricate landscape of AI while safeguarding individual privacy rights. The evolving nature of AI necessitates continuous reevaluation and adaptation of privacy-preserving technologies and regulatory measures to ensure responsible and ethical AI implementation across various fields and user groups.

4. ACCOUNTABILITY AND TRANSPARENCY IN AI DECISION-MAKING

In the realm of AI, the call for transparency and accountability echoes across various scholarly works. Calo (2017) advocates ethical guidelines emphasizing transparency, explaining its role in fostering user trust and understanding AI operations. Bostrom (2018) identifies AI autonomy as a challenge and proposes strategies such as certification and adaptive regulations to enforce accountability. Mittelstadt (2021) underscore the need for clear explanations within AI algorithms, advocating interpretable machine learning to bridge the gap between complexity and user comprehension. Expanding into robotics, Wachter et al. (2017) emphasize understandable decision-making mechanisms in AI systems for human interaction, highlighting the significance of transparent explanations to strengthen user trust. Moreover, Mittelstadt et al. (2016) focus on audits and ethical practices promoting fairness and responsibility within AI systems. This foundational research paves the way for responsible AI innovation. It stresses the criticality of transparency, explainability, and accountability not only for user trust but also for ethical AI deployment. O'Neil et al. (2019) study shifts focus to AI's impact on labor markets, exposing socioeconomic disparities. Several subsequent studies echo the need to unravel the 'black box' of AI decision-making to enhance accountability and legitimacy, aligning with de Fine Licht et al. (2020) and Kim et al. (2020) call for increased transparency.

Efforts to embed transparency in AI face practical complexities, echoed by Felzmann et al. (2020). Challenges in algorithmic decision-making underscore the necessity for transparency, resonating with De Laat (2018) and Diakopoulos (2020), who highlight barriers impeding accountability. The pursuit of accountability is recurrent throughout the literature. Some studies focus on technological aspects, like Gualdi et al. (2021), while others compare AI transparency to human decision-making standards, as seen in Zerilli et al. (2019) work. Efforts to ensure accountability often involve providing explanations within AI systems, as indicated by Shin (2020) and Ehsan et al. (2021), highlighting the significance of explanations in user perceptions and social transparency. Similarly, Smith (2021) explores the opacity of clinical AI and associated responsibilities and liabilities. The discourse surrounding explainability in AI systems is prominent in various studies, including Doshi-Velez et al. (2017) research, investigating the role of explanations in enhancing accountability.

These studies collectively underline the intricate relationship between transparency, accountability, and explainability in AI decision-making. They underscore the challenges and complexities inherent in ensuring responsible and ethical AI deployment across various domains, necessitating ongoing exploration and adoption of transparency and accountability measures in AI systems. The synthesis of research on AI decision-making consistently underscores the pivotal role of transparency and accountability.

5. SOCIETAL IMPACT ON WORKFORCE

Undoubtedly, research on the impact of artificial intelligence (AI) on the job market and society offers a clear view of the scale of this phenomenon. The rise of AI presents promising opportunities but also carries risks for employment and social equality. Workforce displacement caused by automation leads not only to job loss but also generates a range of socio-economic problems, such as income disparities, decreased job security, or potential social unrest. Particularly, industries relying on routine tasks face the risk of job displacement, eroding stable employment opportunities for a significant portion of the population. Studies such as Russell et al. (2010) underscore the necessity of a human-centric approach in addressing the challenges linked with AI deployment. They advocate for AI systems that complement human capabilities, emphasizing the need to redefine the human-machine relationship. Authors emphasize that AI can enhance human skills rather than replace them outright. Other works, such as Autor et al. (2003) research on labor market polarization due to technological progress, illustrate that technological advancements benefit certain highly skilled occupations while increasing vulnerability for lower-skilled jobs. This polarization deepens existing socio-economic disparities, necessitating tailored interventions to overcome the growing skills gap. Frey et al. (2017) analysis demonstrates that various professions are susceptible to automation driven by AI to different extents. Their conclusions highlight the percentage of jobs at risk due to AI progress. Such insight requires preemptive measures to mitigate the negative impacts on sectors threatened by disintegration.

To tackle these challenges, comprehensive policy interventions are necessary. These encompass the development of robust educational and training programs aimed at elevating worker qualifications to adapt to changing job demands. Additionally, policies that encourage entrepreneurship, job creation in new industries, and income support mechanisms play a crucial role in navigating this transformative period. Collectively, these scientific contributions point to the multifaceted implications of AI for society, workforce dynamics, ethics, and decision-making. They all underline the need for ethical development, workforce skill adaptation, and a deeper understanding of the complex relationships between AI and social structures. All these studies focus on transforming the dynamics of the labor market and the role AI plays in shaping employment, as highlighted by Zhang et al. (2023), Wright et al. (2018), and Whittlestone et al. (2019). These works delve into the complex relationships between AI industries and employment structures, advocating for ethical framing and strategic investments in workforce development. Furthermore, studies by Pereira et al. (2023), Frank et al. (2019), and Makridakis (2017) emphasize the importance of understanding the impact of AI on skills, decision-making, and social structures. They emphasize the necessity of adapting workforce skills, navigating ethical considerations, and understanding the broader societal implications of AI integration.

Ethical issues related to the introduction of AI and its impact on society and employment are a consistent topic of scientific discussions, emphasized by researchers like Susar et al. (2019), Walsh et al. (2019), and Tschang et al. (2021). These studies advocate for responsible AI development, ethical frameworks, and corporate social responsibility to address implications for society and the job market. Additionally, the scientific discourse underscores the need to adapt to changes arising from AI, strengthen privacy protection, and understand socio-economic consequences, as demonstrated in works such as Naidu (2019), Ertemel et al. (2021), and Pavaloiu (2016). They emphasize the importance of considering AI's broad impact on economic systems, social structures, and labor market dynamics, urging for preventive actions and ethical considerations.

In summary, the body of research indicates a significant impact of AI on various societal areas. These studies stress the need for ethical AI development, strategic workforce adaptation, and a comprehensive understanding of the complex relationships between AI and social as well as economic structures, enabling responsible and effective adjustment to AI integration.

6. ETHICAL DILEMMAS OF AUTONOMOUS SYSTEMS

The integration of autonomous systems, particularly in military contexts, introduces intricate ethical complexities demanding comprehensive governance. Russell and Norvig (2016) underline the urgent requirement for ethical frameworks in AI's military applications, emphasizing human oversight to prevent ethical breaches and unintended harm. This assertion highlights the imperative of preserving human agency and accountability in crucial decision-making processes. Expanding on this quandary, Helkala et al. (2023) analysis investigates the inherent ethical challenges in AI-powered military operations. Their scrutiny accentuates the need for human responsibility and accountability in AI-driven warfare, advocating for human operators' pivotal roles to avert unforeseen consequences. Arkin (2010) work adds depth by emphasizing ethical governance architectures for autonomous weapon systems, ensuring adherence to international laws. This proactive integration of ethical norms into system design prevents inadvertent harm, aligning with ethical standards. Winfield et al. (2014) outline fundamental principles guiding ethical design for robots and autonomous systems, emphasizing the significance of proactive ethical foresight in averting potential ethical quandaries.

The evolving landscape of autonomous systems and AI technologies presents a complex array of ethical challenges, as illuminated by Charisi et al. (2017). They spotlight the potential for AI misuse, shedding light on moral dilemmas within society. Meanwhile, Winfield et al. (2019) emphasize the assumption that autonomous systems, especially driverless cars, should navigate ethical dilemmas. Leikas et al. (2019) propose an ethical framework for designing intelligent systems, addressing ethical issues from their inception to their practical use. Ethical autonomy, as explored by Mohammed (2023), investigates the potential threat AI poses to human autonomy, especially in autonomous systems. McDermid et al. (2019), and Bodenschatz et al. (2021), navigate ethical challenges across various autonomous systems, from self-driving cars to weaponized robots, uncovering moral dilemmas inherent in their operation. Advancements in human-centered AI, as highlighted by He et al. (2021), reveal ethical, societal, regulatory, and educational challenges, urging a deeper understanding of ethical AI deployment. Gill (2021) discusses AI ethics in human judgment, while Werkhoven et al. (2018) stress the necessity of AI systems performing as intended. Anderson et al. (2021) aim to determine ethical principles to resolve dilemmas, while Nichols et al. (2020) address industry-specific ethical and moral issues related to autonomous systems and AI. Dignum (2017) advocates for responsible AI development to tackle ethical dilemmas, and Reuel et al. (2022) propose adaptive stress testing for identifying ethical paths in autonomous systems. Martinho et al. (2021) compile ethical issues within the autonomous vehicles industry, considering various guidelines. Studies by Wang et al. (2020), Etienne (2021), and others explore ethical decision-making in autonomous vehicles and emerging technologies, advocating for ethical AI methods to mimic human ethics while addressing privacy concerns.

Collectively, these studies underscore the critical need for ethical frameworks and responsible development in the realm of autonomous systems and AI. They emphasize intricate ethical considerations and decision-making paradigms that define their integration into societal structures. The multifaceted nature of these challenges necessitates proactive ethical foresight, human oversight, and robust governance mechanisms to ensure their ethical deployment and mitigate potential risks.

7. CONCLUSION

The comprehensive review delves into the ethical intricacies accompanying the rise of Artificial Intelligence (AI) in various sectors, highlighting the imperative for a nuanced understanding and proactive measures to address these ethical quandaries. It traces recent scholarly contributions from 2020 onwards, offering a holistic view of the evolving ethical landscape surrounding AI applications.

The exploration within this review encompasses diverse themes, unraveling the multifaceted facets of ethical concerns inherent in AI integration. It illuminates pivotal issues such as bias and discrimination embedded within AI algorithms, privacy infringements, the need for accountability and transparency in AI decision-making, societal impacts of AI-driven automation on the workforce, and the ethical implications of autonomous systems, particularly in military contexts. Studies across various sectors and disciplines emphasize the urgency of rectifying biases within AI systems, particularly evident in criminal justice settings. These works underscore the imperative of ensuring fair and unbiased AI applications to guarantee equitable outcomes, irrespective of race or ethnicity. Efforts advocate for transparent AI processes, thorough audits, and comprehensive ethical frameworks to rectify systemic biases and ensure ethical integrity across various domains.

The review extends its focus to privacy and data protection concerns in AI applications, stressing the ethical responsibility of developers to safeguard sensitive data. Scholars emphasize the integration of privacy by design principles and encryption techniques to ensure ethical handling of personal data. Discussions on AI's practical implications in education, banking, healthcare, and beyond highlight the importance of user data protection across diverse domains. Moreover, the discourse surrounding accountability and transparency in AI decision-making resonates throughout the literature. Scholars advocate for understandable decision-making mechanisms, ethical guidelines, and explanations within AI systems to foster user trust and mitigate opacity surrounding AI algorithms. Efforts aim to enhance user comprehension and societal transparency while holding AI systems accountable for their decisions. The impact of AI on the workforce unveils promising opportunities but also raises concerns regarding job displacement and socio-economic disparities. Studies emphasize the need for a human-centric approach, redefining the human-machine relationship to complement human capabilities rather than replacing them. Efforts highlight the necessity of policy interventions, robust educational programs, and income support mechanisms to navigate the transformative impact of AI on the job market.

The integration of autonomous systems, particularly in military contexts, introduces complex ethical challenges necessitating comprehensive governance. Works underscore the importance of ethical frameworks, human oversight, and proactive governance to prevent unintended harm and ensure adherence to international laws. Collectively, these studies underscore the intricate relationship between AI and ethical considerations, urging responsible AI development, ethical frameworks, and proactive governance across diverse domains. The evolving nature of AI demands continuous exploration and adaptation of ethical measures to ensure its responsible and ethical integration into societal frameworks. These discussions echo beyond academia, emphasizing the real-world implications and calling for collaborative efforts from policymakers, industry leaders, and AI developers to prioritize ethical deployment for societal well-being.

In summary, future work in the realm of AI ethics must prioritize collaborative efforts across academia, industry, and policymakers. It should focus on addressing biases, enhancing privacy, fostering accountability, supporting workforce adaptation, and governing autonomous systems. Continuous exploration and adaptation of ethical measures are imperative to ensure responsible and ethical integration of AI into societal frameworks.

REFERENCES

Anderson, S. L., & Anderson, M. (2021). AI and ethics. AI and Ethics, 1, pp. 27-31.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. Ethics of data and analytics, Auerbach Publications, pp. 254-264. Retrieved from https://www.taylorfrancis.com/books/edit/10.1201/9781003278290/ethics-data-analytics-kirsten-martin

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. Journal of Military Ethics, 9(4), pp. 332-341.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly journal of economics, 118(4), pp. 1279-1333.

Bodenschatz, A., Uhl, M., & Walkowitz, G. (2021). Autonomous systems in ethical dilemmas: Attitudes toward randomization. Computers in Human Behavior Reports, 4, 100145.

Bostrom, N. (2018). Strategic implications of openness in AI development. In Artificial Intelligence Safety and Security, pp. 145-164. Chapman and Hall/CRC.

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. UCDL Rev., 51, 399.

Castro-Toledo, F. J., Miró-Llinares, F., & Aguerri, J. C. (2023). Data-Driven Criminal Justice in the age of algorithms: epistemic challenges and practical implications. Criminal Law Forum, pp. 1-22. Dordrecht: Springer Netherlands.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., ... & Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint arXiv:1703.04741.

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., & Yang, Q. (2021). Secureboost: A lossless federated learning framework. IEEE Intelligent Systems, 36(6), pp. 87-98.

Cowgill, B., & Tucker, C. E. (2019). Economics, fairness and algorithmic bias. preparation for: Journal of Economic Perspectives.

Dash, B., Sharma, P., & Ali, A. (2022). Federated learning for privacy-preserving: A review of PII data analysis in Fintech. International Journal of Software Engineering & Applications (IJSEA), 13(4).

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. AI & society, 35, pp. 917-926.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability. Philosophy & technology, 31(4), pp. 525-541.

Diakopoulos, N. (2020). Accountability, Transparency, and Algorithms. The Oxford handbook of ethics of AI, 17(4), 197.

Dignum, V. (2017). Responsible autonomy. arXiv preprint arXiv:1706.02513.

Dignum, V., & Marchiori, E. (2021). Privacy and AI: Tensions and potential trade-offs. Minds and Machines, 31(1), pp. 57-71. https://doi.org/10.1007/s11023-020-09541-7

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.

Duy, P. T., Hien, D. T. T., & Pham, V. H. (2020). A survey on Blockchain-based applications for reforming data protection, privacy and security. arXiv preprint arXiv:2009.00530.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-19.

Ertemel, A. V., Karadayı, T., & Makaritou, P. (2021). Investigating the Socio-Economic Consequences of Artificial Intelligence: A Qualitative Research. Journal of International Trade, Logistics and Law, 7(1), pp. 75-89.

Etienne, H. (2021). The dark side of the 'Moral Machine'and the fallacy of computational ethical decision-making for autonomous vehicles. Law, Innovation and Technology, 13(1), pp. 85-107.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and Engineering Ethics, 26(6), 3333-3361.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Rossi, F. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Atomium. European Institute for Science, Media and Democracy: Brussels, Belgium.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., ... & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. Proceedings of the National Academy of Sciences, 116(14), pp. 6531-6539.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation. Technological forecasting and social change, 114, pp. 254-280.

Gerke, S., & Rezaeikhonakdar, D. (2022). Privacy aspects of direct-to-consumer artificial intelligence/machine learning health apps. Intelligence-Based Medicine, 6, 100061.

Gill, K. S. (2021). Ethical dilemmas: Ned Ludd and the ethical machine. AI & society, 36(3), pp. 669-676.

Gualdi, F., & Cordella, A. (2021). Artificial intelligence and decision-making: The question of accountability.

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. IEEE Transactions on Cognitive and Developmental Systems, 14(4), pp. 1398-1412.

Helkala, K. M., Lucas, G., Barrett, E., & Syse, H. (2023). Ethical challenges in AI-enhanced military operations. Frontiers in Big Data, 6, 1229252.

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics, 24, 1521-1536.

Jain, B., Huber, M., Fegaras, L., & Elmasri, R. A. (2019). Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 599-607.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), pp. 389-399.

Jobin, A., Ienca, M., & Vayena, E. (2020). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2(9), 389-399. https://doi.org/10.1038/s42256-020-0214-1

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), pp. 305-311.

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. Decision Support Systems, 134, 113302.

Kuner, C., Cate, F. H., Lynskey, O., Millard, C., Ni Loideain, N., & Svantesson, D. J. B. (2018). Expanding the artificial intelligence-data protection debate. International Data Privacy Law, 8(4), pp. 289-292.

Lai, S. T., Leu, F. Y., & Lin, J. W. (2019). A banking chatbot security control procedure for protecting user data security and privacy. In Advances on Broadband and Wireless Computing, Communication and Applications: Proceedings of the 13th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2018), pp. 561-571. Springer International Publishing.

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical framework for designing autonomous intelligent systems. Journal of Open Innovation: Technology, Market, and Complexity, 5(1), 18.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, pp. 46-60.

Malek, M. A. (2022). Criminal courts' artificial intelligence: the way it reinforces bias and discrimination. AI and Ethics, 2(1), pp. 233-245.

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. Transport reviews, 41(5), pp. 556-577.

Mazurek, G., & Małagocka, K. (2019). Perception of privacy and data protection in the context of the development of artificial intelligence. Journal of Management Analytics, 6(4), pp. 344-364.

McDermid, J., Müller, V. C., Pipe, T., Porter, Z., & Winfield, A. (2019). Ethical issues for robotics and autonomous systems.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp. 1-35.

Milaninia, N. (2020). Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications. International Review of the Red Cross, 102(913), pp. 199-234.

Mittelstadt, B. (2021). Interpretability and Transparency in Artificial Intelligence. The Oxford Handbook of Digital Ethics (online edn, Oxford Academic, 10 Nov. 2021), https://doi. org/10.1093/oxfordhb/9780198857815.013

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Mohammed, S. (2023). Ethical Dilemmas in Autonomous Systems: Impacts on Human Autonomy in the Era of Artificial Intelligence (Doctoral dissertation, Department of Philosophy, Jahangirnagar University).

Mühlhoff, R. (2023). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. Big Data & Society, 10(1), 20539517231166886.

Naidu, A. (2019). Impact of Artificial Intelligence on Society. Indian Institute of Science, 1-13.

Nichols, R. K., Mumm, H. C., Lonstein, W. D., Ryan, J. J., Carter, C., & Hood, J. P. (2020). Counter unmanned aircraft systems technologies and operations. New Prairie Press.

O'Neil, D. S., Chen, W. C., Ayeni, O., Nietz, S., Buccimazza, I., Singh, U., ... & Cubasch, H. (2019). Breast cancer care quality in South Africa's public health system: An evaluation using American Society of Clinical Oncology/National Quality Forum measures. Journal of global oncology, 5, pp. 1-16.

Osoba, O. A., Welser IV, W., & Welser, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Pavaloiu, A. (2016). The impact of artificial intelligence on global trends. Journal of Multidisciplinary Developments, 1(1), pp. 21-37.

Pereira, V., Hadjielias, E., Christofi, M., & Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. Human Resource Management Review, 33(1), 100857.

Reuel, A. K., Koren, M., Corso, A., & Kochenderfer, M. J. (2022). Using Adaptive Stress Testing to Identify Paths to Ethical Dilemmas in Autonomous Systems. In SafeAI@ AAAI.

Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach. London.

Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information. Available at SSRN 4454761.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551.

Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. Phylon (1960-), 55(1 & 2), pp. 9-37.

Smith, H. (2021). Clinical AI: opacity, accountability, responsibility and liability. Ai & Society, 36(2), pp. 535-545.

Stahl, B. C., & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing responsible research and innovation. IEEE Security & Privacy, 16(3), pp. 26-33.

Susar, D., & Aquaro, V. (2019, April). Artificial intelligence: Opportunities and challenges for the public sector. In Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance, pp. 418-426.

Thapa, C., & Camtepe, S. (2021). Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in biology and medicine, 129, 104130.

Timan, T., & Mann, Z. (2021). Data protection in the era of artificial intelligence: trends, existing solutions and recommendations for privacy-preserving technologies. In The Elements of Big Data Value: Foundations of the Research and Innovation Ecosystem, pp. 153-175. Cham: Springer International Publishing.

Tschang, F. T., & Almirall, E. (2021). Artificial intelligence as augmenting automation: Implications for employment. Academy of Management Perspectives, 35(4), pp. 642-659.

Van den Hoven van Genderen, R. (2017). Privacy and data protection in the age of pervasive technologies in AI and robotics. Eur. Data Prot. L. Rev., 3, 338.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science robotics, 2(6), eaan6080.

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. M. (2019). The effective and ethical development of artificial intelligence: an opportunity to improve our wellbeing. Australian Council of Learned Academies.

Wang, H., Huang, Y., Khajepour, A., Cao, D., & Lv, C. (2020). Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller. IEEE transactions on vehicular technology, 69(8), pp. 8164-8175.

Werkhoven, P., Kester, L., & Neerincx, M. (2018, September). Telling autonomous systems what to do. In Proceedings of the 36th European Conference on Cognitive Ergonomics, pp. 1-8.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation.

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. In Advances in Autonomous Robotics Systems: 15th Annual Conference, TAROS 2014, Birmingham, UK, September 1-3, 2014. Proceedings 15, pp. 85-96. Springer International Publishing.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. Proceedings of the IEEE, 107(3), pp. 509-517.

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. Business Horizons, 61(6), pp. 823-832.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. European Journal of criminology, 18(5), pp. 623-642.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard. Philosophy & Technology, 32, pp. 661-683.

Zhang, D., Peng, G., Yao, Y., & Browning, T. R. (2023). Is a college education still enough? The IT-Labor relationship with education level, task routineness, and artificial intelligence. Information Systems Research.

BRINGING ETHICAL VALUES INTO AGILE SOFTWARE ENGINEERING

Olaf Zimmermann, Mirko Stocker, Stefan Kapferer

OST Eastern Switzerland University of Applied Sciences (Switzerland)

itolz@bluewin.ch; stefan.kapferer@ost.ch; mirko.stocker@ost.ch

ABSTRACT

In principle, it is well understood how software engineers should behave; codes for ethics and professional conduct collect principles providing related guidance (ACM (2018)). However, these codes do not translate seamlessly into tangible advice for software engineering routines on development projects, for instance those applying agile principles. Value statements and principles in documents can easily be ignored, e.g., by busy engineers. Conflicts arise in practice, for instance, between public and commercial interests and between stakeholder groups. To improve the situation, we investigate three research questions: 1. How can ethical awareness be stimulated and integrated into agile software practices? 2. How can ethical concerns be actively identified and weighted against other requirements? 3. How can methods and tools trigger, assist, and validate ethical behaviour on agile projects? To answer these three questions, we propose Ethical Software Engineering (ESE) as an active, integrated approach to value-based software engineering advancing the existing passive, retrieval-based state of the art. In this paper, we report on first method engineering results and outline our plans for future work on ESE.

KEYWORDS: Agile Software Development, Design Decisions, IEEE 7000, Moral Values, Normative Ethics, Requirements Engineering, User Stories, Value-Based Systems Design.

1. INTRODUCTION AND BACKGROUND INFORMATION

An ethical value is a “value in the context of human culture that supports a judgment on what is right or wrong” (IEEE 2021). Ethics should concern all project stakeholders, in particular software engineers as initial creators of possibly harmful software. Acting ethically is not a binary, absolute virtue but a multi-faceted, relative, and highly context-depending effort (Ozkaya (2019), Spiekermann (2019)). Stakeholder concerns differ across business sectors, application genres, and organizational units; tradeoffs between entrepreneurial goals and human values must be found (Whittle (2019)).

Professional societies describe the behavior they expect from their members in terms of ethics and professionalism in codes of conduct. The Association for Computing Machinery (ACM), the IEEE Computer Society, and other organizations have issued such codes. To give an example, general principle 1.6 in the ACM code is “respect privacy” and professional responsibility principle 2.9 is “design and implement systems that are robustly and usably secure” (ACM (2018)). It is worth noting that not only engineers but also the software they develop should behave ethically.

Agile practices became popular after the above-mentioned codes of conduct were published; for example, predecessors of the current ACM code (ACM (2018)) were released in 1966, 1972, and 1992. Agile practices bring novel challenges; some of them emphasize early and continuous delivery, which may contradict or hinder careful ethical thinking, planning, and execution (Spiekermann (2019), Gibson et al. (2022)). Certain agile practices, however, might be well-suited to identify potential issues; for instance, having business representatives and end users work with the development team on a daily base reduces the risk of misunderstanding and failing to meet their expectations. Ethics are not mentioned explicitly but touched upon in the “Manifesto for Agile Software Development” from 2001, which is based on four value statements itself; technical excellence is established as one of twelve

principles in the Manifesto. Working software is the primary measure of progress and success, not its ethical properties.[1]

2. CURRENT STATE OF RESEARCH AND PRACTICE

2.1. State of the Art in Academia

Many researchers highlight the relevance of ethics in software engineering and the threats posed by recent developments in related fields such as artificial intelligence, big data, and Web development. An IEEE Software editorial positioned ethics as a "software design concern" (Ozkaya (2019)). Hole (2019) called for five principles: "ensure openness, avoid lock-in, pay for user information, provide multiple solutions with similar services, and combine minds and machines." Safety and privacy as well as robustness have received more attention than other values so far (IEC (2000), GDPR (2016)). Application domains differ in their adoption and maturity w.r.t. these values and qualities; e.g., medical device controllers can be expected to do better than situational apps for leisure and entertainment.

Few research projects address the problem domain from a method engineering or design science point of view; managing ethical values and risks on agile projects has received little attention so far. Issues have been reported (Gregory and Taylor (2013), Dindler (2022)) and the connection between technical debt and ethics has been identified (Gibson et al. (2022)). Economics researchers define digital value systems (Spiekermann (2019), Diethelm and Sennhauser (2019)).

2.2. State of the Practice in Industry

In many countries, ethics education receives increasing attention in computer science and software technology curricula (Dodig-Crnkovic and Feldt (2009)). The Software Engineering Body of Knowledge (SWEBOK)[2] references the ACM and IEEE codes (ACM (2018, IEEE Computer Society (2013)) in its Chapters 1 and 10. While it clearly emphasizes the importance of ethically responsible behavior, it does not provide related adoption and application advice in the form of practices.

The gray literature raises awareness. An online article points at general ethics decision making guidelines,[3] an industry thought leader points out that software engineers are "not just code monkeys" (according to M. Fowler in his OOP 2014 keynote)[4], and practitioners launch initiatives to collect and share more concrete guidance, for instance Code:Ethics in the United Kingdom.[5] An example of valid but rather generic and abstract advice to practitioners is to focus on service delivery quality (of people) and "act with integrity" and value "respect, trust, responsibility" (Hall (2009)).

The IEEE standard 7000-2021, "Standard Model Process for Addressing Ethical Concerns during System Design", defines five analysis and design processes to support this advice; it also suggests (but does not norm) an initial value catalog. IEEE 7000-2021, which we refer to as IEEE Std. 7000 from now on, "aims to support organizations in creating ethical value through system design. Creating ethical value is a vision for organizations that recognizes their central role in society as shapers of well-being and carriers of societal progress that benefits humanity. Implementing IEEE Std. 7000 can help them to strengthen their value proposition and avoid value harms. It is applicable to all kinds of products and services" (IEEE (2021)). Key concepts in IEEE Std. 7000 are (in alphabetical order):

[1] https://agilemanifesto.org/

[2] https://www.computer.org/education/bodies-of-knowledge/software-engineering

[3] https://www.infoq.com/articles/ethical-software-engineer

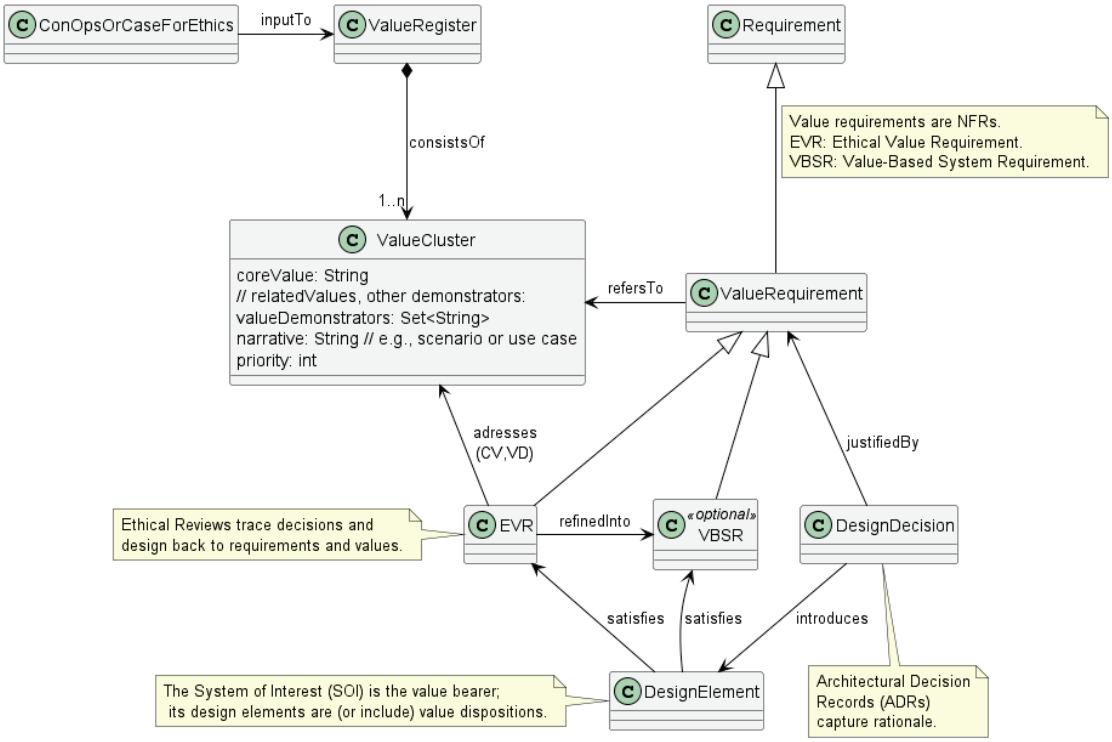
[4] https://martinfowler.com/tags/technical%20leadership.html

[5] https://www.codeethics.org/

- Concept of Operations (ConOps), from ISO/IEC/IEEE15288: 2015: “Verbal and/or graphic statement, in broad outline, of an organization’s assumptions or intent in regard to an operation or series of operations.”
- Ethical Value Requirement (EVR): “Organizational or technical requirement catering to values that stakeholders and conceptual value analysis identified as relevant for the SOI.” (SOI: System of Interest)
- Value: “A conception that influences the selection from available modes, means and ends of action.”
- Value-Based System Requirement (VBSR): “System requirement that is traceable from ethical value requirements, value clusters, and core values.”
- Value Cluster: “Group containing one core value and several values instrumental to, or related to, the core value.” IEEE Std. 7000 defines: “An information store created for transparency and traceability reasons, which contains data and decisions gained in ethical values elicitation and prioritization and traceability into ethical value requirements.” (IEEE (2021))

IEEE Std. 7000 advises how to go from context to value to requirements to design; Figure 1 illustrates its core concepts and their relations (note that ADRs are not mentioned explicitly in it).

Figure 1. IEEE Std. 7000 artifacts addressing ethical concerns (UML class diagram).



Source: self-elaboration based on IEEE Std. 7000 (2021).

2.3. Open problems

The following deficits in the current state of the art must be overcome to advance from opportunistic, passive knowledge sharing to active guidance that respects context and conflicts:

1. Existing knowledge and advice are comprehensive but not always actionable in a given project context and not integrated into methods and tools applied in practice. It must be looked up (*pulled*) and consulted on demand (Berenbach and Broy (2009)).

2. Methods for dealing with the "dual use dilemma" (i.e., most projects and most software can do good, but also be harmful) are missing; it is hard to deal with conflicts such as "the right to privacy vs. the need to protect vulnerable user groups" (Rashid, Weckert, and Lucas (2009)).
3. Tools to promote and ensure ethically responsible, trustworthy behavior of software engineers and the software they produce are missing. It is not clear yet whether ethical behavior can be expected to be understood and demonstrated by software at all; hence, such tools might be desirable but neither theoretically nor practically feasible – and ethically acceptable (Spiekermann (2019)).

We propose to overcome these deficits by integrating ethical values into contemporary agile development routines. We do so in the form of novel and extended/enhanced agile practices.

3. RESULTS OVERVIEW

As explained in Section 2, existing work has focused on creating awareness. It followed a passive, document-oriented approach requiring project teams to pull knowledge and advice from the literature; methods and tools to stimulate ethically responsible behavior are missing. In contrast, we propose to overcome these deficits by integrating ethical values into contemporary agile development routines. We do so in the form of an extended set of agile practices. We contribute an active push approach that makes the elicitation and prioritization of ethical values mandatory, effectively bringing value-based design into development workflows.

Our contributions fall in three categories: knowledge, methods, and tools. Our method, called *Ethical Software Engineering (ESE)*, balances both human values such as fairness and diversity with agile values such as customer collaboration and responding to change. We inject value-based ethical engineering in the agile software development mainstream by way of a novel approach to method engineering and tool design. ESE is released publicly via a git repository that renders Markdown pages to HTML; it is available at: https://github.com/ethical-se/ese-practices.

Knowledge. We compiled a set of essential questions to ask on agile development projects, derived and distilled from existing software engineering codes of ethics and professionalism as well as related sources on value-based software engineering and agile coaching (Agile Alliance (2022), IEEE (2021)). This compilation is disseminated in the form of two novel agile practices called *Story Valuation* and *Ethical Review*; the existing practice of user storytelling is amended with value information complementing the business benefits in the "so that [benefit]" part of the story template (that also has "As a [role]" and "I want to [capability]" parts).

Methods. In line with (Spiekermann (2019)), we propose a decision support and tradeoff method for value-based resolution of conflicts between ethical and other design concerns. Our method adopts the process defined in IEEE 7000 (IEEE (2021)) and complements them with agile practices, working with its ConOps, value register, Ethical Value Requirement (EVR) and Value-Based System Requirements (VBSRs) artifacts. Existing agile practices for requirement prioritization, project planning, and reflection (e.g., definition of ready, definition of done, retrospective) are updated; we also integrate the existing agile concepts of product backlog, sprint planning, and acceptance testing. Each ethically desired behavior on projects is derived from a) the existing body of knowledge (methods, guidelines, codes of conduct) and b) current project context and requirements. Values and resulting requirements are articulated in several different formats inspired by the agile user story template, including value narratives, value weightings and decision-oriented "context-criteria-options" triples. Such template-based value statements help to raise awareness for ethical concerns and make it harder to behave unethically because they force certain questions. To stimulate ethical thinking even further, we also

envision concrete, actionable conflict resolution advice that leaves professional responsibility with the engineer (where it belongs) but moderates the decision-making process.

Tools. We experimented with a demonstrator of a continuous ethics linter as a first tool that actively places ethical awareness in the development mainstream. This tool looks for ethical smells (i.e., suspects that a value might be harmed), inspecting source code and supplemental artifacts in project repositories. A first, basic, text-based prototype of such a linter tool showed technical feasibility but also unveiled ethical concerns; further research is required to set an adequate direction here.

4. SELECTED METHOD AND KNOWLEDGE RESULTS

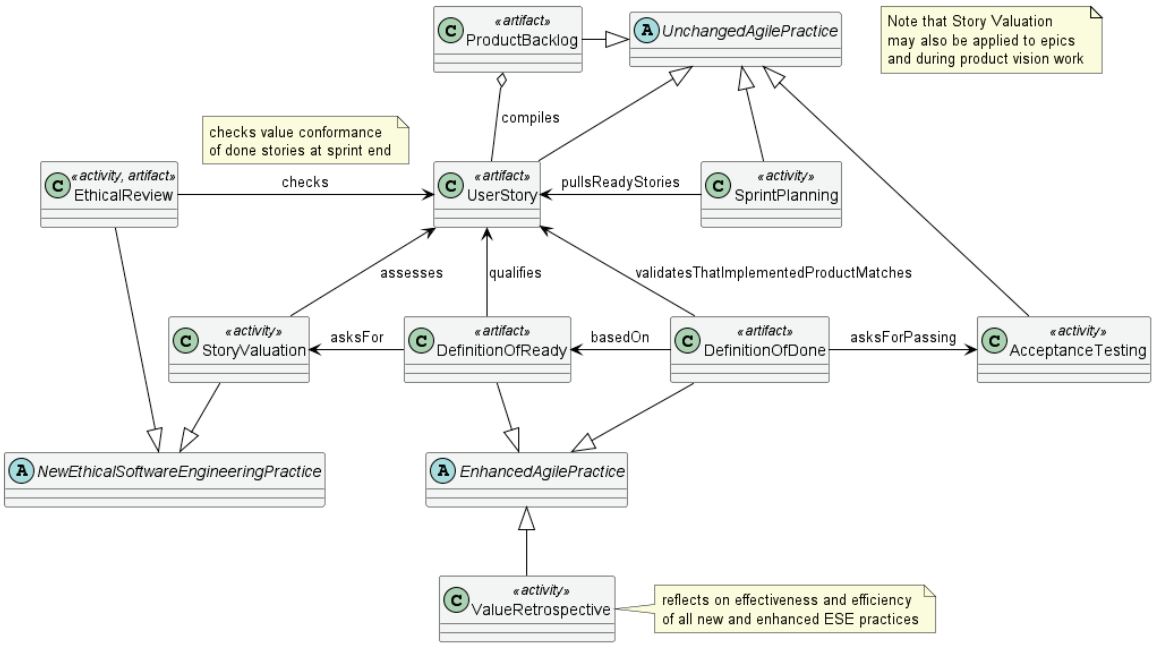
In our method design, we use the term *practices* as a generalization of artifacts and activities here; artifacts serve as input to and output from activities. Our roles originate from IEEE Std. 7000 (IEEE (2021)) and are presented in a rather short/dense/terse form in Release 1.0 of ESE.[6]

The remainder of this section is structured in the following way: Section 4.1 gives an overview of ESE. Section 4.2 maps its agile practices to the processes in IEEE Std. 7000 (and back); Section 4.3 covers Story Valuation, Section 4.4 Ethical Review. Section 4.5 lists additional repository content.

4.1. Practices Overview

ESE includes a total of nine agile activities and artifacts: (new) *Story Valuation* activity, the entry point for ESE usage and (new) *Ethical Review* activity and artifact; (extended artifact) *Definition of Done* and (extended artifact) *Definition of Ready*; (extended artifact) *Value Retrospective*; (unchanged activity) Acceptance Testing, (unchanged activity) Product Backlog, (unchanged activity) Sprint Planning, (unchanged artifact) User Story. The Agile Alliance Glossary provides compact and concise reference information for the mentioned existing practices (Agile Alliance). Figure 2 gives an overview of our practices and where they come from.

Figure 2. Overview of new, enhanced, unchanged ESE practices (UML class diagram).



[6] https://github.com/ethical-se/ese-practices/tree/v101/roles

4.2. Mappings between Agile Practices and IEEE Std. 7000

Table 1 shows how agile teams can integrate concepts from IEEE Std. 7000 into their practices, including those featured in the previous Section 4.1. The mapping is the result of a literature research and reflection of own practices by one author of this paper; the other two authors then reviewed and commented on their own experiences, which led to two incremental refinements of the mapping.

Table 4. Mapping from Agile Practices to IEEE Std. 7000 concepts.

Agile Practice	Related Concept in IEEE Std. 7000	Comments
User Story (as Product Backlog Item)	Functional system requirement	see ESE FAQs[7]
Sprint Planning	no direct mapping	parts of all processes executed in each iteration
Definition of Ready	EVRs elicited (Clause 9)	V in INVEST widened
Definition of Done	Design artifacts produced and reviewed (Clause 10)	from Scrum, many templates/criteria
Sprint Review	Verification and validation activities in all processes	Novel practice in ESE: Ethical Review
Retrospective	no direct mapping, contributes to Transparency Management Process	many variations

Source: own presentment (ESE repository, Background Information page)

Table 2 maps in the opposite way and suggests agile practices to adopters of IEEE Std. 7000. It came to be in the same way as Table 1.

Table 2. Mapping from IEEE Std. 7000 processes to Agile practices.

IEEE Std. 7000	Related Agile Practice	Comments
Clause 7: Concept of Operations (ConOps)1 and Context Exploration Process 2	no direct pendant; epics as input	Concepts in HCI/UX community such as personas, Context Diagram in DPR
Clause 8: Ethical Values Elicitation and Prioritization Process	User Stories and related practices (e.g., backlog refinement/grooming)	Novel practice in ESE: Story Valuation
Clause 9: Ethical Requirements Definition Process	User Stories and related practices	Mapping, splitting; estimation/planning games such as planning poker
Clause 10: Ethical Risk-Based Design Process	n/a (implicit, evolutionary/emerging)	Related literature: “Domain-Driven Design” (E. Evans), “Just Enough Architecture” (G. Fairbanks)
Clause 11: Transparency Management Process	no direct mapping	Ethical Review Meeting/Report, in ESE design decision logs from DPR

Source: own presentment (adopted from ESE repository).

[7] https://github.com/ethical-se/ese-practices/blob/v101/ESE-FAQ.md

4.3. Story Valuation: Notations and Techniques

We introduce this practice in the form of a user story, the agile practice for requirements engineering and iteration scoping (rationale: user stories are also suited in method engineering):

As a responsible software engineer, I want to craft working software that delivers value to users and other stakeholders while not harming any individuals, society and/or the planet. I also want to identify goal conflicts so that adequate trade-offs can be found.

We now provide usage instructions, discuss notations briefly and list techniques to perform the activity. The online ESE method repository also specifies input and output, covers notation in depth, describes the techniques in detail and provides examples and pointers to the literature.

Instructions. The practice description in Release 1.0 of ESE advises to: "Add individual, societal, and environmental values to the business and user values in the "so that" part of epics, user stories or other types of product backlog items. Do so from the perspective of different stakeholder groups; compare and prioritize their value clusters and derive value requirements from them. Start this activity in Product Vision (or Sprint 0 or Minimum Viable Product development); return to it and resume valuation in each sprint/iteration as/if needed. Apply one of the techniques in ESE to do so and record your results in one of the proposed notations; alternatively, work with your own (or other recognized) techniques and notations; ESE is suggestive and not normative in this regard."

Notations. ESE does not mandate a certain format for the IEEE Std. 7000 Value Register; overview figures and comparison tables can be well suited. That said, ESE still suggests three novel formats: Value Epic, Value Weighting and Value Narrative; see ESE repository for notation templates as well as examples.[8] EVRs may take this form:

> As a [role]
>
> I want to [action/feature]
>
> so that [benefit] is achieved
>
> and that [values a, b, c] are promoted,
>
> accepting that [values x, y, z] are reduced.

The first three clauses ("As-a", "I want to", "so that") are well-established convention/template in the Agile community specifying role, feature, and benefit of the story; the new clauses "and that" and "accepting that" add positive and negative values and other ethical consequences of an implementation of the feature described by the story.

Techniques. The valuation techniques proposed in ESE are a) *Goals and Vision First*: Question-Based Elicitation, b) *User Requirements First*: Story-Driven Value Jam and c) *Individual Values First*: Catalog-Guided Value Mapping.

When applying technique a), Question-Based Value Elicitation (Goals and Vision First), you may want to ask the following questions when developing and hardening the product vision to identify ethical values and elicit EVRs/VBSRs:

- "How does the system (product, service) under construction make the world a better place?
- Even if there were no explicit, external stakeholder goals and business drivers, why is it still a good idea to develop the system? Which positive ethical values does it promote?

[7] https://github.com/ethical-se/ese-practices/blob/v101/practices/ESE-StoryValuation.md#notations

- Which positive ethical values are degraded by any realizations of the envisioned functionality? Which negative values are promoted? What are the related elements of risk, those with high probability and huge impact in particular?
- How do positive and negative values, as well as benefits and harms, relate to each other? What is their relative and/or absolute weight?
- Which resources will the system consume, and can this consumption be justified by the business/user and ethical values that it delivers?"

The ESE technique b), User Requirements First (Story-Driven Value Jam), asks the following value-related questions:

- "What are the individual and collective values of the persona/role in the "As a" part of the story (aka stakeholder groups)?
- Which responsibilities do their user interfaces have that may promote or degrade ethical values (both positive and negative ones)?
- Which ethical values are affected when executing the program code realizes the "I want to" part of the story?
- What is the desired and, presumably, actual impact (good and bad) of the "so that" part of the story on individuals, society, and planet?"

It also asks the following more technical questions:

- "Which values are affected positively or negatively when end user input is received and validated, and when computation and query output is displayed?
- How does the data processing (application/business/domain logic) do w.r.t values?
- How do data access components and data storage (persistence mechanisms) behave w.r.t. values?
- What is the ethical risk of APIs and other transport channels, as well as shared services such as loggers?"

Technique c), Individual Values First (Catalog-Guided Value Mapping), works with Appendix G of IEEE Std. 7000. It suggests the following three steps:

1. "Pick 2-3 core values from the table. Use the information in the columns Related value and Opposing value in Table G.1 to make them more complete, concrete, and tangible.
2. Explain the relevance of each core value in the given project/product context by way of example and/or refinement. You might want to tell a story and/or point out the positive and negative consequences of this value in the form of a narrative or demonstrator (see column in table above). This information may come from the Value Register of the project/product development effort when IEEE Std. 7000 is followed (if no Value Register exists yet, it is a good time to create one now).
3. Prioritize the value(s), either absolutely or relatively. For instance, you may want to use a writing style akin to that used in the Agile Manifesto ('We value mm over nn')."

4.4. Ethical Review (Meeting and Report)

We follow a similar presentation structure here as in the previous section.

An Ethical Review Report captures the results from an ethical values-enhanced (or -centric) sprint/project Ethical Review Meeting. It may include mere meeting notes and/or an assessment with follow up actions (aka recommendations and findings).

Instructions. The practice description in Release 1.0 of ESE advises to: "Inspect the user story implementations with regards to the prioritized Value Cluster demonstrators/narratives and Value Requirements from Story Valuation in a meeting. Capture the meeting outcome in a report.

Record answers to the following questions during the review meeting:

- Have the Ethical Value Requirements (EVRs) been included in the acceptance testing and did these tests pass?
- Has any feature been introduced that stands in conflict with the EVRs as well as the team's values and personal beliefs of each team member? If yes, have tradeoffs and mitigation tactics been discussed?
- Would all team members use the new features themselves, and let their closest relatives use them (assuming that these individuals are in the target audience of the system under construction)?
- Has the behavior of the team and its members been in line with the eligible Codes of Conduct (CoCs), for instance the ACM Code of Ethics and Professional Conduct? Does it respect the values of the involved organizations (sponsor, clients, business partners)?
- Do any of the IEEE Std. and ESE artifacts (i.e., ConOps, Value Register, EVRs) require further updates, caused by the review results?

There are no easy answers to many of these questions typically; this is inherent, discussing them (as well as follow-on questions) is as important as the outcome of this activity to create awareness, for instance when planning the next sprint/iteration."

Notations. The review results can be recorded in free form in a new text document or added to the reviewed artifacts.

4.5. Other Repository Content

The ESE repository also contains Markdown templates for important artifacts specified in IEEE Std. 7000 (e.g., Case for Ethics), application hints per practice, and pointers to the literature including standards and books (e.g., on material ethics).

5. VALIDATION

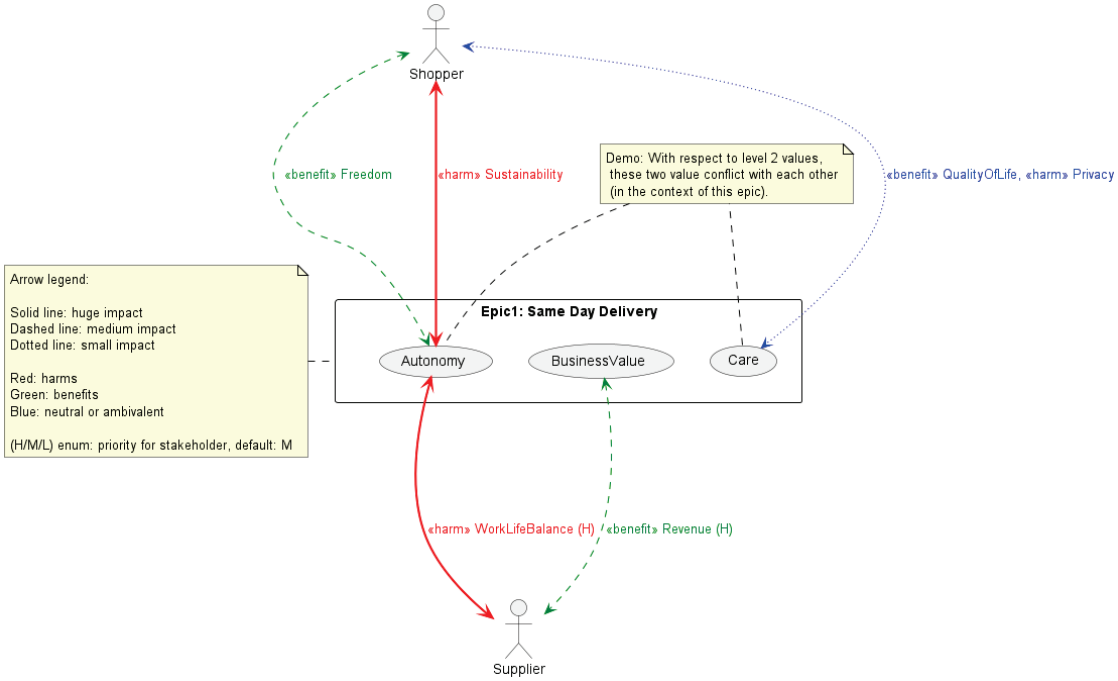
We validated our method engineering results in action research and self-experiments so far. Two of the authors of the paper that did not author the first version of the method content, reviewed an intermediate version of it and applied it to two of their own real-world projects. This early feedback led to a substantial revision of the entire method content and the introduction of two additional Story Valuation techniques. The three authors then designed a fictitious but realistic "same day delivery" product vision for an online shop and applied the revised second version of the method content in a joint half-day workshop. General feedback on readability, applicability, and usefulness was positive; scalability, time management, dealing with conflicting stakeholder interests, and visualization challenges were identified and partially addressed (see Section 7 for future work about addressing these challenges). Reviewers commented that more examples would be welcome, on all levels of

analysis and design refinement (i.e., values, value requirements, and architectural decisions addressing these requirements) to make the method even more accessible and useful.

An external reviewer who contributed to the standard appreciated the diagrams and the method organization but requested some terminology clarification so that utilitarian values (harms and benefits) are not confused with positive and negative values from any ethics school (and their consequences). The reviewer suggested to not only focus on stories and epics but also on business-level ideation activities such as product envisioning on Scrum; we also discussed whether the Value Lead role from IEEE Std. 7000 should be taken by a single person or be a shared responsibility of the development team (both positions have pros and cons attached).

Figure 3 shows an example from the validation workshop. Autonomy and care are core values from Annex G of IEEE Std. 7000 (IEEE (2021)); freedom and sustainability, quality of life and privacy are listed as related values in "Table G.1—Typical ethical values for systems design" of the standard.

Figure 3. Values and conflicts in the Same Day Delivery example (UML use case diagram).



Online experiment instructions and review questions are available online. They can be found at https://github.com/ethical-se/ese-practices/tree/main/experimentation. We invite readers to participate.

6. DISCUSSION OF DESIGN DECISIONS

We reviewed the OMG SPEM metamodel for methods and methods engineering (OMG (2008)) as well as other inputs before compiling the practice templates for our Design Practice Repository (Zimmermann and Stocker (2021)), now adopted and further refined for ESE.

Revisiting the research questions and contribution types from Sections 2 and 3, we decided to focus on method engineering primarily and less on tool development because of a) risk-benefit issues and b) importance and complexity of method engineering (early feedback, see below), and c) existence of IEEE Std. 7000 and supporting literature.

We decided not to feature the entire standard due to its size and general-purpose nature (i.e., its usage is not limited to software systems). Furthermore, we made several standards concepts optional (e.g.,

VBSRs) to achieve the goal of being attractive to our target audience, agile software development teams. The Frequently Asked Questions (FAQ) page in the ESE repository provides further information.

7. CONCLUSION AND OUTLOOK

Ethical Software Engineering (ESE) integrates ethical values and IEEE Std. 7000 with agile software development practices. ESE introduces two new practices, extends three existing ones, and reuses four existing ones without changing them. In this paper, we motivated the need for such methods in the state of the art and practice, provided a method overview, featured the two new/novel practices in some more detail, and then discusses early validation feedback as well as critical success factors for a broader adoption. The full ESE method is available at https://github.com/ethical-se/ese-practices, a public open-source repository under the Creative Commons Attribution 4.0 International License.

In our future work, we consider including pre-defined value catalogs and assessments of their relevance w.r.t. project phases and architectural layers (presentation, business logic, data access and storage) into our approach. We also consider developing additional templates and notations, emphasizing usability, scalability, and conflict management in our method engineering. Other directions for future work include tool support for the method and its content, for instance in the form of a questions-answers moderation or conflict visualization and resolution support. Another area that we consider is starting even earlier – evaluating and assessing product visions with respect to their ethical ramifications. Such early start would support "go"-"no go" decisions for software development before it even starts.

ACKNOWLEDGEMENTS

This research was supported by the Hasler-Foundation. Bärbel Bohr answered our many questions about IEEE Std. 7000 and reviewed parts of the ESE repository.

REFERENCES

ACM (2018). "ACM Code of Ethics and Professional Conduct 2018."

ACM, and IEEE Computer Society (2013). Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science. New York, NY, USA: ACM.

Agile Alliance (2022). "Code of Ethical Conduct for Agile Coaching, Version 2.0." https://www.agilealliance.org/agile-coaching-code-of-ethical-conduct/

Agile Alliance. "Agile Glossary" https://www.agilealliance.org/agile101/agile-glossary/

Berenbach, B., and M. Broy (2009). "Professional and Ethical Dilemmas in Software Engineering." Computer 42 (01): 74–80.

Diethelm, C., and P. Sennhauser. (2019). "Digitale Ethik, HWZ Whitepaper."

Dindler, C., P. G. Krogh, K. Tikær, and P. Nørregård (2022). "Engagements and Articulations of Ethics in Design Practice." *International Journal of Design* 16 (2): 47–54.

Dodig-Crnkovic, G., and R. Feldt. (2009). "Professional and Ethical Issues of Software Engineering Curriculum Applied in Swedish Academic Context." In HAoSE 2009 First Workshop on Human Aspects of Software Engineering. Online Proceedings.

GDPR. (2016). "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC." *OJ L 119/1*.

Gibson, J. P., M. Narouwa, D. Gordon, D. O'Sullivan, J. Turner, and M. Collins (2022). "Technical Debt is an Ethical Issue." Proc. of ETHICOMP 2022.

Gregory, P., and K. Taylor. (2013). "Social and Communication Challenges for Agile Software Teams.", Proceedings of ETHICOMP 2013, 186–191.

Hall, D. (2009). "The Ethical Software Engineer." *IEEE Software* 26 (4): 9–10.

Hole, K. J. (2019). "Dominating Software Systems: How to Overcome Online Information Asymmetry." *IEEE Software* 36 (4): 81–87.

IEEE. (2021). "IEEE 7000 Standard Model Process for Addressing Ethical Concerns During System Design."

IEC. (2000). "IEC 61508: Functional safety of electrical/electronic/programmable electronic safety related systems."

OMG. (2008). Software & Systems Process Engineering Metamodel Version 2.0.

Ozkaya, I. (2019). "Ethics Is a Software Design Concern." *IEEE Software* 36 (3): 4–8. https://doi.org/10.1109/MS.2019.2902592

Rashid, A., J. Weckert, and R. Lucas. 2009. "Software Engineering Ethics in a Digital World." Computer 42 (6): 34–41.

Spiekermann, S. (2019). *Digitale Ethik: Ein Wertsystem für das 21. Jahrhundert*. Droemer Verlag.

Spiekermann, S. (2023). *Value-Based Engineering: A Guide to Building Ethical Technology for Humanity*. De Gruyter.

Whittle, J. (2019). "Is Your Software Valueless?" *IEEE Software* 36 (3): 112–115.

Zimmermann, O., and M. Stocker (2021). *Design Practice Repository*, LeanPub.

THE DEMOCRATIZATION OF OUTER SPACE: ON LAW, ETHICS, AND TECHNOLOGY

Eleonora Bassi, Ugo Pagallo

Polytechnic of Turin (Italy)

eleonora.bassi@polito.it; ugo.pagallo@unito.it

ABSTRACT

The paper addresses the challenges brought forth by projects and investments of private companies on mass space exploration, space tourism, and scientific research in outer space missions. Such projects and investments are examined with the case study of the regulatory framework for the Columbus Laboratory in the International Space Station. The aim is to illustrate the limits of traditional approaches in the field of space law, and how ethics and moral arguments help filling the gaps of current legal regulations. The quest for the democratization of outer space casts light on the democratic deficit of such institutions, as the European Union, vis-à-vis current trends on the privatization of outer space.

KEYWORDS: Artificial Intelligence; Democracy; Human Rights; International Space Station; Robotics; Space Law.

1. INTRODUCTION

Over the past decades, scholars have dissected the manifold ways in which private companies have disrupted the legal framework set up with the United Nations (UN) treaties and conventions on outer space from the 1960s and 1970s. The overall architecture of space law as a branch of public international law should be understood in accordance with a state-centric approach that revolves around the powers of sovereign states and the responsibilities and duties they have under the UN legal framework. Whereas Art. VI of the 1967 Outer Space Treaty establishes that states are internationally responsible for national activities in outer space, Art. VII determines the international liability of the states that launch – or procure the launching of – an object in outer space for the damages that such object may provoke to other states or to their natural and legal persons. Further, the 1972 Liability Convention distinguishes two kinds of liability for whether a space object causes damage on the Earth's surface or flying aircrafts (Art. II), or not (Art. III). Liability of Art. II is strict: the launching state shall demonstrate its lack of responsibility due, for example, to the gross negligence of the claimant state under Art. VI of the Convention. Liability of Art. III, vice versa, hinges on the fault of the launching state or of the persons for whom such state is internationally responsible (Dennerley 2018).

Against this framework, scholars have stressed the growing shortcomings of the law in tackling activities of private companies (Deem 1983; Ernest 1990; Ziemblicki and Oralova 2021). The general rule remains that of the UN treaties and conventions, according to which duties and obligations of private companies depend on the responsibilities and accountability of sovereign states. For example, it can be tricky to determine what is the proper jurisdiction when a private company, incorporated in one country, launches a spacecraft in a different country (Freeland and Ireland-Piper 2022). Moreover, some wonder about the level of legal protection that private companies should guarantee in outer space vis-à-vis the safeguard of basic rights of individuals (Freeland and Jakhu 2014; Lim 2020). In addition, there is a persisting thorny set of legal issues that regard whether and to what extent sovereign states can adopt legislations that set up business incentives with the promise of property rights, e.g., the U.S. Space Resource Exploration and Utilization Act from 2015 that establishes the right of U.S. citizens to all the resources they can obtain (Foster 2016).

Another set of legal issues related to the increasing role of private companies in outer space has to do with current promises of mass space exploration, space tourism, or even space settlers. The CNN announced on 2nd May 2022 the first space hotel scheduled to open in 2025.[1] These scenarios of the democratization of outer space – as a result of current trends on privatization – suggest a new generation of tortious claims and contractual issues of space law that do not only affect the kinds of rights to be protected, or the kinds of damage to be covered, vis-à-vis the lack of enforcement mechanisms to hold private actors accountable for every contravention of the law (Isnardi 2019). In our view, the next generation of space explorers and scientists, tourists and settlers will raise a fundamental question of every democracy: the right of individuals to have a say in the decisions affecting them, at least through their institutional representatives, dealing with the activities of private companies (Bobbio 1987). This "broken problem of democracy," according to the jargon of the Italian philosopher, is also traditionally at stake with all debates on the 'democratic deficit' of the European Union (Follesdal 2006). On the one hand, we may wonder about the extent to which market competition related to outer space missions and activities is going to be good enough to deal with the protection of individual rights; on the other hand, "it will not be legally or morally satisfactory that corporations providing such services require the individual participants to sign away their rights through contractual exclusion of liability clauses" (Freeland and Ireland-Piper 2022).

Experts of space law have examined issues of accessibility and legal certainty, equality and fair power, protection and dispute resolution, procedure and compliance, namely, all the elements that the common law tradition conceives of as the fundamental "ingredients" of the rule of law (Bingham 2010). The shortcomings of today's legal framework for outer space activities do not only depend, however, on flaws of traditional categories of public international law and their critical assessment in the field of political philosophy, or moral theory. The quest for the democratization of outer space entails another layer of complexity. The quest shall also be examined in connection with the normative challenges of Artificial Intelligence (AI) and other emerging technologies. Most experts would agree that AI systems and smart robots will increasingly play a crucial role in outer space (Martin and Freeland 2021); yet, a consistent amount of research has eviscerated over the years the multiple ways in which computers, the internet, smartphones and social networks have shaped human societies, and still, work on how AI systems and smart robots may shape today's space race and its privatization is just moving its first steps (Pagallo et al. 2023).

One of the features of AI that has most attracted the attention of scholars and the public at large is the 'autonomy' of such systems and of the robots equipped with them. Autonomy means that "AI systems and smart robots can change their inner states or properties without external stimuli, therefore exerting control over their actions without any direct intervention from humans" (Pagallo 2013). The mounting use of AI systems that augment or replace analysis and decision-making by humans, making sense of huge streams of data, or defining and modifying decision-making rules autonomously, has a price: AI systems can be overused, or misused in space missions, devaluing human skills, removing human responsibility, reducing human control, or eroding human self-determination (Floridi et al. 2018). The interactivity, opacity, and unpredictability of autonomous 'space objects' may challenge space law regulations on whether and to what extent the decisions of these 'objects' fall under the fault of persons for whom a state is liable (Bratu et al. 2021). These liability issues also concern the use of AI systems for space-based services, such as AI systems using Global Navigation Satellite System (GNSS) signals to support emergency-response services, autonomous vehicles, unmanned aircraft systems, and more (Bratu 2021).

[1] See at https://edition.cnn.com/travel/article/space-hotel-orbital-assembly-scn/index.html.

The information revolution and the disruption of AI in outer space since the mid 2010s raise formidable problems. In 2022, the Director of the new heavyweight aerospace contractor SpaceX, i.e., Benji Reed declared: "We want to make life multi-planetary, and that means putting millions of people in space."[2] This scenario entails fascinating problems of political philosophy and legal theory on how to rule millions of people in space, namely, according to the classical tripartition of Aristotle, wondering on what constitution (*politeia*), law (*nomos*), and both a public habit and an individual attitude (*ethos*) shall govern this brave new world. The focus of this paper is restricted to the current legislative efforts of technological regulation in EU law to properly address the challenges of AI systems in outer space. Although most legislative initiatives of the European Commission do not refer specifically to the use of AI systems and robots in outer space, attention should be drawn to fundamentals of space law complemented with current provisions of data protection and privacy, cybersecurity and machinery regulation, down to tortious liability and consumer law. This legal framework should provide the elements of the quest for the democratization of outer space in accordance with the *politeia*, *nomos*, and *ethos* of current EU law.

Accordingly, the paper is divided into four sections. Next, we focus on the dramatic decreasing costs for space missions and spacecrafts as a main driver of the new space economy. In Section 3, the analysis dwells on the legal core of the privatization of outer space, the very appropriability of space resources and its impact on international law. Section 4 investigates how EU law regulates the status of potential outer space temporary inhabitants, as space tourists or explorers, with the case study of the regulatory framework for the Columbus Laboratory in the International Space Station (ISS). On this basis, Section 5 scrutinizes the drawbacks of this regulatory framework concerning current efforts of lawmakers to tackle the normative challenges of AI in such fields, as cybersecurity, machinery safety, consumer law, and more. Drawing on basic tenets of space law, philosophy of technology, ethics, and technological regulation, the conclusions of the investigation stress the relevance of the issue, i.e., the 'democratization of space' and why the speed of technological advancements and innovation together with human ingenuity will increasingly put this topic in the spotlight.

2. LEAVING EARTH PRIVATELY

In its 2018 annual report, the United Nation's office for outer space affairs predicted that space business could generate revenues of 1.1 up to 2.7 trillion dollars by 2040 (UNOOSA 2019, at 38). Around 70% of space activity was already driven by the private sector in the late 2010s, and it is likely that such figures will keep growing over the years. In May 2021, Forbes, the American business magazine, released an article, drawing on a report from SpaceTech Analytics, in which the number of space-focused companies was esteemed over 10 thousand globally. More than half of them – namely, 5582 space-focused companies – had their headquarters in the U.S.A., almost ten times more than the next country, the UK, with 615.[3] Spanning across twenty different business sectors, private companies have disrupted a market traditionally populated by defense contractors. It is noteworthy that some of the most known companies in the field have been launched by four multibillionaires: Elon Musk, Jeff Bezos, Paul Allen, and Richard Branson. Between the early 2000s and 2011, they founded SpaceX, Blue Origin, Stratolaunch, and Virgin Galactic, respectively.

The overall figures of the space economy are staggering. In addition to the 10,000 + companies around the world, Forbes esteems 5,000 significant investors; 150 research and development hubs; and 130

[2] Reed's announcement at https://www.space.com/spacex-launches-inspiration4-civilian-orbital-mission.

[3] See https://www.forbes.com/sites/johnkoetsier/2021/05/22/space-inc-10000-companies-4t-value--and-52-american/?sh=3fee48a755ac.

state or governmental organizations. The largest business sectors regard navigation and mapping (2,820 companies); cloud solutions (2,406); and manufacturing (1,048). Costs are dramatically decreasing. Whereas the orbital delivery costs of NASA's Space Shuttle are $20,000/kg, the costs of SpaceX's Starship are estimated $500/kg. The startup companies of the last two decades have not only shown they can fully compete against the traditional powerful aerospace contractors, such as Lockheed and Boeing, but they are doing so by exponentially reducing the costs of space activities. Current prices for suborbital flights of space tourists – up to $50 million per seat for a full orbit flight – will all but go down.

This crucial role of private actors has been dubbed as the "New Space" (Vernile 2018). From a legal viewpoint, the side effects of this new role of private companies in space-related activities have been dissected through the manifold issues brought forth by the privatization of outer space (Deem 1983; Ernest 1990; Ziemblicki and Oralova 2021; etc.). Such problems include issues of damages and accountability of private companies and their jurisdiction (Freeland and Ireland-Piper 2022), fault and liability (Bratu 2021), and the protection of basic rights of individuals (Freeland and Jakhu 2014, Lim 2020). The crux of such work on the privatization of space has arguably to do with the incentives that many State Parties of the UN outer space treaties and conventions have endorsed with their acts and laws. As shown by the 2015 U.S. Space Act, mentioned above in the introduction, the privatization of space goes hand-in-hand with states that grant the right of their citizens and companies to all the asteroid and celestial resources they can obtain. After the U.S.A., other countries have followed suit, such as Luxembourg with its Law on the Exploration and Use of Space Resources from 2017, or Japan with the Law Concerning the Promotion of Business Activities Related to the Exploration and Development of Space Resources from 2021. To attract business, many jurisdictions include tax exemption of insurance contracts that cover the space objects, and tax credits for investments made by operators of such space objects. It is worth noting that Luxembourg, although one of the smallest countries in Europe, is one of the largest satellite operators in the world.

Scholars have debated the core of these legislations on the appropriability of space resources. As a result of the privatization of outer space, this kind of debate is relevant to put the further quest for the democratization of outer space in proper light. Most discussions have revolved around how to interpret the first two Articles of the UN-sponsored Outer Space Treaty. Whilst Art. I of the 1967 Treaty establishes that exploration and use of outer space should benefit all countries, Art. II repudiates national appropriation over outer space resources. It may seem obvious that all legislations, such as the 2015 U.S. Space Act, the 2017 Luxembourg Law on Space, or the 2021 Japanese Law on Space Resources would blatantly repudiate two pillars of the moral grounds on which public international law lies: the principle of beneficence and that of outer space as a province of mankind, a global commons. Yet, against this debate on whether and to what extent there is a legal contrast between the 'global commons' of outer space in public international law and the proprietary rights of individuals and corporations over celestial resources in national law, the aim of this paper is not to take sides in this kind of debate. Rather, the debate provides the necessary backdrop for the analysis on the most common legal source of regulation in the field of space business, that is, contracts. How such contracts relate to principles and rules of public international law sheds light on the moral and legal challenges triggered by the privatization of outer space and its by-product, that is, the quest for the democratization of outer space.

3. THE TROUBLES WITH INTERNATIONAL LAW

The rush of Luxembourg to space activities and the use of space resources in U.S. or Japanese law have been extensively debated in connection with the claim of such states that space resources are appropriable (Foster 2016; Steele 2021). Scholars have often supported these provisions, asserting

that there is no incompatibility between international law, e.g., Art. I and II of the Outer Space Treaty, and the national claims of legislators because such appropriations, as extracting golden nuggets from an asteroid, would be in line with the wording of the UN treaties (Tepper 2019). Others have stressed that such legislations may draw on a basic principle of old French property law, regarding the separable legal nature of mines or the surface of land (Su 2017). Apart from business incentives with the promise of property rights, the appropriation of resources may in fact be deemed as legitimate according to the classic argument of John Locke, developed in chapter 16 of the Second Treatise of Government, concerning that which humans own, because of their work, efforts, and ingenuity. We should indeed distinguish between the territorial expansion of states as appropriation and the property rights of individuals (Locke 1689).

Remarkably, this kind of argument is at stake with all legislations that establish the right of their citizens to all the space resources they can obtain. Rather than an overt violation of Art. II of the Outer Space Treaty, for example, the U.S. Space Law "should be seen as a valid interpretation of Article II given the numerous ambiguities inherent in the article itself" (Blount and Robison 2016). By distinguishing between the territorial expansion of states and the rights of individuals to possess, own, transport, use, and sell either asteroid resources, or space resources, the U.S. act would not assert any sort of sovereignty over celestial bodies. On the contrary, the distinction would cast light on how the legal status of outer space as a global commons does not preclude the commercial extraction of space resources, as much as occurs in other fields of international law, such as the Law of Sea, or of Antarctica (Feichtner 2019). The very fact that no specific clause of international space law prohibits the extraction of space resources could be interpreted as the legal basis for their legitimate extraction. It is worth noting that the International Institute of Space Law supported this view with a report of Marcia S. Smith in 2015.

These trends of national space law draw the attention to traditional well-known problems of coordination and potential antinomies between public international law and national space law (Gabrynowicz 2010). Although a certain degree of flexibility stems from the open clauses of international space law that may accommodate multiple diverse applications at the national or regional level, some insist that such legislations – as the U.S. 2015 Space Act or the Luxembourg 2017 Law on Space Resources – represent a breaking point in international space law (Tronchetti 2015). Problems of appropriability of space resources, occupation of celestial bodies, and the repudiation of outer space as a global commons call into question the overall sustainability of the field and its environmental impact both in space and on Earth, e.g., the use of kerosene and liquid oxygen powering the rockets of SpaceX's Falcon 9 that inject black carbon into the upper atmosphere (Viikari 2008; Loder 2018; Hoffmann and Bergamasco 2020).

From these shortcomings, however, it does not follow the legal paralysis of the system, nor the impossibility to make business as usual. The amazing figures and statistics illustrated above in Section 2 substantiate this view. Space business and space economy thrive notwithstanding some open issues and drawbacks of today's space law with its moral dilemmas on appropriability, occupation, and sustainability (Jessen 2017). This success entails its own price. The more everyday people are involved in missions and activities on Moon hotels, celestial settlements, or stations for scientific purposes, the less current provisions on the protection of rights and interests of such individuals appear as satisfactory. Such protection includes the right to "a high level of environmental protection and the improvement of the quality of the environment," in the phrasing of Art. 37 of the EU Charter of Fundamental Rights. To support this claim, the analysis of how EU law regulates and disciplines its activities in the International Space Station (ISS) is particularly fruitful: on the one hand, the ISS case study sheds light on how missions and activities are legally regulated today, despite all troubles with the interpretation of current public international space law; on the other hand, the case study helps

determine the reasons why national regulations may fall short in addressing the challenges of the privatization of space, and in perspective, its democratization. The next section is devoted to the illustration of the case study; the subsequent section provides the assessment.

4. THE INTERNATIONAL SPACE STATION: A CASE STUDY

The International Space Station (ISS) was established – and launched – in 1998. The International Space Station Intergovernmental Agreement, or IGA, is an intergovernmental, rather than international agreement, because only fifteen governments were involved in the project. From a scientific viewpoint, ISS represents today's largest modular station in low Earth orbit that functions as a research laboratory for physics and astronomy, astrobiology and meteorology, etc. The ISS also tests the resilience of spacecraft systems and the equipment that will be required in deep space missions. From a legal viewpoint, according to Art. 1 of the IGA, the aim is to create "a long term international cooperative framework on the basis of genuine partnership, for the detailed design, development, operation, and utilization of a permanently inhabited civil Space Station for peaceful purposes, in accordance with international law."

A Memoranda of Understandings involves five space agencies, i.e., in addition to the coordinator, the US Space Administration (NASA), four cooperating space agencies: the European Space Agency (ESA), the Canadian Space Agency (CSA), the Russian Federal Space Agency (Roscosmos), and the Japan Aerospace Exploration Agency (JAXA). Such agencies shall determine responsibilities and their role concerning the design, development operation and effective use of the space station divided into two areas. The condominium, so to speak, regards the Russian Orbital Segment (ROS) with six modules, and the eight modules of the US Orbital Segment (USOS). USOS supports roughly three quarters of its services for NASA, around 10% for JAXA and ESA, and around 2/3% for CSA.

The regulation of this new space of human interaction, i.e., a civil space station permanently inhabited, rests on the principle of national sovereignty. This means that the main partners of the project – the USA, Russia, Japan, Canada, and the European Union (EU) – are not only liable for what they provide for in the project. In addition, in the wording of Art. 5 of the IGA, "each partner shall retain jurisdiction and control over the elements it registers and over personnel in or on the Space Station who are its nationals." As a corollary of the principle of sovereignty, the new legal space of permanently inhabited space stations is thus governed according to the principle of extra-territoriality that determines the laws applicable for activities occurring in or on the station. Each partner of the project applies its own national laws of intellectual rights protection, tortious liability, or criminal responsibility, depending on the nationality of the personnel and the elements registered by the launching state. Although the EU is not to be considered as a sovereign entity, or even some sort of federal state, the EU member states wisely chose to be represented in the IGA as a single entity.

The extension of national jurisdictions over activities in or on space stations should be strictly understood according to the general principles and rules of the UN space treaties. The IGA, after all, regards activities in or on the space station, rather than how astronauts and spacecrafts reach that place. In other words, tortious liability for damages caused by space objects, whether or not on Earth surface or to aircrafts in flight shall be covered by the 1972 Liability Convention; damages caused by national personnel, or by the elements registered by the partners of the ISS project have to be assessed in accordance with the extra-territorial effects of a certain jurisdiction, for example, EU law in the case of the European Columbus Laboratory. This twofold level of legal safeguards and space regulations fits the dualistic model of international law endorsed by the IGA. The latter complements the 1972 Liability Convention with its own provisions on liability. Art. 16 of the IGA sets up a "cross-waiver of liability" which prohibits all launching states and their related entities – that is, contractors and sub-contractors,

users or costumers – to claim damages one against each other. Moreover, all the ISS partners should implement this clause in their contracts with users, costumers, and contractors. It is up to each partner and its jurisdiction, however, how to apportion responsibility within its own legal boundaries.

In particular, regarding EU law, claims arising between ESA and one of its users can be covered by contractors or sub-contractors that do not concern other international partners. As the ESA website is keen to inform us: "Space Station users will be asked to agree to an interparty waiver of liability as part of their contract with the European Space Agency, stating that each party will not bring claims in arbitration or sue the other party as a result of International Space Station activities. The applicable law for disputes and the detailed procedures in case of arbitration will be decided mutually by the Space Station users and European Space Agency. The contract will specify the country where the Arbitration Tribunal shall sit, normally in the country where the user has his legal seat."[4]

The limits of ESA's arbitration clauses and cross-waivers of liability seem obvious once confronted with the SpaceX Director's vision of "putting millions of people in space." Arbitration clauses would be in this case either too expensive or ineffective. Moreover, ESA's contract does not consider – but should presuppose – all the legal constraints that govern an outer space mission under EU law and the law of its Member States. Some of such legal constraints regard, for example, the interaction with robots equipped with AI systems that will likely populate the next generation of spacecrafts with tourists, explorers, or scientists. This is the case of the Crew Interactive Mobile Companion (CIMON) that enables voice-controlled access to media and documents, navigation through operating and repair instructions, or planetary exploration, especially in conditions too dangerous or prohibitive for humans. Long space flights will arguably require robots to face the mental and emotional challenges of deep space missions (Pagallo et al. 2023). As a result, what laws and regulations should be applied for, e.g., the malfunctioning of our new AI assistant and the corresponding damage occurred in outer space?

The ESA's contract is instructive for what it establishes, but also for what it omits, or presupposes. These omissions, or presuppositions can be divided into two sets. In addition to the problems of space law with the privatization of space – as stressed time and again in the previous sections of this paper – attention should be drawn to the legal challenges brought forth by the further democratization of outer space and Benji Reed's "millions of people." We stressed that the aim of this paper is not to flesh out what constitution, laws, and *ethos* should govern the interplay of such million people in deep outer space missions and next generation colonies. The focus is rather on the fields of EU law that complement the clauses of the ESA's contracts, and the reasons why the scale of the problems matters. We should not wait for millions of individuals interacting and having fun out there, to admit that the law will be affected by the growing numbers of individuals in space, say, 1,000 by 2030; 10,000 by 2040; etc. At the time of this writing, there were less than 10 astronauts in space. The aim of the next section is thus to illustrate why these figures matter vis-à-vis current efforts of legislators on the regulation of technology. The speed of technological innovation raises some unique challenges that were sci-fi scenarios only a few years ago. Summed up with the quest for the democratization of outer space, *nomoi* and *ethos* of current EU law provide the necessary normative backdrop for the examination of these challenges.

[4] https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/International_Space_Station/International_Space_Station_legal_framework

5. THE SCALE OF THE PROBLEMS

The list of open problems in the ESA's contract is the delight of lawyers. We may even concede that the waiver of liability will resist short-time challenges, but lessons learned from the impact of AI systems here down on Earth suggest inspecting ESA's contract considering further fields of EU regulation that are not traditionally covered by discussions on space law, such as norms on cybersecurity and machinery safety of robots, consumer law and personal data protection. All such provisions of EU law reasonably apply to a new generation of 'outer space contracts.' Even Art. 5 of the IGA recognizes the jurisdiction of all partners and their control over the elements they register or over personnel who are their nationals. The good news of this statement is that most of such fields of regulation that may be at work within the Columbus Laboratory, fall under the regulatory powers of EU law under the principle of subsidiarity (Pagallo 2022), thus preventing many risks of fragmentation among the 27 Member States. The bad news is that many of such regulations and principles of EU law are either under revision, or still represent a work in progress.

First, as regards the field of machinery products in EU law, the report of the Commission's Regulatory Fitness and Performance Programme (REFIT), from 2018, stressed certain shortcomings in the enforcement of the EU machinery directive 2006/42/EC. All in all, it "found that despite its technology-neutral design, the directive might not sufficiently cover new risks stemming from emerging technologies (in particular robots using artificial intelligence technologies)." The Commission presented a proposal for a new regulation on machinery products that entered into force in July 2023 as Regulation (EU) 2023/1230, applying to the design of smart robots as regards safety and security of the machinery in human-robot interaction. Although such provisions do not consider the challenges of outer space specifically, they could reasonably be expanded to outer space missions and activities.

Second, the Commission issued a new proposal for the amendment of the 1985 Product Liability Directive, or PLD regime, in September 2022.[5] Several crucial definitions of the old legal framework regarding software and digital products, or causal relationships between defects and damages, fell short in tackling the challenges of AI (Barfield and Pagallo 2020). The new PLD regime, complemented with the provisions of the new AI liability directive,[6] can be understood as a regulatory minimum for the protection of space explorers and settlers that count on the help, assistance, or companionship of AI robots (Pagallo et al. 2023). The new PLD does refer to 'navigation services,' but the reference in Whereas no. 15 of the Act triggers the old EU legal issue on the role of such referrals and their non-binding nature.

Third, it is worth mentioning the set of duties and obligations for designers, manufacturers, and end-users of high-risk civilian AI applications in EU law, as established in the Artificial Intelligence Act (AIA) just approved in December 2023. Only time could tell us how this regulation will impact the design and use of many AI systems mentioned throughout this paper, e.g., robots equipped with AI systems for healthcare in outer space (Pagallo et al. 2023).

Fourth, it is all about cybersecurity: the EU Regulation 2019/881, which establishes the new European Union Agency for Cybersecurity (ENISA), shall be complemented with further sets of provisions on the network and information systems directives (NIS), i.e., the NIS2 directive 2022/2555, and Regulation (EU) 2023/588, which add cybersecurity of space activities to the list of sectors – such as health and water, energy and transport, banking and financial markets, or digital infrastructures – all considered "of high criticality." This regulatory framework must be further integrated with the governance of cybersecurity in outer space for defense and military purposes (Pagallo 2015; Falco 2019).

[5] See the European Commission's proposal COM(2022) 495.

[6] See the European Commission's proposal COM(2022) 496.

To many of these 'vertical' or 'sectorial' problems we should add problems of EU law that are 'horizontal' and triggered by the advancements of technology, such as the 'autonomy' of AI systems and smart robots. Such problems either regard basic notions of the law, e.g., fault and foreseeability of what the state-transition system of an AI application 'decides' to do in outer space, or how to cover damages that fall within the loopholes of current regulations. Cases of hacking show how current rules of tort law may prove insufficient to defend victims of cyberattacks. It can be impossible for the individual victim of a cyberattack to flesh out a human tortfeasor either in outer space or here down on Earth. Scholars and expert groups set up by the European Commission have recommended the adoption of new standards that regard duties of care and information, presumptions, and burdens of proof, as well as solutions that often step away from traditional approaches of tort law, as the compensation funds for victims of cyberattacks (HLEG 2019; Pagallo et al. 2023). The European Commission has partially adopted these proposals with the new AI Liability directive, complementing the AIA and accompanying the new PLD regime. Most legal balls are yet up in the air.

The issues of space law overlapping with regulations on machinery safety, consumer law, cybersecurity, environmental law, etc., will be exacerbated by the scale of the problems. A new generation of space tourists, space explorers, space scientists and space settlers will increasingly put the troubles of the law with the regulation of this new legal space in the spotlight, sparking the debate on which laws and which courts should govern space activities in the foreseeable future. The quest for the democratization of outer space – opening up space activities or missions for laymen and rich people to thousands of individuals with no particular economic resources – raises new problems of its own. Some of them used to be niche problems for space law experts only a few years ago, discussing damages to be covered for the malfunctioning of an AI system (Pagallo 2013; Freeland and Jakhu 2014; Lim 2020; Bratu et al. 2021). The 30 years old history of cyberspace and its regulations tell however a cautionary tale about why the scale of the problems matters (Pagallo 2015). That which used to be the exception and cabined off in a corner of the legal system becomes the rule, e.g., the extra-territorial effects of national or regional regulations. The legal technique of exerting control and jurisdiction beyond its own territorial boundaries often collides with the right of individuals to have a say in the decisions affecting them. This lack of autonomy may regard liability rules and matters of enforcement and jurisdiction, or how technology is actually designed and regulated.

EU scholars dealing with the constitution, laws, and *ethos* of their own legal system show that these problems – often summarized as the 'democratic deficit' of the EU institutions – are not new (Follesdal 2006). Still, this section has stressed that the next generation of space pioneers with their AI assistants and companions will inevitably add some issues of their own, regarding a full array of unique problems on (i) contractual clauses and tortious liability claims that arbitration and waivers of liability will likely find it difficult to address; (ii) the international law duties of private corporations to protect the fundamental rights of individuals, e.g., environmental protection; and, (iii) the behaviour of AI systems and smart robots with their cybersecurity issues that often make it hard to flesh out a tortfeasor, if something goes wrong.

It is too early to say whether EU law will successfully meet these challenges (Pagallo 2022). We noted that the success of every regulatory effort does not only depend on provisions and principles of the legal system, e.g., the protection of fundamental rights, but also, on the moral grounds on which such principles and provisions rest (Rogerson 2022). Growing concerns for the extra-territorial effects of regulation and the protection of a new generation of space settlers, explorers, scientists, and tourists can widen the gap between the EU institutions and the new EU space citizens. The gap has been so far filled by the *ethos* of few European astronauts working in and on the ISS. We should be attentive to the risk that current drawbacks of contracts and waivers of liability ruin this *ethos*.

6. CONCLUSION

The paper has examined the quest for the democratization of outer space through the lens of the regulations and *ethos* of EU law. Multiple problems related to the challenges of how to govern work and lives of a growing number of tourists, scientists, explorers, or even settlers remain quite open. Some of these problems were investigated vis-à-vis contractual issues and tortious liability claims that arbitration clauses and waivers of liability will likely find it difficult to address. A further set of issues has been stressed with the cautionary tale of cyberspace on the extra-territorial effects of legislation that also the IGA embraces. On top of that, the legal problems depended on the uniqueness of the challenges brought forth by AI systems and smart robots, whereas some of these challenges are unique when displayed in outer space. The combined effect of all these three different sets of issues may result in moral disruption and the degree of (dis)agreement in the community regarding values and principles that are at stake with the norms of legislators. The paper summed up this disruption with the quest for the democratization of space to stress the concerns for the autonomy and protection of a new generation of mass space settlers, explorers or tourists, which should be bolstered by the constitution and rules of every democratic institution, including the European Union.

How to tackle such risks for human autonomy and the protection of fundamental rights will increasingly be the subject of scholarly debate and institutional proposals also but only in the EU. The quest for the democratization of outer space will cast light on the democratic deficit of such institutions against current trends on the privatization of space. Leaving aside the promise, or the menace of Space X's Director on "millions of people in space," it seems fair to admit that the law should be ready to properly tackle the challenges of the next generation of humans that will leave Mother Earth. Moral arguments on the autonomy of individuals and the protection of their fundamental rights should provide a guide for the new laws of outer space.

REFERENCES

Barfield, W., & Pagallo, U. (2020). *Advanced introduction to law and artificial intelligence*. Edward Elgar Publishing.

Bingham, T. (2011). *The rule of law*. Penguin Uk.

Blount, P. J., & Robison, C. J. (2016). One small step: The impact of the US Commercial Space Launch Competitiveness Act of 2015 on the exploitation of resources in outer space. *NCJL & Tech.*, *18*, 160.

Bobbio, N. (1987). The Future of Democracy. Minnesota: University of Minnesota Press.

Bratu, I. (2021). Blaming Galileo: Liability for Damage Caused by Artificial Intelligence Operating Based on GNSS. *Proceedings of the International Institute of Space Law*, *64*(6).

Bratu, I., Lodder, A. R., & van der Linden, T. (2020). Autonomous Space Object and International Space Law: Navigating the Liability Gap. *Indonesian J. Int'l L.*, *18*, 423.

Dennerley, J. A. (2018). State liability for space object collisions: The proper interpretation of 'fault'for the purposes of international space law. *European Journal of International Law*, *29*(1), 281-301.

Deem, C. L. (1983). Liability of Private Space Transportation Companies to Their Customers. *BYU L. Rev.*, 755.

Ernest, V. C. (1990). Third Party Liability of the Private Space Industry: To Pay What No One Has Paid Before. *Case W. Res. L. Rev.*, *41*, 503.

Falco, G. (2019). Cybersecurity principles for space systems. *Journal of Aerospace Information Systems*, *16*(2), 61-70.

Feichtner, I. (2019). Mining for humanity in the deep sea and outer space: The role of small states and international law in the extraterritorial expansion of extraction. *Leiden Journal of International Law*, *32*(2), 255-274.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, Ch. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke and E. Vayena (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, *Minds and Machines*, 28(4): 689-707.

Follesdal, A. (2006). The legitimacy deficits of the European Union. *Journal of Political Philosophy*, *14*(4), 441-468.

Foster, C. (2016). Excuse me, you're mining my asteroid: Space property rights and the US space resource exploration and utilization act of 2015. *U. Ill. JL Tech. & Pol'y*, 407.

Freeland, S., & Jakhu, R. (2014). What's human rights got to do with outer space?: everything!. In *Our World Needs Space: Proceedings of the 65th International Astronautical Congress (IAC 2014), 29 September-3 October 2014, Toronto, Canada* (pp. 10376-10384).

Freeland, S., & Ireland-Piper, D. (2022). Space law, human rights and corporate accountability. *UCLA J. Int'l L. Foreign Aff.*, *26*, 1.

Gabrynowicz, J. I. (2010). One half century and counting: The evolution of US national space law and three long-term emerging issues. *HARv. L. & PoL'Y REv.*, *4*, 405.

HLEG (2019). *Liability for Artificial Intelligence and Other Emerging Technologies*, Report from the European Commission's Group of Experts on Liability and New Technologies, available at https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608.

Hofmann, M., & Bergamasco, F. (2020). Space resources activities from the perspective of sustainability: Legal aspects. *Global sustainability*, *3*, e4.

Isnardi, C. (2019). Problems with enforcing international space law on private actors. *Colum. J. Transnat'l L.*, *58*, 489.

Jessen, D. (2017). Modern Ethical Dilemmas Stemming from Private One-Way Colonisation of Outer Space. *J. Space L.*, *41*, 117.

Lim, J. (2020). Charting a human rights framework for outer space settlements 71st.

Locke, J. (1689, ed. 2013). *Two Treatises of Government*, P. Laslett (Ed.), Cambridge UP (24th printing).

Loder, R. E. (2018). Asteroid mining: ecological jurisprudence beyond Earth. *Virginia Environmental Law Journal*, *36*(3), 275-317.

Martin, A. S., & Freeland, S. (2021). The advent of artificial intelligence in space activities: New legal challenges. *Space Policy*, *55*, 101408.

Pagallo, U. (2013). *The Laws of Robots: Crimes, Contracts, and Torts*, Springer.

Pagallo, U. (2015). Good onlife governance: On law, spontaneous orders, and design. *The onlife manifesto: Being human in a hyperconnected era*, 161-177.

Pagallo, U. (2022). The Politics of Data in EU Law: Will It Succeed?. *Digital Society*, *1*(3), 20.

Pagallo, U., Bassi, E., & Durante, M. (2023). The Normative Challenges of AI in Outer Space: Law, Ethics, and the Realignment of Terrestrial Standards. *Philosophy & Technology*, *36*(2), 23.

Rogerson, S. (2022). *Ethical Digital Technology in Practice*. CRC Press.

Steele, J. (2021). Luxembourg and the Exploitation of Outer Space. *Nottingham LJ*, *29*, 32.

Su, J. (2017). Legality of unilateral exploitation of space resources under international law. *International & Comparative Law Quarterly*, *66*(4), 991-1008.

Tepper, E. (2019). Structuring the discourse on the exploitation of space resources: Between economic and legal commons. *Space Policy*, *49*, 101290.

Tronchetti, F. (2015). The Space Resource Exploration and Utilization Act: A move forward or a step back?. *Space Policy*, *34*, 6-10.

UNOOSA (UN Office for Outer Space Affairs) (2019). *Annual Report 2018*, United Nations, June2019. Retrieved from https://www.unoosa.org/documents/pdf/annualreport/UNOOSA_Annual_Report_2018.pdf.

Viikari, L. (2008). *The environmental element in space law: assessing the present and charting the future* (Vol. 3). Brill.

Vernile, A. (2018). The Rise of Private Actors in the Space Sector, Springer.

Ziemblicki, B., & Oralova, Y. (2021). Private Entities in Outer Space Activities: Liability Regime Reconsidered. *Space Policy*, *56*, 101427.

INCOPORATING EXPERIENTIAL LEARNING PLATFORM FRAMEWORK FOR AN ONLINE GRADUATE CLASS

Shalini Kesar, Ashely Tyler

Southern Utah University, (USA)

Kesar@suu.edu; ashleytyler@suu.edu

ABSTRACT

This paper is part of the collaborative on-going research between the author and co-author (PI of an educational grant in the partnership with Forest Inventory and Analysis. The initial research began a few years ago with the idea of developing or modifying the design curriculum to provide an educational experiential learning using the right ethical practices of National Society for Experiential Education (NSEE, 2009) framework (Kesar and Pollard, 2020, 2021). It started with focusing on undergraduate students in STEM field (computer science, information systems, cybersecurity and technology). As the research has progressed, the context has shifted to an online cybersecurity graduate class. This paper sheds light on how author collaborated with the co-author to design the class. It also discusses how the NSEE framework proved beneficial in creating team building project in cybersecurity as well as adding value to the on- going research collaboration.

KEYWORDS: Cybersecurity, experiential learning, online class, collaboration

1. INTRODUCTION

This paper is part of the collaborative on-going research between the author and co-author (PI of (PI of an educational grant in the partnership with Forest Inventory and Analysis). The initial research began a few years ago with the idea of developing or modifying the design curriculum to provide an educational experiential learning using the right ethical practices of National Society for Experiential Education (NSEE, 2009) framework (Kesar and Pollard, 2020, 2021). It started with focusing on undergraduate students in STEM field (computer science, information systems, cybersecurity and technology). As the research has progressed, the context has shifted to online graduate students in cybersecurity. This paper specifically focuses on the designing of the cybersecurity assignment using the NSEE framework.

Founded in 1971, the Society for Experiential Education (SEE) is the premier, nonprofit membership organization composed of a global community of researchers, practitioners, and thought leaders who are committed to the establishment of effective methods of experiential education as fundamental to the development of the knowledge, skills and attitudes that empower learners and promote the common good (NSEE, 2023). The framework consists of eight principals linked with good practices. In this paper, the project conducted with graduate cybersecurity students is discussed that was designed by the instructor (author) as part of an experiential learning activity. The goal was that this experience and the learning will add value to the fundamental of creating an online cybersecurity training as part of a group project. While the authors (instructor and client) collaborated and designed the training, it is hoped that all the parties are empowered to use the right principals mentioned in the framework. Consequently, ensuring both the quality of the learning experience and of the work produced by the students, and in building an assignment that underlie the pedagogy of experiential education. Although the NSEE framework was used, the main thought process was very different when designing the project. It considered the framework as well the research regarding the team projects and importance of training in cybersecurity. This is because this style of pedagogy will provide an experiential learning education environment, which will better prepare the student to face challenges

in the ever-evolving cybersecurity field. While developing the team project curriculum, various studies were considered, including author's previous published research. Best standards and Guiding Principles of Ethical Practice by the National Society for Experiential Education (NSEE) were used.

2. NSEE PRINCIPLES

The NSEE Guiding Principles of Ethical Practices are used to develop the pedagogy to teach ethics and professional as part of an experiential education. This paper describes how the how instructor included ethics and professionalism in this team project. The eight principals are exampled below. It outlines the project and how the NSEE framework's principles were modified or/and revisited to cultivate an empathic learning pedagogy for mostly non-traditional students pursuing their master's degree in cybersecurity. Finally, it discusses some concluding remarks including lessons learned. The contribution of this paper is that these suggestions can be used by other instructors while designing experiential classrooms. This is important and has significantly impacted on how educators as well students view what constitutes an experiential classroom environment proved to be useful in developing the assignment in cybersecurity class.

2.1. Intention

Intention: In this principal, all parties must outline a clear vision on the reason which this experience is chose and the why experience is the chosen approach to the learning. The motivation for this on-going research started with the intention that core skills such as teamwork, communication, professionalism and ethics are part of experiential learning pedagogy. These skills in turn prepare the students to deal with challenges faced in today's technology related businesses. As part of the on-going research, this paper reflects on the using NSEE framework for an online cybersecurity graduate class assignment where students develop training sessions for the client. The principal of intention in general focuses on the purposefulness that enables experience to become knowledge and, as such, is deeper than the goals, objectives, and activities that define the experience. The cybersecurity online graduate students compromise of both traditional and non-traditional students. The intention of a real business setting was to allow the students to apply what they have learned in cybersecurity topics to develop training sessions for the client. The main intention of this assignment was to enhance both technical and soft skills about writing, communication and presentation in the context and cybersecurity working environment where through training best practices of cybersecurity is incorporated within organizations.

The client included a set of employees, part of a research team in the university, who worked in partnership with Forest Inventory and Analysis (FIA). Most of the research team were part-time employees and full-time students at the institute where the author was the instructor. The research team is a part of a small group for the Forest Inventory and Analysis (FIA). They use Design and Analysis Toolkit for Inventory and Monitoring (DATIM) which is an application being developed by the U.S. Forest Service. 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 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 DATIM core team is comprised of both R&D and NFS staff from resource inventory and forest planning programs.

2.2. Preparedness and Planning

Based on the first principle above, the second step ensures participants enter the experience with sufficient foundation to support a successful experience. In this step, it is important that intentions identified are aligned with the goals and objectives of the assignment using the NSEE framework. When preparing and planning, the authors (instructor and the client) discussed the requirement, possible outcomes and expectations from the training session. The lessons learned from the earlier projects that used NSEE framework were considered. In the planning phase the main objective of this principal was to ensure the students have a group project experience where each member of the project has a successful experience from the earliest stages of the experience/program. As mentioned earlier, the project was to design a cybersecurity training for the client's team. The training should include topics that are considered as best practices of cybersecurity. While preparing and planning the assignment, the student learning outcomes were kept in mind: 1) Understand what the key aspects of cybersecurity training are; 2) Research existing training programs in context of cybersecurity field; and 3) Work as a team to research and examine different perspectives of training.

2.3. Authenticity

In this principal it is important the students have an experience while developing the cybersecurity training for the client as part of their annual mandatory training. To provide a real-world experience, students also met with the client to share their progress and ask questions to clarify any doubts they may have while developing the training. The three groups of students compromising of three to four members who divided the tasks and created a process that included research, tracking progress, etc. The three main topics chosen by the students was Phishing, Social Engineering, and Passwords.

2.4. Reflection

NSEE refers to Reflection as an element that transforms simple experience to a learning experience. With this principal in mind, the assignment was designed students conduct research, test assumptions and examine the hypotheses about the outcomes of decisions and actions taken in context of cybersecurity training. It also gave them an opportunity to weigh as well as reflect the outcomes against past learning and future implications. This reflective process in the assignment comprised of a report writing and presentations at conference and as a final exam. In addition to this, the students were required to present to their client with lessons learned and share the suggestion on the training. This, according to NSEE, 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.

2.5. Orientation and Training

This principle 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. The students were required to discuss and show the training they had developed to the client and their team members. This not only prepared them work as a team but also experience and learn about each other and about the context and environment in which the training will be presented to the small division of the FIA team. The training included assumptions; identify at least two basic causes of the problem; and detailed training session with topics/aspects as well as the training itself.

2.6. Monitoring and Continuous Improvement

It is important to note that any learning activity designed should be dynamic and changing. In addition, the instructor (author) outlined the assignment with the student learning outcomes that included

reports and presentation that provided the richest learning possible to the students. Students also had to write a self-reflection on their own progress as well as their team members. This feedback process relates to learning intentions and quality objectives. Consequently, this allows the structure of the experience to be sufficiently flexible that permitted changes in response to what that feedback suggests. Subsequently, monitoring and continuous improvement represent the formative evaluation tools.

Assignment was built into the curriculum that monitored each student and the team's progress, their challenges, and how they overcame them. For example, group project members on a weekly basis were required to answer the questions in the build-in template: 1) Was your project on schedule? (If not, what and how will you make adjustment(s) to meet required timelines?); and 2) How have you dealt with the new issues? (e.g., what solutions/work-around have you explored or adopted?). In the weekly report, a few more questions were added to specifically address the geographical location of the students and how they meet as a team. Students discussed their results, challenges, and weekly progress on the remote Zoom meetings.

2.7. Assessment and Evaluation

Assessment is a means to develop and refine the specific learning goals and quality objectives identified during the planning stages of the experience. Whereas evaluation provides comprehensive data about the experiential process as a whole and whether it has met the intentions which suggested it. Based on the NSEE definitions, the outcomes and processes of assignments of the project included systematically reports, presentations, and self-reflection that were linked with the initial intentions. Students were required to complete the assessment templates provided them which included: 1) Charter template; 2) Organization structure; 3) Basics assumptions of the training session; 4) Planning of the training session with resources. At the end of the project, the assignment also included that students recognize the lessons learned, recognition of learning and impact occur throughout the experience by way of the reflective and monitoring processes and through reporting, documentation and sharing of accomplishments.

3. CONCLUDING REMARKS

The cybersecurity training session assignment pedagogy was modified based on previous year's findings of the on-going research. The main intent was to provide students a real business setting where they appreciated the depth of responsibility, accountability and soft skills needed to work in a team. Students also appreciated the different perspectives and team members from different work experience. Although literature suggest that team-work and collaboration skills are crucial when preparing students facing global challenges in the work field, with this being an 100% online class where students met weekly via Zoom did pose some challenges in coordinating meetings. Having said that the instructor was able to cultivate an experiential learning pedagogy by revisiting the eight principles of NSEE's framework. Consequently, an informed learning context that fostered students' insight into understanding the pragmatic challenges and finding solutions that address the client's requirements. During re-planning and modifying the assignment, the NSEE learning framework significant contributed to designing an assignment where students learned skills beyond the classroom curriculum.

ACKNOWLEDGEMENTS

Both formal and informal acknowledgments were designed into this classroom. Acknowledgment section in the report and an informal suggestions and comments about the project during the breakout

room on Zoom were provided. As this is on-going research, a special acknowledgment is also given to James Pollard, a former research fellow with whom the author (instructor) started collaboration and the research using NSEE framework to provide students with experiential learning classroom environment.

REFERENCES

Jervis, K. J., and Hartley, C. A. (2005). Learning to design and teach an accounting capstone. Issues in Accounting Education, 20 (4), 311-339.

Kesar, S and Pollard, J. (2021) "Cultivating an Empathic Learning Pedagogy: Experiential Project Management", in Normal Technology Ethic Proceedings of the ETHICOMP* 2021, Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata, Ana María Lara Palma, Universidad de La Rioja, Spain, Universidad de La Rioja, Spain. pp. 257-259. https://dialnet.unirioja.es/servlet/libro?codigo=824595

Kesar, S. and Pollard, J. (2020). Project Management: Experiential Learning Pedagogy". In "Societal Challenges in the Smart Society", Oliva, M., Borondo, J., Murata, K., and Palma, A., Universidad de La Rioja, Spain. pp. 147-151.

Kesar, S., (2016). 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). https://doi.org/10.1145/2874239.2874303

The National Society (2009). Guiding Principles of Ethical Practices. http://www.nsee.org

ADVOCATE TO INCREASE WOMEN IN CYBERSECURITY

Shalini Kesar

Southern Utah University (United States)

kesar@suu.edu

ABSTRACT

This paper discusses the author's outreach projects designed for rural high schools in remote locations. The strategy along with the hands-on cybersecurity activities to increase awareness about education and career opportunities for young women is outlined in this paper. The goal is to share the possible methods that can be incorporated in different contexts when trying to reduce the diversity gap in the cybersecurity field. It also discusses how the author mentors her undergraduate interns (especially females) to participate in outreach activities for young girls for high school. This ongoing research proposes some pragmatic steps towards creating a cybersecurity pipeline with diversity including women and underrepresented groups.

KEYWORDS: Women in Cybersecurity, rural areas, diversity, hands-on activities, role model.

1. INTRODUCTION

Various research and articles shed light on the underlying factors on the lack of diversity in STEM field including cybersecurity and the impact of stereotypes and gender bias. Hundreds of studies have conducted research on, for example, the power of stereotypes to influence performance through a phenomenon known as "stereotype threat.", disengagement from fields in which women are negatively stereotyped, such as computing and cybersecurity (solving the equation). The good news is that practitioners, academia and community as whole acknowledges that there are concerns in the talent pool or the lack of talent pool that goes beyond just demand and supply gap numbers in the cybersecurity area. The Economics Forum stated two major issues: 1) The global cybersecurity skills gap; and 2) The lack of diversity in the cybersecurity workforce. Studies in SANS (ICS, 2023) highlight that 3.4 million people are needed to fill the global cybersecurity workforce gap. Consequently, a survey by the World Economic Forum (2022)) that 59% of businesses would find it difficult to respond to a cybersecurity incident due to the shortage of skills. It also states that the cybersecurity sector needs 3.4 million people to fill its workforce gap.

There are many articles that reflect on how to create a pipeline starting from kindergarten school. This can be empowering and can perhaps impact the shift in the mindset of a field itself. For example, the author the author's white paper, collaborative paper with Microsoft (2018) highlights how engaging young girls creating a role and having hands-on activities can spark interest in cybersecurity and STEM fields. For example, "STEM clubs and activities also correlate with a girl's likelihood of pursuing STEM and computer science later in her education. Seventy-four percent of middle school girls who participate in these activities say they are likely to study computer science in high school, compared to only 48 percent of those who don't participate. (Microsoft and Kesar, 2018).

This paper focuses on some of the pragmatic ways outreach activities can be designed for high schools to motivate them to explore the education and career opportunities in cybersecurity. This paper also can be useful in understanding how to retain young women in cybersecurity who have decided to explore this field in college and university. Consequently, it can contribute to changing structures and environments with increase in women's representation and underrepresented group. This paper uses the "9 Strategies to Improve Gender Diversity in the Security Workforce" Security Intelligence, 2020)

article as a starting point to highlight some examples when the author designed outreach activities and mentor female undergraduate interns in her university. The strategies include: 1. Support Competitions and Scholarships Specifically for Women; 2. Set Up Internship Opportunities; 3. Use Inclusive Language in Hiring Efforts; 4. Involve Women in Recruitment; 5. Provide Opportunities for Lateral Growth; 6. Enable Employees to Pursue External Certifications; 7. Consider Women Who Are Rejoining the Workforce; 8. Offer Fair and Equitable Compensation; and 9. Organize Pathways for Advancement. Support Competitions and Scholarships Specifically for Women. These are explained below.

Although women are aware of cybersecurity, yet there is a perception that awareness of cybersecurity is low among women. It was found that the opposite to be true: 82% of survey respondents said they had some or a lot of knowledge of cybersecurity; 2) Women have access to cybersecurity education. Another perception: low participation of women in cybersecurity because they lack access to cybersecurity education. Our survey indicated otherwise. Specifically, 58% of respondents said they had access to cybersecurity education, and 68% had already taken a cybersecurity-related course; 3) Role models and senior encouragement are critical. That's what anecdotal evidence suggested, and our survey validated the hypothesis. Role models played an important factor to avoid the negative perceptions of cybersecurity as a career choice. The top three priorities for women in choosing a job are contributing to society, earning a high salary and having a good work-life balance. However, 37% of respondents regard cybersecurity as a field where achieving that balance is difficult; 4) Lack of awareness also had a negative perception with a mindset that in cybersecurity is that it's often regarded as a "boys' club".

In light of the above, many articles highlight the importance of reducing the gap in cybersecurity field. For example, in the Cybercrime Magazine, Osborne (2022) highlight the women will hold 30 Percent of Cybersecurity jobs globally by 2025 and female representation expected to reach 35 percent by 2031. Furthermore, it has been shown in a survey of 2,000 female STEM undergraduate students in 26 countries spanning six regions conducted by BCG (Panhans et al, 2022) indicates "Solving both of these cybersecurity challenges—the staffing shortfall and the gender-based inequity—begins with opening STEM doors to women and girls. But the effort can't stop at early-stage access. It must gain breadth and depth as women advance in the field so that they can fully participate in cybersecurity throughout a career trajectory".

2. WOMEN IN CYBERSECUIRTY OUTREACH

The categories, as mentioned above, include:1) Support Competitions and Scholarships Specifically for Women; 2. Set Up Internship Opportunities; 3. Use Inclusive Language in Hiring Efforts; 4. Involve Women in Recruitment; 5. Provide Opportunities for Lateral Growth; 6. Enable Employees to Pursue External Certifications; 7. Consider Women Who Are Rejoining the Workforce; 8. Offer Fair and Equitable Compensation; and 9. Organize Pathways for Advancement. Support Competitions and Scholarships Specifically for Women. These are explained below in the context of the author's outreach activities.

2.1. Support Competitions

This refers to various scholarships or events that are inclusive to young women. For example. host a security-focused hack-a-thon or a capture the flag competition specifically for women that focuses on hands-on security skills, teamwork and applications to real-world cybersecurity challenges. It is also a good idea to share opportunities about scholarships and competition that are women centric conferences. For example, Women in Cybersecurity annual conference or The Women's Society of

Cyberjutsu (WSC), a 501(c)3 non-profit, is dedicated to raising awareness of cybersecurity career opportunities and advancement for women in the field, closing the gender gap and the overall workforce gap in information security roles.

As part of the internship programs with undergraduates, the author mentored and encouraged students (high school and undergraduate) to participate in competitions and present at conferences where outreach activities conducted in high school were presented. Some of the competitions included women in cybersecurity conference, National Centre for Women & IT Aspirations competition for young girl sin high school. Supporting students interns in competitions also benefited the high school young girls as the undergraduate interns unintentionally became their role models. This also to some extent broke the stereotypes perspective that that this field is for only men since men are better suited to technical skills and pursuits in general. The author's goal to support competition was in two folds: 1) Through mentorship and involving both young women and men interns for high school outreach not only broke the stereotype barrier but also "showed" the audience it is possible to explore the opportunities in both education and career in cybersecurity; 2) Given that the audience rural schools lacked resources and awareness, it was challenging for young girls to consider these field because they were no training, classes, and opportunities that discussed the opportunities in education and cybersecurity careers.

2.2. Set Up Internship Opportunities

Student internships were offered to undergraduate students for outreach activities linked with young girls in high schools. The internships were part state and regional grants awarded to the author. The grants supported the paid intern and travel for activities when visiting schools. The internship included designing activities to: 1) Provide more exposure to high school girls to role models and mentors they could aspire to explore a cybersecurity education; 2) Demonstrate a path forward in terms of turning an interest in STEM and computer science into success in school and in a career; 3) Support extracurricular STEM activities that teach girls how to create and build confidence; 4) Provide hands-on experiences and real-world examples; 5) Emphasize the creative aspects of STEM and computer science; 6) Listen to what girls say about their challenges and desires (Microsoft & Kesar, 2018).

2.3. Use Inclusive Language in Hiring Efforts and Involve Women in Recruitment

The use of inclusive language in hiring efforts and involving women in recruitment process are important factors to encourage women to consider cybersecurity career. For example, advertisements in cybersecurity positions with language and images that are inclusive of all applicants can motivate women to apply for employment. When posting the internship job post, the author worked with HR closely to ensure advertise cybersecurity positions with language and images that are inclusive of all applicants. This is important to fill the existing gap in the cybersecurity workforce. For example, Women held 25 percent of cybersecurity jobs globally in 2022, up from 20 percent in 2019 and around 10 percent in 2013. Cybersecurity Ventures predicts that women will represent 30 percent of the global cybersecurity workforce by 2025, increasing to 35 percent by 2031 (Osborne, 2023).

As a senior-level women, the author directly interviewed and was part of the recruiting processes so applicants are aware early on that there are other women at the university work in this field as well as opportunities for mentorship in education and career questions.

2.4. Provide Opportunities

Providing opportunities for lateral growth is a strategy is to improve gender diversity in the security workforce. The student interns had an opportunity to start as basic interns and then grow to a senior

position where they had more responsibilities in the sustaining the outreach activities. This helped the students to be exposed to different soft skills need for the workforce in the long term. In their survey, Economic Forum suggested it's important to engage girls in computing including cybersecurity early. Their research confirmed this hypothesis that a majority – 78% – of our respondents said that they had first developed an interest in STEM in middle school or high school (Economic Forum, 2022). Another strategy is to enable employees to pursue external certifications. The author, as lead of outreach activities, provided support for her intern students and young high school girls to engage in external training and certification programs related to STEM and security, such security pro, Fundamentals of Security etc.

2.5. Other Considerations as part of the Strategies

Other considerations are to include women in recruitment, rejoining the workforce and design programs to offer fair and equitable compensation. Design a program to recruit women who are re-entering the workforce or pursuing a change in career so they can receive the necessary training and start working in the field immediately. Compare salaries across cybersecurity roles to ensure that women are not being paid less than men for the same job. On average, according to (ISC)2, women working in cybersecurity have higher levels of education than their male colleagues and are still paid lower salaries. Organizing pathways for advancement is important to host opportunities for women in cybersecurity to network with higher-level executives and managers within the organization to create pathways for advancement and promotion. The author conducted webinars and in person seminars where undergraduate student interns participated and shared their experience in cybersecurity. The targeted audience included young girls from high schools, educators and parents.

Finally, it is important to organize pathways for advancement. The author has experienced that first hand as she has conducted many outreach projects for high and middle school girls that are linked to STEM including cybersecurity subjects. In the article "Empowering women can help fix the cybersecurity staff shortage" (2022) published in The Economic Forum states that Our survey corroborated some traditional thinking – but refuted other key, long-held hypotheses: 1) It's important to engage girls in STEM early. Their research confirmed this hypothesis as a majority – 78% – of our respondents said that they had first developed an interest in STEM in middle school or high school.

3. CONCLUDING REMARKS

This paper discusses and possible solutions to reduce the diversity gap in the computing field including cybersecurity. It presents the reality check of cybersecurity that clearly highlights lack of women and underrepresented groups in this field. It also reflects on the various efforts proposed or/and implemented to address this concern. This is ongoing research where the author is passionate and motivated to address as well as propose some pragmatic steps towards creating a cybersecurity pipeline with diversity including women and underrepresented groups.

REFERENCES

ICS (2022). Five Startling Findings In 2023's ICS Cybersecurity Data. Retrived from https://www.sans.org/blog/five-startling-findings-2023-ics-cybersecurity-data/

Kesar, S (2018). Closing the STEM Gap Why STEM classes and careers still lack girls and what we can do about it. Retrieved from https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE1UMWz

Microsoft and Kesar, S. (2018). "Closing the STEM Gap Why STEM classes and careers still lack girls and what we can do about it", retrieved from https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE1UMWz

Osborne, C (2023). Women To Hold 30 Percent Of Cybersecurity Jobs Globally By 2025. Cybercrime Magazine, London, retrieved from https://cybersecurityventures.com/women-in-cybersecurity-report-2023/

Panhans, D, Hoteit, L., Yousuf, S., Breward, T., Wong, C., AlFaadhel, A., AlShaalan, B. (2022). Empowering Women to Work in Cybersecurity Is a Win-Win. BCG Report. Retrieved from https://www.bcg.com/publications/2022/empowering-women-to-work-in-cybersecurity-is-a-win-win

Women in Cybsersecurity (2022), retrieved from http://www.wicys.org

World Economic Forum (2022). Empowering women can help fix the cybersecurity staff shortage. Retrieved from https://www.weforum.org/agenda/2022/09/cybersecurity-women-stem/

ETHICS IN INTERNET OF THINGS SECURITY: CHALLENGES AND OPPORTUNITIES

Sabina Szymoniak, Mariusz Kubanek

Department of Computer Science, Czestochowa University of Technology, Poland

sabina.szymoniak@icis.pcz.pl; mariusz.kubanek@icis.pcz.pl

ABSTRACT

The Internet of Things is a network of physically interconnected devices that employ sensors to gather environmental data and communicate with one another online. These devices are utilized in many facets of human life. However, data security and user and device security are the two biggest problems facing Internet of Things solutions. These devices communicate using specifically created security protocols to raise security standards. Every security protocol should adhere to the CIA triad that guarantees that data is accurate, undamaged, and accessible, that information is safeguarded from unauthorized access, and that customers may obtain the appropriate information when needed. However, Security procedures may be open to intrusions by malevolent users, posing a severe risk to the systems mentioned. The Internet of Things offers both benefits and challenges. They keep private information vulnerable to theft, exploitation, and data leaking. Communication protocols must take ethics into account, especially while preventing cyberattacks. In addition to summarizing the research findings, this work attempts to highlight risks, vulnerabilities, ethical issues, and recommendations for ethics in Internet of Things systems security. In order to safeguard people and devices against cyberattacks, it underlines the necessity of adequate security and ethical regulation in Internet of Things networks.

KEYWORDS: Internet of Things, ethics, security, attacks, vulnerabilities.

1. INTRODUCTION

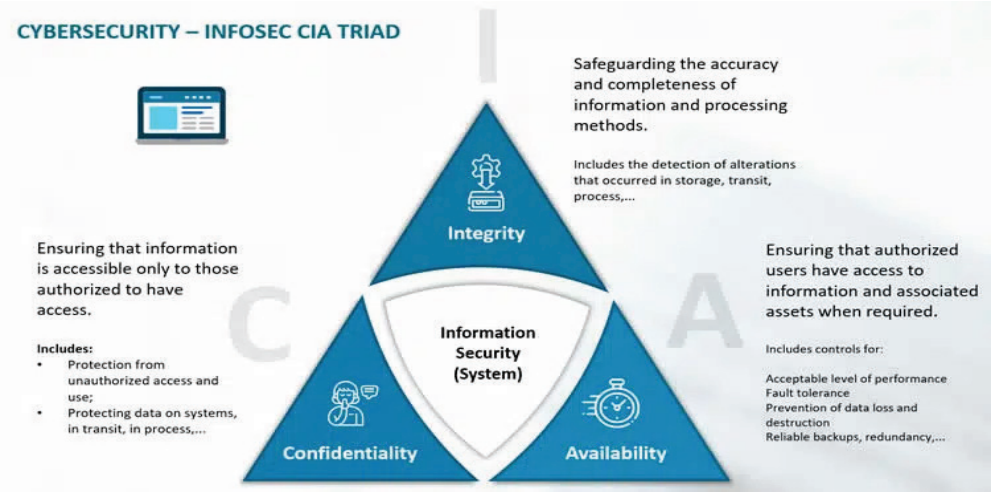
With each passing day of the 21st century, we increasingly recognise the impact and role of the Internet of Things in our lives. The Internet of Things (IoT) is a network of connected physical devices like smart TV sets, smart light bulbs, smart washing machines, etc. Connected devices have the proper sensors to measure their work environment. For example, these things use sensors to regulate a building's lighting or water heating intelligently. Moreover, they exchange data between themselves using the Internet. Also, their users can manage and operate them via the Internet from different geographic localisations. Thus, IoT connects devices like home appliances, vehicles, sensors or smartphones to the Internet network to control and manage some areas (Szymoniak & Kesar, 2022).

We can find IoT devices and connections in many areas, mainly when they use Artificial Intelligence solutions. Medical IoT uses smart devices to manage the critical functions of patients with chronic illnesses, testing blood glucose levels in people with diabetes, alerting doctors when a patient needs medication, and promptly delivering it to the patient (Singh et al., 2022). In healthcare solutions, IoT devices improve, avoiding potentially fatal scenarios, so athletes use them to regulate vital processes and performance (Zhou et al., 2021). The tracking sensors can safeguard our security (Khan et al., 2022; Alsaeed & Nadeem, 2022). Also, IoT devices can be used to warn people about the potential for an earthquake (Sivakumar et al., 2022).

The IoT solutions meet with two main challenges. First of them is the security of data, users and devices. IoT devices use the Internet to communicate. Depending on the system's architecture, they often use wireless data transmission, for example, WiFi and LTE / 5G (Imam-Fulani et al., 2023), as secure channels supported by secure cryptographic protocols like SSL/TLS (Paris et al., 2023). Additionally, to improve security levels, IoT solutions use specially designed security protocols. The security protocols define the order in which messages must be sent and the cryptographic techniques

used. Usually, they are designed and dedicated to the specific solutions. We can find different security protocols for different solutions, such as fog or edge processing (Pardeshi et al., 2022), medicine or healthcare (Rasslan et al., 2022), (Masud et al., 2022), meetings security (Szymoniak & Siedlecka-Lamch, 2022), industry (Yi et al., 2022) or suitable for many domains (Yan et al., 2022).

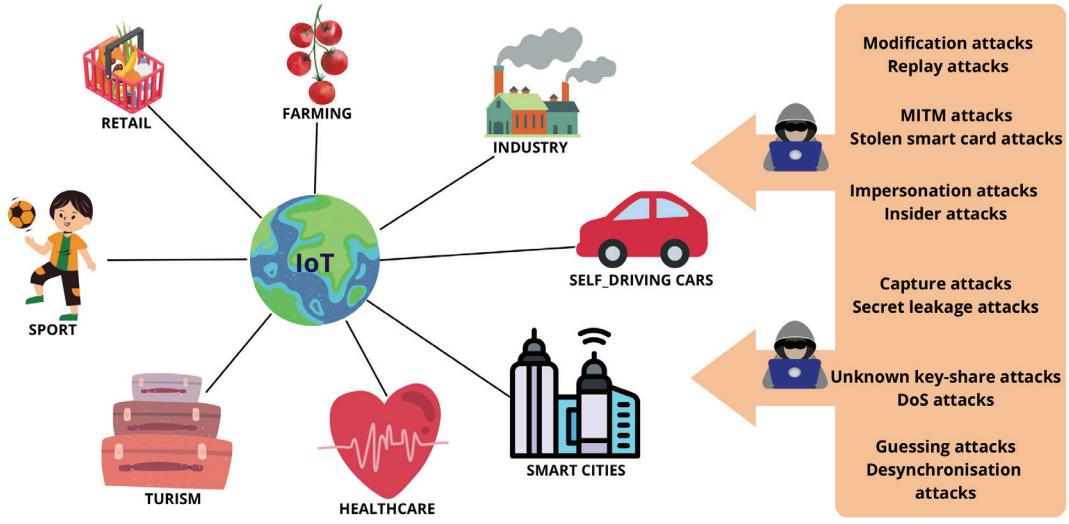
Figure 1. CIA triad.



Source: https://www.i-scoop.eu/cybersecurity/cia-confidentiality-integrity-availability-security/

Security is a very complex challenge. We send many different data during communication between devices using different protocols and communication channels. The fundamental IT security principle, the CIA triad, should be implemented in every security protocol. CIA triad means confidentiality, integrity, and availability. Information is protected from illegal access and is only accessible to authorized users, thanks to confidentiality. The integrity guarantees that data will be correct, undamaged, and unchanged, and it prevents unauthorized users from altering or deleting data. The availability makes sure that consumers may access the desired information when they need it. This objective is increased by using backup tools and sensible hardware safeguards. Figure 1 presents CIA triad features. Also, depending on the protocol's destination, it should satisfy other security features like mutual authentication, anonymity, secrecy or untraceability ((Szymoniak & Kesar, 2022), (Kubanek et al., 2022)).

Figure 2. IoT solutions and typical cyberattacks on IoT systems.



Unfortunately, the security protocols can be vulnerable to many attacks by malicious users who search for vulnerabilities in such systems and try to break into them, for example, to take control of other devices in the network or the whole Smart Home (Szymoniak et al., 2021). Figure 2 summarises IoT solutions and typical cyberattacks on IoT systems.

The second challenge is ethics. Devices and servers store different and sensitive data. Data can be stolen from devices or servers and used. Also, there is the risk of data leakage from IoT systems. It is necessary to consider the ethics of communication on such systems, mainly from cyberattacks perspective. We must know security protocol should realize and solve ethical issues connected with data. Also, we must investigate how security protocols deal with mentioned earlier security features, what communication elements make vulnerabilities, and how to protect IoT systems and their users against cyberattacks.

As mentioned, the Internet of Things systems surround us from every side. We use smart devices that send a lot of data and users' sensitive data. So, such systems require appropriate security and ethical regulation. In this paper we will identify and discuss challenges and opportunities of IoT systems. We will focus on the opportunities these systems offer us, challenges connected with network security and ethical issues related to data storage and communication via network.

The rest of this paper is organised as follows. In the second section, we will describe threats and vulnerabilities that may occur in IoT systems. The third section will discuss ethical problems connected with IoT systems security. In the next section, we will present and discuss recommendations for ethics in IoT systems security. In the last section, we will summarise this article and present findings from the research and our plans for the future.

2. INTERNET OF THINGS THREATS AND VULNERABILITIES

The IoT systems bring many benefits but also carry usage risks due to threats and vulnerabilities. Vulnerabilities affect the occurrence of threats, especially those related to security. Some vulnerabilities are related to the lack of standardization and the diversity of IoT devices. Therefore, vulnerabilities are related to various protocols, software and security measures. Both manufacturers and their users do not regularly update many IoT devices. For this reason, they may contain known security holes that cybercriminals can exploit.

Mainly, the threats in IoT systems are associated with the communication process. IoT devices use different security protocols to communicate in the network. Depending on the cryptographic techniques and protocol structure, they can offer different security levels. For example, one of the first security protocols, the Needham Schroeder Public Key protocol (Needham & Schroeder, 1978), was broken very simply because it was vulnerable to replay attacks. Modern security protocols use more advanced and performance-demanding cryptographic techniques like hashing functions, pseudonymity, elliptic curve cryptography (Ullah et al., 2023).

It is also worth noting that many manufacturers and users of IoT devices do not pay enough attention to the security of their devices. Default passwords and weak encryption algorithms are often used, making devices vulnerable to attacks. Many users are unaware of the risks associated with IoT devices or do not take appropriate measures to protect them. The lack of proper authorization may allow unauthorized users to access these devices.

The most dangerous threat are cyber attacks. Cybercriminals can target IoT devices to gain access to our home or business network. They can use these devices to spread malware, steal data, or perform other complex attacks. IoT devices can be vulnerable to malware injection, which can be used for various purposes, such as sending spam, manipulating surveillance cameras or accessing control

systems. Some attacks may target IoT infrastructure, such as communication networks or device management servers. This can lead to service interruptions or loss of control over devices.

In the case of typical attacks on IoT systems, we can indicate a stolen attack, during which the attacker can guess or steal the verifier (for example, the smart card). The attacker can use such a stolen verifier directly to impersonate the authorized participant of the communication (Kumari et al., 2020). Also, the attacker executes guessing attacks, during which they try to iteratively guess a password or other login details to impersonate the user (Guan et al., 2022). The guessing attacks are dangerous when the user does not change a default password, or the changed password is very simple. Moreover, the attackers can execute a dictionary (checking a predefined word list) or brute force (creating all character combinations in a specified range and length) attack (Alkhwaja et al., 2023). During a capture or cloning attack, the attacker hijacks a sensor node or IoT device to take over the network. In the next step, he removes the hijacked node from the network and redeploys it as a malicious node (Hameed et al., 2022).

IoT system security breaches and cyberattacks can lead to privacy breaches (Atlam et al., 2020). Internet of Things systems often collect large amounts of data about users, their habits and preferences. Also, many IoT systems use the cloud to store and process data. Threats in this area can lead to data leakage or loss of data control. Consequently, such data may be leaked, and the user may be discredited or his belongings lost if the thief learns all the information necessary to break into the user's home.

Thus, the lack of effective monitoring and analysis of administrative data from IoT devices can make it difficult to detect and respond to attacks and vulnerabilities. A well-prepared network threat detection and response system should be part of every network. Logs are necessary for the proper operation of such a system, thanks to which the network can automatically block dangerous network traffic.

It is worth mentioning that each technology, including IoT systems, can cause social problems such as extremism, polarization, misinformation, and Internet addiction (Ahmad et al., 2022). For this reason, it is necessary to consider whether technology realizes ethical rules regularly.

3. ETHICS IN INTERNET OF THINGS SECURITY

The mentioned threats cause ethical problems and challenges in many planes. Most of these problems and challenges are strictly connected with IoT systems security.

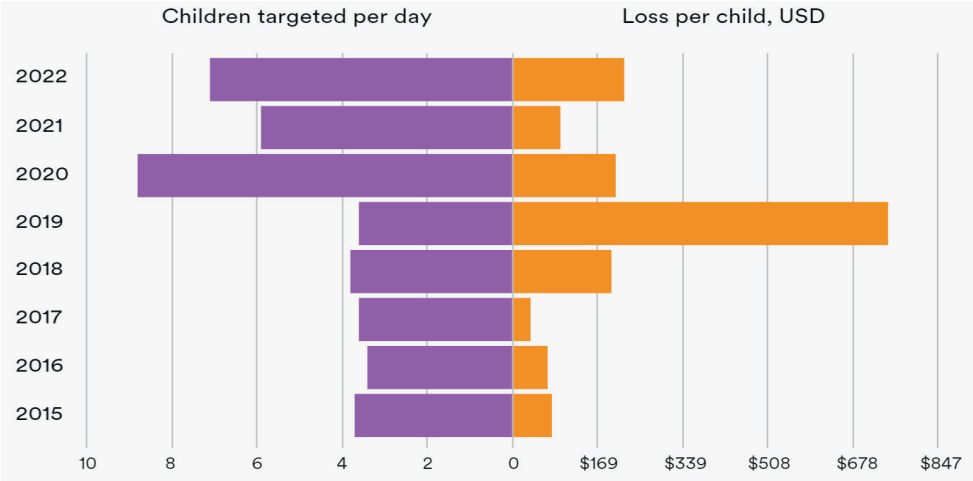
The first problem is data collection ethics. The decision about IoT data collection can be problematic. Often, users must be aware of what data is collected and how it is used. There is a need for clear information, consent to data collection, and the possibility of their removal. The decision on which data should be collected and how long they should be stored can also be problematic. This problem is mainly noticeable in the case of medical or healthcare data (Sholla et al., 2017), (Parthasarathy et al., 2023)).

In the case of medical or healthcare data leakage, they can be made public or used. Published data give us information about someone's health, diseases and drugs, operations, etc. This data leakage may lead someone to depression or even suicide. Moreover, before the publication of data, the attackers can try to blackmail someone to receive some money. The medical or healthcare data will be published if someone does not pay. Similarly, the attacker can prepare a fake fundraising for costly treatment or medical hardware. The people will see the real person with real health problems and needs, but the money from the fundraising will go to someone else. Furthermore, medical or healthcare data may be modified, including medical test results. If the attacker breaks into the

database with prescribed medications or breaks into the IoT device that controls drug delivery, he may modify recommendations. In effect, the patient's health may get worse, or the patient may die.

Also, our sports activity data are dangerous. Devices like smartwatches or bands collect information about our activity, including geographical coordinates. So, for example, the smartwatch "knows" when we run, how often we do this, how long we leave our home, and where our activities occur. When the attacker obtains such information, he may break into our home because the attacker will know when and how long there is no one at home. Also, the attacker can cause physical harm, knowing our activity preferences. Similarly, sharing such information on social media may have an identical effect when the attacker performs an OSINT investigation.

Figure 3. Cybercrimes against children between 2015-2022.



Source: https://www.digitalinformationworld.com/2023/05/cybercrime-against-children-is-on-rise.html

In the case of data storage, there are also issues related to privacy and data protection. IoT systems collect vast amounts of data from various devices and sensors that can violate users' privacy. Here, we can indicate monitoring cameras that have not been adequately secured and transmit the image from the monitored area using a public IP address. A list of such cameras can be found using the Shodan Eye tool (Alsmadi et al., 2022). In addition, IoT devices may collect private data related to our network activity. The lack of proper protection and management of the collected data may result in their use for unauthorized purposes, such as behaviour monitoring, the sale of personal data or use without users' consent. The development of the IoT can lead to constant surveillance and monitoring, which undermines the privacy and freedom of individuals. The ethical questions are who has access to this data, for what purpose it is used, and the consequences for individuals.

Children, i.e. the youngest users of IoT systems, are particularly at risk of privacy. Children who are not always aware of their online privacy often use devices. Moreover, children often install insecure applications or click on insecure links on their smartphones. This type of action may cause hacking of the device by the attacker and the use of it for illegal purposes like network control or performing a Distributed Denial of Service attack (de Neira et al., 2023). Therefore, parents and guardians need to know what data is collected, what risks may occur, and the ability to choose data to save and delete. Also, parents should pay special attention to children's Internet activity (Lonergan et al., 2023). Figure 3 shows the number of victims and financial losses during cybercrimes involving children in the USA between 2015 and 2022.

Cyberattacks using IoT systems can also have a much more comprehensive range than just a single device, network or some data. These attacks can also affect critical infrastructures such as energy or transport networks. Consequently, these attacks can interrupt the supply of electricity, water, gas,

public transport or Internet access. This can disrupt people's daily lives and lead to social disorganization. In some cases, attacks on critical infrastructures can threaten people's health and lives. For example, a power outage in a hospital can disrupt medical equipment and put patients at risk. Service interruptions can affect the functioning of businesses and enterprises, leading to economic losses, job losses and economic slowdown.

An essential ethical challenge for the security of IoT systems is the evolution of machine morality. The development of artificial intelligence methods has made them an indispensable part of our lives. IoT devices are often programmed to make moral decisions, for example, in the case of autonomous cars and property protection systems. Ethical concerns are raised by the question of who is responsible for determining the rules of conduct of devices and what values should guide their actions in order to ensure an appropriate level of security for people (especially in the case of medical systems), their data and property (Zhang & Zhang, 2023). Moreover, when users depend on IoT devices to manage various aspects of their lives, they may lose control of their data and decisions. This can lead to the excessive impact of technology on people's lives. Seymour et al. in (Seymour et al., 2023) noticed the ethical problem with AI-driven domestic voice assistants integrated into IoT devices. The voice assistants have privacy implications of having devices because they are always listening. Also, such application may spread misinformation (Edu et al., 2023).

Ethical issues can also be considered from the perspective of increasing consumerism and generating more electronic waste (e-waste), which has a negative impact on the environment.

4. RECOMMENDATIONS FOR ETHICS IN IOT SYSTEMS SECURITY

Ensuring ethics in the security of Internet of Things systems is extremely important to protect user privacy, integrate moral principles into projects and responsibly use IoT technology. This section will present our recommendations for ethics in IoT systems security, considering the challenges described in the previous section.

First, solving the ethical issues surrounding IoT requires collaboration between industry, regulators, the scientific community, and users to develop rules, standards, and policies that protect privacy, ensure security, and promote responsible use of the technology. Legal and ethical issues related to liability require clear regulations and standards, for example General Data Protection Regulation (Regulation, 2018) and industry standards (e.g. ISO 27001).

Second, we should know that ethical practices are essential for protecting user privacy, data integrity and the general well-being of society. Thus, we should consider these challenges from the beginning of system design and incorporate security measures from the outset when designing IoT devices and systems. In the case of user privacy, we should implement data minimization techniques to collect only necessary data to achieve specific goals. Excessive amounts of data can unnecessarily burden devices. Data anonymization or pseudonymization will reduce the risk of personal identification. Administrators should communicate to users how their data will be collected, processed and shared. Also, they should allow users to easily consent to data collection and sharing and users' consent should be explicit, informed and obtained before collecting their data. Users may withdraw consent if it is necessary.

It is worth mentioning that ethical data processing excludes the use of data for discriminatory, malicious or harmful purposes. Thus, each administrator needs to establish guidelines and policies for ethical data handling in his network, ensure data fairness and transparency, and avoid discriminatory profiling using data analytics in IoT.

Furthermore, each IoT system device needs regular updates and, patches on IoT devices and software to eliminate security vulnerabilities. IoT devices with secure default configuration settings will minimize security vulnerabilities. Also, long-term support for IoT devices will give users access to the latest security patches. The clear end-of-life plan for devices will minimize security risks when they become obsolete. Thakral et al. (Thakral et al., 2022) highlighted problems with old, not updated platforms, programming, and infrastructure layers that may significantly impact security.

Also, risk assessment and management are considerable challenges of IoT systems. Comprehensive risk assessments will identify potential threats and vulnerabilities in IoT systems and implement risk mitigation strategies such as intrusion detection systems and firewall protection. Each IoT system should be regularly audited. Regular security audits and penetration tests are necessary to identify and address vulnerabilities in IoT systems, track new threats, and adapt security measures as the situation changes.

No system is entirely immune to hacking. For this reason, each administrator should accept responsibility for possible security breaches and take quick steps to correct problems. Also, they should cooperate with law enforcement and regulators in the event of serious incidents. Moreover, administrators should develop a robust incident response plan to respond to security breaches quickly and ethically. Also, data analysis processes should be audited for ethical implications.

The following recommendation for IoT systems concerns users' education. Users should be regularly educated about the potential threats of IoT devices and how to protect themselves from these threats. Also, users should receive clear instructions on how to safely configure and use IoT devices, including encouraging users to change default passwords and settings during setup and information about IoT risks and their security rights and responsibilities. The staff responsible for managing IoT systems need training in the ethical and responsible use of technology.

As mentioned, many IoT systems use an Artificial Intelligence methods. These techniques always raise questions about their fairness, unbiasedness and transparency. Thus, if AI is involved, we must make sure ethical AI principles such as honesty, transparency and accountability are followed. We must be beware of algorithmic biases that can lead to unfair or discriminatory results.

The following recommendation refers to the youngest users of IoT systems. If we consider smartphones part of IoT systems or networks, we will observe how many dangerous situations are waiting for their users. First, there will be problems with privacy and data storage because these devices store a lot of unnecessary data about users. Second, the youngest users usually do not know how the Internet and smartphones work, so they grant too much permission to applications, and these applications can share users' data. Each data leakage (even permitted by user) for the youngest users can cause depression, self-harm and even suicide in a child or teenager. For this reason, parents should implement appropriate controls on children's devices, including installed applications, granted permissions, websites visited, etc.

However, communication between devices is the most challenging process in an IoT system. The communication process touches many IoT systems' issues and is responsible for them. Each data leakage is associated with its transfer via the network using the security protocol. We can indicate many types of protocols, dividing them into uses, for example, protocols that establish connections in the network, protocols that only send data between devices or protocols that establish security keys or login parameters. We must remember that if the protocol is not secured, attackers can break it and steal data.

The security of protocols is based mainly on encryption algorithms (symmetric or asymmetric) and other techniques like hashing functions, pseudonymity, or elliptic curve cryptography. The appropriately designed security protocol will improve data security in IoT systems.

Regarding secure communication and security protocols, we should choose the appropriate protocol for our solution, considering the security needs and the whole system's performance. We cannot use demanding algorithms if our system or devices have limited computing power. Also, we should consider data stored in our system or sent between devices. In most typical IoT systems, we can significantly minimize the amount of stored and transmitted data only to those necessary for the system's functioning. In the case of mentioned earlier medical systems, primarily, we cannot minimize the amount of data because, in the medical history, each piece of information can be critical. So, when choosing a security protocol for an IoT solution, it is necessary to identify needs and requirements connected with users' data and privacy, security and system performance.

Ensuring ethics in the security of Internet of Things systems is critical because IoT significantly impacts our everyday lives and can potentially violate people's privacy, security and rights. Ensuring ethics in the security of IoT systems requires cooperation between manufacturers, users, government organizations and society. It is crucial to evolve these standards and practices as IoT technologies evolve to ensure they continue to meet the highest ethical and security standards.

Let us summaries considered opportunities, challenges and recommendations.

Collecting, processing and storing IoT data should fully respect users' privacy. Personal data should be appropriately secured and protected against unauthorized access (for example, using specially designed security protocol). Users should be aware of what data is collected by IoT devices and for what purposes it is used. Companies should provide clear privacy policies. Manufacturers of IoT devices should ensure a high level of security, which means regular software updates, patching vulnerabilities and using strong authentication mechanisms. IoT devices should be configured securely by default, and users should be encouraged to change default passwords and settings. IoT device design should consider ethical aspects such as minimizing collected data, preventing discrimination and respecting consumer rights. Companies should avoid designing devices that can be used for unethical purposes, such as unauthorized monitoring or spying. Education should be provided to IoT consumers and users about security and privacy risks in IoT and how to protect themselves against them. Organizations and manufacturers should also invest in IoT security research and share information on best practices. Government authorities should create appropriate regulations regarding security and ethics in IoT and monitor companies' compliance with them. Companies that do not comply with regulations should face appropriate consequences. Also, Tzafestas in (Tzafestas, 2018) highlighted that ethics and legislation of Internet of Things should regularly evolve and adapt to changing trends and technological developments.

By following these recommendations, we can create and maintain IoT systems that prioritize ethics, privacy, and security, strengthening trust among users.

5. CONCLUSION

The Internet of Things is a network of connected physical devices that use sensors to measure their environment and exchange data using the Internet. These devices are used in various sectors. IoT solutions also use tracking sensors to protect security and warn people about potential earthquakes. However, IoT solutions face two main challenges: data, user and device security.

In this paper, we discussed the challenges and opportunities of IoT systems, focusing on their potential benefits, network security challenges, and ethical issues related to data storage and communication.

Also, it emphasized the need for ethical regulation in the context of cyberattacks and the need for security protocols that address vulnerabilities and protect users from cyberattacks.

First, we identified that the ethical issues surrounding IoT systems security are primarily related to data collection, storage, and privacy. Data collection ethics require clear information, consent, and potential removal. Medical and healthcare data can be particularly problematic, as it can reveal sensitive information about patients' health, leading to potential depression or suicide. Data storage issues are also related to privacy and protection, as IoT systems collect vast amounts of data from various devices and sensors, potentially violating users' privacy. Monitoring cameras and IoT devices may collect private data, leading to unauthorized use, behaviour monitoring, or sales of personal data without user consent. Also, children are at risk of privacy breaches due to their unawareness to protect their online activities.

Second, we focused on recommendations for IoT systems' security. It is necessary to consider ethical practices from the beginning of system design, including data minimization storage techniques, data anonymization, and user consent. Administrators should establish guidelines for ethical data handling, ensure data fairness and transparency, and avoid discriminatory profiling. Also, they should pay attention to regular updates and patches on IoT devices and software, user education, regular audits, and cooperation with law enforcement and regulators.

At last, we noticed that communication between devices is a challenging process in IoT systems, with data leakage associated with network transfers using security protocols. Such protocols use encryption algorithms and other techniques that can improve data security in IoT systems. When choosing a security protocol, consider the system's performance and the needs of users' data. Minimizing data storage and transmission is crucial, especially in medical systems.

The development of IoT can lead to constant surveillance and monitoring, undermining privacy and individual freedom. Ethics in the security of Internet of Things (IoT) systems is crucial for protecting user privacy, integrating moral principles, and responsibly using technology. To address ethical issues, collaboration between industry, regulators, the scientific community, and users must develop rules, standards, and policies that protect privacy, ensure security, and promote responsible use of IoT technology. Presented recommendations for IoT systems will prioritize ethics, privacy, and security, strengthening user trust.

After analyzing the current state of knowledge in the field of ethics in IoT systems' security, we set ourselves further research goals. We will focus on designing and creating a secure, ethical communication protocol for IoT solutions. We will consider the mentioned recommendations to ensure ethics in IoT systems security.

REFERENCES

Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir, J., & Al-Fuqaha, A. (2022). Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. *Computer Science Review*, *43*, 100452.

Alkhwaja, I., Albugami, M., Alkhwaja, A., Alghamdi, M., Abahussain, H., Alfawaz, F., ... & Min-Allah, N. (2023). Password Cracking with Brute Force Algorithm and Dictionary Attack Using Parallel Programming. *Applied Sciences*, *13*(10), 5979.

Alsaeed, N. H., & Nadeem, F. (2022). Authentication in the Internet of Medical Things: Taxonomy, Review, and Open Issues. *Applied Sciences*, *12*(15), 7487. https://doi.org/10.3390/app12157487

Alsmadi, I., Dwekat, Z., Cantu, R., & Al-Ahmad, B. (2022). Vulnerability assessment of industrial systems using Shodan. *Cluster Computing*, *25*(3), 1563-1573.

Atlam, Hany F., and Gary B. Wills. "IoT security, privacy, safety and ethics." *Digital twin technologies and smart cities* (2020): 123-149.

Edu, J., Such, J., Suarez-Tangil, G., Bispham, M., Sattar, S. K., & Zard, C. (2023, May). Misinformation in Third-party Voice Applications. In *ACM conference on Conversational User Interfaces*. ACM.

Guan, A., & Chen, C. M. (2022). A Novel Verification Scheme to Resist Online Password Guessing Attacks. *IEEE Transactions on Dependable and Secure Computing*, *19*(6), 4285-4293.

Hameed, K., Garg, S., Amin, M. B., Kang, B., & Khan, A. (2022). A context-aware information-based clone node attack detection scheme in Internet of Things. *Journal of Network and Computer Applications*, *197*, 103271.

Imam-Fulani, Y. O., Faruk, N., Sowande, O. A., Abdulkarim, A., Alozie, E., Usman, A. D., ... & Taura, L. S. (2023). 5G Frequency Standardization, Technologies, Channel Models, and Network Deployment: Advances, Challenges, and Future Directions. *Sustainability*, *15*(6), 5173.

Khan, F., Xu, Z., Sun, J., Khan, F. H., Ahmed, A., & Zhao, Y. (2022). Recent Advances in Sensors for Fire Detection. *Sensors*, *22*(9), 3310. https://doi.org/10.3390/s22093310

Kubanek, M., Bobulski, J., & Karbowiak, Ł. (2022). Intelligent Identity Authentication, Using Face and Behavior Analysis. *ETHICOMP 2022*, 42.

Kumari, A., Kumar, V., Abbasi, M. Y., Kumari, S., Chaudhary, P., & Chen, C. M. (2020). Csef: cloud-based secure and efficient framework for smart medical system using ecc. *IEEE Access*, *8*, 107838-107852.

Lonergan, A., Moriarty, A., McNicholas, F., & Byrne, T. (2023). Cyberbullying and internet safety: a survey of child and adolescent mental health practitioners. *Irish journal of psychological medicine*, *40*(1), 43-50.

Masud, M., Gaba, G. S., Kumar, P., & Gurtov, A. (2022). A user-centric privacy-preserving authentication protocol for IoT-AmI environments. *Computer Communications*, *196*, 45–54. https://doi.org/10.1016/j.comcom.2022.09.021

de Neira, A. B., Kantarci, B., & Nogueira, M. (2023). Distributed denial of service attack prediction: Challenges, open issues and opportunities. *Computer Networks*, *222*, 109553.

Needham, R. M., & Schroeder, M. D. (1978). Using encryption for authentication in large networks of computers. *Communications of the ACM*, *21*(12), 993-999.

Pardeshi, M. S., Sheu, R., & Yuan, S. (2022). Hash-Chain Fog/Edge: A Mode-Based Hash-Chain for Secured Mutual Authentication Protocol Using Zero-Knowledge Proofs in Fog/Edge. *Sensors*, *22*(2), 607. https://doi.org/10.3390/s22020607

Paris, I. L. B. M., Habaebi, M. H., & Zyoud, A. M. (2023). Implementation of SSL/TLS Security with MQTT Protocol in IoT Environment. *Wireless Personal Communications*, 1-20.

Parthasarathy, S., Panigrahi, P. K., & Subramanian, G. H. (2023). A framework for managing ethics in data science projects. *Engineering Reports*, e12722.

Rasslan, M., Nasreldin, M., & Aslan, H. K. (2022). Ibn Sina: A patient privacy-preserving authentication protocol in medical internet of things. *Computers & Security*, *119*, 102753. https://doi.org/10.1016/j.cose.2022.102753

Regulation, P. (2018). General data protection regulation. *Intouch*, *25*, 1-5.

Seymour, W., Zhan, X., Cote, M., & Such, J. (2023, August). A systematic review of ethical concerns with voice assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131-145).

Sholla, S., Naaz, R., & Chishti, M. A. (2017, July). Incorporating ethics in Internet of Things (IoT) enabled connected smart healthcare. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 262-263). IEEE.

Singh, S., Nandan, A. S., Sikka, G., Malik, A., & Vidyarthi, A. (2022). A secure energy-efficient routing protocol for disease data transmission using IoMT. *Computers & Electrical Engineering*, *101*, 108113. https://doi.org/10.1016/j.compeleceng.2022.108113

Sivakumar, P., Sandhya Devi, R.S., Ashwin, M., Rajan Singaravel, M.M. & Buvanesswaran, A.D. (2022). Protocol Design for Earthquake Alert and Evacuation in Smart Buildings. In: Rani, S., Sai, V., Maheswar, R. (eds) IoT and WSN based Smart Cities: A Machine Learning Perspective. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-84182-9_1

Szymoniak, S., & Kesar, S. (2022). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. *Applied Sciences*, *13*(1), 404. https://doi.org/10.3390/app13010404

Szymoniak, S., & Siedlecka-Lamch, O. (2022). Securing Meetings in D2D IoT Systems. *ETHICOMP 2022*, 31.

Szymoniak, S., Siedlecka-Lamch, O., Zbrzezny, A. M., Zbrzezny, A., & Kurkowski, M. (2021). Sat and smt-based verification of security protocols including time aspects. *Sensors*, *21*(9), 3055.

Thakral, M., Singh, R. R., & Kalghatgi, B. V. (2022). Cybersecurity and ethics for IoT system: A massive analysis. In *Internet of Things: Security and Privacy in Cyberspace* (pp. 209-233). Singapore: Springer Nature Singapore.

Tzafestas, S. G. (2018). Ethics and law in the internet of things world. *Smart cities*, *1*(1), 98-120.

Ullah, S., Zheng, J., Din, N., Hussain, M. T., Ullah, F., & Yousaf, M. (2023). Elliptic Curve Cryptography; Applications, challenges, recent advances, and future trends: A comprehensive survey. *Computer Science Review*, *47*, 100530.

Yan, D., Luo, Y., Chen, X., Tong, F., Xu, Y., Tao, J., & Cheng, G. (2022). A Lightweight Authentication Scheme Based on Consortium Blockchain for Cross-Domain IoT. *Security and Communication Networks*, *2022*, 1–15. https://doi.org/10.1155/2022/9686049

Yi, F., Zhang, L., Xu, L., Yang, S., Lu, Y., & Zhao, D. (2022). WSNEAP: An Efficient Authentication Protocol for IIoT-Oriented Wireless Sensor Networks. *Sensors*, *22*(19), 7413. https://doi.org/10.3390/s22197413

Zhang, J., & Zhang, Z. M. (2023). Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*, *23*(1), 7.

Zhou, H., Wang, Z., Zhao, W., Tong, X., Jin, X., Zhang, X., Yu, Y., Liu, H., Ma, Y., Li, S., & Chen, W. (2021). Robust and sensitive pressure/strain sensors from solution processable composite hydrogels enhanced by hollow-structured conducting polymers. *Chemical Engineering Journal*, *403*, 126307. https://doi.org/10.1016/j.cej.2020.126307

HOW CAN BEST PRACTICES OF CYBERSECURITY INCLUDE ARTIFICIAL INTELLIGENCE WITHIN SMART CITIES

Sabina Szymoniak, Shalini Kesar

Czestochowa University of Technology (Poland), Southern Utah University (USA)

sabina.szymoniak@icis.pcz.pl; kesar@suu.edu

ABSTRACT

This research primarily aims to guide researchers and industries on the importance of including artificial intelligence (AI) when developing best cybersecurity practices for smart cities. No or limited research has been conducted in combining AI and cybersecurity tools in the upcoming smart cities, where technologies are collaborated and interconnected. This paper provides a significant contribution by focusing on developing: 1) theoretical assumptions for a framework that aims to improve the identification and protection of situations in healthcare; 2) specifically where threat actors threaten the health and well-being of inhabitants in a Smart city; 3) Can be used as a starting point for other contexts. Smart cities are metropolitan areas that utilise digital technology and data-driven solutions to enhance people's efficiency, sustainability, and overall quality of life. These cities employ many technologies and data sources to optimise the efficiency of infrastructure, transportation, public services, and other domains. Smart cities utilise these technologies to improve many aspects of their infrastructure. Communication in smart cities entails employing various technologies and strategies to maximise the flow of information and improve people's overall efficiency and quality of life. Smart cities ought to employ rigorous cybersecurity measures and communication protocols to bolster the security of data, users, and gadgets. Smart city efforts utilise various security measures customised to tackle unique challenges. Moreover, there is an expected increase in cellular connectivity for the Internet of Things initiatives in smart cities. The developers of AI-based systems have prioritised establishing trust in technology as a fundamental goal. The TAI-equipped systems must be assessed according to the requirements and parameters defining their features or characteristics. It begins by reviewing current literature on AI algorithms and security solutions, followed by the authors' assumptions regarding AI and smart cities. It concludes by outlining a theoretical framework that employs a network of Internet of Things (IoT) devices to establish an intelligent system for assisting ambulance services. The architecture presented here can serve as a foundation for various services aimed at mitigating hazards to human life in the context of the growing prevalence of smart cities in urban areas.

KEYWORDS: Smart cities, AI, ambulance services, risk and security.

1. INTRODUCTION

Smart cities employ many technologies and data sources to optimise the efficiency of infrastructure, transportation, public services, and other domains. Smart cities utilise these technologies to improve many aspects of their infrastructure. Communication in smart cities entails employing various technologies and strategies to maximise the flow of information and improve people's overall efficiency and quality of life. Smart cities ought to employ rigorous cybersecurity measures and communication protocols to bolster the security of data, users, and gadgets. Smart city efforts utilise various security measures customised to tackle unique challenges. In addition, Edwards & Veale (2017) state that one of AI's most prominent ethical issues with immediate ramifications is its potential to discriminate, perpetuate biases, and exacerbate existing inequalities. Because algorithms are trained on existing data, they can end up replicating unwanted patterns of unfairness due to the data they have ingested.

Moreover, there is an expected increase in cellular connectivity for the Internet of Things initiatives in smart cities. The developers of AI-based systems have prioritised establishing trust in technology as a fundamental goal. The TAI-equipped systems must be assessed according to the requirements and parameters defining their features or characteristics.

The following section reviews the current literature on AI algorithms and security solutions in the context of smart cities. Although the solutions mentioned by various researchers are effective and efficient, there seem to be gaps when it comes to developing solutions for smart cities. The authors have outlined some assumptions as a starting point when presenting the framework based on the gaps in the existing solutions, literature on the Internet of Things, and Smart Cities. The outline of the theoretical framework, the first phase of this research, employs a network of Internet of Things (IoT) devices to establish an intelligent system for assisting ambulance services. The architecture presented here can serve as a foundation for various services aimed at mitigating hazards to human life in the context of the growing prevalence of smart cities in urban areas.

The rest of this paper is organised as follows. Section 2 presents the gaps in the existing literature of security in AI within Smart cities. In Section 3 we present a theoretical framework assumptions, including network and software assumptions. Also, we will present assumptions of security protocol as a part of the proposed framework. In the third section, we will analyse and discuss the contributions of the proposed framework. The last section describes our conclusions.

2. GAPS IN THE EXISTING LITERATURE OF SECURITY SOLUTIONS IN AI WITHIN SMART CITIES

Smart cities are urban areas that use digital technology and data-driven solutions to enhance their residents' efficiency, sustainability, and overall quality of life. These cities leverage various technologies and data sources to optimise infrastructure, transportation, public services, and more. One of the key aspects and components of smart cities is infrastructure and connectivity. Smart cities use advanced infrastructure, including high-speed broadband, 5G networks, and IoT sensors. The mentioned technologies enable real-time data collection and communication among various city systems ((Joshi et al., 2016), (Internet of Things, 2022), (Tura & Ojanen., 2022)).

Smart cities use these technologies for many facilities. They include gathering sensor data (for example, about air quality) to derive insights and inform decision-making, real-time traffic monitoring, smart parking solutions, promoting renewable energy sources, or public services. Furthermore, smart cities employ surveillance and monitoring systems to enhance safety. These systems include video analytics, gunshot detection, and emergency response optimisation. Also, with increased connectivity, cybersecurity becomes a paramount concern. Smart cities must protect sensitive data and infrastructure from cyberattacks and ensure the privacy of their citizens ((Steingartner et al., 2022), (Su & Fan, 2023)).

Communication in smart cities encompasses various technologies and strategies to improve information flow and enhance residents' overall efficiency and quality of life. As communication networks collect vast amounts of data, ensuring the privacy and security of residents' information is paramount. Smart cities must implement robust cybersecurity measures and communication protocols to improve data, user and device security. Smart city approaches may use many different security protocols and are specially designed for specific solutions (Szymoniak & Kesar, 2022). We can find security protocols for specific smart cities' solutions, such as fog or edge processing (Pardeshi et al., 2022), medicine or healthcare (Rasslan et al., 2022), (Masud et al., 2022), meetings security (Szymoniak & Siedlecka-Lamch, 2022), industry (Yi et al., 2022) or suitable for many domains ((Yan et al., 2022), (Singh et al., 2023)).

Moreover, smart devices employ artificial intelligence (AI) methods to analyse collected data, causing the device to operate without human intervention or predict some situations like weather or climate changes (Rehman et al., 2023), traffic (Chen et al., 2023) or earthquake (Bhatia et al., 2023). Dependency on data and technology in smart cities will continue to increase. A recent article in IT Magazine (2023) states that data by smart cities is expected to grow by more than 140% between 2023 and 2027. More so, there will be more cellular connections in the Internet of Things projects in smart cities, which are expected to increase at a compound annual rate of 17.9% between 2022 and 2027, reaching a plateau of more than 122 million, with exceptionally high growth in the next two years. As a result, there will also be more risks from data breaches to fatal accidents on the road to minimise, manage, mitigate and transfer risks. Rescue services dependency will increase to help the injured and secure the area around the incident. Also, city dwellers can witness situations that may turn into dangerous situations, for example, when a group of people argue. In some cases, the argument may turn into a fight.

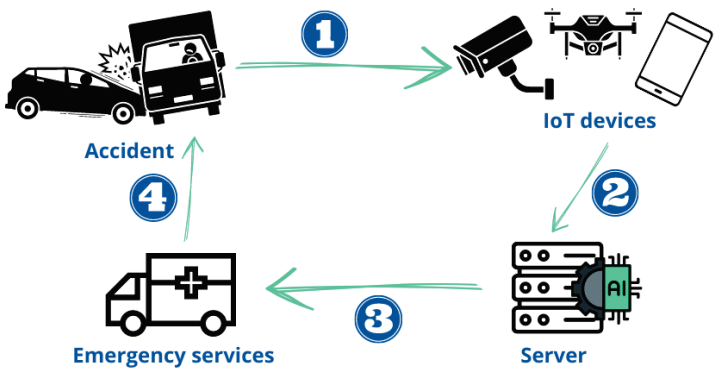
Considering that our daily lives cannot function without smart devices, we decided to use them to improve security in smart cities. This paper proposes a proof of concept as a framework that focuses on using a network of IoT devices as an intelligent system to support ambulance services, which can minimise fatality. This framework will support notifying emergency services within smart cities. After evaluating the issues with the current artificial intelligence frameworks, we provide a theoretical framework that uses AI to address such issues and offer a reliable mechanism for ambulance services in smart cities. Given the growing popularity of smart cities in metropolitan areas, this framework can serve as the foundation for several services that can aid in mitigating, minimising, managing, and transferring hazards related to human life.

3. THEORETICAL FRAMEWORK

Based on the gaps identified above, the assumption for the security protocol, part of the theoretical framework for smart cities include: 1) Based on the current technology trends, smart cities will continue to use IoT and AI for sustainability. If there are no solution for managing, mitigation, minimizing or transferring risks, the misuse of technology will create more harm than safety in the medical field; 2) IoT devices and networks are already part of monitoring systems monitor for smart city. Current research indicates that for best cybersecurity solutions it is important to combine both trusted sever's software and communication; The proposed framework's purpose is to communicate with various devices that will collect information related to the safety of smart city residents.

In light of the above, Figure 1 shows the proposed system's concept, which consists of four ingredients.

Figure 1. The architecture of the proposed system.



The network can support any number of smart devices. The trusted server's capacity or the area that is patrolled may place restrictions on this number. We assume that the municipal authorities of smart cities want to improve the safety of their inhabitants. Therefore, they configure a network of interconnected IoT devices, a trusted server and a selected device from the Emergency Notification Center (ENC). The smart city authorities may define the specific area of the city or the whole city as a monitored space. Also, they may define the specific monitoring time that will depend, for example, on crime or accident rates.

Depending on different scenarios, the system will respond accordingly. An example of an accident in a smart city and the systems's step/process is explained below.

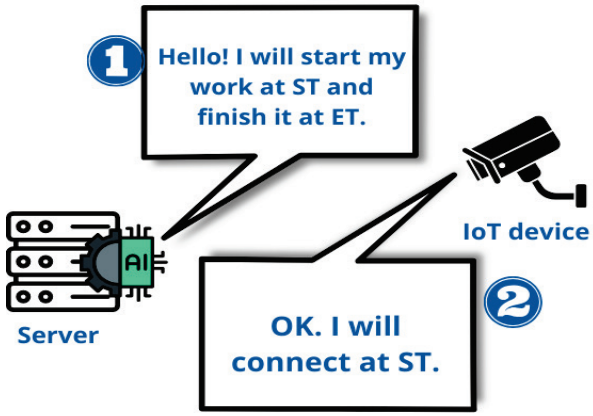
When an "accident" (see Figure 1) occurs, the first step is when the IoT device records an image of the event. In the second step, the device sends the captured photo to the trusted server via the Internet connection. The trusted server processes the obtained photo using AI-based software. Then it compares the image data to identify the level of the "dangerous" situation. This identification process is completed within minutes due to the AI tools support, which is a great advantage in saving lives. Based on the "dangerous" level, the server then notifies the Emergency Notification Center as part of the third step. The Emergency Notification Center's employees will review the image provided and allocate the right rescue services where needed in the first place (see Figure 1). This framework not only includes the gaps in current security solutions but also considers the emerging IoT tools and devices that will make the society dependent on technology for economic growth and sustainability.

Consequently, if tools and framework do not assume the unimaginable risks, the purpose of smart cities will be defeated. Subsequently, this framework includes two critical aspects: 1) Server's software and 2) Server's communication support task. The server's software allows management of the monitored territory and cameras, data collecting, communication support and verifying received photos. The server has high-performance computing power to solve each task. Managing the monitored territory and cameras requires software-enabling operations such as adding, deleting, switching on or switching off cameras. Data collecting requires a database storing information about dangerous situations in smart cities. Verifying received photos requires a specially designed artificial intelligence-based method to decide if the received photo captured a dangerous situation. The method for recognising dangerous situations will be based on Convolutional Neural Networks. The method will assess to what extent the photo shows a dangerous situation (accuracy). If the photo has a high accuracy value, the information about the situation will forwarded to the Emergency Notification Center. Whereas the second significant aspect is the server's communication support task. Each device connected to the framework will communicate with the trusted server. Also, the server will communicate with the Emergency Notification Center. Both communications should be two-way among others for this reason so that the devices know that the messages have reached the recipient correctly. Such communication should be appropriately secured, so it is necessary to prepare and use a security protocol that will be suitable for each device. The security protocol is the second significant framework component.

3.1. Protocol's broadcast phase

The broadcast phase will notify IoT devices that the framework will start its work. Depending on the circumstances or number of threats and dangerous situations in a smart city, the framework may operate for the whole day or a few hours, for example, at night. Also, the framework may monitor the whole city or only specific areas of the smart city. Thus, the trusted server must notify each device that the framework will start its work.

Figure 2. The broadcast phase of the protocol.



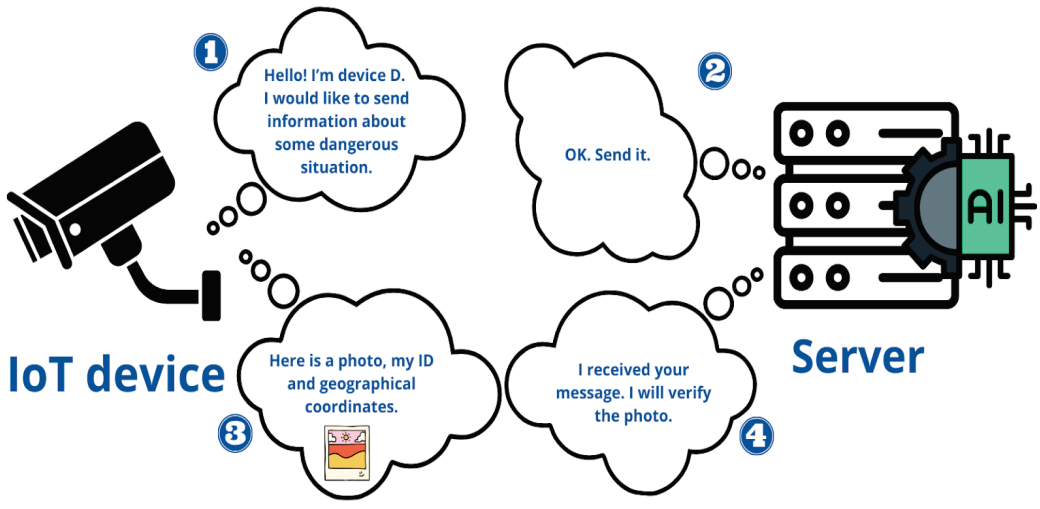
In this case, the framework will execute the broadcast phase of the protocols. This phase consists of two steps. In the first step, the trusted server sends to each device two timestamps defining the starting (ST) and ending (ET) time of the framework's operation. As a response, each device (in the second step) should resend the timestamp defining the starting time to the server. The messages in both steps are encrypted by a symmetric key shared between the device and the trusted server. Figure 2 shows the message flow in the broadcast phase.

3.2. Protocol's contact phase

The contact phase establishes a connection between the device and the trusted server. Figure 3 shows the message flow in the contact phase.

This phase consists of four steps. In the first step, the device tries to authenticate to the trusted server and report the possibility of a dangerous situation. For this reason, it sends its identifier and newly generated timestamp to the server. The trusted server checks the device's identifier in its database. If the identifier is correct, the trusted server sends the device's timestamp to inform the device that communication is possible (the second step). The messages in both steps are encrypted by a symmetric key shared between the device and the trusted server. In the third step, the device sends the captured photo, device identifier and geographical coordinates to the trusted server. Also, this message is encrypted by a symmetric key shared between the device and the trusted server.

Figure 3. The contact phase of the protocol.



Additionally, this same key encrypts the photo file and geographical coordinates to avoid cyberattacks. The trusted server confirms receiving the message from the device's timestamp. The trusted server may perform and verify the received photo and decide to notify the Emergency Notification Center if necessary.

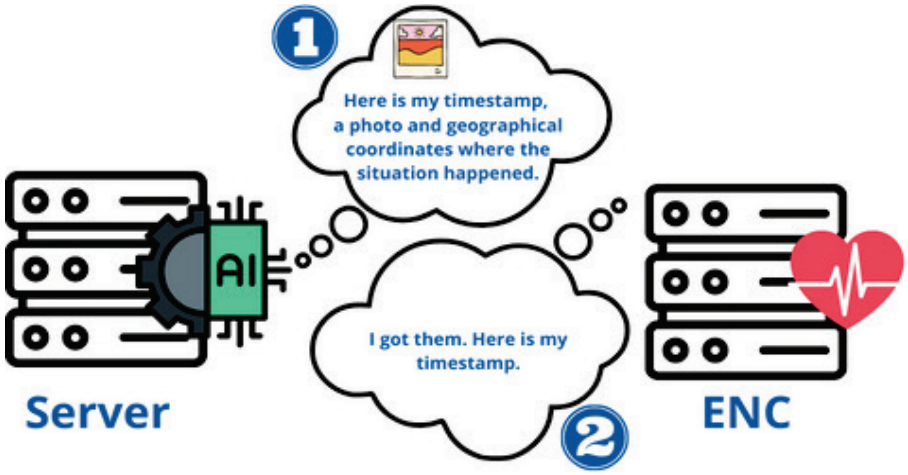
3.3. Protocol's ENC notification phase

If the trusted server decides to notify the Emergency Notification Center, it will execute the ENC notification phase of this protocol. The last phase consists of two steps.

In the first step, the trusted server sends the captured photo and the situation's geographical coordinates to ENC. Both elements are encrypted by a symmetric key shared between the ENC and the trusted server. Also, in this message, the trusted server includes its timestamp. In the second step, the Emergency Notification Center confirms receiving the message and sends the newly generated timestamp to the trusted server. Figure 4 shows the message flow in the contact phase.

Next, the Emergency Notification Center's employees will look at the photo and decide about sending rescue services to help injured victims or secure the area.

Figure 4. The ENC notification phase of the protocol.



4. CONTRIBUTION OF THE THEORETICAL FRAMEWORK

There are many contributions that are both practical and theoretical.

First, the security protocol execution in the smart city has devices as part of the hardware that has not been considered in other frameworks. For example, there is an added camera to the system and database (including its geographical coordinates) for the framework's administrator. It generates the necessary cryptographic objects like the device's identifier or a symmetric key shared between the device and the trusted server.

Second, the framework executes communication using previously prepared information. Adding each camera to the database will avoid authentication and logging processes because the system will know all devices.

Third, from a practical point of view, this framework can improve the identification and protection of situations in healthcare, especially the ones that are considered "dangerous" and required immediate medical care.

Fourth, this framework can be used as a starting point for other contexts. Smart cities are metropolitan areas that utilise digital technology and data-driven solutions to enhance people's efficiency, sustainability, and overall quality of life, specifically where threat actors threaten the health and well-being of inhabitants in a Smart city.

Fifth, this in unique framework framework were both servers software and communication for cybersecurity solutions where AI, IoT, smart cities, gaps in the current frameworks, and authors' experience in cybersecurity is taken into account to develop a framework that can minimize, mitigate, manage risks that are current or/and will be in the future.

Finally, this is phase one of this kind of research that other researchers and industries can use or/and modify for other contexts too.

The author's next phase of this continuing research will address other important factors that warrant attention. For example, they will address how while designing this framework, the ethical and privacy aspects were kept in mind. With data collection, the authors will also examine how challenges and lessons learned from the data collection phase were addressed.

5. CONCLUSION

Thus far, scientific inquiry has only concentrated on road traffic safety. Nevertheless, the occurrences close to the streets are equally crucial. Frequently, perilous circumstances arise in densely populated areas that jeopardise the well-being or survival of their residents. Primarily, it could be either syncope or altercations. Frequently, in such circumstances, medical assistance is delayed to the point where it is unable to prevent the loss of human life, mainly when such a scenario arises without the involvement of individuals who could alert the relevant authorities. The project aims to develop a framework for secure communication and timely warning of emergency services in hazardous circumstances. A recently devised security protocol will guarantee the confidentiality and integrity of communication, while a suitably equipped artificial intelligence algorithm will handle the verification process. Implementing the advanced security protocol and artificial intelligence technique will make a substantial and innovative contribution to enhancing the overall understanding of safeguarding electronic communications and utilising artificial intelligence methods to enhance the well-being and security of urban populations. The research outcomes will facilitate comprehension of the effects of integrating diverse technologies into a unified computing environment on the safety of all inhabitants of urban agglomerations.

In future work, we will continue implementing and testing our framework. Also, we will consider adding regular smartphone users to our framework, which will improve the protocol's complexity. During the framework tests, the correct operation of its two most essential elements will be checked, i.e. the system for recognising dangerous situations and the security protocol. The correct operation of the framework will be checked in the following steps, during which the number of devices connected to the network will increase: an autonomous vehicle with an installed server will move around the urban agglomeration and record dangerous situations using cameras in the vehicle, the autonomous vehicle will be joined by drones, which will also record dangerous situations using the cameras they are equipped with while flying on a specific route, the system will start registering regular users and accepting both correct and incorrect reports from them.

REFERENCES

Bhatia, M., Ahanger, T. A., & Manocha, A. (2023). Artificial intelligence based real-time earthquake prediction. Engineering Applications of Artificial Intelligence, 120, 105856.

Chen, Y., Wang, W., & Chen, X. M. (2023). Bibliometric Methods in Traffic Flow Prediction Based on Artificial Intelligence. Expert Systems with Applications, 120421.Internet of Things. (2022). In Transactions on Computer Systems and Networks. Springer Nature. https://doi.org/10.1007/978-981-19-1585-7

Joshi, S., Saxena, S., & Godbole, T. (2016). Developing smart cities: An integrated framework. Procedia Computer Science, 93, 902-909.

Lilian E. & Veale, M. 'Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?', IEEE Security & Privacy, 2017, p 2.

Masud, M., Gaba, G. S., Kumar, P., & Gurtov, A. (2022). A user-centric privacy-preserving authentication protocol for IoT-AmI environments. Computer Communications, 196, 45–54. https://doi.org/10.1016/j.comcom.2022.09.021

Pardeshi, M. S., Sheu, R., & Yuan, S. (2022). Hash-Chain Fog/Edge: A Mode-Based Hash-Chain for Secured Mutual Authentication Protocol Using Zero-Knowledge Proofs in Fog/Edge. Sensors, 22(2), 607. https://doi.org/10.3390/s22020607

Rasslan, M., Nasreldin, M., & Aslan, H. K. (2022). Ibn Sina: A patient privacy-preserving authentication protocol in medical internet of things. Computers & Security, 119, 102753. https://doi.org/10.1016/j.cose.2022.102753

Rehman, A., Tariq, S., Farrakh, A., Ahmad, M., & Javeid, M. S. (2023, March). A Systematic Review of Machine Learning and Artificial Intelligence Methods to Tackle Climate Change Impacts. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-7). IEEE.

Singh, A. K., Nayyar, A., & Garg, A. (2023). A secure elliptic curve based anonymous authentication and key establishment mechanism for IoT and cloud. Multimedia Tools and Applications, 82(15), 22525-22576.

Su, Y., & Fan, D. (2023). Smart cities and sustainable development. Regional Studies, 57(4), 722-738.

Steingartner, W., Možnik, D., & Galinec, D. (2022, November). Disinformation Campaigns and Resilience in Hybrid Threats Conceptual Model. In 2022 IEEE 16th International Scientific Conference on Informatics (Informatics) (pp. 287-292). IEEE.

Szymoniak, S., & Kesar, S. (2022). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. Applied Sciences, 13(1), 404. https://doi.org/10.3390/app13010404

Szymoniak, S., & Siedlecka-Lamch, O. (2022). Securing Meetings in D2D IoT Systems. ETHICOMP 2022, 31.

Tura, N., & Ojanen, V. (2022). Sustainability-oriented innovations in smart cities: A systematic review and emerging themes. Cities, 103716.

Voigt, P., & Von Dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). In Springer eBooks. https://doi.org/10.1007/978-3-319-57959-7

Yan, D., Luo, Y., Chen, X., Tong, F., Xu, Y., Tao, J., & Cheng, G. (2022). A Lightweight Authentication Scheme Based on Consortium Blockchain for Cross-Domain IoT. Security and Communication Networks, 2022, 1–15. https://doi.org/10.1155/2022/9686049

Yi, F., Zhang, L., Xu, L., Yang, S., Lu, Y., & Zhao, D. (2022). WSNEAP: An Efficient Authentication Protocol for IIoT-Oriented Wireless Sensor Networks. Sensors, 22(19), 7413. https://doi.org/10.3390/s22197413

"DARK PARTNERS": TRANSPARENCY OBLIGATIONS AGAINST DECEPTION IN VIRTUAL INFLUENCER MARKETING

Jacopo Ciani Sciolla

Università degli Studi di Torino (Italy)

jacopo.cianisciolla@unito.it

ABSTRACT

The paper examines the creation of realistic and visually appealing virtual influencers that take the form of hyper-realistic characters who can be nearly impossible to distinguish from real-life influencers. Since only a few national regulators provide a duty to disclose the influencer's real nature, consumers may falsely believe they are engaged in communications with humans. This information is valuable for consumers who are more likely to rely on recommendations from individuals with views and beliefs similar to their own. Based on Habermas' theory of communicative action and Kantian ethics, the overall aim is to suggest that brands must be transparent and should not engage in marketing communications referring to any virtual testimonial of products because such practice would not be based upon any bona fide use, nor personal opinions, beliefs, or experiences of human fellows.

KEYWORDS: digital marketing, dark patterns, virtual influencer, unfair commercial practices, transparency, ethics.

1. INTRODUCTION

The online environment is populated by internet companies exploiting users' psychological vulnerabilities thanks to the use of AI. The aim is to maximise profits by nudging or deceiving users into making decisions that, if fully informed, they might not make.

Such practices normally fall under the "dark patterns" umbrella term, which refers to manipulative interface design choices that negatively impact the user's decision-making, leading users to act against their interests,

The distinctive feature of dark patterns is the practice of user experience (UX) and user interface (UI) design to deceive users into accepting either unwanted purchases or subscriptions that depend on increasing levels of anxiety due to time limits and social pressure. UX and UI are conceptual design disciplines that revolve around the interaction between users and machines to shape systems and computer interfaces that address the user's experience when using a platform (Dove et al. 2017). Good UX aims to provide people with interactions that are seamless, enjoyable, and intuitive. However, UX is a tool that can be used both for good and for evil. Dark patterns are one such category of evil design and given their growing use and the ease with which they can be added to platforms (i.e., dark patterns as a service), the focus is increasingly on the understanding of these practices, consequent harms, and potential countermeasures (EU Commission 2022; OECD 2022; BEUC 2022).

Dark patterns come in many different shapes, may employ different kinds of design-based elements, and can intervene at different stages of a transaction, such as the advertising one. New forms of dark patterns are constantly emerging, with new technologies and new kinds of user interfaces. Various

regulatory measures to respond to dark patterns have been proposed or implemented and calls among the UI/UX design community to adopt ethical standards have increased[1].

This paper aims to study a phenomenon that has the same distinctive character of dark patterns and a very similar influence on end-users from a legal and socio-ethical perspective, but, until now, has not received the same attention by policymakers: the creation of realistic and visually appealing virtual influencers making consumers falsely believe that they are engaged in communications with humans. I refer to them as "dark partners" because they partner in commercial communications with brands, advocating for their products or services, without being transparent about their virtual identity.

Accordingly, the paper is structured as follows. Next, the focus is on the virtual influencer concept and some basic distinctions with real-life ones. The section illustrates why brands are increasingly using them. Section 3 highlights the risk that consumers may confuse virtual influencers with human beings and clarifies why this could be relevant for the success of an ad strategy. Section 4 analyses current law and stresses that only a few national regulators have established a duty to disclose the influencer's virtual nature. Section 5 draws on Habermas' theory of communicative action and Kantian ethics to assess this practice. The conclusions provide recommendations for policymakers.

2. VIRTUAL INFLUENCER MARKETING: THE REASONS FOR A SUCCESSFUL STORY

Influencer marketing involves the promotion of specific brands or products through influencers using the positive impact they are likely to have on consumer perceptions.

There is no legal definition of an "influencer" enshrined in EU law. However, the UCPD guidelines (EU Commission 2021) specify that "an influencer is generally described as a natural person or virtual entity who has a greater than average reach in a relevant platform" (sec. 4.2.6).

Influencers are usually known for being experts in particular topics (e.g., travel, lifestyle) and for creating content carrying different values for consumers (Audrezet et al. 2020), including educational, entertainment, or advertising, often simultaneously. In addition to their content-making skills, influencers' popularity is further driven by their ability to build strong connections and trust with consumers by revealing personal information that consumers can relate to (Penttinen et al. 2022); in a word, by making their life 'transparent' or 'public'.

Traditionally, brands collaborate with real-life influencers (i.e., humans living in a physical world) who can make their own decisions regarding sponsored collaborations with brands and form opinions about the products and services they promote.

With recent technological developments, brands increasingly started to work with virtual influencers.

2.1. The notion of "virtual influencer"

Virtual influencers are non-human digitally created characters sharing social media content and engaging in interactive communications to obtain influential status among consumers. Within this wide category, experts distinguish between those created with computer-generated imagery technology (CGI influencers) and AI influencers that rely on AI technologies in creating content and interacting with consumers.

Virtual influencers can have different forms. Some authors developed a taxonomy based on their similarity to human appearance, also known as anthropomorphism (Mende et al. 2019), and their placement on the reality-virtuality continuum (Hudson et al. 2019), ranging from unimaginable

[1] https://www.design.org.au/code-of-ethics/dia-code-of-ethics.

characters to hyper-realistic characters that can be nearly impossible to distinguish from humans (Mouritzen et al. 2023). Like real-life influencers, hyper-realistic human virtual influencers share content about their personal and social lives, which often features them in the physical world performing human tasks, including attending fashion shows and commercial photoshoots. Consider Lil Miquela who claims to be a 19-year-old AI robot with a passion for social justice, fashion, music, and friendship. Currently, Miquela has over 190,000 monthly listeners on Spotify and gives interviews at major events (Savageaux 2022). She has been featured in campaigns by Calvin Klein and Prada.

Accordingly, virtual influencer marketing can also be classified as mixed reality because it allows to mix objects from both physical and virtual worlds and makes the boundaries between the real and virtual world blurred.

It is worth noting that some digital characters appear to exist only in a virtual world, while some others are avatars of real-life celebrities (Arsenyan-Mirowska 2021). In September 2023, Meta launched 28 AI-powered chatbots featuring Kendall Jenner (Billie), Paris Hilton (Amber), and Snoop Dogg (Dungeon Master). Currently, they are only available for testing in the US but AI shall make celebrities, shortly, omnipresent, since they can penetrate every market and format at any time. Most of the time AI clones of celebrities, grabbing user's attention on YouTube, are just scams relying on AI voice cloning paired with decontextualized video of the celebrity[2].

Even if the language in advertising distinguishes between influencers, endorsers, celebrities, or ambassadors, for the sake of simplicity, I will use the term virtual influencers to refer in general to digital characters that represent brands in digital advertising, irrespective of the degree of reputation they enjoy with the public. Therefore, I will leave aside from my analysis virtual assistants and chatbots acting on the brand's website to provide consumers support: they are more readily recognisable as such.

2.2. Unique features and advantages

Similar to real-life ones, virtual influencers can establish relationships and engage with a large number of consumers on social media (Hugh et al. 2022). Customization allows for designing virtual influencers with attributes that appeal to specific target consumers or fit the values and image of the promoted brands (Conti et al. 2022). Notably, their visual appearances and behaviours can be modified following changes in market trends and evolving consumer preferences.

Marketing collaborations with virtual influencers are also likely to require less time and financial resources in comparison to working with humans (Arsenyan-Mirowska 2021). Because of their digital nature, virtual influencers do not have physical constraints and can be anywhere, anytime (Conti et al. 2022). This means that consumers can potentially interact with the same virtual influencer across different digital platforms simultaneously.

A key issue is that they are not subject to reputational damage. Just in case, they can simply be deleted, and companies can create a new one. Thus, in comparison to working with real-life influencers, collaborations with virtual influencers assume lower risks related to involvement in scandals and unethical behaviours (Guthrie 2020).

[2] https://www.404media.co/joe-rogan-taylor-swift-andrew-tate-ai-deepfake-youtube-medicare-ads/?mc_cid=f250e2b063&mc_eid=f720a42bfb.

2.3. The social response theory

Even though virtual influencers do not exist in real life, several studies show they are perceived as authentic, regarding their physical appearance, personality and behaviour (Moustakas et al. 2020). This is coherent with the social response theory (Festinger 1954), according to which when consumers come across virtual influencers, they engage with them as they do with real-life ones, by applying the same social rules of interactions with humans, though they know virtual influencers are not humans (Moon 2003). Hence, as long as consumers will respond to virtual influencers as they do to real-life ones, it is not surprising that the former are capable of being preferred to humans.

3. RISKS OF CONSUMER DECEPTION AND MISINFORMATION

Virtual influencers may have many attractive features for brands, but they also raise some concerns.

3.1. Distinguishing virtual from real-life influencers

As virtual influencers are designed to have human-like features and behaviours, these brand's commercial partners might be particularly difficult to distinguish from real-life ones. Much like deepfakes, the rise of "dark partners" highlights our inability to distinguish reality from fabrications.

Many warn of the serious consequences when we can no longer trust (Durante 2011) any of the information we consume. The prevalence of fake presences may eradicate our sense of reality in the virtual realm. AI-powered virtual influencers may be purposefully designed for or tricked into (e.g., by untruthful or low-quality online data) spreading misinformation and other unethical communications (Mustak et al. 2022). The background stories of virtual influencers, the content they share, and, most importantly, their visual appearance can create false representations in society, like unrealistic perceptions of beauty standards (Gill 2023). This can be problematic, i.e., consumers having difficulties distinguishing virtual from human influencers (Franke et al. 2023), as these consumers do not realize that they are comparing themselves to a non-human and may feel anxious about the way they look, to the point of inhibiting their ability to live well (Deng-Jiang 2023).

This risk gets worse when influencers are involved in marketing activities. As consumers are more likely to rely on recommendations from individuals that have views and beliefs similar to their own, making consumers falsely believe they are engaged in communications with humans, might suspend consumer's abilities to identify and critically evaluate persuasive marketing tactics.

3.2. The commercial risk of disclosing virtuality

Some virtual influencers are transparent about their virtual identity. However, this is not always the case. The reason is that disclosing it may negatively affect the effectiveness of the communication.

The source credibility model shows that influencers' perceived characteristics may impact their trustworthiness, expertise, and attractiveness and affect the desired results of their messages (Ohanian 1991). Having low source credibility, influencers will lose the ability to engage consumers with sponsored posts. In this regard, knowing exactly the human or non-human nature of an influencer is pretty relevant. The outcomes of marketing research show that anthropomorphism increases brand liking and purchase intentions, while disclosing virtuality may lead people to feel uncomfortable or become more suspicious of persuasion attempts (Woodroof et al. 2020).

Thus, disclosing virtuality in commercial communications may lead to lowered brand trust and attitudes, lower purchase intentions (van Reijmersdal et al. 2016), and engagement (Boerman 2020).

4. A REVIEW OF THE LEGAL FRAMEWORK

Currently, there are very few legislations providing guidance on what brands should do to avoid misleading consumers about the real nature of influencers. While there has been some initial research on virtual influencers in law and ethics, such studies are largely descriptive, mostly documenting the existence of these practices. The main objective of this section is to address this gap and design a legal and ethical benchmark to support a clear understanding of the lawful or unlawful nature of virtual influencers' marketing.

4.1. Jurisdictions specifically addressing the virtual influencers practice.

The first and, at the time of writing, only jurisdiction establishing that virtual influencers "must additionally disclose consumers that they are not interacting with a real human being" is India. Following the Consumer Protection Act of 2019, the Advertising Standards Council of India (ASCI) in 2021, became the first national regulator to require an "upfront and prominent" disclosure of this kind (Patnaik 2021)[3].

India has been followed by France. The Influencers Act, which came into effect on 1 June 2023, supplements the pre-existing regulations on advertising establishing that content with altered or artificially intelligent images must be accompanied by statements such as "virtual images" in order to limit the psychological impact on the public.

In the U.S., the Federal Trade Commission released an updated version of the Endorsement Guides[4], which makes clear that brands may be held liable for virtual influencer's unfair commercial practices as it happens for human endorsers. This means that virtual influencers should avoid making statements implying their humanity or a personal experience with the product.

Notwithstanding that, the Guides do not provide for any duty of disclosing virtual identity and some scholars have already recommended to fill this gap (Masteralexis 2021).

It is also noteworthy to say that the state of California recently introduced a ban from using avatars in political communication, but this is valid only "within 60 days of an election"[5].

4.2. Virtuous examples of self-regulation

Platforms are also taking matters into their own hands. For example, TikTok updated its platform guidelines to require that synthetic or manipulated media that shows realistic scenes be clearly disclosed. This can be done using a sticker or caption, such as "synthetic", "fake", "nor real", or "altered". The guidelines require disclosure to be directly in the videos, not just in the virtual influencer's bio[6].

4.3. Omission of material information under the Unfair Commercial Practices Directive

Obligations to disclose virtual identity do not exist at the EU level. However, any BtoC practice that materially distorts or is likely to distort the economic behaviour of an average consumer normally amounts to a misleading practice, as regulated by the Directive 2005/29/EC (Art. 6) concerning unfair

[3] ASCI (2021). The Code for self-regulation of advertising content in India. See Article 1.4 of the Guidelines for influencer advertising in digital media providing that "A virtual influencer must additionally disclose to consumers that they are not interacting with a real human being. This disclosure must be upfront and prominent".

[4] 16 CFR Part 255: Guides Concerning the Use of Endorsements and Testimonials in Advertising.

[5] A.B. 730, 2019 Leg., Reg. Sess. (Cal. 2019).

[6] https://www.tiktok.com/creators/creator-portal/en-us/community-guidelines-and-safety/ai-generated-content-label/

commercial practices (UCPD). Misleading practices could be either by actions or by omissions. Articles 7(1) and (2) establish a positive obligation on traders to provide all the 'material information' that the average consumer needs to make an informed purchasing decision.

The UCPD does not define 'material information'. However, by way of interpretation, it is possible to argue that the virtual nature of an endorser should be considered as such.

Relevant to this purpose is Article 7.4 which mandates to disclose "whether the third party offering the products is a trader or not" even if limited to the specific case of an 'invitation to purchase' on online marketplaces. Furthermore, in the Wathelet case[7], the Court stressed that "it is essential that consumers are aware of the identity of the seller". The same principle should be extended to an influencer as a person 'acting in the name of or on behalf of a trader'. Indeed, the EU Commission clarified that for the purposes of the UCPD, an influencer may be qualified as a 'trader' (EU Commission 2021). Consequently, the obligation to be clear about the identity concerns directly all persons that carry out promotional activities towards consumers on behalf of a trader.

4.4. The duty to disclose the commercial intent of a commercial practice

Other arguments in favour of a disclosure duty may be derived by analogy with other information requirements established directly by EU law[8].

First, Article 6(a) of the e-Commerce Directive[9], Articles 9, 10 and 28(b) of the Audiovisual Media Services Directive[10], similarly to Article 7(2) UCPD, establish that failing to identify the commercial intent of a practice is regarded as a misleading omission.

Coherently, the driving force behind all influencer marketing regulation adopted by advertising self-regulation authorities is the principle that influencers must disclose when they have a material connection with brands they promote through clear and understandable disclaimers such as #ad or #sponsored (Ciani-Tavella 2017).

Second, the UCPD prohibits as misleading by default "falsely claiming or creating the impression that the trader is not acting for purposes related to his trade, business, craft or profession or falsely representing oneself as a consumer".

The purpose of this information requirement is to make sure that consumers always understand the very nature of the communication and know with whom they are interacting online. The same transparency interest exists when consumers are facing virtual influencers.

4.5. Transparency obligations under the Consumer Rights Directive and the Digital Services Act

EU law is generally averse to any form of hidden marketing.

Article 8(5) of the Consumer Rights Directive[11], in the case of telemarketing, held the trader to "disclose the identity and, where applicable, the identity of the person on whose behalf he makes that call, and the commercial purpose of the call" (Luzak 2015).

[7] EUCJ, 9 November 2016, Sabrina Wathelet v Garage Bietheres & Fils SPRL, C-149/15, para 37.

[8] Article 7(5) UCPD clarifies that 'information requirements established by EU law in relation to commercial communication, including advertising', shall be regarded as material information by "default".

[9] Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

[10] Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation, or administrative action in Member States concerning the provision of audiovisual media services.

[11] Directive 2011/83/EU on consumer rights.

In the same direction, social media platforms have recently seen their transparency obligations reinforced by the Digital Services Act[12] (Cauffman-Goanta 2021). Recital 68 states that "providers of online platforms should…be required to ensure that the recipients of the service have certain individualised information necessary for them to understand when and on whose behalf the advertisement is presented". Article 26 establishes that for any advertisement presented, the recipients should be "able to identify, in a clear, concise and unambiguous manner and in real time, the following:…(b) the natural or legal person on whose behalf the advertisement is presented; (c) the natural or legal person who paid for the advertisement if that person is different from the natural or legal person referred to in point (b)"[13].

Based on these principles, it is not surprising that the European Consumers Organisation (BEUC 2023) recently supported the introduction of two disclosure obligations regarding "edited" or "altered" content (e.g. when a picture has been photoshopped), and "virtual picture" or content for virtually created images (via AI for instance).

4.6. Transparency obligations under the Proposal for an Artificial Intelligence Act

In all the AI ethics charters and guidelines (Jobin et al. 2019) it is stressed the need to understand the decision-making processes of AI (Floridi et al. 2018) and the importance of a transparency principle (Pagallo-Durante 2022), articulated as the duty to make an object or entity knowable (Hayes (2020).

Following these recommendations, specific obligations to disclose AI influencers (see supra para. 2.1) shall enter into EU law when the AI Act is adopted, after the provisional agreement between the Parliament and the Council reached on 9 December 2023[14].

The new rules establish the general principle according to which users should be made aware when they are interacting with AI, including systems that generate or manipulate image, audio, or video content, for example, deepfakes. In particular, Article 52(1) establishes that "providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system"[15]. The same is valid under para 3 for "an AI system that generates or manipulates image, audio, or video content that appreciably resembles existing persons…and would falsely appear to a person to be authentic or truthful (deep fake)". As clarified by Recital 70, the purpose of such obligation is to "*take account of the specific risks of manipulation*".

The user is requested to "disclose that the content has been artificially generated or manipulated", but this is not an absolute requirement. Natural persons should be notified that they are interacting with an AI system "unless this is obvious from the circumstances and the context of use".

This approach is however problematic because it leaves any evaluation to the user and adds a layer of subjectivity and uncertainty. Then the exception does not seem consistent with the rationale followed by the same EU legislator when establishing the duty to disclose the advertising nature of ad-contents (see *supra* para. 4.4). In that case, the notice that a content has been sponsored must be given even if it could be inferred by the context, for example because the sponsor company name is part of the message. The European Committee of the Regions pointed out this aspect in its Opinion over the AI

[12] Regulation (EU) 2022/2065 on a Single Market For Digital Services.

[13] Providers of online platforms are also requested to "provide recipients of the service with a functionality to declare whether the content they provide is or contains commercial communications".

[14] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain union legislative acts, COM/2021/206 final.

[15] This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate, and prosecute criminal offences, unless those systems are available for the public to report a criminal offence.

Act proposal[16], inviting to remove the exception on the grounds that "natural persons should always be duly informed whenever they encounter AI systems and this should not be subject to interpretation of a given situation. Their rights should be guaranteed at all times in interactions with AI systems". Even if the final wording of the article has been modified after the amendments approved by the Parliament[17], the exception is likely to be maintained.

At least, the definitive text is expected to better clarify what "notice" actually means. According to the revised text of para. 3 "Disclosure shall mean labelling the content in a way that informs that the content is inauthentic and that is clearly visible for the recipient of that content. To label the content, users shall take into account the generally acknowledged state of the art and relevant harmonised standards and specifications". These references shall probably consent to adopt disclaimers in the form of hashtags followed by the notice, as widely used for disclosing the commercial nature of ad-contents, as well as to apply existing advertising self-regulations code of conduct, particularly the European Advertising Standard Authority (EASA) Best Practice Recommendation on Influencer Marketing and the similar provisions contained in local codes and national guidelines.

This obligation does not wave the EU institutions to set a similar obligation also for virtual influencers which does not technically amount to an AI system, but still may perfectly resemble a human and pose specific risks of deception.

5. LANDMARKS FOR AN ETHICAL ASSESSMENT

The assumption for establishing transparency obligations for virtual influencers is that virtual reality is fictional or illusory so what goes into it is not truly real. Many philosophers embrace this so called virtual fictionalism (Bateman 2011), according to which things that are supposed to happen in virtual words do not happen in reality.

The opposite virtual realism theory (Chalmers 2017) claims that virtual experiences are as valuable as non-virtual practices and are non-illusory. On this basis, the virtual realism theory is irreconcilable with any kind of discriminatory obligation imposed only on virtual influencers and not also on real-life ones.

This does not mean that virtual reality should be confined to something we can only perceive and describe. It is primarily something we can interact with and based on this, artificial agents can produce real consequences, including negative ones, to the human agents with which they interact (Durante 2020).

This work provides a concrete example and application of this theory, showing how the interaction between virtual influencers and consumers in digital environments, if does not transparently take place, is morally objectionable (Carson et al. 1985). Habermas' critical social theory and Kantian ethics provide marketing communication researchers with a valuable theoretical perspective to impose information duties on such practices.

The basic idea of Habermas' pragmatic theory of meaning is that "we understand a speech-act when we know what makes it acceptable" (Habermas 1984): to understand what a speaker means the hearer has to have access to the reasons for the speaker's utterance. The idea is that agents, through their

[16] Opinion of the European Committee of the Regions — European approach to artificial intelligence — Artificial Intelligence Act (2022/C 97/12).

[17] The final draft should be written as follow "Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that the AI system, the provider itself or the user informs the natural person exposed to an AI system that they are interacting with an AI system in a timely, clear and intelligible manner, unless this is obvious from the circumstances and the context of use".

speeches, make validity claims, i.e. a claim to truth, to rightness, and to truthfulness, sincerity, or authenticity, on which the meaning of their statements depends. These claims are inherent in all communicative actions. Therefore, the meaning of an utterance should not be found in the sentence itself, as per the encoding/decoding paradigm of language, but in the intent of the speaker and the reconstruction by the receiver, as per the intentionalist and dialogic paradigms (Kraus 1998). By loading language with intent, the speaker exercises an illocutionary force that can effectively change the hearer's mind (Austin 1962) and eventually attain cooperation based on a shared understanding of reality. This process of interaction between subjects capable of speech and action that establishes interpersonal relations is that which Habermas calls communicative action. Such action can be triggered by potentially many different types of illocutionary forces, but especially by virtue of "shared knowledge, mutual trust, and accord with one another".

Virtual marketing challenges this view because the utterances of virtual influencers do not imply validity claims, nor are trustworthy. The truthfulness and authenticity of recommendations shared by these influencers are highly questionable (Lou et al. 2023). In the end, virtual influencers can neither have actual experiences with products and services nor form their own opinions (Conti et al. 2022). Therefore, their endorsement of the product or service is in no way based upon its *bona fide* use, nor it is based upon personal opinions, beliefs, or experiences with or about the product or service because the virtual influencer has never used them.

As a result, without duties of disclosure, users are not enabled to detect and analyze distorted communications. This lack of disclosure impacts the "public sphere" (Habermas 1962; 1992), namely, the public arena wherein every democratic discourse is institutionalised. Since such a role, today is arguably played in large part by the Internet and social media, it seems fair to admit that placing boundaries to misleading communications plays a crucial role in the healthy functioning of a democracy.

Kantian ethics leads to a very similar conclusion. Within a Kantian system, businesses should not be able to use promotions that are dishonest or hide key facts (Perni 2023). The assumption hinges on Kant's hypothetical publicity test in the second appendix to his Perpetual Peace. Even if specifically applied to the political sphere, the formula "all actions that affect the rights of other human beings, the maxims of which are incompatible with publicity, are unjust" (Kant 1795) has a normative and transcendental structure valid in any context concerning people interactions and coexistence (Pirni 2022). On this basis, we can critically assess current regulations and determine whether and to what extent they should be amended in accordance with duties of disclosure, publicity, and transparency.

6. CONCLUSIONS

Based on this ethical evaluation and review of the current legislation, I suggest that brands need to be transparent about using virtual characters in their communications through disclaimers. I advise that when using virtual influencers in advertising, brands should disclose this information because material to the purposes of the UCPD.

The question now is whether such disclosure is enough. My opinion is that whereby commercial communications contain a testimonial or endorsement of a product or service by the virtual influencer, a mere disclosure of virtuality does not prevent consumers' deception. Indeed, in this case, the communication would not be, by definition, genuine, verifiable, and relevant, for the same reasons as those for which the communication lacks truth and authenticity under the Habermas' validity claim test.

It is relevant in this connection that the EU Omnibus Directive prohibits fake reviews and endorsements (such as 'likes' on social media) of products and requires platforms to verify their authenticity and take reasonable and proportionate steps to ensure that these reviews are genuine and reflect the experience of real consumers. The Directive also establishes that traders giving access to such reviews should clearly state how the reviews are obtained and checked, and how they ensure that these come from consumers who have used or purchased the product (Durovic 2022).

The same approach guided advertising self-regulation authorities[18] to require that marketers must held documentary evidence that a testimonial or endorsement used in a marketing communication is genuine, i.e. that the quote is from a real person, and it reflects what this said.

On these grounds, testimonials or endorsements by virtual influencers should be banned because of their lack of authenticity, while, in any other cases, the law should require a mandatory disclosure of their virtual nature. This kind of obligation is already placed by the AI Act Proposal for virtual influencers which do technically amount to an AI system and should be complemented by a similar duty of disclosure for those who are not AI-generated. These may still perfectly resemble a human and pose similar risks of impersonation or deception.

REFERENCES

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human-Computer Studies*, 155(102694).

Audrezet, A., de Kerviler, G., & Moulard, J. G. (2020). Authenticity under threat: When social media influencers need to go beyond self-presentation. *Journal of Business Research*, 117, 557-569.

Austin, J.L. (1962). How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955, Harvard University Press.

BEUC (2023). From Influence to resposibility. Time to regulate influencer marketing, BEUC-X-2023-093.

BEUC (2022), "Dark Patterns" and the eu consumer law acquis. Recommendations for better enforcement and reform.

Bateman, C. (2011). Imaginary Worlds. Zero Books.

Boerman, S. C. (2020). The effects of the standardized Instagram disclosure for micro- and meso- influencers. *Computers in Human Behavior*, 103, 199-207.

Callahan, K.. (2021). Cgi social media influencers: are they above the ftc's influence?. *Journal of Business and Technology Law*, 16(2), 361-386.

Carson, T.L.; Wokutch, R.E.; Cox, J.E (1985), An Ethical Analysis of Deception in Advertising, *Journal of Business Ethics*, 1985, 4(2), 93-104.

Cauffman, C.; Goanta, C. (2021). A new Order: The Digital Services Act and Consumer Protection. European Journal of Risk Regulation, 12(4), 758-774.

Chalmers D.J. (2017). The Virtual and the Real. *Disputatio* 9(46), 309-352.

Ciani, J., Tavella, M. (2017) La riconoscibilità della natura pubblicitaria della comunicazione alla prova del digital: native advertising tra obbligo di disclosure e difficoltà di controllo. *Informatica e diritto*, 2(1), 485.

Conti, M., Gathani, J., & Tricomi, P. P. (2022). Virtual Influencers in Online Social Media. *IEEE Communications Magazine*, 60(8), 86-91.

[18] As the ICC under Article 13 Advertising and Marketing Communications Code or the UK Advertising Standard Authority – ASA under rule 3.45 of the UK Code of Non-broadcast Advertising and Direct & Promotional Marketing - CAP Code.

Deng, F., & Jiang, X. (2023). Effects of human versus virtual human influencers on the appearance anxiety of social media users. *Journal of Retailing and Consumer Services*, 71, 103233.

Dove, G., Halskov, K., Forlizzi, J., Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17. New York, NY: Association for Computing Machinery. pp. 278–288.

Durante, M. (2011). The Online Construction of Personal Identity through Trust and Privacy. *Information*, 2, 594-620.

Durante, M. (2020). Technology and the Ontology of the Virtual. In S. Vallor (ed.) *The Oxford Handbook of Philosophy of Technology* (pp. 318–340). New York, NY: OUP.

Durovic, M., Kniepkamp, T. (2022). Good advice is expensive – bad advice even more: the regulation of online reviews. Law, innovation and technology, 14(1), 128-156.

EU Commission (2021). Guidance on the interpretation and application of Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market, C/2021/9320.

EU Commission (2022) Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation. Final Report

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, 7(2), 117- 140.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E. (2018). AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds and Machines, 28, 689-707.

Franke, C., Groeppel-Klein, A., Müller, K. (2022). Consumers' Responses to Virtual Influencers as Advertising Endorsers: Novel and Effective or Uncanny and Deceiving? *J. of Advertising*, 1-17.

Gill, R. (2023). Perfect. Feeling Judged on Social Media. New York, NY: Polity.

Gunawan, D.D., Huarng, K.-H. (2015). Viral effects of social network and media on consumers' purchase intention. *Journal of Business Research*, 68, 2237-2241.

Guthrie, S. (2020). Virtual influencers: More human than humans. In S. Yesiloglu & J. Costello (Eds.), Influencer Marketing. Building Brand Communities and Engagement (pp. 271- 285). Routledge.

Habermas, J. (1992). Further reflections on the public sphere. Habermas Public Sphere, 428.

Habermas, J. (1962). The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society, MIT Press.

Habermas, J. (1984), The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society, Beacon Press, Boston, MA, 297.

Hayes, P. (2020). An ethical intuitionist account of transparency of algorithms and its gradations. *Business Research*, 13, 849–874

Hudson, S., Matson-Barkat, S., Pallamin, N., & Jegou, G. (2019). With or without you? Interaction and immersion in a virtual reality experience. *Journal of Business Research*, 100, 459-468.

Hugh, D. C., Dolan, R., Harrigan, P., & Gray, H. (2022). Influencer marketing effectiveness: The mechanisms that matter. *European Journal of Marketing*, 56(12), 3485-3515.

Jobin, A., Ienca, M., Vayena, E. (2019). The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389–399.

Kant, I. (1795), Zum ewigen Frieden, in Kants Werke, Berlin: Prussische Akademie Ausgabe, vol. VIII, 1923, 381.

Krauss, R.M. & Chiu, C.-Y. (1998). Language and Social Behavior. In Handbook of Social Psychology 4th edn.,Vol. 2, McGraw-Hill.

Lou, C., Kiew, S.T.J., Chen, T., Lee, T.Y.M., Ong, J. E. C., & Phua, Z. (2022). Authentically Fake? How Consumers Respond to the Influence of Virtual Influencers. *Journal of Advertising*, ahead-of-print, 1-18.

Luzak, J. (2015) Online Disclosure Rules of the Consumer Rights Directive: Protecting Passive or Active Consumers?. *J. Eur. Consumer & Mkt. L.*, 4(3), 79.

Masteralexis, J., McKelvey, S., Keevan, S. (2021). #IAMAROBOT: Is It Time for the Federal Trade Commission to Rethink Its Approach to Virtual Influencers in Sports, Entertainment, and the Broader Market?, *Harvard Journal of Sports & Entertainment Law*, 2021, 12, 353-386, 376-377.

Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, 56(4), 535–556.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology*, 13(1-2), 125-137.

Moustakas, E., Lamba, N., Mahmoud, D., & Ranganathan, C. (2020). Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers. *International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020*, 1-6.

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154.

Mouritzen, S. L. T., Penttinen, V., & Pedersen, S. (2023). Virtual influencer marketing: the good, the bad and the unreal. *European Journal of Marketing,* Vol. ahead-of-print No. ahead-of-print.

OECD (2022). Paper on Dark Commercial Patterns. Digital Economy Papers No. 336.

Ohanian, R. (1991). The impact of celebrity spokespersons' perceived image on consumers' intention to purchase. *Journal of Advertising Research*, 31(1), 46-54.

Pagallo, U., Durante M. (2022). The Good, the Bad, and the Invisible with Its Opportunity Costs: Introduction to the 'J' Special Issue on "the Impact of Artificial Intelligence on Law", *J*, 5(1), 139-149.

Patnaik, P. (2021) Regulations for Social Media Influencers and Celebrity Endorsement, *Indian Journal of Law and Legal Research*, 3(1), 1-13.

Penttinen, V., Ciuchita, R., & Čaić, M. (2022). YouTube It Before You Buy It: The Role of Parasocial Interaction in Consumer-to-Consumer Video Reviews. *Journal of Interactive Marketing*, 57(4), 561-582.

Perni, R. (2023). Pubblicità, educazione e diritto in Kant, Firenze University Press.

Pirni, A. (2022). At the roots of transparency: a public-ethics perspective. *Etica Pubblica*, 2, 15-27.

Savageaux, L. (2022, August 18). Virtual Influencers: Harmless Advertising or Dystopian Deception?, *De Pauw The Prindle Institute for Ethics*.

Woodroof, P. J., Howie, K. M., Syrdal, H. A., & VanMeter, R. (2020). What's done in the dark will be brought to the light: Effects of influencer transparency on product efficacy and purchase intentions. *Journal of Product & Brand Management*, 29(5), 675-688.

CYBERSECURITY - THE BEST LIFE PATH FOR EVERYONE

Aleksandra Pyrkosz, Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology (Poland)

aleksandra.pyrkosz@pcz.pl; sabina.szymoniak@icis.pcz.pl

ABSTRACT

Cybersecurity is a rapidly evolving field that involves protecting information, maintaining customer trust, and ensuring the stability of a company's operations in the era of universal digitisation. It requires constant monitoring, adapting strategies, and protective measures to the changing threat landscape. Researchers are crucial in this field, designing algorithms and techniques for security improvement in computer systems exposed to cyberattacks. They propose new security protocols using encryption, timestamps, pseudonymity, or hashing functions, which must be constantly verified to ensure an appropriate security level. Intrusion Detection and Prevention Systems (IDPS) are tools to monitor computer networks and detect and respond to suspicious or harmful activities. However, IDPSs have limitations, such as not detecting new or advanced attacks, and can generate false positives. Cyber attackers constantly explore computer systems, improving their knowledge and hacking skills. This diverse field offers numerous opportunities for personal and professional development. This paper discusses the authors' experiences with cybersecurity, their chosen path in life, and their achievements. Both authors have worked as students, teachers, thesis promoters, and co-organizers of cybersecurity events. Their insights and tips clarify doubts about choosing cybersecurity as a life path.

KEYWORDS: cybersecurity, way of life, scientific work, experiences in cybersecurity.

1. INTRODUCTION

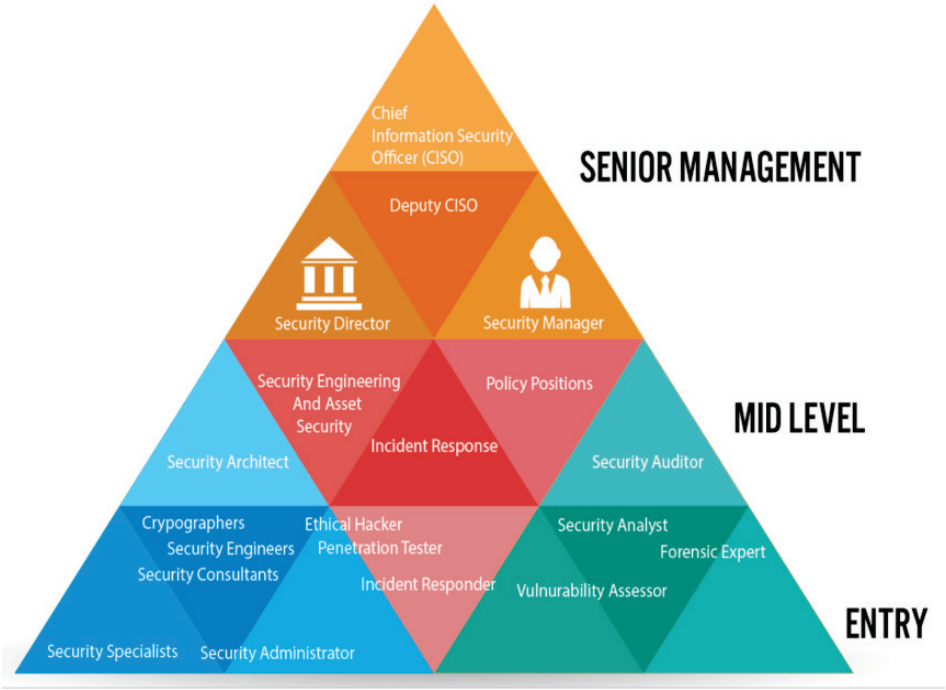
Cybersecurity is one of the most exciting areas of work and science. It is evident in the digital sector, where practically all businesses of every size can access the Internet. Numerous benefits are presented by network access, but there are also new difficulties. One of them is taking adequate care of online security. Cybersecurity, or information security or cybersecurity, refers to the activities, processes, and practices designed to protect computer systems, networks, data, and information from cyber-related threats. Cybersecurity is vital for individuals and organisations as more and more data is stored and processed online, and cyber threats become more advanced and numerous.

In tandem with the escalating significance and utilisation of digital technologies, the quantity of dangers that organisations must mitigate likewise experiences a corresponding rise. Using more sophisticated attack techniques by cybercriminals has necessitated evolving and enhancing security measures. Implementing robust security measures in cyberspace is of utmost importance in safeguarding information, upholding consumer confidence, and ensuring the continued stability of organisational operations within the era of pervasive digitisation. Cybersecurity management entails an ongoing and iterative procedure that necessitates the perpetual surveillance, adjustment of plans, and implementation of protective measures in response to the evolving landscape of threats. The procedure under consideration is intricate, although several options in the market offer potential enhancements while ensuring a suitable level of safety (Steingartner et al., 2022; Nwankpa & Datta, 2023). Figure 1 provides a comprehensive overview of the career trajectory within the field of cybersecurity, illustrating the various actions undertaken by professionals specialising in different areas of cybersecurity.

Furthermore, the field of cybersecurity presents an ideal domain for researchers. Computer systems vulnerable to cyberattacks necessitate the implementation of specifically tailored algorithms and

procedures to enhance their security. Secure communication between network nodes is a necessary need for such systems. Therefore, researchers must develop security protocols that delineate the specific order of operations in communication processes while incorporating security measures such as encryption, timestamps, pseudonymity, and hashing functions. Security protocols can be developed to cater to either cross-domain or customised solutions. Additionally, it is imperative to regularly assess the efficacy of security policies to ascertain their ability to maintain an adequate level of security (Bartłomiejczyk et al., 2022; Szymoniak, 2021).

Figure 1. Cybersecurity career path.



Source: https://www.spiceworks.com/tech/it-careers-skills/articles/cybersecurity-career-path/

Furthermore, the field of cybersecurity presents an ideal domain for researchers. Computer systems vulnerable to cyberattacks necessitate the implementation of specifically tailored algorithms and procedures to enhance their security. Secure communication between network nodes is a requisite for such systems to function properly. Therefore, researchers must put forth novel security protocols that establish a specific sequence of steps for communication and incorporate security approaches such as encryption, timestamps, pseudonymity, and hashing functions. Security protocols can be devised to cater to either cross-domain or customised solutions. Additionally, it is imperative to regularly assess the efficacy of security measures to determine whether they can provide a suitable level of security (Bartłomiejczyk et al., 2022; Szymoniak, 2021).

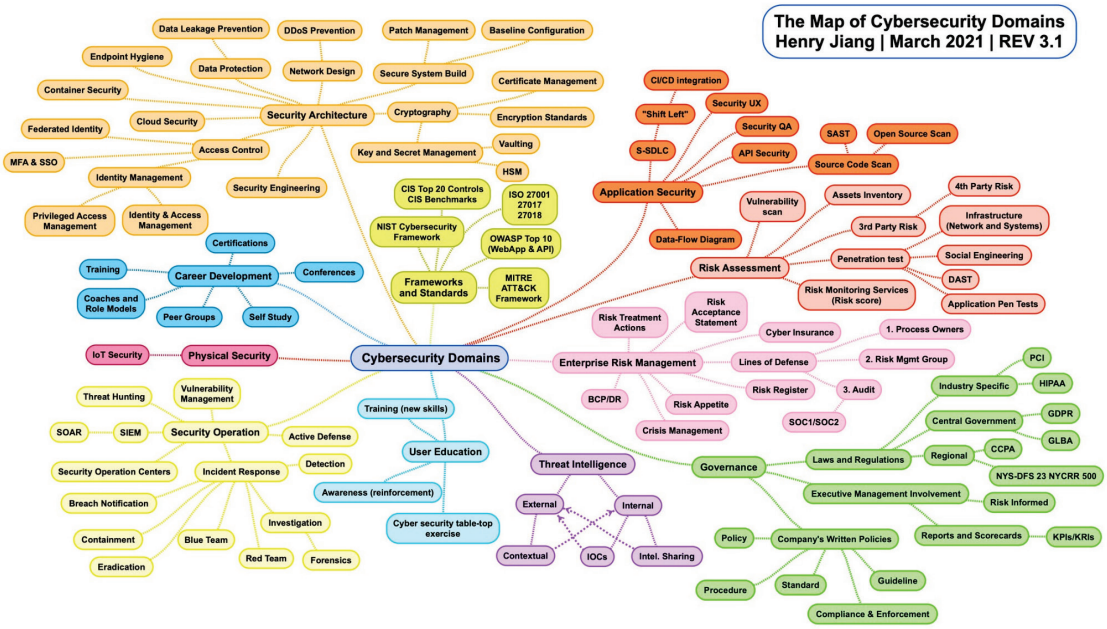
Furthermore, computer systems require specialised tools for assessing the vulnerabilities inside the system and implementing Intrusion Detection and Prevention Systems (IDPS). Intrusion Detection and Prevention Systems (IDPSs) actively monitor computer networks to identify and promptly address potentially malicious or detrimental actions. The researchers examine network data and discern patterns that could signify unauthorised access attempts, attacks, or other undesirable activities. Intrusion Detection and Prevention Systems (IDPSs) are designed to monitor network traffic, analyse packet signatures, identify deviations from standard patterns, and promptly respond to detected security incidents. The tools primarily employ artificial intelligence methodologies and approaches in their operations (Apruzzese et al., 2023).

It should be noted that Intrusion Detection and Prevention Systems have inherent limitations. The potential limitation of their detection capabilities lies in their inability to identify and respond to novel or sophisticated threats that need to be established signatures. Consequently, Intrusion Detection and Prevention Systems must undergo regular updates and maintain access to current signature databases. Moreover, Intrusion Detection and Prevention Systems have the potential to provide erroneous positive alerts, particularly in intricate network environments, necessitating additional scrutiny and examination by network administrators.

Furthermore, the assailants are not inactive. They engage in ongoing exploration of computer systems to identify novel access points. They enhance their understanding of various systems and refine their abilities in hacking. Various cyberattacks can be identified, such as spoofing, the exploitation of known session-specific transitory information or replay assaults. Moreover, the executed attacks may combine various conventional attack methods. The impacts of cyber attacks can vary depending on factors such as the specific attack type, its intended objectives, and the methods employed, encompassing activities such as data theft, sabotage, and privacy breaches (Szymoniak & Kesar, 2023).

The diverse range of potential methods for enhancing computer system security presents individuals with numerous chances to shape their career trajectories. Figure 2 depicts the cartographic representation of cybersecurity domains. This diagram provides a concise overview of the various opportunities and interests that are inherent in the field of cybersecurity. This expansive domain offers ample opportunities for individuals to find suitable professional and personal growth avenues.

Figure 2. Cybersecurity domains.



Source: https://www.linkedin.com/pulse/cybersecurity-domain-map-ver-30-henry-jiang

The authors of this research have also selected cybersecurity as their chosen career path. Security concerns during undergraduate study marred the initial encounter. The author's interest in cybersecurity was piqued by the subject matter of security and networks, prompting them to pursue further studies in this specialised area. The author of the Master's thesis examined the process of security simplification in large enterprises, focusing on maintaining an adequate level of security. Additionally, the second author is employed in cybersecurity as a specialist.

The second author encountered security-related challenges during their doctoral study. The author examined the process of verifying security protocols and the influence of time on executing these procedures. The author has acquired numerous compelling findings and effectively disseminated them through various scholarly publications. Additionally, the second author proposes novel security procedures for implementing Internet of Things solutions in scientific research. The author of this instructional piece also emphasises security concerns in a broader scope. The courses address many areas of computer system security.

The two authors initially encountered each other in various academic capacities, including as students and instructors and during their respective roles as graduate students and thesis advisors. Subsequently, they collaborated as co-organisers of cybersecurity events. The individual expresses their intention to share their personal experiences related to the field of cybersecurity and to inspire others to pursue a similar career trajectory.

This study aims to present the authors' experiences in cybersecurity, acquired through their academic pursuits and professional engagements. The individuals in question will elucidate the rationale behind their decision to pursue a particular trajectory in life, specifically within cybersecurity. Furthermore, they will expound upon cybersecurity concerns' captivating and stimulating aspects. They will present their most notable accomplishments. Moreover, individuals will consider the obstacles encountered in prior endeavours and contemplate the potential for growth and the acquisition of stimulating experiences presented by the several avenues within the field of cybersecurity. By sharing our experiences, ideas, and suggestions, we provide clarity to individuals uncertain about their decision to pursue a career in cybersecurity.

The rest of this paper is organised as follows. Section 2 will present the first author’s path to cybersecurity. Also, the first author will present her scientific interest associated with disinformation in cybersecurity, social engineering and APT groups. Section 3 will present the second author’s path to cybersecurity. Also, the second author will present her scientific interest associated with security protocols, the Internet of Things and attack detection. In the last section, both authors will conclude and present their insights and tips to clarify doubts about choosing cybersecurity as a life path.

2. MY PATH TO CYBERSECURITY - ALEKSANDRA

The author's interest in computers began in primary school during the first computer science classes. Initially, she focused on graphic editor programs. As her education progressed, her interest in computer science grew. She chose a technical high school specialising in computer science, and her favourite subjects were computer networks, especially their security. Then, she attended various conferences related to cybersecurity, which only strengthened her belief that this was the right path for a career.

In 2018, Aleksandra started her studies at the Faculty of Mechanical Engineering and Computer Science at Częstochowa University of Technology. The topic of her engineering thesis was building web services based on Docker software. Later, she continued her master's studies at the same faculty, writing a thesis about simplifying security in large organisations while maintaining an adequate level of safety.

Meanwhile, Aleksandra was employed in an automotive industry company, working in the team responsible for the security of the company's client workstations. However, beyond her work, she also wanted to continue her development at the university. In 2023, she began her PhD studies at the same faculty, where she completed her engineer’s and master's degrees. Aleksandra intends to dedicate her time in the PhD to exploring the impact of disinformation on cybersecurity.

In addition to her interest in disinformation, she is passionate about social engineering in the context of cyberattacks and the activities of APT (Advanced Persistent Threat) groups.

2.1. Disinformation in cybersecurity

Disinformation, frequently regarded as a subset of misinformation, involves intentionally disseminating deceptive or untrue information. The underlying intent distinguishes disinformation from misinformation, the primary motivator for the source, typically originating from human actors (Caramancion, 2020).

Disinformation boasts robust historical traditions dating back to ancient times, and its legacies persistently extend into the modern era, particularly with the advent of new technological opportunities such as the Internet, social media, and artificial intelligence. These novel forms of communication and media have ushered in a new dimension of threat (Mareš & Mlejnková, 2021).

The shift of disinformation to the online realm has been facilitated by the substantial increase in the popularity of digital information channels as news sources over the past decade. The primary advantage of these platforms is their ability to create and disseminate information swiftly. However, this also introduces pressure, accelerating editorial tasks, fact-checking, and assessing source credibility.

2.2. Social engineering

Social Engineering is commonly recognised as influencing or manipulating individuals to disclose sensitive information or provide access to restricted areas (Uebelacker & Quiel, 2014). In cybersecurity, its primary application is to prompt individuals to disclose confidential information or undertake actions that violate security protocols, leading to the inadvertent infection of systems or the disclosure of classified information (Breda & Barbosa & Morais, 2017). Social engineering is at the core of one of the attack techniques known as phishing.

Phishing attacks focus on exploiting vulnerabilities within systems attributable to the human factor. Numerous cyber attacks propagate through mechanisms that take advantage of weaknesses present in end users. This places users at the forefront of the security chain, making them the most susceptible element in the defence against cyber threats. The challenge posed by phishing is multifaceted, encompassing a range of deceptive tactics employed by malicious actors. Cyber adversaries' complexity and evolving tactics necessitate a dynamic and multifaceted approach to cybersecurity. (Khonji & Iraqi & Jones, 2013)

Consequently, organisations and individuals employ proactive and reactive strategies to mitigate specific phishing attacks. Educational initiatives, user awareness training, and the implementation of robust security protocols are essential components of a comprehensive defence strategy. By fostering a culture of cybersecurity awareness, organisations can empower users to recognise and resist phishing attempts.

2.3. APT Groups

The Advanced Persistent Threat, commonly called APT, represents well-funded and organised groups systematically compromising government and commercial entities. APTs exhibit high levels of sophistication, managing to circumvent virtually all "best practice" cybersecurity programs to establish a long-term presence within networks. These attacks are characterised by stealth, precision targeting, and a focus on data, setting them apart from traditional worms or viruses.

APTs are orchestrated by well-organised entities, often foreign adversaries, to gather specific information from an organisation. Their strategy involves obtaining immediate data and ensuring

prolonged access for future extraction at their discretion. APTs defy conventional attack patterns by adapting techniques continually, utilising users as entry points, and carefully concealing their tracks (Code, 2012).

A comprehensive approach that combines advanced technology, continuous monitoring, and collaborative efforts is crucial for effectively safeguarding against the persistent and sophisticated threats posed by APTs.

3. MY PATH TO CYBERSECURITY - SABINA

When the first author finished primary school, her class had workshops to show them how to find their path or profession. One of the exercises that the students did was to pass around the class a piece of paper with the name and surname of the person, and their classmates were to write down the profession for which the person fits. The profession of a teacher appeared most often in the first author's paper. The boys joked about entering the profession of a car mechanic. However, once upon a time, someone entered the word IT specialist. So, many years later, the first author chose computer science studies, and after graduating with a Ph.D., she started working as an academic teacher in computer science.

Sabina started her studies in 2007 at the Faculty of Mechanical Engineering and Computer Science of the Częstochowa University of Technology. Her favourite subject was databases, so she wrote her master's thesis. After that, she started her PhD at this same faculty. During these studies, she considered issues with modelling and verifying security protocols and network delays. In 2017, Sabina started working as an academic teacher in computer science at her Alma Mater. During her previous work, she conducted lectures and laboratories connected with the security of computer and network systems. She also promoted thirty students' theses and the first author's thesis. Also, she published many conferences and journal articles on security, mainly associated with security protocols.

The department authorities entrusted her with the duties of a tutor for the cybersecurity speciality, bearing in mind her didactic and scientific interests. She is enthusiastic about working with students of this speciality. They are very interested in cybersecurity and willing to work.

Sabina's detailed scientific interests are security protocols, the Internet of Things security and cyberattack detection. She will describe these interests in the following subsections.

3.1. Security protocols

A *security protocol* is an algorithm that guarantees the achievement of essential security objectives during communication. Frequently, cryptographic algorithms and techniques are employed for this objective. When communicating, the primary objectives for ensuring data security encompass authorisation, safeguarding of information, maintaining integrity, and facilitating cryptographic key distribution. Furthermore, implementing security protocols has been identified as crucial in ensuring the integrity and confidentiality of Internet of Things (IoT) systems. The effective management of cryptographic primitives, such as keys and passwords, is crucial, particularly in ensuring their timeliness and monitoring their validity throughout their lifespan (Steingartner et al., 2021; Thakur et al., 2023).

The first security protocols were created in the 1970s. In the earliest versions, they were straightforward algorithms that used only encryption for protection. The Internet was used only for military or scientific solutions, so no one thought better security would be necessary. The most characteristic example is the Needham Schroeder Public Key protocol (Needham & Schroeder, 1978). This protocol uses asymmetric cryptography, so it seems secure because only the private key’s owner may decrypt the message addressed to him. The Needham Schroeder Public Key protocol was widely

used to secure the authentication process for almost twenty years. In the mid-1990s, Gavin Lowe (Lowe, 1995), using formal modelling and implementing the automatic security protocols verification tool - FDR (Roscoe, 995), discovered the possibility of attacking this protocol simply.

We use security protocols in each area associated with an Internet connection. The main goals of security protocols may be authentication or key distribution. Also, we can find many other protocols' solutions, for example, for fog or edge processing (Pardeshi et al., 2022), for industry (Tanveer et al., 2023), for medicine or healthcare solutions (Rasslan et al., 2022), (Wu et al., 2023), meetings security (Szymoniak & Siedlecka-Lamch, 2022) or suitable for many domains (Yan et al., 2022). Most of them work in the Internet of Things solutions.

3.2. Internet of Things security

The Internet of Things (IoT) refers to a network of interconnected physical devices, vehicles, buildings, and other objects embedded with sensors, software, and network connectivity, allowing them to collect and exchange data. IoT enables these things to communicate with each other and central control systems over the internet, enabling automation, remote monitoring, and data-driven decision-making (Sarker et al., 2023).

IoT networks consist of different types of devices with different software and hardware equipment. Thus, the security protocols for IoT meet with some requirements. Minimising the computational burden of IoT devices during communication is crucial to ensure optimal operational efficiency and timely performance for both devices and their users. The computations in cloud or fog environments promote the attainment of reduced transmission delays and optimal utilisation of bandwidth resources. Moreover, we should take into consideration the cross-platform compatibility. Internet of Things protocols must exhibit cross-platform compatibility. Several methods examined in this publication are particular to certain applications, such as medicine. When formulating a security protocol for IoT, it is prudent to contemplate a broader range of applications, thereby enabling the implementation of a single authentication or key agreement and distribution protocol across multiple solutions (Szymoniak & Kesar, 2022).

The security of IoT devices or users is crucial because devices exchange data between themselves using the Internet, sometimes without user knowledge. Both protocols and IoT networks that also use protocols are vulnerable to cyberattacks.

3.3. Cyberattack detection

A cyberattack refers to malicious activities conducted over the internet with the intent to compromise the confidentiality, integrity, or availability of computer systems, networks, or data. These attacks are typically carried out by individuals or groups of hackers who aim to gain unauthorised access, steal information, disrupt services, or cause damage. Cyberattacks can take many forms and evolve as technology advances (Szymoniak & Kesar, 2022).

IoT environments (including security protocols) may fall victim to attackers. The target of attacks may be individual IoT devices, their users, and the entire network. Moreover, the type, tools and techniques depend on the attacker's intentions. The attacker may want to steal the user's data or data from the network. Next, the attacker may use them for other unethical activities. Also, attackers may use one device or whole network to perform other, more complex attacks (Szymoniak & Kesar, 2022).

One of the typical cyberattacks in IoT environments is the Denial of Service (DoS) attack. This cyber attack aims to disrupt the availability of a computer system or network by overwhelming it with a flood of illegitimate requests or excessive traffic. A sinkhole attack occurs when an attacker disseminates

modified routing information, leading to the diversion of network traffic towards their malicious infrastructure. This traffic diversion might afterwards enable the attacker to initiate additional attacks (Attkan & Ranga, 2022). An impersonation attack refers to a malicious act in which an attacker assumes the identity of another user, such as a user, server, gateway, node, or IoT device (Nyangaresi et al., 2022). A capture attack refers to an attacker seizing control of a sensor node or IoT device to gain control over the network. In this assault, the attacker removes the node from the network and redeploys it as a malicious node. Another type of attack, a cloning attack, involves the attacker replicating a sensor node or IoT device to create unauthorised copies that can be used for harmful purposes (Attkan & Ranga, 2022).

4. CONCLUSIONS

This paper presented our firsthand encounters with cybersecurity, our selected path in life, and our notable accomplishments. Both authors described their experience in various roles, such as students, teachers, thesis supervisors, and co-organisers of cybersecurity events. Their profound understanding and valuable advice dispelled any uncertainties regarding the decision to pursue a career in cybersecurity.

The first author described disinformation in cybersecurity, social engineering and APT groups. Disinformation, a subset of misinformation, involves intentionally disseminating deceptive or untrue information. Its long history has its legacies extending into the modern era, particularly with new technological opportunities like the Internet, social media, and artificial intelligence. The popularity of digital information channels as news sources has facilitated the shift of disinformation to the online realm. Social engineering, a form of phishing, is at the core of this attack, influencing individuals to disclose sensitive information or provide access to restricted areas. Cyberattacks exploit vulnerabilities within systems, making users the most susceptible element in defence against cyber threats. Proactive and reactive strategies are employed to mitigate specific phishing attacks, including educational initiatives, user awareness training, and robust security protocols. *Advanced Persistent Threat groups* are well-funded and organised groups that systematically compromise government and commercial entities. A comprehensive approach combining advanced technology, continuous monitoring, and collaborative efforts is crucial for effectively safeguarding against these persistent and sophisticated threats.

The second author described security protocols, the Internet of Things and attack detection. Security protocols are algorithms that guarantee the achievement of essential security objectives during communication, often using cryptographic algorithms and techniques. The primary objectives for ensuring data security include authorisation, safeguarding of information, maintaining integrity, and facilitating cryptographic key distribution. Implementing security protocols is crucial in ensuring the integrity and confidentiality of Internet of Things (IoT) systems. IoT security is crucial for interconnected physical devices, vehicles, buildings, and other objects embedded with sensors, software, and network connectivity. Cyberattacks are malicious activities conducted over the Internet to compromise the confidentiality, integrity, or availability of computer systems, networks, or data. IoT environments may fall victim to attackers, targeting individual IoT devices, their users, and the entire network. Cyberattacks can take many forms and evolve as technology advances. By considering a broader range of applications and implementing a single authentication or key agreement and distribution protocol across multiple solutions, security protocols can help protect IoT systems and users from cyberattacks.

Cybersecurity professionals are essential to ensuring that our increasingly integrated and automated digital infrastructure remains secure, reliable and protected from ongoing threats. Cybersecurity professionals are essential for many reasons, especially in today's digital world. As technology

becomes more advanced, cyber threats become more complex. Cybersecurity specialists help protect organisations against hacking, ransomware, phishing and DDoS attacks. In the digital age, data is a valuable asset. Cybersecurity specialists ensure that this data is safe from theft, loss or unauthorised access. Many jurisdictions have data protection and privacy laws, such as GDPR in Europe. Cybersecurity professionals help organisations comply with these regulations. Attacks on critical infrastructure such as power plants, water systems, and transportation networks can seriously affect society. Cybersecurity specialists work to protect these key systems from potential threats. Cybersecurity professionals are crucial in educating employees and communities about cyber threats and security best practices. Cybersecurity professionals can develop new tools, techniques, and defensive strategies by continuously researching and analysing new cyber threats. As the Internet of Things, artificial intelligence, and other emerging technologies evolve, the potential for cyber threats also increases. Cybersecurity professionals ensure that these new technologies are designed and implemented securely.

It is worth mentioning that there is an ever-increasing demand for cybersecurity professionals worldwide. This means excellent job prospects and career development opportunities. The field of cybersecurity is exceptionally diverse and dynamic. Every day can bring new challenges, which makes work exciting and rewarding. Working in cybersecurity directly impacts the protection of people, organizations, and critical infrastructure from cyber threats. This gives us the feeling that our work has a real and positive impact on society. The field of cybersecurity is constantly developing and evolving. This means that professionals can constantly learn new technologies, tools and techniques, making the work exciting and challenging. Many cybersecurity professionals enjoy attractive salaries and benefits packages, making this a financially attractive career path. A cybersecurity education and experience opens doors to various positions and career paths, from security analysts to incident response specialists to IT security leadership. In the era of globalization and digitalization, cybersecurity specialists can work on both the local and international markets, cooperating with organizations from various industries and regions. To sum up, working in the field of cybersecurity offers not only attractive professional and financial prospects but also the opportunity to constantly develop, learn and influence the digital security of society. Therefore, it can be very encouraging for people interested in technology, security and solving complex problems.

REFERENCES

Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The role of machine learning in cybersecurity. *Digital Threats: Research and Practice*, *4*(1), 1-38.

Attkan, A., & Ranga, V. (2022). Cyber-physical security for IoT networks: a comprehensive review on traditional, blockchain and artificial intelligence-based key-security. *Complex & Intelligent Systems*, *8*(4), 3559-3591.

Bartłomiejczyk, M., El Fray, I., Kurkowski, M., Szymoniak, S., & Siedlecka-Lamch, O. (2022). User Authentication Protocol Based on the Location Factor for a Mobile Environment. *IEEE Access*, *10*, 16439-16455.

Breda F., Barbosa H., Morais T. (2017). Social Engineering and Cyber Security. In 11th International Technology, Education and Development Conference. INTED2017 Proceedings (pp. 4204-4211). https://doi.org/10.21125/inted.2017.1008

Caramancion K. M. (2020). An Exploration of Disinformation as a Cybersecurity Threat. In 2020 3rd International Conference on Information and Computer Technologies (ICICT). IEEE. https://doi.org/10.1109/ICICT50521.2020.00076

Code E,. Advanced Persistent Threat, Understanding the Danger and How to Protect Your Organization. 1st Edition. Amsterdam: Elsevier. (2012).

Khonji M., Iraqi Y., & Jones A. (2013). Phishing Detection: A Literature Survey. IEEE Communications Surveys & Tutorials, 15(4). https://doi.org/10.1109/SURV.2013.032213.00009

Lowe, G. (1995). An attack on the Needham− Schroeder public-key authentication protocol. Information processing letters, 56(3).

Mareš M. & Mlejnková P. (2021). Challenging Online Propaganda and Disinformation in the 21st Century (pp.75-103). Political Campaigning and Communication. Springer. https://doi.org/10.1007/978-3-030-58624-9

Needham, R. M., & Schroeder, M. D. (1978). Using encryption for authentication in large networks of computers. *Communications of the ACM*, *21*(12), 993-999.

Nwankpa, J. K., & Datta, P. M. (2023). Remote vigilance: The roles of cyber awareness and cybersecurity policies among remote workers. *Computers & Security*, *130*, 103266.

Nyangaresi, V. O., Rodrigues, A. J., & Abeka, S. O. (2022). Secure Algorithm for IoT Devices Authentication. In *Industry 4.0 Challenges in Smart Cities* (pp. 1-22). Cham: Springer International Publishing.

Pardeshi, M. S., Sheu, R., & Yuan, S. (2022). Hash-Chain Fog/Edge: A Mode-Based Hash-Chain for Secured Mutual Authentication Protocol Using Zero-Knowledge Proofs in Fog/Edge. *Sensors*, *22*(2), 607. https://doi.org/10.3390/s22020607

Rasslan, M., Nasreldin, M., & Aslan, H. K. (2022). Ibn Sina: A patient privacy-preserving authentication protocol in medical internet of things. *Computers & Security*, *119*, 102753. https://doi.org/10.1016/j.cose.2022.102753

Roscoe, A. W. (1995, June). Modelling and verifying key exchange protocols using CSP and FDR. In *Proceedings The Eighth IEEE Computer Security Foundations Workshop* (pp. 98-107). IEEE.

Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. (2023). Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions. *Mobile Networks and Applications*, *28*(1), 296-312.

Steingartner, W., Galinec, D., & Kozina, A. (2021). Threat defense: Cyber deception approach and education for resilience in hybrid threats model. *Symmetry*, *13*(4), 597.

Steingartner, W., Možnik, D., & Galinec, D. (2022, November). Disinformation Campaigns and Resilience in Hybrid Threats Conceptual Model. In *2022 IEEE 16th International Scientific Conference on Informatics (Informatics)* (pp. 287-292). IEEE.

Szymoniak, S., & Siedlecka-Lamch, O. (2022). Securing Meetings in D2D IoT Systems. *ETHICOMP 2022*, 31.

Szymoniak, S. (2021). Amelia—a new security protocol for protection against false links. *Computer Communications*, *179*, 73-81.

Szymoniak, S., & Kesar, S. (2023). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. Applied Sciences, 13(1), 404. https://doi.org/10.3390/app13010404

Tanveer, M., Badshah, A., Alasmary, H., & Chaudhry, S. A. (2023). CMAF-IIoT: Chaotic map-based authentication framework for Industrial Internet of Things. *Internet of Things*, *23*, 100902.

Thakur, G., Kumar, P., Jangirala, S., Das, A. K., & Park, Y. (2023). An effective privacy-preserving blockchain-assisted security protocol for cloud-based digital twin environment. *IEEE Access*, *11*, 26877-26892.

Uebelacker S. & Quiel S., The Social Engineering Personality Framework (2014). Workshop on Socio-Technical Aspects in Security and Trust, Vienna, Austria, 2014, pp. 24-30, http://doi.org/10.1109/STAST.2014.12

Wu, T. Y., Wang, L., & Chen, C. M. (2023). Enhancing the Security: A Lightweight Authentication and Key Agreement Protocol for Smart Medical Services in the IoHT. *Mathematics*, *11*(17), 3701.

Yan, D., Luo, Y., Chen, X., Tong, F., Xu, Y., Tao, J., & Cheng, G. (2022). A Lightweight Authentication Scheme Based on Consortium Blockchain for Cross-Domain IoT. *Security and Communication Networks*, *2022*, 1–15. https://doi.org/10.1155/2022/9686049

HIGHLIGHTING ETHICAL DILEMMAS IN SOFTWARE DEVELOPMENT: A TOOL TO SUPPORT ETHICAL TRAINING AND DELIBERATION

Pak Hei Li, Dharini Balasubramaniam

University of St Andrews (United Kingdom)

issacli7401@gmail.com; dharini@st-andrews.ac.uk

ABSTRACT

Ethical dilemmas in software development occur when professionals face conflicting moral obligations, or their actions may have unintended consequences. These dilemmas may arise throughout software development, from the commissioning of a system to its maintenance and retirement. Identifying and addressing potential ethical dilemmas are important steps in producing ethical software systems. However, currently there is a lack of practical support for such ethical deliberation in software engineering. This paper aims to partly address this gap by creating an extensible training resource, and a custom agile project management tool that highlights ethical dilemmas in software development.

We review existing ethical frameworks for software development in the context of the principles of ACM Code of Ethics and Professional Conduct incorporate the Code of Ethics into agile project management. An interactive training curriculum, incorporating text adventure scenarios, chatbot recommendations, and pop-ups that explain ethical principles, was implemented to explain the Code of Ethics. A Kanban board was also built to highlight ethical dilemmas during development. The proof-of-concept tool includes compliance checklists, ethics self-assessment, and logic-based semantic flagging that highlights dilemmas. The usability of the tool was evaluated and received positive feedback. This paper thus contributes to raising awareness of ethical principles and providing a proof-of-concept tool to highlight ethical dilemmas.

KEYWORDS: ethical dilemma, ethical deliberation, software ethics.

1. INTRODUCTION

Ethics is an important facet of software development and a key concern for all stakeholders of software systems. Codes of ethics can serve as a foundation for ethical decision-making, focusing on resolving ethical issues, rather than following a prescriptive algorithmic approach to ethics (Gotterbarn et al., 2017). However, codes of ethics are usually expressed as abstract principles, and software professionals are not typically offered training or tools to support ethical deliberation and decision making in practice.

In attempting to address this gap, the main contributions of this paper are an extensible and customisable training resource and an agile project management tool that highlights ethical dilemmas. The training resource focuses on illustrating the principles of the ACM and IEEE Software Engineering Code of Ethics (SWECOE) interactively via text adventure scenarios and chatbot recommendations. The project management tool focuses on highlighting ethical concerns and resolutions to ethical dilemmas in software development decision-making. As an exemplar for agile software development, a Kanban board is incorporated into the tool. The highlighted dilemmas are related to feature issues or tasks in the product backlog. Tasks are visually represented on the Kanban board, allowing engineers and project managers to see the state of each task at any time to promote ethical awareness and foster a culture of ethical responsibility in software development.

2. BACKGROUND AND RELATED WORK

2.1. Software Engineering Code of Ethics

Software engineers have responsibilities to both their profession and society, which include adhering to the SWECOE. The ACM and IEEE SWECOE principles state that software engineers must commit to respecting their profession throughout the development process, and ensure that their software benefits society without causing harm (Gotterbarn et al., 1997). Gotterbarn and Miller (2009) have also shown that having a Code of Ethics can assist software engineers in making complex decisions and resolving conflicts that may arise during software development. Studies on the IEEE Code of Conduct and topics such as privacy, malware, and net neutrality have identified obligations for software engineers to act ethically (Evans, 2012). These codes require engineers to consider both their clients and the wider society during development (Godfrey, 1996). Continuing research is required in the field of SWECOE as it plays a crucial role in determining the quality of software and its impact on our lives (Gotterbarn et al., 1999; Bynum, 2000).

Nevertheless, the ACM and IEEE SWECOE presents a significant challenge for practitioners due to the lack of clarity and mechanisms for adherence to best practice. To tackle this problem, it is necessary to comprehend how engineers perceive and enforce ethical guidelines (Lurie & Mark, 2016). Further research is needed to examine the challenges in enforcing ethical standards in software development. One constraint of the ACM and IEEE SWECOE is its generality, necessitating its application at each phase of the software development process, ethical issues that may arise in each stage can be addressed and software quality improved. Thus, establishing a categorisation of the SWECOE grounded on the development process could help engineers implement ethical practices while developing software. Guidelines, policy checklists, and tools can be designed to improve the standard of the software development process by utilising a structured categorisation (Karim et al., 2017).

2.2. The Ethics-Driven Software Development Framework

Building on the established ACM and IEEE SWECOE, Lurie & Mark (2016) proposed the Ethical-Driven Software Development (EDSD) framework. In contrast to the ACM and IEEE SWECOE, which views the Code of Ethics and practical protocols as separate independent entities, the EDSD framework advocates for the integration of ethical standards into the daily work of software engineers, viewing them as practical tools. This integration occurs throughout the development process, resembling a sub-system design with abstract and concrete classes and interfaces. The framework is intended to serve as a means for engineers to adhere to development regulations, simultaneously establishing ethical boundaries and limitations within the realm of applied ethics (Sommerville, 2011). The EDSD framework consists of a set of “yes/no” questions that aim to increase ethical awareness throughout software development. All major stakeholders must be familiar with these questions before commencing any development stage. The framework is implemented using EDSD Index Cards, which are stage-specific and colour-coded. Importantly, these cards were designed to raise awareness of ethical considerations and do not provide model answers. They facilitate the creation of tailor-made ethical questions, with flexibility for adjustment based on development process requirements, life cycle, team dynamics, leadership, and complexity of requirements (Lurie & Mark, 2016). However, this framework is not integrated into standard development tools.

2.3. Ethical Deliberation in Agile Software Development

Agile methodologies contain features to develop software that is informed by ethical considerations. These methodologies emphasise the importance of continuous feedback and communication with all relevant stakeholders. This emphasis can support the integration of ethical considerations into the

software development process from conception. By addressing stakeholders' concerns and values throughout the development cycle, ethical dimensions can be embedded into software creation. To effectively integrate ethical deliberation into software development processes and foster responsible engineering practices, alignment with the practices employed by engineers is imperative. Failure to integrate ethical deliberation may lead to neglect or superficial application of ethics. This necessitates the training of software engineers in effective ethical communication skills to proficiently address ethical issues arising from technical designs. While technical solutions may resolve some ethical questions associated with information technology, normative deliberation should be systematically incorporated into development processes and seamlessly documented as part of quality assurance practices (Judy, 2009). This comprehensive approach should ideally permeate the entire software lifecycle and align with agile ethics in DevOps methodologies (Ebert et al., 2016). Notably, various methodologies, including Values in Design and Ethics by Design (Simon, 2016), have emerged to facilitate the deliberation of embedded values in technology. These frameworks assert the significance of individual autonomy in controlling technology and its consequences, underscoring the ethical impact of integrated values on real-life structures, thereby subjecting them to normative evaluation (Floridi, 2008; van den Hoven et al., 2015).

2.4. Decision-Making Models for Ethical Dilemmas

Ethical deliberation necessitates a structural integration to facilitate ethical decision-making. The decision-making process entails the analyses of various options and the determination of responses to opportunities and threats based on individual or organisational goals and values. Selecting an alternative in line with personal goals, preferences, and values is integral to this process. According to Jones (1991), the initial stage of ethical decision-making involves recognising the moral dilemma in an act or omission. Failing to recognise a moral choice takes one's actions out of the ethical decision-making process. After identifying a moral dilemma, the next step is to make a moral judgment and establish a moral intent. An ethical dilemma occurs when individuals face a difficult decision where they must choose between taking an action that may benefit someone else but may be against their own interests. It constitutes a complicated situation involving a conflict between moral principles, where adherence to one principle implies a violation of another (Ali et al., 2012). Illustrative instances of ethical dilemmas encompass scenarios where resources are scarce, necessitating a decision on allocation, or situations where the determination of whether and under what circumstances to withdraw treatment from a patient is at hand. In the process of ethical decision-making, having a guiding model is crucial as it facilitates the identification of ethical dilemmas and separates them from factual issues. Models typically consist of interrelated steps, and the completion of one step can reveal inadequacies in previous steps that may require revision. These steps can be used iteratively and sometimes do not necessarily have to be implemented chronologically, depending on the presence of specific facts.

Rest (1986) proposed a four-step model for ethical decision-making, delineating a sequential process. The initial step involves the identification of the moral dilemma, followed by the formulation of a judgment regarding the appropriate course of action. Subsequently, the third step involves committing to a morally sound course of action, and the fourth step requires taking action to address the ethical concerns. Rest's model treats each step as distinct and standalone, emphasising that the outcome of one step does not guarantee seamless progression to the next. In contrast, Trevino (1986) introduced an interactionist model that focuses on the recognition of ethical dilemmas and proceeds to the cognitive stage. Conversely, Ferrell & Gresham (1985) presented a contingency framework for ethical decision-making in the marketing domain, highlighting the emergence of ethical issues or dilemmas within the cultural environment. These models are derived from research and when integrated with

practical experience and common sense, they help to establish objectives, generate innovative alternatives, make informed trade-offs, resolve ambiguities, and assess risks. Through continuous practice, such models can become ingrained and familiar to software professionals, enabling them to navigate effortlessly without explicit reference to each step (Ali et al., 2012). Nonetheless, decision-making models have limitations, as they overlook the emotional and human dimensions of decision-making. While an ethical decision-making model requires setting aside emotions and following a structured approach, human decision-makers do not consistently operate rationally. The alignment of the "best option" with the ethically "right" one becomes intricate when ethical considerations and the interests of various stakeholders are at play. Individuals exhibit varying degrees of innate decision-making abilities, with some relying on decision-making models to enhance the quality of their judgments. Nevertheless, the efficacy of decision models is not uniform across all users, given the diverse array of questions and problems they must address (Schuelke-Leech et al., 2018). Despite these shortcomings, decision-making models remain valuable tools for guiding moral decisions, fostering critical thinking, encouraging deliberative discussions, and prompting a spectrum of options.

3. DESIGN AND IMPLEMENTATION

To address the challenges ethical deliberation in software development, this paper proposes a novel ethics-centred project management tool, integrating ethical considerations into agile processes. The tool, developed with features including a Kanban board which highlights ethical dilemmas, ethics self-assessment, and compliance checklists, provides tailored ethics training and guidelines for professionals and other stakeholders. The web application employs a client-server architecture, utilising the MERN stack (MongoDB, Express.js, React.js, Node.js) for efficient implementation.

3.1. Ethical Dilemma Text Adventure

The ethical dilemma text adventure game presents players with a range of morally complex scenarios where there are no clear "right" or "wrong" answers and where the consequences of their decisions are not immediately apparent. The game provides a range of options that reflect different ethical principles and perspectives, encouraging players to consider different ways of thinking about ethical issues via hypothetical scenarios (Figure 1). Additionally, the game prioritises accessibility and clarity in its user interface, with simple and clear language that is easy for players to understand.

Finally, the game includes feedback mechanisms that allow players to reflect on their decisions and learn from their experiences, promoting a deeper understanding of the complexities of ethical decision-making in the software engineering profession. The game offers a realistic and engaging experience, using real-life examples to challenge players with ethical decisions in software engineering contexts. Scenarios cover various ethical aspects, including data privacy, security, and transparency, prompting players to assess multiple options and consequences for stakeholders. The game's design, presented as a text adventure, provides intuitive and engaging gameplay with clear instructions and decision-making options. Feedback on player decisions aligns with ACM's ethical principles and software engineering guidelines, fostering an understanding of decision implications. The game features multiple endings based on the player's choices, mirroring realistic outcomes in the real world, and enhancing awareness of ethical consequences. By utilising real-world examples and adhering to established ethical principles, the game facilitates learning about ethical considerations.

3.2. Ethical Dilemma Game Tree

Game trees, as a directed graph structural representation, are employed for the text adventure game. Nodes within the tree denote game positions and store potential outcomes for each decision in the

ethical dilemma scenario text adventure game. This design aids players in comprehending the consequences of their choices, guiding them towards ethical decisions. Multiple options are available for each scenario state, and the tree branches out to depict decisions that result in varied outcomes.

The game tree has a hierarchical structure, starting from the initial scenario state and branching out to depict possible decisions and outcomes. The first level represents the scenario state (e.g., data breach), and subsequent levels denote player decisions, ensuring each decision offers ethically diverse options as per ACM and IEEE SWECOE. Ethically effective options yield positive outcomes, while less ethically effective options lead to negative outcomes (e.g., reporting a breach versus ignoring it). Decisions vary in complexity and associated risks to enhance engagement and challenge. For example, reporting a data breach may involve obtaining higher permission and considering reputational impact, while ignoring it may lead to severe legal consequences (Figure 1). The game tree prioritises replay value, enabling players to restart and make alternative decisions, fostering critical thinking and exploration of different possibilities. Overall, the hierarchical game tree incorporates diverse ethical decision-making options, varying complexities, and risks, promoting replay value for critical thinking.

3.3. Ethical Framework Infographics

The interactive ethical framework infographics page is an educational resource for software professionals to learn about the core principles of the ACM Code of Ethics and Professional Conduct. Based on user-centred design principles, it prioritises meeting the needs of software professionals seeking to understand ethical principles. Infographics are used as they are effective in conveying information, combining visuals and text to enhance comprehension (Figure 2).

The minimalist design of the infographics minimises visual clutter, emphasising crucial information with bullet points. The page is responsive, ensuring accessibility across different devices and a consistent user experience. The user-friendly and accessible design introduces the ACM Code of Ethics and Professional Conduct to software professionals in an engaging and comprehensible manner.

Figure 1. Text adventure example depicting scenarios of ATM data breaches along with game tree.

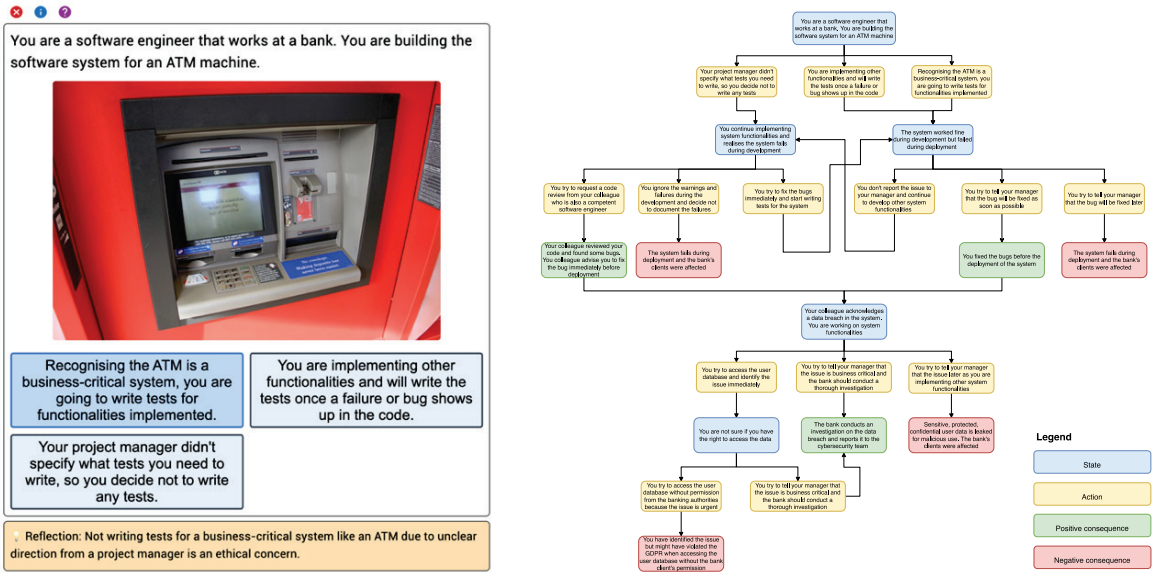
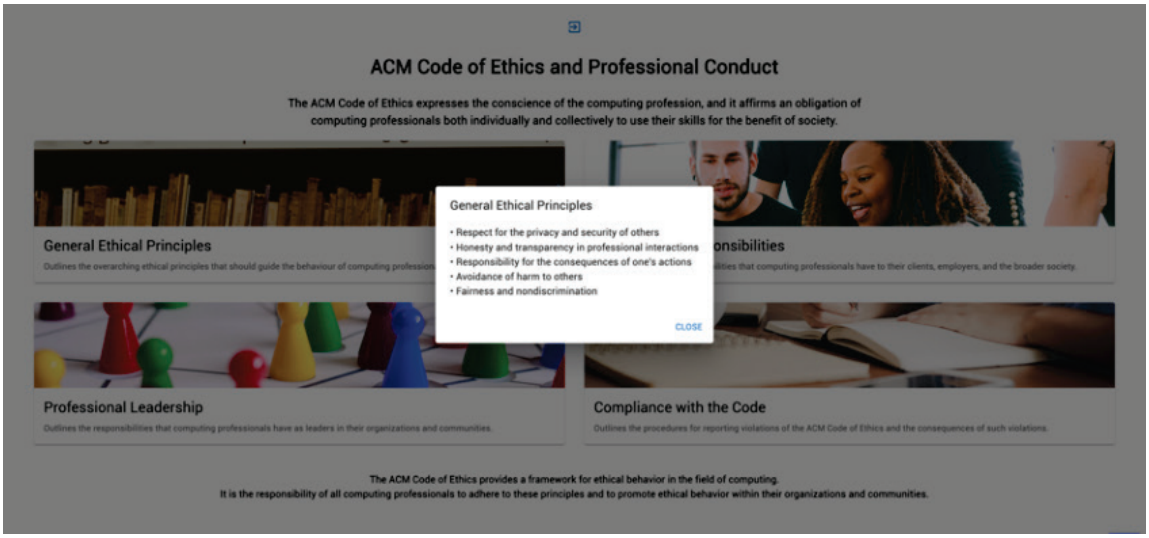


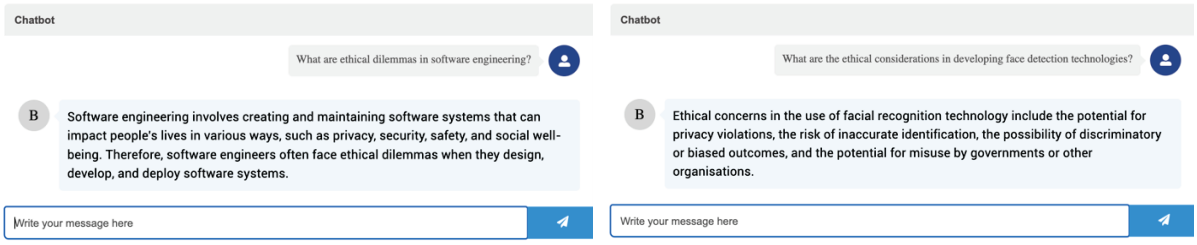
Figure 2. Infographics modal showing responsibilities based on the selected Code of Ethics.



3.4. Chatbot Recommendation

The chatbot recommendation in the project management tool offers users a responsive way to obtain information about the ACM Code of Ethics and Professional Conduct, ethical principles, and resolutions to ethical dilemmas. Designed with logic-based semantic matching, the chatbot accesses a knowledge base for user interaction. Functioning as a recommendation tool, it provides continuous support and ethical advice. Its user-centric design ensures ease of use and relevance in responses, and it offers an onboarding tour for user familiarisation. Programmed to logically respond to common queries about the ACM Code of Ethics, the chatbot aids users in understanding ethical principles. It does not store conversations, as it employs a knowledge base for responses, avoiding the need for advanced AI techniques (Figure 3).

Figure 3. Chatbot answers two example queries including “What are ethical dilemmas in software engineering?” and “What are the ethical considerations in developing face detection technologies?” using explanations of ethical dilemmas and considerations in software engineering. The chat history is not stored, so the chatbot initiates a new chat each time the system is used.

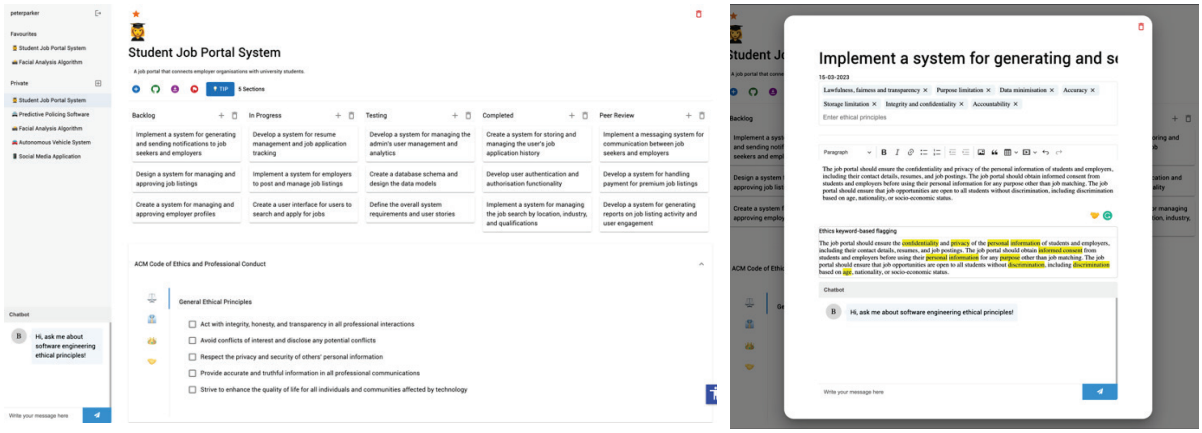


3.5. Ethics-Centred Kanban Board

The ethics-centric Kanban board enhances agile project management by facilitating the integration of ethical considerations into software development processes (Figure 4). To reduce the tendency of users to overlook ethical concerns in software development, the board highlights keywords related to ethical concerns and offers resolutions for ethical dilemmas. The Kanban board promotes ethical awareness among software professionals, cultivating a culture of responsibility and accountability. It is also useful for project managers, enabling them to monitor ethical considerations across the software development lifecycle and ensuring early identification and resolution of ethical issues,

ultimately saving time and resources. The Kanban board adapts to specific development workflows, which enhances its customisability and increases the user's adoption and productivity.

Figure 4. Example project board for student job portal system development with sections for software tasks. Icons: plus for adding task, grey bin for deleting section, red bin for deleting board.



3.6. Ethics Keyword Flagging System

The ethics keyword flagging system enables users to identify ethical concerns in their software design decisions on the Kanban board (Figure 4). Utilising semantic matching, a natural language processing technique, the system matches a predefined list of ethical keywords against task descriptions. This feature aims to aid software professionals in recognising ethical considerations, crucial in projects involving complex ethical dilemmas. The flagging system effectively alerts users to potential ethical dilemmas, fostering informed decision-making in software development (Figure 5). Complementarily, the chatbot recommendation assists users in resolving flagged ethical dilemmas by providing guidance and suggesting solutions (Figure 3). The ethics keyword flagging feature enhances ethical awareness, enabling users to identify and address ethical concerns in software development decisions.

3.7. Ethical Compliance Checklists

The ethics compliance checklists facilitate software professionals in self-assessing their adherence to ethical guidelines and data protection regulations. Incorporating essential frameworks such as the ACM Code of Ethics, General Data Protection Regulation (GDPR), and ethical AI system design, these checklists enable a comprehensive review of the software development process against ethical considerations (Figure 6). They prompt users to address privacy, security, and data protection factors, ensuring alignment with principles outlined in the aforementioned frameworks. Additionally, the checklists document compliance and serve as tools for transparency and accountability, demonstrating alignment with ethical principles to stakeholders. Their systematic application throughout the software development process can prevent oversight of ethical issues and foster ethical awareness. Integrated into the ethics project management tool, users can customise checklists based on their project domains, offering a practical approach which ensures ethical compliance in software engineering.

Figure 5. Ethics keyword flagging highlights terms and principles in the Code of Ethics. The system compares an example task description with an ethics keyword collection using semantic matching.

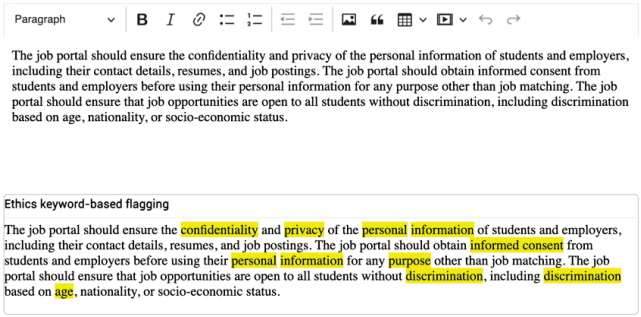
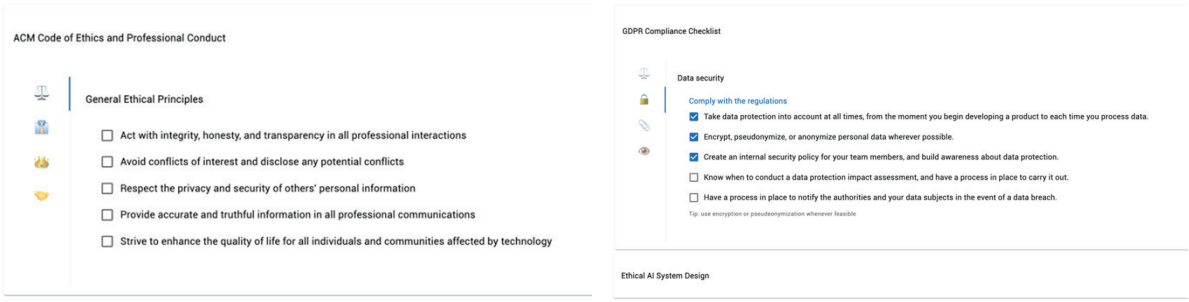


Figure 6. ACM Code of Ethics, GDPR Compliance and Ethical AI System Design checklists have a uniform compact display which allows users to switch between lists by clicking on the emoji icons.



3.8. Ethical Self-Assessment Form and Ethics Documentation

The ethical self-assessment form, within the ethics-centred project management tool, offers a structured approach for software professionals to evaluate their project's adherence to the Code of Ethics. Customisable for specific projects, the form streamlines the assessment of ethical practices, covering areas like data privacy, security, and algorithmic bias. The form's questions guide users in considering engineering practices and ensuring ethical alignment in software development. The downloadable self-assessed ethics document serves as a project record, beneficial for audits or demonstrating ethical commitments to stakeholders. This document can also be shared externally, fostering transparency with end users or regulatory bodies. The self-assessment and its documentation provide a systematic means to assess and record a project's ethical adherence.

4. TOOL EVALUATION SURVEY

Following ethics approval obtained from the authors' institution, an anonymous questionnaire was employed to evaluate the tool with users. The user evaluation questionnaire, conducted using Qualtrics, aimed to gauge the tool's efficacy in supporting ethical practices in software development. Additionally, it explored the support provided by the tool for ethical training, guideline customisation, identification and mitigation of potential ethical risks, and the promotion of ethical awareness in decision-making. The questionnaire contained 19 statements which evaluated the interface clarity, ease of use, progress tracking, and integration of ethical guidelines of the tool during software development. Additionally, questions addressed the tool's effectiveness in identifying and mitigating ethical dilemmas, providing resources for addressing dilemmas, and integrating with existing project management tools. Participants were instructed to indicate whether they agreed or disagreed with each statement and to what extent. Qualtrics was chosen for designing the questionnaire as it allowed

intuitive customisation and offered data analysis tools. It was used to facilitate post-collection analysis through data visualisations, as presented in the results.

The questionnaire, employing opportunistic sampling, gathered feedback from 21 students on the Computer Science programme at the authors' institution. They used the prototype tool before completing the survey.

5. RESULTS

The evaluation of the tool yielded positive overall feedback, highlighting its user-friendly interface and efficacy in assessing ethics practices in software development. Users commended its customisability and support for training software engineers on ethical matters. They indicated that engaging in ethical scenarios and adventure games facilitated interactive learning, while the ethics checklist effectively assessed project ethical practices. Users identified the limitations of the recommendation system, and suggested additional features including a glossary for ethics terms in software development, version control, a peer review system, push notifications, reflections on ethical dilemma decisions, and guidelines for non-agile processes. Highlighting ethics keywords on the Kanban board, push notifications and player reflection were implemented in the next version of the tool. User feedback also underscored the tool's effectiveness in facilitating ethical decision-making, with praise for the ethics checklist and suggestions for further features. The ethics dilemma adventure games received a commendation for their intuitiveness, though a query was raised about project management alterations under a different development process. Despite some users finding the recommendation system limited, the consensus was that the tool possesses a user-friendly interface.

Figure 7. Number of participants in agreement that "the tool supports the training and education of software engineers on ethical dilemmas and best practices of software development ethics".

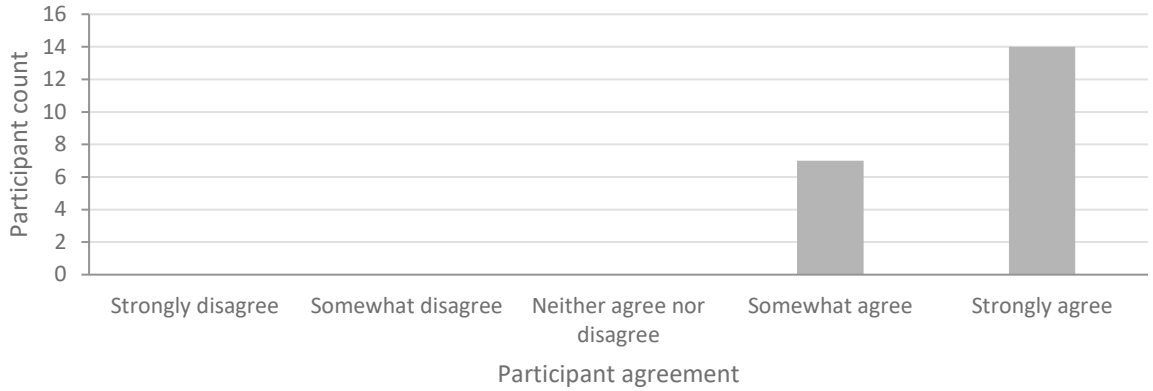
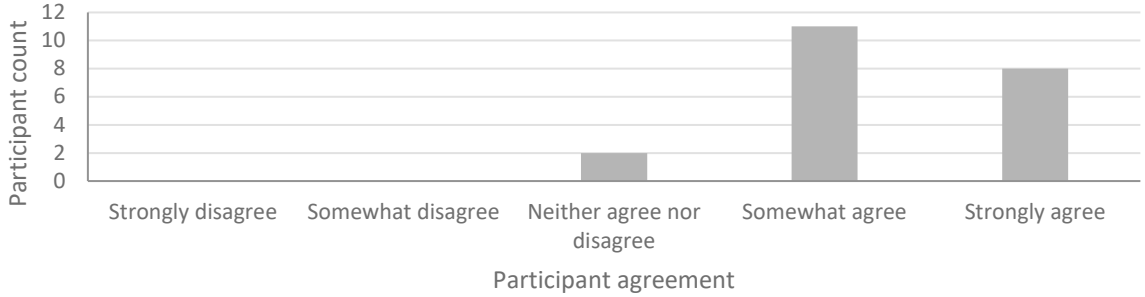


Figure 8. Number of participants in agreement that "the tool provides resources and support for addressing ethical dilemmas and challenges that may arise during software development".



Two of the 19 survey results are illustrated in Figures 7 and 8. The evaluation results offer useful insights into the tool's usability and effectiveness in promoting ethical deliberation. Among 21 participants, 66.67% of users strongly agree and 33.33% somewhat agree that the tool supports the training and education of software engineers on ethical dilemmas and best practices (Figure 7). Regarding the tool's ability to provide resources and support for addressing ethical dilemmas, 38.10% strongly agree, 52.38% somewhat agree, and 9.52% neither agree nor disagree (Figure 8).

6. STRENGTHS AND LIMITATIONS OF FINDINGS

A review of prevailing ethical frameworks, including the ACM Code of Ethics and Professional Conduct/IEEE Code of Ethics, forms the foundation for a robust ethics-driven development framework. While leveraging established frameworks ensures alignment with industry standards, the limitation lies in potential inflexibility in addressing unique ethical scenarios. Implementing an engaging training curriculum, featuring text adventure games and chatbot recommendations, enhances ethical awareness and retention rates. However, this approach may not cater to all learning styles, potentially impacting engagement with gamified ethical dilemmas. Integrating pop-up tips in the training curriculum provides immediate user feedback, strengthening the learning process. Nonetheless, some users may find these intrusive, which may be resolved via an option to dismiss reminders. Additionally, the tool evaluation assesses the training curriculum's usability, offering useful feedback for improvement, though limitations may exist in response honesty. An agile project management tool, integrating a Kanban board to highlight ethical concerns in the development lifecycle, aligns the process with ethical practices. Implementing an ethics self-assessment form and documentation system streamlines project ethics assessment, generating downloadable documentation for auditing ethical software development. The chatbot recommendation and ethics keyword flagging, based on natural language processing, identify ethical dilemmas. However, the absence of advanced logic poses limitations in the accuracy of recommendation, leading to potential false positives and negatives in chatbot responses.

7. DISCUSSion: Challenges and opportunities

The adaptation of ethical frameworks for software development is intricate due to diverse stakeholder perspectives, necessitating a comprehensive literature review for synthesising ethical considerations. Stakeholder involvement in the development process can provide valuable insights into unique ethical concerns. Designing an interactive training curriculum and project management tool for complex ethical issues presented obstacles, mitigated by strategies like incorporating real-life ethical scenarios, gamification, and offering diverse training formats to sustain user engagement. The intricate task of balancing ethical considerations with various project requirements (such as time and budget constraints) is acknowledged, prompting the development of decision-making frameworks explicitly considering ethics to prevent oversight or undervaluation. Regular reviews are emphasised to ensure the ethical framework's ongoing relevance.

Regarding opportunities, the development of an ethics-centred training curriculum and project management tool presents an opportunity to elevate ethical awareness amongst stakeholders. Integrating ethics compliance checklists ensures project compliance with ethical guidelines and government data protection regulations. The implementation of chatbot recommendation and ethics keyword flagging enhances the user experience by highlighting ethical dilemmas and allowing for the customisation of ethical protocols. Furthermore, the incorporation of an ethics self-assessment form and documentation system distinguishes the software from existing project management tools, potentially providing a competitive advantage in evaluating ethical software development.

8. CONCLUSION

In conclusion, this paper contributes to the software engineering community by introducing an extensible training resource and a customisable agile project management tool that prominently highlights ethical dilemmas. The proposed tool aims to addresses the gap in ethical training and tools available for software professionals. By integrating ethical deliberation in project management, this proof-of-concept tool serves as a catalyst for enhancing ethical practices within the software development community. Moving forward, the integration of these tools into industry practices has the potential to foster a culture of ethical responsibility, ensuring that software professionals are well-equipped to navigate and address ethical dilemmas effectively. This work lays a foundation for future advancements in ethical considerations within software development, promoting a holistic and principled approach in the ever-evolving landscape of technology and software engineering.

REFERENCES

Ali, B. A., Ul Haq, M. M. A., Al-Rebh, A. A., Al-Qahtani, M., & Al-Qurashi, T. (2012). Decision making in ethical dilemma. *2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies*, 589–599.

Bynum, T. W. (2000). The foundation of Computer Ethics. *ACM SIGCAS Computers and Society*, 30(2), 6–13. https://doi.org/10.1145/572230.572231

Ebert, C., Gallardo, G., Hernantes, J., & Serrano, N. (2016). DevOps. *IEEE Software*, 33(3), 94–100. https://doi.org/10.1109/ms.2016.68

Evans, S. (2012). Ethics in software engineering, net neutrality and badware - what obligations rest on software engineers to act ethically in their profession? *Proceedings of the 2012 4th IEEE Software Engineering Colloquium (SE)*, 4–6. https://doi.org/10.1109/se.2012.6242348

Ferrell, O. C., & Gresham, L. G. (1985). A contingency framework for understanding ethical decision making in marketing. *Journal of Marketing*, 49(3), 87–96. https://doi.org/10.1177/002224298504900308

Floridi, L. (2008). Foundations of Information Ethics. *The Handbook of Information and Computer Ethics*, 1–23. https://doi.org/10.1002/9780470281819.ch1

Godfrey, R. (1996). The Compleat Software Engineering Professional-doing the right thing as well as doing it right: Five steps on the road to an ethics curriculum. *Proceedings 1996 International Conference Software Engineering: Education and Practice*, 26–32. https://doi.org/10.1109/seep.1996.533977

Gotterbarn, D., Bruckman, A., Flick, C., Miller, K., & Wolf, M. J. (2017). *ACM Code of ethics. Communications of the ACM*, 61(1), 121–128. https://doi.org/10.1145/3173016

Gotterbarn, D., & Miller, K. W. (2009). The public is the priority: Making decisions using the Software Engineering Code of ethics. *Computer*, 42(6), 66–73. https://doi.org/10.1109/mc.2009.204

Gotterbarn, D., Miller, K. W., & Rogerson, S. (1997). Software engineering code of ethics. *Communications of the ACM*, *40*(11), 110–118. https://doi.org/10.1145/265684.265699

Gotterbarn, D., Miller, K. W., & Rogerson, S. (1999). Computer Society and ACM approve software engineering code of ethics. *Computer*, 32(10), 84–88. https://doi.org/10.1109/mc.1999.796142

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, 16(2), 366–395. https://doi.org/10.5465/amr.1991.4278958

Judy, K. H. (2009). Agile principles and ethical conduct. *2009 42nd Hawaii International Conference on System Sciences*, 1–8. https://doi.org/10.1109/hicss.2009.53

Karim, N. S., Ammar, F. A., & Aziz, R. (2017). Ethical software: Integrating code of ethics into software development life cycle. *2017 International Conference on Computer and Applications (ICCA)*, 290–298. https://doi.org/10.1109/comapp.2017.8079763

Lurie, Y., & Mark, S. (2016). Professional Ethics of Software Engineers: An ethical framework. *Science and Engineering Ethics*, 22(2), 417–434. https://doi.org/10.1007/s11948-015-9665-x

Rest, J. R. (1986). Moral research methodology: James R. Rest. *Lawrence Kohlberg*, 454–468. https://doi.org/10.4324/9780203823781-44

Schuelke-Leech, B.-A., Leech, T. C., Barry, B., & Jordan-Mattingly, S. (2018). Ethical dilemmas for engineers in the development of Autonomous Systems. *2018 IEEE International Symposium on Technology and Society (ISTAS)*, 49–54. https://doi.org/10.1109/istas.2018.8638282

Simon, J. (2016). Values in design. *Handbuch Medien- Und Informationsethik*, 357–364. https://doi.org/10.1007/978-3-476-05394-7_49

Sommerville, I. (2011) Software Engineering. 9th Edition, Pearson.

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. *Handbook of Ethics, Values, and Technological Design*, 1–7. https://doi.org/10.1007/978-94-007-6970-0_40

Trevino, L. K. (1986). Ethical decision making in organizations: A person-situation interactionist model. *Academy of Management Review*, 11(3), 601–617. https://doi.org/10.5465/amr.1986.4306235

ADDRESSING THE AI RESPONSIBILITY GAP WITH THE ACM CODE OF ETHICS

Don Gotterbarn, Marty J. Wolf

East Tennessee State University (USA), Bemidji State University (USA)

don@gotterbarn.com; Marty.Wolf@bemidjistate.edu

ABSTRACT

The Responsibility Gap arises in Artificial Intelligence (AI) development due to breaks in the accountability causal chain between an event and a responsible agent. We argue that by taking a broader view of responsibility that includes both backward looking accountability and forward looking positive responsibility, the AI responsibility gap can be avoided. Positive responsibility requires a shared set of values among AI developers. We argue that the ACM Code of Ethics and Professional Conduct reflects such a set of values and by using tools and techniques designed to embrace and support positive responsibility, less harmful AI becomes more likely. Further, AI for good--AI that addresses existing societal harms--is also a more likely outcome.

KEYWORDS: Responsibility, Responsibility Gap, Positive Responsibility, Artificial Intelligence Responsibility, ACM Code of Ethics, Proactive CARE.

1. INTRODUCTION

Like many types of technology, early computing technology was developed, at least in part, to support and advance military goals. For some, this put the technology on morally shaky grounds. As computing technology advanced and became more generally accessible, it increasingly was used by bad actors to intentionally cause harm outside the military realm. Others used the technology for seemingly innocuous purposes and caused harm that was unseen by them, but felt by others. Recognition of these situations sometimes led to an “ethical hysteria” where people used ethical expressions in unhelpful ways or developed tools that only partially addressed the harms or, worse, went down a path unrelated to addressing the harms.

An early attempt to address harm caused by computing was the introduction of the notion of “Software Engineering” in 1968 at the first NATO Software Engineering Conferences. The goal was to establish the professional technical skills that were needed to solve the “software crisis” of failed software (Naur & Randell 1968). Solutions tended to be technical, centered on the processes used to develop software; that is, solutions align the software development process with engineering models. Naur and Randell defined “software engineering techniques” to resolve the “software crisis” (1968). As new software life cycles and case tools were developed, they were supposed to help create better, more reliable, and less harmful software. The emphasis on “reliable” meant consistent performance that conformed to specified requirements. Developing more reliable software was also an impetus for improving programming languages. Languages were designed to reduce programming errors and make the translation of a design into code less likely to introduce errors.

Later, documentation and close adherence to formal development processes were emphasized. They were prevalent in the way computer science and programming were taught. Some saw the U.S. government’s adoption of the DOD-STD-2167A standard (2167a) as requiring that military software be developed using the waterfall development model. This led many to view ethics as merely compliance with the standard. There was an attitude that professional responsibility amounted to merely ensuring that the proper software development techniques were used and then closely adhered to. The need to help practitioners recognize and mitigate potential ethical difficulties was largely ignored in many

sectors of education and industry. Professional responsibility was increasingly understood as producing a reliable product that met specifications in a timely fashion. Gotterbarn was among those who argued that "computer ethics" was merely a professional ethics devoted to the development and advancement of standards of good practice (Gotterbarn 1991).

The focus on these technical solutions did not eliminate the software crisis. A well-known example is the case of the Therac-25, a radiation therapy machine where the safety of the systems was primarily left to computer controls. The machine overdosed multiple patients causing severe radiation burns and death in several cases (see Levenson and Turner (1993) for a full description). The analysis of these incidents pointed to many sources for blame including poor design, obscure error messages, and coding errors. Some (Leveson and Turner 1993) argued the contribution of "many hands" to these events made it impossible and pointless to try to assign accountability. Nissenbaum rejected treating situations with a complex network of causes and decisions as "mere accidents" (1996, p. 31). Many hands, she claims, does "obscure accountability," but it does not "follow that the harms were mere accidents." She thinks that treating such incidents as accidents means accepting them as agentless mishaps, rather than as accountable events (1996, p. 32). She advocates assigning accountability of varying degrees to those involved in the design, implementation, and use of the device. In her analysis of the Therac 25 case, the focus was on the technical and procedural aspects of the case and a concern for a lack of quality in those aspects.

In 1992, in the shadow of Therac-25 like events, the ACM revised its Code of Ethics and Professional Conduct. Like the analysis of Therac-25, it was still primarily about computing professionals building systems that were consistent with stated requirements. The guidance for Imperative 2.1 of the 1992 ACM Code tells readers that "[t]he computing professional must strive to achieve quality and to be cognizant of the serious negative consequences that may result from poor quality in a system" (ACM 1992). That code included an obligation to identify requirements at the beginning of the development process and later test that those requirements were met:

> **3.4 Ensure that users and those who will be affected by a system have their needs clearly articulated during the assessment and design of requirements; later the system must be validated to meet requirements.**
>
> Current system users, potential users and other persons whose lives may be affected by a system must have their needs assessed and incorporated in the statement of requirements. System validation should ensure compliance with those requirements. (ACM 1992)

This imperative reflects the steps of the waterfall model of software development: gather requirements, design, develop, test, deploy. However, best practice in software engineering had moved on to iterative processes that allowed for or even required revisiting requirements and design for correction and enhancement. Later the essential importance of both ethical and technical issues in addressing the software crisis was formalized when the IEEE-CS and the ACM published software engineering ethics standards in the IEEE/ACM Software Engineering Code of Ethics (Gotterbarn et al. 1999). The Preamble establishes a bias toward the well-being and quality of life of the public: "In all these judgments concern for the health, safety and welfare of the public is primary; that is, the 'Public Interest' is central to this Code" (Gotterbarn et al. 1999).

The Internet revolution had similar missteps in holding too narrow of a focus. A significant ethical problem in the early days of the Internet was pornography, especially child pornography, which was initially addressed in the U.S. with the enactment of the Child Pornography Prevention Act of 1996. At the 1998 SIGCSE Technical Symposium, during a discussion of Carswell's paper on the then new idea of delivering classes over the Internet (Carswell 1998), several audience members claimed that the

Internet had nothing to do with ethics, it was just about communicating information. These examples demonstrate that a narrow focus on technical issues rather than impacts computing has on a range of stakeholders was common among computing professionals, as was the assumption of the ethical neutrality of computing.

Thus, the computing profession has a track record whereby each new computing ethics awakening repeats mistakes of an earlier era and goes down the wrong path, at least initially. Last century when software development was working toward becoming an "organized" and "professional" discipline many difficulties were blamed on a lack of technical skills while at the same time not acknowledging any ethical responsibility. The way to become a computing professional was to be certified in particular skills that were needed to implement system requirements exactly as stated. That implementation was the full measure of the computing professional's ethical responsibility. Indeed, formal structures were put in place to reflect this value. In 1993 a group of eight professional computer societies founded the Institute for the Certification of Computing Professionals to certify skills in information processing (ACE). The certification of computing skills has become an industry with globally defined skills and competency standards (SFIA 2023).

Examining the current ethical awakening surrounding AI, especially machine learning (ML) and generative AI, we note some important differences from the earlier software crisis. First, and importantly, there is a broad range of people who are at the table discussing the ethics and social implications of AI. The different perspectives brought by philosophers, ethicists, linguists, sociologists, computer scientists, mathematicians, data scientists, humanities scholars, and so many more, all come to bear on identifying actual and potential harms of this technology. Further, they offer suggestions on different ways to prioritize mitigating those harms. Another major change is that there is a greater misalignment between corporate interests and "good AI" (AI that does good for society) than with previous ethical awakenings. Developing software that met specification was good for business. Having developers understand the same software lifecycles added efficiencies to software development and contributed to the corporate bottom line. Reducing the availability of harmful content is valued by society, allowing companies to steer clear of bad public relations.

Bender et al. (2021) layout many pressing concerns about the development of large language models (LLMs). To address those concerns and more general concerns about ML, scholars have developed tools such as Datasheets (Gebru et al. 2021), Model Cards (Mitchell et al. 2019), Data Statements (Bender & Friedman 2018), and Method Cards (Adkins et al. 2022). Unlike attempts to address the software crisis, these tools explicitly build in ethical consideration. Yet, like those earlier attempts, the tools address the *processes* by which AI is developed. It appears that we are having yet another computing ethical awakening that is potentially repeating the mistakes from earlier eras.

In this work, we argue that one mistake is understanding responsibility in a narrow sense--responsibility as accountability. John Ladd argues for a dual notion of responsibility that includes both backward-looking and forward-looking aspects. Ladd (1991) calls the backward-looking aspects of responsibility that involve accountability and causal chains, "negative responsibility." Assigned after the fact, negative responsibility locates the individual who was the direct cause of an unfortunate incident. Negative responsibility is also used to excuse people from moral responsibility. This happens when it can be shown that someone is not in the direct causal chain or when there are legal extenuating circumstances granting deniable culpability, thus breaking the direct causal chain. One example of this sense of negative responsibility is legislation, often with vague or undefined terms, like the U.S. Computer Fraud and Abuse Act of 1986 against accessing a computer without authorization.

Ladd's "positive responsibility" includes proactive consideration of what ought to be done. Unlike negative responsibility, which tends to be direct, positive responsibility can be indirect. Ladd argues

that pointing to the technology as a primary source of the harm does not remove from the practitioner this sense of positive responsibility (Ladd 1991, p. 675). Thus, in Ladd's view, a professional is not merely a technician, but someone who additionally is responsible for exercising a higher order of care for those whom their work impacts. This forward-looking responsibility is expressly included in Principle 1.1 of the current ACM Code of Ethics and Professional Conduct, "A computing professional should [c]ontribute to society and to human well-being, acknowledging that all people are stakeholders in computing" (ACM 2018). Additionally, the guidance for Principle 2.5 specifically mentions ML systems and the positive responsibility of their developers to do proactive analysis:

> Extraordinary care should be taken to identify and mitigate potential risks in machine learning systems. A system for which future risks cannot be reliably predicted requires frequent reassessment of risk as the system evolves in use, or it should not be deployed. Any issues that might result in major risk must be reported to appropriate parties. (ACM 2018)

Ladd's notion of positive responsibility and these observations from the ACM Code of Ethics tell us that while the tools being developed for use to address ethical concerns in the ML development process are an important piece of the process, there is still further to go in AI ethics. We will summarize some of the problems with AI in general and ML in particular in section 2. One of those concerns is the "responsibility gap." We analyze tools to address it. In section 3 we argue that the gap can be avoided. Section 4, concludes the paper.

2. CONCERNS WITH ARTIFICIAL INTELLIGENCE

AI is a term that encompasses a broad range of technologies, including expert systems and ML systems. Advances in ML, especially in the context of large language models (LLMs), have led to closer analysis of its workflow. Unlike in traditional rule-based software input data, especially training data, take on a more substantial role in the "programming" process. This transition from causally linked rule-based programs obscures the transparency of the decision process that leads to particular outcomes.

The opaqueness of both how patterns are extracted from raw, unanalyzed data and the resulting logical processes for decisions has led to ongoing concerns with ML and its use in decision support and decision making. Type 1 decision support involves systems that recognize patterns in complex data that help users make judgments about individual behavior. Some systems go beyond presenting patterns and actually make recommendations based on probabilistic outcomes. This is type 2 decision support. The morality of the recommendation is not addressed unless it has been programmed into the AI by careful prior ethical analysis. Type 3 decision support removes or prohibits human intervention. In autonomous decision making, the AI system makes a decision and then "executes it," sometimes parroting words that sound like an apology, "I am sorry Dave."

As far back as 1972, there have been concerns about the roles AI plays in decision making (Dreyfus 1972). We have come a long way since then. On the one hand, there is considerable press coverage of the positive applications of AI. Despite these applications, there are widely known examples of biased decisions, misclassifications, overgeneralizations, exaggerated stereotypes, and unintended consequences. These harms stem from a variety of sources including the lack of contextual understanding by the system (or worse, system developers) and adversarial attacks.

When a human is responsible for these kinds of problems, they are held accountable, blamed as the cause. This is Ladd's sense of negative responsibility. In talking about this aspect of responsibility--accountability--Nissenbaum (1996) identified "four barriers to accountability" for general software. In addition to the Problem of Many Hands mentioned earlier, Bugs are also a barrier in the sense that

they are omnipresent and because of their inevitability offer a convenient way to disclaim blame. The Computer as Scapegoat barrier is the temptation we have to blame the computer itself, rather than those who were responsible for the harm. Ownership without Liability is software developers ability to disclaim warranties for their software. She denied the computer was a moral agent. With advances in AI this may not be all that clear in type 3 decision support.

When an AI system makes type 3 judgments there is clear difficulty in assigning responsibility/accountability to a person for any problems. When a human makes a decision to fire a missile and that action is simply mediated by a computer following a fixed set of rules there is a direct causal connection to the human decision maker. However, when the machine produces a decision, the assignment of responsibility for a bad decision is indirectly related to a person. Given the complexity of generating an AI decision, the causal chain, be it direct or indirect, is obscured by the rules and the data that are used, making the determination of a blameworthy group or individual difficult, getting more difficult as the decision type goes up. Blame requires a complete clear line to the cause of the problem. Adreas Matthias has called these voids in the responsibility causal chain the "responsibility gap" (2004), which results in no one who can be held morally responsible for the harm.

2.1. The Responsibility Gap(s)

Matthias introduced the notion of a responsibility gap in 2004. He identifies how in various forms of AI, "the programmer transfers part of his control over the product to the environment," with this being "particularly true for machines which continue to learn and adapt in their final operating environment" (Matthias 2004, p. 182). He argues that this transfer of control results in situations where it is not possible to hold a particular person (or persons) responsible for the machine's actions.

Since its introduction, many have addressed facets of the AI responsibility gap, including when and whether it exists. Our contention is that arguments surrounding responsibility gaps misdirect discussions about responsibility in the same way that the early computing responsibility discussions focused on a narrow notion of professional responsibility, thus missing the opportunity to more comprehensively improve computing. Next we review some of the recent work that challenges the narrow notion of responsibility taken by Matthias.

In introducing the notion of "agency laundering" Rubel notes "that Matthias's conception of the responsibility gap focuses on an automated system's causal responsibility for some outcome" (2019, p.1035). Rubel's conception of moral responsibility involves the conjunction of three notions of responsibility. Role responsibility comes through the role one is taking on and the usual expectations of someone in that role. Causal responsibility stems from the causal chain from one's action (or inaction) to the event under consideration. Finally, Rubel's moral responsibility requires the capacity to control the causal event under consideration, which also requires "access to relevant information" (2019, p. 1020). Like Matthias, Rubel requires a clear direct causal connection for there to be moral responsibility. This aligns with Ladd's negative responsibility. Interestingly, Rubel shows how an organization might use the responsibility gap to obscure the scope of its causal responsibility and, a fortiori, obscure Rubel's sense of moral responsibility that is tied to causal control.

Santoni de Sio and Mecacci (2021) consider the notion of responsibility more closely in the context of responsibility gaps and identify four different types of responsibility and corresponding gaps: culpability, moral accountability, public accountability, and active responsibility. They identify Matthias' notion of a responsibility gap as a culpability gap. A moral accountability gap arises because "AI may make individual persons less able to understand, explain, and reflect upon their own and other agents' behaviour" (Santoni de Sio & Mecacci 2021, p. 1065). The public accountability gap with AI arises from the black box nature of AI systems and the public, especially governmental, systems that

employ AI in decision making. Discretionary powers and the ability to explain decisions and decision-making processes are hidden inside the AI systems. Like Rubel discussions of causal connections, these three sense of responsibility are primarily backward looking.

Santoni de Sio and Mecacci's fourth type of responsibility looks more like Ladd's positive responsibility. "[A]ctive responsibility is forward-looking and concerns the goals, values, and (legal) norms that professionals such as engineers are supposed to promote and comply with as well as the consequences they need to prevent and avoid" (Santoni de Sio & Mecacci 2021, p. 1066). The active responsibility gap arises when engineers are not aware of "their respective moral and social obligations towards other agents'' (Santoni de Sio & Mecacci 2021, p. 1067) or when they are aware, they are not "*sufficiently able or motivated* to fulfil an obligation" (Santoni de Sio & Mecacci 2021, p. 1068, italics in the original).

Tigard (2021) argues that Matthias and others have a narrow conception of the notion of responsibility. Namely, responsibility is understood "only in the terms of accountability" (Tigard 2021 p. 598) and more narrowly, backward looking accountability. Tigard goes on to argue that there is a richer set of considerations that includes attributability, answerability, and forward accountability. Attributability focuses on the underlying cares and commitments of an actor. Answerability calls on someone to provide explanations. Forward accountability focuses on actors who are responsible for addressing demands about what we would like to have happen in the future. This sort of forward accountability seems to be limited to situations where an act has already occurred and there is interest in changing future behavior.

2.2. Tools for Addressing the Responsibility Gap(s)

We have argued that the historical evidence shows that merely modifying software development processes has not fully addressed ethical problems in software development. Next we consider four tools that change the ML development process and at least partially address the responsibility gap. While they invite their users to become more involved in positive responsibility, we argue below (and this is acknowledged by some authors) that they do not help individuals identify and facilitate ethical goals or make ethical judgements--all part of Ladd's more comprehensive sense of responsibility.

Bender and Friedman (2018) introduced data statements for use in natural language processing in order to reduce bias. They propose a schema in which properties of the input speech data and the intentions of the curators are made explicit. Their system aids in mitigating bias because those deploying systems trained on data sets with data statements "are empowered to assess potential gaps between the speaker populations represented in the training and test data and the populations whose language the system will be working with" (Bender and Friedman 2018). They see data statements (or a similar practice) as a piece of "critical enabling infrastructure" to address bias problems in natural language processing.

While data statements are narrowly focused on the creation of data sets for natural language processing, datasheets for datasets is a more general approach for arbitrary data sets, and it is based on the notion of datasheets that is used in the electronics industry. Introduced by Gebru et al. (2021) datasheets for datasets consists of a sequence of 57 questions split across 6 broad categories. Some of the questions approach technical aspects of the dataset, but others get at important ethical edges. Even so, many are questions that allow for binary answers when typical ethical problems require complex answers.

Model cards are more general in that they cover everything from the dataset to the model. Model cards "aims to standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical,

inclusive, and fair considerations" (Mitchell et al. 2019). Model cards include practical details about the model, the intended uses of the model, as well as factors that might influence future ethical decision making regarding whether to use the model to develop a particular application with the model. It raises questions for consideration in five categories: data, human life, risk mitigations, risks and harms, and use cases.

From our perspective, these techniques might identify some problems without determining if there is a moral obligation to address them and if so, who should be providing ways to meet those obligations. Further, there is no mention of providing criteria for evaluating alternative ways to meet the obligation. Professionalism asks people to bring insights and a higher order of care to their work that goes beyond minimally accomplishing a task. These techniques do not address the professional's obligation to proactively consider how to support ethical opportunities to morally improve ways to to satisfy a systems requirements.

Adkins et al. (2022) note that the above techniques are descriptive techniques. In response, they introduce method cards, which incorporates prescriptive information into the communication process. That is rather than describe features of the dataset or model,

> method creators are expected to provide sufficient instructions and documentation to guide ML engineers. These instructions should help the engineers in choosing appropriate preprocessing steps, model components, and hyperparameters for their task. Furthermore, the instructions should make these engineers aware of how their choices can potentially impact the model's behavior, how to evaluate this impact, and how to handle prediction errors accordingly. Likewise, the instructions should enable defining a suitable evaluation and benchmarking setup for the method. Finally, the instructions should explicitly inform the engineers on responsible usage of the method by addressing potential fairness and privacy concerns. (Adkins et al. 2022)

The evaluation they propose is about the method quality not the "obligation satisfaction quality." Clearly method cards places responsibility for avoiding bad practice with a method at the feet of the method creators. The method creator must first give ethical consideration to the system under development, and then must provide guidance for ML engineers on how to use the system in a way that prevents harm and in a way that is consistent with how the system was developed. This places a greater expectation of ethical excellence on the method creators than any of the three previously discussed methodologies, while seemingly, or at least potentially, reducing the ethical responsibility of the ML engineers. Our concern is that the latter need only follow the prescription to meet their ethical obligations. This is simple causal responsibility.

We raise two concerns with these approaches to the responsibility gap. First, like most responses to the software crisis, these tools focus primarily on improving technical solutions to the backward traceability problem of determining accountability. Unlike the earlier technical solutions, they recognize that something needs to be done about ethical responsibility and draw attention to the importance of everyone in the development process being significant to producing good ethical outcomes. In addition to addressing the accountability causal chain, they offer at least the suggestion that positive responsibility is important. But the second concern with these tools is that they misdirect the approach to addressing ethics and do not center proactive actions to mitigate future problems. Most of the discussion about the responsibility gap results in the same limited view of responsibility found in responses to problems in earlier software development in that they address one portion of the epistemological problem--being aware that there is a moral responsibility. Yet these methods do not provide support to practitioners in understanding responsibility. Those who hold the mistaken belief that technology is ethically neutral leads to the belief that the production of the technology is

not the ethical problem and justifies not “being motivated to fulfil an obligation” (Santoni de Sio & Mecacci 2021). We can see the ghosts of assertions from the 1960s that computers just do arithmetic. The only ethical problem is the bad people who misuse programs. A developer’s responsibility was merely the quality of the software.

3. AVOIDING THE GAP

3.1. Minding The Gap Diverts Attention From Doing Better

Responsibility as accountability was used during the software crisis of the 1960s to blame developers for the failure to develop “reliable” systems. The resulting focus on the software development process, which emphasized finding a program's errors and avoiding those technical errors in meeting requirements, left little room for developers to learn how to anticipate and avoid negative ethical impacts and to maximize positive ethical outcomes. Worse, some developers would dodge responsibility by blaming the client for inadequately specifying requirements. The focus on responsibility as accountability led to significant effort being placed on developing precise technical requirements for a problem solution and adhering to them in the development process. Ethical requirements were understood as technically proficient development that satisfied some list of activities that needed to be checked off.

Developers ignored or were ignorant of the positive responsibility of recognizing and avoiding anticipatable negative ethical impacts of their work. Development techniques did not include suggestions that system requirements or development methodology might be adjusted to improve ethical outcomes for the client and those impacted by the software; they did not address ethical opportunities.

By turning our attention away from the responsibility gap and toward positive responsibility, we can more effectively address problems in AI ethics. We identify two facets of positive responsibility, one technically based and the other is based on values. These two facets of positive responsibility are necessary for effectively approaching the technical facet of professional responsibility. Data statements, model cards, and method cards are focused on the ML development process much like attempts to address the 1960s software crisis. They differ in that they provide opportunities for positive responsibility to be part of the process. They reflect at a minimal level Ladd’s notion of positive responsibility. More philosophically, Santoni de Sio and Mecacci’s active responsibility aligns with Ladd’s notion of positive responsibility in that it is forward looking. Further, it recognizes the problems of computing professionals' ignorance of and lacking motivation to meet their moral obligations.

The second facet of positive responsibility addresses this issue. It starts with actively engaging with the prospect that computing systems will have an impact. Guidance from Principle 2.2 of the ACM Code of Ethics and Professional Conduct is clear: “Professional competence starts with technical knowledge and with awareness of the social context in which their work may be deployed.” Computing professionals are responsible for applying standards within their profession and attempting to avoid anticipatable negative ethical impacts of their work.

Focus on the AI responsibility gap diverts attention from opportunities to engage with positive responsibility. Underlying the AI responsibility gap is an assumption of a causal chain looking for a particular person. Implicitly, this makes standard responsibility denial moves much easier and misses an opportunity to change the behavior of system developers. The discussion of accountability is important but it must not distract us from actively addressing opportunities to practice a higher order of care. Positive responsibility is an essential part of professional responsibility. Professional

responsibility includes knowing about and being concerned for others who are impacted by one's professional actions.

The next step is to encourage and facilitate this facet of positive responsibility. Policy and law are traditionally slow and incomplete ways to do so. Next we suggest some tools that can be used to guide system development so that the resulting system is more likely to reflect shared professional values, such as those found in the ACM Code of Ethics and Professional Conduct, and is more likely to meet the needs of all stakeholders.

3.2. Leveraging Shared Values

Santoni de Sio & Mecacci (2021) identified two difficulties in achieving positive responsibility, the epistemological problem of knowing that there are moral obligations and, once known, that there may be a lack of motivation to try to meet those obligations.

One way to address the epistemological problem begins with values that are broadly shared by the computing community and then incorporating them into the practice of computing more broadly, and AI in particular. Regular discussions among developers of ethical concerns surrounding AI systems and subsequent decision-making about open-ended ethical goals leads to not only better processes, but systems that reflect shared values that prevent harms and advance the public good. Such discussions and transparent ethical decision making raise the visibility of "doing ethics," leading to an environment where developers are well-practiced at both knowing and doing ethics. We point to two tools that embrace positive responsibility. One is the ACM Code of Ethics and Professional Conduct, an aspirational guide that invites professionals to contribute to society in ways that both minimise unintentional ethical mistakes and maximise positive impacts. It encourages and identifies opportunities for positive action by practitioners who can apply these aspirational principles in their work.

The development of the Code and its acceptance by multiple computing societies, including the 54 national and regional members of the International Federation of Information Processing (Kreps 2020) show it has, to some degree, begun to address the epistemological problem of articulating common ethical values. The next step is to incorporate those values into daily professional practice so that practitioners develop opportunities for AI to contribute to society.

A heuristic tool for computing professionals, called Proactive CARE (PCARE) (Gotterbarn et al. 2022), helps translate the values expressed in the Code of Ethics into positive actions. A simple heuristic, Consider, Analyze, Review, Evaluate, guides a discussion and reasoning process to leverage the values in the Code into the daily decision-making process and computing workflow resulting in better computing systems and a better world. PCARE helps to identify ethical opportunities through an iterative consideration of stakeholders and the development and evaluation of alternative approaches to the system under development. "Once those opportunities are identified, integrating the Code, which was designed to guide and inspire, into the practice of computing for in-the-moment analysis, can lead to both products and processes that better integrate ethical values" (Gotterbarn, D., et al. 2022). PCARE calls for more cognitive attention to ethical concerns during the development and decision-making processes. It takes attention to Consider relevant ethical elements, Analyze the impact of those elements, Review pertinent responsibilities, authority, and alternate approaches, and Evaluate those alternatives. Using tools like PCARE facilitates the internalized judgments consistent with the Code and thus expresses a higher order of care for all stakeholders. This internalization of value-guided judgment may help address the motivational problem by helping developers realize their power and their ability to affect others even though one does not intend, or even foresee, the likely

effect. The goals identified by PCARE are primarily calls to action, to promote ethical opportunities that are frequently missed.

3.3. Visual Analytics for Sense-making in Criminal Intelligence Analysis (VALCRI)

Not all decisions AI systems make are bad decisions, and many decisions are designed to mitigate risks through rigorous testing, validation, and ongoing monitoring. However, understanding and addressing potential pitfalls is essential to ensure responsible and ethical AI development and deployment. Attention to ethical values in European technical development projects is not new. Gotterbarn chaired an Internal Ethics Board (IEB) for a European Commission-funded research project to develop a semi-automated decision support system (both types 1 and 2) using visual technologies to aid police intelligence analysts in several EU nations. Rather than addressing ethics after the systems were built, the IEB addressed mitigating ethical issues during VALCRI's design and development. The emphasis was on positive responsibility, with a goal of identifying implicit and explicit values in design choices and the intentional and unintentional value choices made in the technology development of this decision support system.

Potential ethical issues and risks were identified for each element of the system, such as cognitive bias, reasoning reliability, data integrity, and privacy issues (Duquenoy 2018, p. 32). By focusing on the technologies, their impact on society, and the ethical issues with the various components of the system, the IEB, in conjunction with developers and stakeholders, established ways to mitigate each issue before the development of the system.

Built-in ethical safeguards have the system both identify bias and then either prevent or alert users to alter their (unintentional) decision strategies. For example, those using a computer-decision support system designed to aid visual reasoning tend to place too much confidence in the system's results. One method to address this bias is to introduce provenance and explicitly provide reasoning and empirical evidence for a decision (Based 2010). Identifying system bias in decisions and inferences made from the data was done through transparency in the operations of the system. This led to requiring understandable process logs and documentation to provide the rationale for the process and decisions made in the design choices; logging mechanisms that show changes made by a user, who that user was, when the changes were made, and other significant details. To ensure the integrity of the data and increase the reliability of inferences made from it, a "reliability tag" indicating the level of reliability was added to all data. A bias detection mechanism looked at patterns of decisions and judgments made by different users. Decisions deemed by the system as "out of the norm" clearly showed the criteria for that determination. A method of explaining the basis for the result was produced. It provided the rationale and alleviated concerns about discrimination.

4. CONCLUSION

With AI systems, understanding and addressing potential pitfalls associated with human developers is essential to ensuring responsible and ethical AI development and deployment. Our more comprehensive understanding of responsibility moves the discussion away from an emphasis on negative responsibility and blame and toward thinking about AI as an ethical opportunity to contribute to society. This is only part of the answer, but a significant part which has been ignored. The ACM Code of Ethics and Professional Conduct supports this sense of responsibility and advocates avoiding the responsibility gap through a broader understanding of responsibility. Using PCARE is one way to bring the aspirational values of the Code to the attention of AI developers and have those values permeate the development process.

Individual reflections on PCARE questions serve as a common starting point for conversation during all stages of development. These questions help establish a mindset. They are not a checklist of yes or no responses, but they require elaboration and creative thinking. This common starting point is a foundation for positive ethical action and outcomes.

We have argued that failure to recognize moral obligations, which also contributed to the difficulties rule-based computing had, as one of the causes of difficulties in AI ethics. A narrow focus on negative responsibility and technical solutionism is a distraction from a positive, forward-looking sense of responsibility that includes shared values and a search for ethical opportunities. The Code was built based on this more comprehensive sense of responsibility and the use of it through tools like PCARE will mitigate some AI harms.

REFERENCES

ACE https://www.acenet.edu/National-Guide/Pages/Organization.aspx?oid=b3532c35-75c4-ea11-a812-000d3a33232a accessed 01/02/2024

ACM Code of Ethics and Professional Conduct. (2018). https://www.acm.org/code-of-ethics

ACM Code of Ethics and Professional Conduct (1992). https://ethics.acm.org/code-of-ethics/previous-versions/1992-acm-code/

Adkins, D., Alsallakh. B., Cheema, A., Kokhlikyan, N., McReynolds, E., Mishra, P., Procope, C., Sawruk, J., Wang, E., Zvyagina, P. (2022). Method Cards for Prescriptive Machine-Learning Transparency CAIN'22, https://doi.org/10.1145/3522664.3528600

Based, A. (2010). Information handling in security solution decisions. In S. Tarek (Ed.), *Innovations and advances in computer science and engineering*. Dordrecht, Springer.

Bender, E.M. and Friedman, B. (2018). Data statements for nlp: toward mitigating system bias and enabling better science. Transactions of the ACL (TACL).

Bender, E. M, Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Carswell, L. (1998). The "Virtual University": toward an Internet paradigm? SIGCSE Bull. 30, 3 (Sept. 1998), 46–50. https://doi.org/10.1145/290320.283017

Dreyfus, H. (1972). *What Computers Can't Do*. New York: Harper and Row.

Duquenoy, P., Gotterbarn, D., Kimppa, K.K., Patrignani, N., William Wong, B.L. (2018). Addressing Ethical Challenges of Creating New Technology for Criminal Investigation: The VALCRI Project. In: Leventakis, G., Haberfeld, M. (eds) Societal Implications of Community-Oriented Policing and Technology. SpringerBriefs in Criminology. Springer, Cham., pp. 31-39. https://doi.org/10.1007/978-3-319-89297-9_4

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. Commun. ACM 64, 12, 86–92. https://doi.org/10.1145/3458723

Gotterbarn. D, (1991). "Computer Ethics, Responsibility Regained," National Forum: The Phi Beta Kapp Journal, 71:26-31.

Gotterbarn, D., Kirkpatrick, M.S., and Wolf, M.J. (2022). "From the page to practice: Support for computing professionals using a code of ethics," ETHICOMP 2022.

Gotterbarn, D., Miller, K., and Rogerson, S. (1999). Software engineering code of ethics is approved. Commun. ACM 42, 10 (October 1999), 102-107. http://doi.org/10.1145/317665.317682

Kreps, D. (2020). IFIP Adopts New Code of Ethics and Professional Conduct, https://www.ifipnews.org/ifip-adopts-new-code-ethics-professional-conduct/

Ladd, J. (1991). Computers and Moral Responsibility: A Framework for an Ethical Analysis, in: Dunlop, C. and Kling, R. (eds) Computerization and Controversy: Value conflicts and Social Choices,, Academic Press 1991 (664-675) Gould, Carol (ed.) The Information Web: Ethical and Social Implications of Computer Networking, Westview Press.

Leveson, N., and C. Turner. (1993). An investigation of the Therac-25 accidents. Computer 26(7): 18–41.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mitchell, M. Wu S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on Fairness, Accountability, and Transparency. 220–229.

Naur, P. & Randell, B., eds. (1968). Software Engineering: Report on a conference sponsored by the NATO Science Committee. http://homepages.cs.ncl.ac.uk/brian.randell/NATO/nato1968.PDF

Nissenbaum, H. (1996). Accountability in a computerized society. Sci Eng Ethics 2, 25–42. https://doi.org/10.1007/BF02639315

Rubel, A., Castro, C. & Pham, A. (2019). Agency laundering and information technologies. Ethic Theory Moral Prac 22, 1017–1041. https://doi.org/10.1007/s10677-019-10030-w

Santoni de Sio, F., Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos. Technol. 34, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

SFIA. (2023). Skills Framework for the Information Age. https://sfia-online.org/en

Tigard, D.W. (2021). There is no techno-responsibility gap. Philos. Technol. 34, 589–607. https://doi.org/10.1007/s13347-020-00414-7

THE ETHICAL AND LEGAL CHALLENGES OF DATA ALTRUISM FOR THE SCIENTIFIC RESEARCH SECTOR

Ludovica Paseri

Law Department, University of Turin (Italy)

ludovica.paseri@unito.it

ABSTRACT

Scientific research nowadays is increasingly data-driven and therefore a growing amount of data need to be accessible and of high quality. The data altruism mechanism, as regulated in the Data Governance Act (DGA), aims to meet this demand. This paper investigates how data altruism mechanisms apply to the scientific research sector. This mechanism, based on the voluntary release of data, raises several normative challenges. From a legal viewpoint, data altruism in the research sector entails (1) the risk of fragmentation; (2) security concerns; and (3) the duty of control on data altruism organisations. From an ethical perspective, is the challenges regard (1) the very idea of altruism between ethics and infra-ethics; (2) the interplay between public interest, general interest, and common good; and (3) a concern related to the autonomy of both data subjects and data holders. Given the set of challenges, both legal and ethical, and the multiplicity of actors involved in the data altruism mechanism, the intent of the analysis is to provide an assessment on how the data altruism mechanism should be implemented at national level.

KEYWORDS: data governance act, DGA, data governance, data altruism, scientific research, general interest, public interest.

1. INTRODUCTION

Scientific research nowadays is increasingly data-driven and therefore requires a growing amount of data, which need to be accessible and of high quality. The data altruism mechanism, that results as a means to meet this demand, is regulated by the Data Governance Act (DGA, hereinafter). The DGA is a Regulation of the European Union, which is applicable from 23 September 2023[1], that aims to "foster the availability of data for use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the EU", as described in the explanatory memorandum accompanying the proposal for a Regulation[2]. The DGA is a crucial part of the so-called "politics of data" (Pagallo, 2022) developed by the European Commission in 2020[3] and can also be considered as complementary to the Open Data Directive (ODD, hereinafter)[4], integrating the European framework on data sharing and reuse (Ruohonen & Mickelsson 2023). Article 3 of the DGA, which identifies the scope of application, underscores the complementarity between the DGA and the ODD by stating that the DGA provides for the reuse of certain categories of data, such as data held by the public sector that are protected on the basis of commercial confidentiality, statistical confidentiality, protection of third parties'

[1] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), ELI: http://data.europa.eu/eli/reg/2022/868/oj.

[2] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM/2020/767 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020PC0767.

[3] European Commission Communication, A European strategy for data, COM/2020/66 final (2020), ELI: https://eur-lex.europa.eu/legal- content/EN/TXT/?uri=CELEX:52020DC0066.

[4] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), ELI: http://data.europa.eu/eli/dir/2019/1024/oj.

intellectual property rights and protection of personal data. Therefore, the DGA concerns the reuse of the public sector data excluded from the scope of the ODD (Van Eechoud, 2021, p. 376).

Data altruism is defined by the DGA under Article 2(16):

> 'data altruism' means the consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services.

This paper aims to investigate the data altruism mechanism for the scientific research sector[5]. This mechanism, based on the voluntary release of data, raises several normative challenges. From a legal viewpoint, data altruism in the research sector entails the following challenges: (1) the risk of fragmentation; (2) security concerns; and (3) the duty of control on data altruism organisations. From an ethical perspective, starting from the assumption that data altruism is a form of distributed morality (Floridi, 2020), the analysis focuses on three main ethical issues: (1) the very idea of altruism between ethics and infra-ethics; (2) the interplay between public interest, general interest, and common good; and a concern related to the autonomy of both data subjects and data holders with the scope of consent.

The analysis argues that the manifold ethical and legal challenges may jeopardize the implementation of the data altruism mechanism, *a fortiori* considering that is partially delegated to national policies. Next section provides an overview of the data altruism mechanism as designed by the European institutions, focusing on the phases of the process, conditions and actors involved. Section 3 is devoted to the investigation of the legal issues; whilst Section 4 considers the ethical aspects. The conclusions provide some recommendations that should be adopted at the national level to mitigate some drawbacks of the European mechanism of data altruism.

2. DATA ALTRUISM UNDER THE DGA

The mechanism of data altruism involves several actors: (i) the subjects of personal data; (ii) the holders of non-personal data; (iii) the data altruism organisations; (iv) the data users; and (v) the competent authority for registration.

(i) The notion of data subject is indirectly derived from the definition of personal data provided in Article 4(1) of the GDPR. The data subject is the natural person, identified or identifiable, to whom the personal data pertain.

(ii) The data holder according to the Article 2(8) of the DGA is "a legal person, including public sector bodies and international organisations, or a natural person who is not a data subject with respect to the specific data in question, which, in accordance with applicable Union or national law, has the right to grant access to or to share certain personal data or non-personal data".

[5] Consider that the proposal for a Regulation on the European Health Data Space (EHDS) also makes specific reference to the data altruism mechanism: Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, COM/2022/197 final, ELI: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197. The EHDS Proposal does not define the data altruism mechanism but explicitly refers to the DGA definition (Article 2(1)*c* of the EHDS Proposal). The Article 40 of the EHDS Proposal is then specifically dedicated to "Data altruism in health". At the moment, the EHDS Proposal has not yet been adopted. This contribution is limited to the mechanism of data altruism as regulated in the DGA.

(iii) The data altruism organisations are legal entities, specifically registered as such, operating not for profit, independent of any profit-driven data processing activity.

(iv) The data user refers to any "natural or legal person who has lawful access to certain personal or non-personal data and has the right, including under Regulation (EU) 2016/679 in the case of personal data, to use that data for commercial or non-commercial purposes" (Article 2(9) of the DGA).

(v) The competent registration authorities are entities in charge of the data altruism organisation's registration process, designated in every Member States.

Underlying the operations of this multiplicity of actors there are two conditions enshrined in Article 2(16) of the DGA. First, data subjects and data holders must release their data to the data altruism organisation free of charge, not in return for a reward. This condition may be interpreted as a measure to avoid the establishment of a buying and selling of personal data[6].

The second condition is that the reuse needs to be carried out exclusively for general interest purposes. The purposes of general interest are specified in Recital 45, which states: "[S]uch purposes would include healthcare, combating climate change, improving mobility, facilitating the establishment of official statistics or improving the provision of public services. Support to scientific research, including for example technological development and demonstration, fundamental research, applied research and privately funded research, should be considered as well purposes of general interest". The mechanism of the data altruism, based on data "voluntarily made available by individuals or companies" (Proposal DGA, Explanatory Memorandum, 2020, p. 8), may generate a considerable impact on the data management in the scientific research sector.

According to the European institutions, the data altruism mechanism hinges on an articulated process with several phases. Any entity intending to be recognised as a data altruism organisation has to undergo a registration process, and among other information, has to declare "the purposes of general interest it intends to promote when collecting data" (Article 19(4)*h*, DGA).

As mentioned above, the scientific research is considered an example of general interest purposes (Article 2(16) of the DGA). However, indicating scientific research as a general interest purpose is a very broad, uninformative notion that adds to the on-going debate about the information to be provided to the data subject by the data controller under Articles 13 and 14 of the General Data Protection Regulation (GDPR, hereinafter) [7] (Hallinan, 2020, p. 8). Indeed, it is often complicated to specify the aims pursued in a specific scientific research project (Pagallo & Bassi, 2013, p. 183).

Once the requesting entity meets all the requirements laid down by the DGA, it will be included in the national register of data altruism organisations by the competent national authority or authorities, within 12 weeks from the date of application, pursuant to Article 19(5) of the DGA. The competent authority in charge of certifying an entity as a data altruism organisation is designated by each Member State and, under Article 23 of the DGA, is responsible for maintaining the national public register of such data altruism organisations.

The voluntary release of personal data by data subjects to the data altruism organisations is based on consent. The consent needs to be given in compliance with the two conditions described above, i.e.,

[6] Recital 45 DGA specifies that "Data subjects should be able to receive compensation related only to the costs they incur when making their data available for objectives of general interest".

[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) ELI: http://data.europa.eu/eli/reg/2016/679/oj.

no reward and public interest purposes. The registered data altruism organisations provide to several natural and legal persons (i.e., data users) the possibility to process the data they hold, for purposes of general interest, eventually on the basis of a fee. Each data altruism organisation is required to keep accurate records – very similar to the processing register set out in Article 30 of the GDPR – concerning a set of accurate information about the specific data processing activities, based on the data altruism consent.

In addition, any data altruism organisation, pursuant to Articles 20 and 21 of the DGA, has several reporting obligations towards data subjects and data holders. In particular, prior to any processing, entities have to meet two requirements. First, they are compelled to inform data subjects and data holders of "the objectives of general interest and, if applicable, the specified, explicit and legitimate purpose for which personal data is to be processed, and for which it permits the processing of their data by a data user" (Article 21(1)*a* of the DGA). Second, any entity must communicate "the location of and the objectives of general interest for which it permits any processing carried out in a third country, where the processing is carried out by the recognised data altruism organisation" (Article 21(1)*b* of the DGA). In this regard, it is worth noting that the corresponding article in the proposal of the Regulation (Article 19(1)*b* of the DGA Proposal) identified a general duty to communicate "any processing outside the Union". By contrast, in the Regulation in force, the wording seemingly provides that processing in a third country may be carried out only by the data altruism organisation itself.

According to the current structure of the data altruism mechanism, the data altruism organisation is pivotal. Article 19(2) of the proposal of the DGA bestowed a proper function of control on the data altruism organisation by stating that the "entity shall also ensure that the data is not be used for other purposes than those of general interest for which it permits the processing". The DGA currently in force is more softened than the proposal[8]. However, the central role and responsibility of the data altruism organisation in the control over data provided by data subjects and data holders persists, as does its challenging aspects, considering that the very definition of data reuse is intrinsically broad and open to many possible operations (Bassi, 2011, p. 67).

The purpose of the data altruism mechanism, as unfolded in Recital 46[9], is the establishment of data repositories providing pools of data able to generate value serving the general interest. Pursuing this goal, which could represent a "significant step towards decentralization of the web and de-monopolization of data" (Van de Hoven *et al.*, 2021, p. 143), remains problematic. Next section tackles the legal challenges of implementing the data altruism mechanism.

3. THE LEGAL CHALLENGES OF DATA ALTRUISM

In light of the analysis of the data altruism mechanism under the DGA, two crucial actors clearly stand out: the European Commission and the Member States. According to Article 22 of the DGA, some fundamental clarifications are left to delegated acts of the Commission, which are currently still pending. In fact, the European Commission will play a key role in specifying certain aspects of the functioning of the data altruism mechanism, such as information requirements for the provision of consent, the technical and security requirements, the communication duties, and interoperability standards. Furthermore, data altruism is up to the discretionary powers of the Member States: They

[8] See: Article 21(2) DGA.

[9] "The registration of recognised data altruism organisations and use of the label 'data altruism organisation recognised in the Union' is expected to lead to the establishment of data repositories. Registration […] is expected to facilitate […] the emergence of data pools covering several Member States.", Recital 46 DGA.

may (not 'shall') "establish national policies for data altruism", under Article 16 of the DGA (Baloup *et al*., 2021, p. 35).

However, prior to the further intervention of delegated acts of the European Commission and national policies, a few knots need to be untangled. They can be summarised in the following issues, which are investigated below: (1) the risk of fragmentation; (2) security concerns; (3) the duty of control on data altruism organisations.

3.1. Fragmentation

The DGA Regulation poses a crucial problem. On the one hand, the aim is to encourage the sharing of data by facilitating "the emergence of data pools covering several Member States" (Recital 46 of the DGA). In doing so, the registration of the data altruism organisation "is expected to facilitate cross-border data use within the Union" (Recital 46 of the DGA). On the other hand, however, Recital 48 emphasises that the DGA "should be without prejudice to the establishment, organisation and functioning of entities that seek to engage in data altruism pursuant to national law and build on national law requirements to operate lawfully in a Member State as a not-for-profit organisation". Risks of fragmentation follows as a result, given the wide leeway left to the initiatives of the Member States.

Only time will tell whether and how the Member States will implement the data altruism mechanism. So far, not much information is available about national strategies and the topic is under-researched in the literature.

Looking at the list of national data altruism authorities currently designated under the Article 23 of the DGA, there is no common strand in the action of the different Member States[10]. For instance, Spain has involved four entities: the Deputy Directorate General for Digital Society, the Ministry of Economic Affairs and Digital Transformation and two Spanish authorities on Artificial Intelligence (i.e., the Directorate General for Digitization and AI and the State Secretariat for Digitization and AI). Germany has nominated only one authority, i.e., the German Federal Network Agency. Further differences regard the controllers. Lithuania designated the national data protection authority (i.e., the State Data Protection Inspectorate); the Netherlands have opted for the Authority for Consumers and Markets; Finland, the Finnish Transport and Communications Agency Traficom. If, as specified in Recital 46, the registration of an entity as a data altruism organisation is meant to be valid throughout the whole territory of the Union[11], mechanisms of effective communication and cooperation between such different entities will have to be put in place in order to ensure the effective success of this form of data reuse.

In the European institutions' vision, this fragmentation should be overcome through a European form of consent to data processing given by the data subjects and the data holders (Article 25 of the DGA). Considering the difficulties that the choice of the most suitable legal basis for processing personal data for scientific research purposes has generated under the GDPR (Paseri, 2023, p. 11; Hallinan *et al*.,

[10] According to the Article 23(2) of the DGA "Each Member State shall notify the Commission of the identity of their competent authorities for the registration of data altruism organisations by 24 September 2023". However, currently, only 14 Member States have designated their national competent authority (the list is available here: https://digital-strategy.ec.europa.eu/en/policies/data-altruism-organisations). Significantly, very different institutions or authorities have been identified as data altruism competent authorities. At the moment, there is no uniformity of approach among the Member States.

[11] In this regard, the European Commission by means of an implementing act has provided for the design of a logo to identify such organisations unanimously throughout the whole territory of the European Union, as set out in the Article 17 of the DGA.

2023), it is fair to concede that providing a unique European consent form for data altruism requires harmonisation (Lalova-Spinks, Meszaros & Huys, 2023, p. 5; Shabani, 2021). Moreover, this approach allows for the joint collection of both the personal data acquired by consent, and non-personal data acquired by permission. However, this cannot be the only way to harmonise the approach, especially as it raises several ethical challenges concerning the choice of the legal basis of consent, which is discussed below[12].

3.2. Security

The data altruism organisation must also ensure a solid infrastructure system. The goal is the establishment of data pools in order to store, transfer, manage, and share data. This makes the security of infrastructures crucial. The centralisation of data envisages several challenges from a security viewpoint, making those holding the data both very powerful, and at the same time very vulnerable. Very powerful, because the intention is to promote "the emergence of pools of data made available on the basis of data altruism that have a sufficient size in order to enable data analytics and machine learning, including across borders in the Union" (Recital 45, DGA). Highly weak because they are more easily targeted by cyber-attacks and data breaches.

The DGA points out the relevance of security, stating that the "recognised data altruism organisation shall take measures to ensure an appropriate level of security for the storage and processing of non-personal data that it has collected based on data altruism" (Article 21(4) of the DGA). Article 21(5) of the DGA sets out a duty of communication over the data altruism organisation in case of data breaches concerning non-personal data. This provision corresponds to Article 34 of the GDPR for personal data[13]. However, the measures to be taken to ensure an appropriate level of security are not specified in the Regulation. Their definition will be provided in the so-called 'rulebook', pursuant to Article 22(1)*b*, which is a set of delegated acts of the European Commission, still to be issued.

Regarding the security monitoring over the data pools, the Regulation mentions one aspect that is worth emphasising. Recital 46 introduces ethical councils or boards as "oversight mechanisms", which expressly entail "representatives from civil society". The role of these forms of bottom-up control aims to "ensure that the data controller maintains high standards of scientific ethics and protection of fundamental rights" and turns out to be in line with the European policies promoting the involvement of the society at large in the field of scientific research, represented by the phenomenon of the so-called citizen science. This phenomenon describes "the involvement of citizens and the public at large in research projects in various guises" (Paseri, 2022, p. 531). The citizen science phenomenon, expressly mentioned and promoted in the Regulation (EU) 2021/695 establishing the Horizon Europe programme, i.e., the new framework for funding scientific research at the European level[14], can also acquire a renewed role in the dynamics of data altruism by envisaging participatory forms of control and assessment of the security of data pools.

[12] See Section 4.3.

[13] The Article 21(5) of the DGA states that "[T]he recognised data altruism organisation shall, without delay, inform data holders in the event of any unauthorised transfer, access or use of the non-personal data that it has shared". Similarly, the Article 34(1) of the DGA set out that "[W]hen the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall communicate the personal data breach to the data subject without undue delay".

[14] Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013 (Text with EEA relevance), ELI: http://data.europa.eu/eli/reg/2021/695/oj.

3.3. Control

Article 21 of the DGA establishes "[S]pecific requirements to safeguard rights and interests of data subjects and data holders with regard to their data". Crucial here is the role of the data altruism organisation: This entity is in charge of ensuring compliance with the requirements of the law. For instance, the data altruism organisations are obliged to provide a range of information to data subjects and data holders about the processing of their data (e.g. the general interest objectives pursued and the location of the data or the event of a data breach) and are required to take measures to ensure security.

The wording of the DGA proposal is even more explicit and incisive stating that "[T]he entity shall also ensure that the data is not used for other purposes than those of general interest for which it permits the processing" (Article 19(2) of the DGA proposal). The softer wording used in the current text is justified by the fact that such a duty of control over the entire lifecycle of the data held by the data altruism organisation would not have been practically achievable (Veil, 2022).

However, the current Regulation maintains the central role of the data altruism organisations that are in charge of multiple requirements and tasks. Data altruism organisations must ensure that the purpose of general interest is respected. The general interest is a very broad concept *per se*, and the DGA does not provide a definition. In addition, Article 18(1)*a* provides that in order to qualify for registration in a public national register of recognised data altruism organisations, an entity shall "carry out data altruism activities". Nevertheless, the DGA lacks a definition of "data altruism activities", raising difficulties both from a practical and a legal point of view.

The design of data altruism introduced by the DGA seems to align the data altruism organisation with the role of the data controller under the GDPR. In the European Data Protection Regulation, in fact, several tasks and requirements are established for the actor that determines the *purposes* and *means* of processing of personal data, i.e., the data controller (Article 4(7) of the GDPR). In this framework, the principle of accountability (Artilce 5(2) of the GDPR) is a pillar of the Regulation (Pagallo *et al*., 2019, p. 24; Durante, 2021, p. 134), as a meta-principle of the entire approach of the GDPR (Paseri, Varrette, Bouvry, 2021, p. 135; Durante & Floridi, 2022, p. 135).

In the case of the DGA, the control duties, combined with the vagueness of the framing of the role and activities, complicate the context. For this reason, it's challenging to envision which entity would intend to undertake the registration process in order to be identified as a data altruism organisation, especially considering that this entity must "operate on a not-for-profit basis and be legally independent from any entity that operates on a for-profit basis" (Article 18*c* of the DGA).

After the analysis of the legal challenges of data altruism, it is now time to draw the attention to the ethical challenges.

4. THE ETHICAL CHALLENGES OF DATA ALTRUISM

From an ethical perspective, data altruism can be interpreted as an expression of distributed morality (Floridi, 2020). Distributed morality is "the macroscopic and growing phenomenon of global moral actions and non-individual responsibilities, resulting from the «invisible hand» of systemic interactions among multiagent systems (comprising several agents, not all necessarily human) at a local level" (Floridi, 2020, p. 64). In other words, "the voluntary sharing of data" (Article 2(16) of the DGA) at the basis of the data altruism mechanism represents a modular and incremental operation (Benkler, 2006) that finds its morally relevant value in aggregation (Durante, 2007, pp. 248-253). According to the data altruism mechanism, the aggregation of each individual data sharing represents a moral action being "the result of otherwise morally-neutral or at least morally-negligible […] interactions among agents

constituting a multiagent system, which might be human, artificial, or hybrid" (Floridi, 2020, p. 65). The data altruism mechanism under the DGA results in "actions that are morally negligible in themselves", but that "may become morally significant, if properly aggregated" (Floridi, 2020, p. 72).

Accordingly, it is crucial to focus on the forms of such aggregation, ensuring that the actions of individual data subjects or data holders are not nullified, but rather foster positive moral behaviours. Three aspects to be addressed in order to achieve proper aggregation are investigated below, decoding (1) the concept of altruism; (2) the terminological uncertainty; and (3) the individual autonomy and the scope of the consent.

4.1. The Concept of Altruism

Data altruism is not a new phenomenon. The debate about so-called 'data donation' (Skatova & Goulding, 2019; Prainsack, 2019; Bietz, Patrick & Bloss, 2019) or 'data philanthropy' (Kirkpatrick, 2013; Taddeo, 2016; Taddeo 2017; Giannopoulou, 2019) has been going on for rather some time.

These forms of data release entail two moral problems. On the one hand, they may pose a risk to the individual rights: "making personal data available while, at the same time, maximizing their accessibility and use […] highlight a tension between individual rights and data sharing" (Taddeo, 2016, pp. 4-5). On the other hand, these forms of data sharing may lead to a threat to democracy, considering that they "can hinder democratic processes also by facilitating unduly profiling which can then provide the means for unjust discrimination" (Taddeo, 2016, p. 6).

On the contrary, there are compelling reasons for individuals to donate their data in forms of data altruism or data philanthropy, which support the moral desirability of the phenomenon. Three main reasons can be identified. First, individuals, both data subjects and data holders, may be prompted to allow their data to be processed for the benefit of scientific research activities (Pagallo, 2022, p. 74; Ienca, 2023, p. 2), taking part in the formation of the collective and common good. Second, such sharing and the resulting processing, can generate economic value, which might represent indirect forms of self-interest. Third, such forms of sharing or donation may be means to react against forms of data monopolisation (Van de Hoven *et al*., 2021) by large private actors (Prainsack, 2019, p.10), for the benefit of a larger number of actors, both public and private.

Furthermore, Thomas Nagel proposes the interpretation of a 'rational altruism' that "depends on a recognition of the reality of other persons, and on the equivalent capacity to regard oneself as merely one individual among many" (Nagel, 1975, p. 3). In other words, alongside the traditional reasons such as benevolence, indirect self-interest or other subjective factors, there is also a "a motivation available when none of those are, and also operative when they are present, which has genuinely the status of a rational requirement on human conduct" (Nagel, 1975, p. 80). This introduces the interpretation of altruism not as "abject self-sacrifice, but merely a willingness to act in consideration of the interests of other persons, without the need of ulterior motives" (Nagel, 1975, p. 79). The rational altruism of Thomas Nagel is particularly suited to these forms of data release, where the act of sharing carries a limited emotional impact.

In light of the tension between the moral problems and the moral desirability of these forms of data sharing, Mariarosaria Taddeo argues that the "moral ambiguity of data philanthropy, on the one side, and its moral desirability, on the other, unveil the infraethical nature of this phenomenon" (Taddeo, 2016, p. 6). Infraethics is "the not-yet-ethical framework that can facilitate or hinder evaluations, decisions, actions, or situations, which are then moral or immoral" (Floridi, 2017, p. 392). As a result, infraethics is characterised by moral ambiguity. This "moral ambiguity of infraethics is resolved once it is combined with the *right* moral values" (Taddeo, 2016, p. 7). On this basis, Taddeo claims that the "infraethical nature of data philanthropy becomes clear when considering its moral ambiguity and its

potential to foster democratic processes, the advance of scientific knowledge, civic participation" (Taddeo, 2016, p. 7).

By decoding the concept of altruism and determining whether and how to implement the data altruism mechanism at the national level, a twofold challenge should be addressed. First, altruism is also motivated by objective and impersonal reasons and is not just embodied in sentiment. Second, it is crucial to focus on the design of the infra-ethical infrastructure, "which has to be resilient enough to be able to account for the raising of new moral values as well as for the conceptual and practical changes brought about by the information revolution" (Taddeo, 2016, p. 9).

4.2. Terminological Uncertainty

The description of the different phases of the data altruism mechanism according to the DGA, illustrates the relevance of the concept of 'general interest'. However, the DGA doesn't provide a definition of the concept but proposes a non-exhaustive list of purposes included in the general interest. In the phrasing of Recital 45, these objectives "would include healthcare, combating climate change, improving mobility, facilitating the development, production and dissemination of official statistics, improving the provision of public services, or public policy making. Support to scientific research should also be considered to be an objective of general interest".

In particular, Recital 16 states that "[I]n order to facilitate and encourage the use of data held by public sector bodies for the purposes of scientific research, public sector bodies are encouraged to develop a harmonised approach and harmonised processes to make that data easily accessible for the purposes of scientific research in the public interest". The 'public interest' diverges from the concept of 'general interest' and may trigger a considerable debate if conceived according to the GDPR. Among several stances about the interpretation of the concept of public interest, the European lawmakers seem to adopt a practical approach and "under the GDPR public interest can be described as an object worth safeguarding for the needs or interests of the Member States or the EU for the purposes of which a number of specific measures could be taken, including the rights of a data subject could be constrained" (Slokenberga, 2021, p. 23).

On top of that, the memorandum of the DGA proposal also mentions the concept of 'common good', defining data altruism as "data voluntarily made available by individuals or companies for the common good". Referring to 'general interest', 'public interest' and 'common good' generates uncertainty, impacting on the moral problems of these forms of data sharing (Taddeo, 2016). This becomes even more apparent considering that "[M]ost conceptions of the common good define a form of practical reasoning that fits the model of solidarity" (Hussain, 2018), attributing to altruism and data altruism the nature of a morally good action, in contrast to the infra-ethical nature of the phenomenon.

This terminological uncertainty makes the mechanism very flexible. The risk, however, is that this uncertainty encourages the moral ambiguity of data altruism. Once again, therefore, the intervention of the Member States, with their discretionary powers, represents a key factor in the implementation of the data altruism mechanism.

4.3. Autonomy and Consent

Voluntary data sharing under the data altruism mechanism is based on two forms: Permission for non-personal data and consent for personal data. Permission under Article 2(6) of the DGA "means giving data users the right to the processing of non-personal data". Whereas, as specified by Recital 50 of the DGA, consent should be understood in accordance with the provisions of the GDPR, and therefore shall be free, specific, informed, unambiguous, freely revocable and, in addition, obtained in a manner that is clear and understandable to the data subject.

Article 6 of the GDPR provides for a set of mandatory legal bases for the processing of personal data: Consent thus becomes one of the possible legal bases. The reason for this choice made by the European lawmakers in 2016 was to replace a consent-based approach that had proved to be ineffective (Solove 2012; Schermer *et al*. 2014).

In order to allow the joint acquisition of personal data through consent and of non-personal data through permission, Article 25 of the DGA introduces a "European data altruism consent form", that will be adopted by an implementing act of the European Commission. This model "shall allow the collection of consent or permission across Member States in a uniform format" (Article 25(1) of the DGA).

The DGA emphasises the purpose of "building trust among individuals and undertakings in relation to data access, control, sharing, use and re-use" (Recital 5 of the DGA). Given this goal, two considerations stand out, one regarding the autonomy of those who share data, and the other concerning the effectiveness of the unique consent model for data altruism.

As regards matters of autonomy, the strengthening of a trusted environment demands that those who voluntarily share their data, be they natural or legal persons, are involved in the mechanism as "real interlocutors" (Durante, 2015, p. 16; Smichowski, Duch-Brown & Martens, 2021, p. 48). Making these actors capable, to some extent, of shaping the system or at least of actively participating in it is in line with the objective stated in Recital 5 of the DGA mentioned above. This inclusion is by no means easy to achieve through the provision of a unique consensus model. And this leads to the second issue. One of the factors underlying the transition from the previous Directive 95/46/EC[15], based on the "notice and consent" mechanism (Sloan & Warner, 2014), to the GDPR, was the idea to overcome a model of personal data management represented by pointless check lists of activities to be implemented. Developing a European model for the uniform management of data altruism may facilitate the harmonisation of the operation and yet, raise the risk of returning to the previous approach that was intended to be overcome with the GDPR.

Admittedly, as Recital 52 of the DGA points out, such a European model should adopt "a modular approach allowing customisation for specific sectors and for different purposes". Several scholars have stressed the benefits of this modular approach (Pagallo et al., 2019). However, it would be crucial, in national implementation, to envisage further forms of interaction between data subjects, data holders, and data altruism organisations (some examples are given in Smichowski, Duch-Brown & Martens, 2021, p. 49).

5. CONCLUSIONS

Luciano Floridi, in his analysis on distributed morality in the information society, argues that a longstanding discussion exists regarding "incentives and disincentives, which represent the political and legislative side of the ethical discourse", although much work still needs to be done to develop "technological mechanisms that work as «moral enablers»" (Floridi, 2020, p. 72) for harnessing the power of distributed morality. The mechanism of data altruism introduced by the DGA, and the data pools generated as a result, may be fruitful, if further properly implemented at the national level by the Member States.

[15] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, ELI: http://data.europa.eu/eli/dir/1995/46/oj.

In order to do so, a number of conditions should be fulfilled: (i) coordination in application; (ii) security; (iii) practically feasible lifecycle data control mechanisms; (iv) agreement on the concept of altruism; (v) public interest; and, (vi) autonomy in granting consent. The risk of "misuse and moral hazard" (Floridi, 2020, p. 77) is still real. Data may be used by data users to pursue intentions that are not in line with the general interest identified by the European institutions or even to make malicious use of such data. However, the potential benefits, especially for the scientific research sector, are noteworthy. Given the set of challenges, both legal and ethical, and the manifold actors involved in the data altruism mechanism, it is worth focusing on the development of governance mechanisms able to encompass all the interests at stake.

REFERENCES

Baloup, J., *et al*. (2021). White paper on the data governance act. *CiTiP Working Paper 2021*, 1-57.

Bassi, E. (2011). PSI, protezione dei dati personali, anonimizzazione. *Informatica e diritto* 37.1-2, 65-83.

Benkler, Y. (2006). *The Wealth of Networks. How Social Production Transforms Markets and Freedom*, New Haven: Yale University Press.

Bietz, M., Patrick, K., Bloss, C. (2019). Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice* 4.1.

Durante, M. (2021). *Computational power: the impact of ICT on law, society and knowledge*. New York: Routledge.

Durante, M. (2015). The democratic governance of information societies. A critique to the theory of stakeholders. *Philosophy & Technology* 28, 11-32.

Durante, M. (2007). *Il futuro del Web: etica, diritto, decentramento. Dalla sussidiarietà digitale all'economia dell'informazione in rete*, Torino: Giappichelli.

Durante, M., Floridi, L. (2022). A legal principles-based framework for AI liability regulation. *The 2021 yearbook of the digital ethics lab*. Cham: Springer International Publishing, 93-112.

Floridi, L. (2020). Distributed morality in an information society. In Miller, K. W., Taddeo, M. (eds.) *The Ethics of Information Technologies*, London: Routledge, 63-79.

Floridi, L. (2017). Infraethics–on the Conditions of Possibility of Morality. *Philosophy & Technology* 30, 391-394.

Giannopoulou, A. (2019). Access and Reuse of Machine-Generated Data for Scientific Research. *Erasmus Law Review* 2, 155-165.

Hallinan, D., *et al*. (2023). (Un) informed consent in Psychological Research: An empirical study on consent in psychological research and the GDPR. *J. Open Access L*. 11, 1-28.

Hallinan, D. (2020). Broad consent under the GDPR: an optimistic perspective on a bright future. *Life sciences, society and policy* 16, 1-18.

Hussain, W. (2018). The Common Good. In: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2018/entries/common-good/.

Kirkpatrick, R. (2013). A new type of philanthropy: donating data. *Harvard Business Review*. https://hbr.org/2013/03/a-new-type-of-philanthropy-don

Ienca, M. (2023). Medical data sharing and privacy: a false dichotomy? *Swiss Medical Weekly* 153.1, 1-3.

Lalova-Spinks, T., Meszaros, J., Huys, I. (2023). The application of data altruism in clinical research through empirical and legal analysis lenses. *Frontiers in Medicine* 10, 1141685.

Nagel, T. (1975). *The possibility of altruis*m. Oxford: Oxford University Press.

Pagallo, U., Bassi, E. (2013). Open Data Protection: Challenges, Perspectives, and Tools for the Reuse of PSI. In Hildebrandt, M., *et al*. (eds), *Digital Enlightenment Yearbook 2013.* Amsterdam: IOS Press, 179-189.

Pagallo, U. (2020). The Politics of Data in EU Law: Will It Succeed? *Digital Society* 1.3, 1-20.

Pagallo, U., *et al.* (2019). On good ai governance: 14 priority actions, a smart model of governance, and a regulatory toolbox. *AI4People report*, https://ddd.uab.cat/pub/estudis/2019/243144/AI4people_a2019iENG.pdf

Pagallo, U. (2022). *Il dovere alla salute. Sul rischio di sottoutilizzo dell'intelligenza artificiale in ambito sanitario*, Milano-Udine: Mimesis.

Paseri, L. (2023). Open Science and Data Protection: Engaging Scientific and Legal Contexts. *J. Open Access L.* 11, 1-18.

Paseri, L. (2022). From the Right to Science to the Right to Open Science. The European Approach to Scientific Research. *European Yearbook on Human Rights 2022*. Cambridge: Intersentia, 515-541.

Paseri, L., Varrette, S., Bouvry, P. (2021) Protection of Personal Data in High Performance Computing Platform for Scientific Research Purposes. *Annual Privacy Forum*. Cham: Springer International Publishing, 123-142.

Prainsack, B. (2019). Data donation: How to resist the iLeviathan. *The ethics of medical data donation*, 9-22.

Ruohonen, J., Mickelsson, S. (2023). Reflections on the Data Governance Act. *Digital Society* 2.1, 1-10.

Schermer, B. W., *et al*. (2014). The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology* 16.2, 171-182.

Shabani, M. (2021). The Data Governance Act and the EU's move towards facilitating data sharing. *Molecular systems biology* 17.3, e10229.

Skatova, A., Goulding, J. (2019). Psychology of personal data donation. *PloS one* 14.11, e0224240.

Sloan, R. H., Warner, R. (2014). Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L*. 14, 370-413.

Slokenberga, S. (2021). Setting the foundations: Individual rights, public interest, scientific research and biobanking. In: Slokenberga, S., Tzortzatou, O., Reichel, J. *GDPR and biobanking: Individual rights, public interest and research regulation across Europe*. Cham: Springer Nature,11-30.

Smichowski, B. C., Duch-Brown, N., Martens, B. (2021). To pool or to pull back? An economic analysis of health data pooling. No. 2021-06. *JRC Digital Economy Working Paper*, 1-70.

Solove, D. J. (2012) Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review* 126, 1880-1903.

Taddeo, M. (2017). Data philanthropy and individual rights. *Minds and Machines* 27.1, 1-5.

Taddeo, M. (2016). Data philanthropy and the design of the infraethics for information societies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083, 1-12.

Van de Hoven, Je., *et al*. (2021) Towards a digital ecosystem of trust: Ethical, legal and societal implications. *Opinio Juris In Comparatione* 1/2021, 131-156.

Van Eechoud, M. (2021). A Serpent Eating Its Tail: The Database Directive Meets the Open Data Directive. I*IC - International Review of Intellectual Property and Competition Law,* 52, 375-378.

Veil, W. (2022). Data Altruism: How the EU is Screwing up a Good Idea. Discussion paper. *AlgorithmWatch*, 1-8.

TRUSTWORTHY AND USEFUL TOOLS FOR MOBILE PHONE EXTRACTION

Anna-Maria Piskopani, Helena Webb, Liz Dowthwaite, Chris Hargreaves, Nicholas FitzRoy-Dale, Quentin Stafford-Fraser, Christos Nikolaou, Derek McAuley

University of Nottingham (United Kingdom), University of Nottingham (United Kingdom), University of Nottingham (United Kingdom), University of Oxford (United Kingdom), Telemarq (United Kingdom), Telemarq (United Kingdom), Telemarq (United Kingdom), University of Nottingham (United Kingdom)

anna-maria.piskopani@nottingham.ac.uk; helena.webb@nottingham.ac.uk; liz.dowthwaite@nottingham.ac.uk; christoper.hargreaves@cs.ox.ac.uk; nicholas@telemarq.com; quentin@telemarq.com; christos@telemarq.com; derek.mcauley@nottingham.ac.uk

ABSTRACT

Data collected from mobile phones can be a valuable resource in the investigation of crime. However, the processes of collecting, handling and analysing such data raise ethical concerns. In this position paper we focus on mobile phone extraction (MPE) in the context of the UK criminal justice system. We describe how it is a phenomenon of interest to the ethics of computing and then outline the aims and emerging findings of our own research in this area. We point to the potential for digital tools to address some of problems currently associated with MPE but also highlight the existence of wider contextual constraints that limit the impact technical interventions alone can make.

KEYWORDS: criminal justice system, policing, data protection, privacy.

1. INTRODUCTION

The UK criminal justice system increasingly deploys Mobile Phone Extraction (MPE), a practice in which police collect data from mobile phones belonging to suspects, witnesses, or complainants in a crime and analyse it to aid their investigations. The data – or subsets of the data – may then be made available to prosecutors, the Crown Prosecution Service (who determine whether the case will proceed to court), disclosed to the defence (who may also conduct their own analysis of it), and used in court as evidence. This process of acquiring and analysing data can reveal highly valuable evidence. However, it also has drawbacks. The ubiquity of smartphones in our contemporary digital society means that phones typically have very large amounts of data on them. Searching through this data to identify what might be relevant can be extremely time consuming. Furthermore, individuals tend to have large amounts of personal data on their phones, including data which has no relevance to the case being investigated but which may nevertheless be collected and analysed. This could amount to an unlawful invasion of privacy.

This position paper describes the context, aims and early findings of an ongoing research project on the use of MPE in the UK criminal justice system. Section 2 describes MPE as a phenomenon of interest to the ethics of computing. We note recent controversies that have created a crisis of trust and practice in MPE in the UK, discuss its legal framework, and note the potential for digital tools to enhance the process. In Section 3 we introduce our project, 'Trustworthy and Useful Tools for Mobile Phone Extraction'. We describe its aims to address the crisis of trust and practice in MPE and our particular focus on digital tools. We outline the development of our own digital tool, RIME: Responsible Investigation of Mobile Environments and discuss emerging findings from a set of stakeholder workshops and interviews. These findings highlight pressure points throughout the MPE landscape that can problematise the consent for MPE, the acquisition of data, and data analysis. We conclude

that whilst digital tools have the potential to enhance MPE, they cannot alone address all the ethical issues associated with it.

2. MPE AS A PHENOMENON OF INTEREST TO THE ETHICS OF COMPUTING

2.1. Mobile Phone Extraction (MPE): a crisis of trust and practice

Mobile Phone Extraction (MPE) is a process in which data from a mobile phone is collected for the purpose of analysis. This may include telecommunications data (e.g. text messages), GPS data, online browsing history, and data collected from apps. Since this data can offer valuable information in the investigation of a crime, MPE is increasingly used in criminal investigations. In the UK context, data may be collected from the phone belonging to a suspect, witness, or complainant for this purpose. Whilst MPE can be as simple as collecting photographs/screenshots of a phone's digital display, most often it involves extraction of data from the device, either triggering backup mechanisms built into the device or, through the use of exploits in device hardware or software to allow higher privilege access to the device, facilitating access to other data not included in a device backup. Once the data is acquired, the full dataset or subsets of interest are retained for analysis. Analysis may be conducted manually or using digital tools. The data and/or analytic results may then be used as the case progresses. This can provide important evidence to exonerate or incriminate suspects, corroborate other evidence etc. so therefore can be vital to case outcomes. However, in recent years various concerns have been raised in legal reports, governmental reports, media reports, and interest group campaigns regarding practices of MPE in the UK. Although MPE is used across a wide range of criminal investigations, these concerns are often discussed in reference to sexual violence cases, as these particularly highlight the sensitivities involved.

The first concern is that MPE is inefficient. Lack of suitably trained police personnel can mean that it takes a long time for the initial data acquisition to begin and phone owners are left waiting for their devices to be returned. Due to the large volumes of data involved, analysis is also time-consuming and resource intensive. This creates substantial case backlogs (HM Crown Prosecution Service Inspectorate, 2019). Additionally, there are known instances in which inefficient analysis has undermined a case or threatened to cause a miscarriage of justice. In the case R v Allan (Metropolitan Police and CPS, 2018) a man was arrested for rape and sexual assault. The complainant handed over her phone and the police obtained over 57,000 lines of message data for analysis. The officer who conducted the analysis of this data did not record the methods used and when the police made their disclosure to the defence, the existence of messages indicating a consensual sexual relationship between the complainant and suspect was not included. When the trial began, the prosecution counsel provided the defence with a copy of the phone data. This revealed the existence of the messages; the case and charges were dropped soon after (but nearly two years after the original complaint). In a review of the case the police officer who conducted the analysis of the mobile phone data was quoted as saying: '*I had told CPS and the original prosecution counsel... that I had looked through [the data] and identified everything that was relevant. I can only read from this that because of the volume of analysis of phone downloads I deal with, I had wrongly assured myself I had looked through this entire downloa*d'(Metropolitan Police and CPS, 2018, pp5-6).

The second concern is that retaining and reviewing an individual's mobile phone data can represent an excessive invasion of privacy. Until recently the default has been for police to retain and (potentially) review all the data from a phone, even where this may be regarded as not necessary or proportionate. This practice has been described as a 'digital strip search' by campaign organisations such as Privacy International (2018), Big Brother Watch (Big Brother Watch, 2019) and the Centre for

Women's Justice (Centre for Women's Justice, 2020). They highlight that the collection of this volume of data far exceeds that which might be collected in a physical search and typically includes information which is not relevant to the case being investigated. They further note that this potential invasion of privacy can deter victims from reporting crimes – in particular, in cases of sexual violence where they fear irrelevant information may be used to undermine their credibility. Some complainants have said they found the process of discussing MPE with police to be coercive and retraumatising (Centre for Women's Justice, 2020) as they have been told their cases would be dropped unless they handed over their phones. In a 2019 report by Big Brother Watch one complainant described reporting to the police that she had been attacked by a group of strangers. The police asked her to provide 7 years of phone data and dropped her case when she refused. She said: '*My phone documents many of the most personal moments in my life and the thought of strangers combing through it...makes me feel like I am being violated once again. This isn't about trying to stop the police from putting together facts of the case...This is about objecting to the police downloading seven years of information that pre-dates the event and therefore has zero relevance*.' (Big Brother Watch, 2019, p48).

In 2020 the Information Commissioner's Office – an independent UK body set up to uphold personal information rights – produced a report on MPE (ICO, 2020). This report, and its follow up in 2021 (ICO, 2021) recognised the growing concerns over intrusions of privacy in MPE and concluded that action was necessary to ensure the legality of the process. These reports have helped to motivate changes in legislation and guidance, as discussed next.

2.2. The legal framework for MPE

The legal landscape for MPE is highly complex, with different legislation often covering the different processes of acquiring a mobile phone, and the processing of the data, e.g. data access and analysis. Although MPE is used in police forces across the four countries of the UK, Scotland and Northern Ireland have their own legal systems so our description in this section relates specifically to England and Wales. Limitations of space allow us to highlight only a few key pieces of this wide-ranging legislation, correct as of December 2023. Police have powers (e.g. under the Police and Criminal Evidence Act (PACE, 1984) to seize devices, including mobile phones, where they have reasonable grounds to believe they are evidence of an offence and also to prevent evidence being destroyed. However, they may be more likely to pursue a consent process when seeking to acquire the phones belonging to complainants or witnesses. The Criminal Procedure and Investigations Act (CPIA, 1996) and its code of practice oblige the police to pursue all reasonable lines of enquiry in an investigation, whether they point towards or against the guilt of a suspect. The same Act obliges prosecutors to disclose evidentiary material to the defence, including material which might be considered capable of undermining the case against the accused. Following high-profile discussions about police practices and privacy concerns, the Attorney General produced revised Guidelines on Disclosure in 2020, further updated in 2022 (AGO, 2022).

Data processing related to MPE is chiefly covered in the UK GDPR (2016) and the Data Protection Act (DPA, 2018). The DPA covers data processing for the purposes of criminal investigations and sets out the obligation of competent authorities to comply with the following data protection principles: data processing should be lawful and fair; the purposes of processing should be specified, explicit and legitimate; to collect personal data that are adequate, relevant and not excessive; to collect personal data that are accurate and kept up to date; to keep personal data for no longer than is necessary; personal data should be processed in a secure manner. Police forces are also obliged to conduct data protection impact assessments before processing personal data, plus keep logs of data processing actions including collection, alteration, consultation, disclosure, combination and erasure.

The 2020 Information Commissioner's Office report mentioned above, describes the acknowledged concerns over MPE, relating in particular to inconsistencies of police actions and an overly wide approach to extracting data. The report stated that even when investigators follow a consensual approach to MPE, the consent of the phone owner does not often meet the legal requirements set out by the DPA. It also observed that data acquired and processed was often excessive – even considering the CPIA obligation to pursue all reasonable lines of enquiry. It therefore argued that the police need to do more to justify the strict necessity and proportionality of the processing. Furthermore, it highlighted that a typical mobile phone is likely to contain special category data (i.e. information revealing racial or ethnic origin, health issues, sexual orientation and life, political opinions, philosophical beliefs, genetic and biometric data) and that therefore an appropriate policy document describing how this data is handled and what safeguards are applied must be in place. The report's further recommendations included a call for: a new code of practice to improve police forces' compliance with data protection law; a national training standard for MPE; and data protection by design and default approach in MPE tools, with constant review of the software used. The report had a substantial impact. The CPIA code of practice was revised in 2020 and new case law supported the ICO's legal reasoning; e.g. Court of Appeals judgment (Bater-James & Anor v R. [2020] EWCA Crim 790) on reasonable lines of enquiry and strict proportionality of data processing. The College of Policing (2021) produced new operational guidance on the extraction of material from digital devices and new PACE codes of practice were issued in 2023.

The Police Crime Sentencing and Courts Act (PCSC, 2022) provides a clearer statutory basis for the police to extract information from mobile devices and sets out a new code of practice. This code of practice relates specifically to MPE when a user voluntarily hands over their phone (with provision for some other scenarios) and states that consent must be genuine, not coerced, and that phone owners must be separated from their phones for a minimal amount of time. It also refers to the ICO's reports and states that voluntary provision in the PCSC Act does not equal 'consent' as defined under the Data Protection Act 2018. It makes an explicit move towards selective extraction of mobile phone data. It states that extraction must meet tests of strict necessity and proportionality and that due to the large volume of data collected, it is highly unlikely that a full extraction from a device will ever meet these tests. Therefore, selective - rather than full – extraction is the default. In fact, there is no presumption that information will be extracted from a device in any given case. Any irrelevant confidential information collected must be destroyed or redacted immediately, and police must consider rights to privacy of third parties captured in the data. The ICO's update report on MPE (ICO, 2021) included concerns that the PCSC may raise data protection and human rights issues.

2.3. Digital tools for MPE

Both the ICO report and the code of practice relating to the PCSC Act note the potential for digital forensics techniques and tools to improve MPE practices. The ICO report (ICO, 2020, p51, section 3.5) states '*specific hardware and software tools offered by MPE vendors [have] capabilities designed to minimise intrusion and maximise privacy (e.g. by allowing focused extraction of specific pieces of data)*'. Section 90 of the code of practice (Home Office, 2022) also advises privacy protection via selective extraction and suggests that: '*this should include use of appropriate technologies to support selective extraction and use of targeted key words, date ranges or other specifics to identify necessary information.*' In recent years new advances in digital forensics techniques for mobile phones have emerged in areas including, data visualisations (Tassone et al., 2016) and presentation of digital data (Osborne et al., 2010) timeline analysis and reconstruction (Hargreaves and Patterson, 2012), communication pattern analysis (Spranger et al., 2020) and key word searches (Lin et al., 2018).

As observed by the ICO, a number of digital tools to support police MPE practices already exist. These can perform extraction by enabling a download of data from a mobile device, and can also effectively replicate a selective extraction through the filtering of extracted data and identification of datasets to retain from that initial download. They can also provide features for analysis such as the techniques listed above. Many of these tools are commercial products marketed towards law enforcement professionals as well as digital forensics professionals. For instance, Cellebrite, Magnet Forensics, and MSAB all provide a range of solutions for digital data acquisitions (from cloud and other devices in addition to mobile phones) and data analysis. Open-source tools also exist: libimobile device allows backup extractions from iOS devices, and for analysis, primarily iLEAP for iOS devices and ALEAP for Android devices can be used.

Within the UK police, some recent initiatives include efforts to enhance MPE practice, including in the use of digital tools. Once again, these focus on sensitive investigations regarding sexual violence. Operation Soteria is a police and CPS programme (NPCC, nd) to develop new models for the investigation and prosecution of rape. Academics on the programme have conducted deep dive studies into police force practices (Stanko, 2022). Funding from the UK Home Office has supported Project Odyssey (OPCC, 2023), the development of a privacy preserving workflow for MPE in sexual violence and domestic abuse investigations. Pioneered within the Gloucestershire Police, the workflow supports selective extraction from the phones of complainants and follows a consent process that is legally compliant and ensures that phone owners are not separated from their phones for any lengthy period of time.

These examples demonstrate that digital forensics tools for MPE present a promising opportunity to improve practice. They can support selective extraction and speed up the process of searching through the huge volumes of data extracted from mobile phones. It is anticipated that they will increasingly incorporate state-of-the-art AI techniques (Constantini et al., 2019; Du et al., 2020). However, they also have drawbacks. Some commercial tools may be prohibitively expensive for some police force budgets. Furthermore, although an open market for digital forensics exists in the UK, it has historically been constrained by some entrenched monopolies, with small providers in particular finding it difficult to make an impact (House of Commons Science and Technology Committee, 2005). According to the ICO the 'individual implementations' (ICO, 2020, p51) that some forces use instead of commercial tools may lack key capabilities, such as support for selective extraction. In addition, constant updates to phone operating systems, changes in file formats, and availability of new apps means that tools can quickly become out of date (Garfinkel, 2010). Accessibility can also be an issue: some tools (including open-source ones) can require a level of technical proficiency that prevents their wide use amongst police personnel, who may lack specific training and digital literacy (HMICFRS, 2020).

As a relatively new discipline, digital forensic science still faces challenges regarding the standardisation of its processes and outcomes so that they can be regarded as trustworthy and admissible in court (Garfinkel, 2010). The fast pace of technological advance and change can make data acquisition, preservation, analysis, and presentation difficult to standardise (Grobler, 2012; Edward and Ojeniyi, 2019). The problem is compounded by a frequent lack of transparency around commercially available tools and a historically unregulated market for digital forensics products in the UK (House of Commons Science and Technology Committee, 2005) Furthermore, there is a recognised lack of consistency in the conduct of (digital) forensics within the police. The UK has a Forensic Science Regulator whose role is to ensure forensic science services meet scientific quality standards. Their 2017 report (Tully, 2018) stated (page 55) '*if quality of forensic science provision is of insufficient priority to enable risks to be managed effectively and quality standards to be achieved, the logical result is that it will become unsustainable for any forensic services to be managed within some police forces.*" In 2021 the Forensic Science Regulator Act 2021 was accompanied by a code of practice (updated in 2023)

(Forensic Science Regulator, 2023) specifying quality management standards for units of forensics practice (such as the police). This includes provisions for standards of conduct and practice in relation to data capture, processing and analysis from digital storage devices, such as mobile phones. As part of these quality management standards, forensic units are expected to be in compliance with an internationally recognised standard. The ISO 17025 standards are named as the most appropriate for digital forensics. Specifically, ISO/IEC 17025 (ISO/IEC, 2017) regulates laboratories and defines validation as a key requirement to deliver meaningful and consistent forensic results.

3. THE 'TRUSTWORTHY AND USEFUL TOOLS FOR MOBILE PHONE EXTRACTION' PROJECT

3.1. Project Overview

As described already in this paper, there is a crisis of practice and trust in MPE in the UK. High-profile instances of police errors investigating/handling mobile phone data and widely expressed concerns over intrusion of privacy have led to severe consequences such as the collapse of legal cases and the reluctance of victims to report their experience of crime. We are conducting a research project that seeks to take steps to address this crisis. The 'Trustworthy and Useful Tools for Mobile Phone Extraction' project is funded by the UKRI Trustworthy Autonomous Systems Hub and brings together a multi-disciplinary group of researchers and practitioners with expertise in law, social science, psychology, computer science, digital forensics, human computer interaction and software development. We draw on these disciplines to forge a broad and inclusive understanding, and our work is underpinned by principles of responsible research and innovation (RRI) (Owen et al., 2012). One area of work in the project concerns the ongoing development of a digital forensics tool, RIME, which we are developing as a means to investigate different ways in which MPE practices can be both trustworthy and useful. In another area of work, we explore the wider landscape of MPE and identify responsibility practices through which citizens can feel confident of the effective and appropriate use of their mobile phone data in the criminal justice system.

In addition to the development of RIME (described further below), we are conducting a series of interlinked activities, each exploring a dimension relevant to the trustworthiness and usefulness of MPE in the criminal justice system. 1) In order to navigate the legal framework for MPE, we have been conducting a documentary analysis of relevant academic literature, case law, reports by competent authorities and developments in legislation. Parts of this analysis are included in section 2.2. 2) We are conducting interviews and workshops to capture the perspectives of stakeholders across the entire MPE landscape and will draw on these to identify opportunities safeguarding mechanisms in digital forensics tools. 3) We are reviewing existing standards for digital forensics and will synthesise them with our emerging findings to identify quality markers for RIME and other MPE tools. 4) We are conducting analysis to identify user requirements for MPE tools and mechanisms to optimise their usability.

3.2. RIME: Responsible Investigation of Mobile Environments

As mentioned above, one key aim of our project is to support the ongoing development of the RIME (Responsible Investigation of Mobile Environments) tool. Its initial development was funded by a TAS Hub award, and development continues in the current project, enabling us to explore how responsibility mechanisms such as privacy protection and accessibility can be embedded within MPE tools. RIME is designed around principles of reproducibility, transparency, portability, interoperability, and usability. It is intended to be used alongside other tools rather than to displace them. A dataset collected within RIME can be made available to third parties for analysis using their own tools. This

recognises (as identified in our emerging findings) that in some cases both the prosecution and defence may seek to conduct their own analysis of a dataset, and in some instances the dataset may also be made available to an independent authority. As well as supporting analytic reproducibility, this also serves as a transparency measure.

RIME is designed to expose a subset of the contents of a phone for investigation, rather than a complete capture of all data on the device. This complies with the recent moves towards selective extraction in MPE in the UK and guards against overly invasive 'fishing expeditions' through a full dataset. RIME users (such as police personnel) may be authorised to see "all communications between three persons of interest using apps A, B, and C during the last 3 months". That authorisation can be fed into the RIME tool, which then returns only the relevant data. In its current stage of development, RIME is able to collect a data backup from an Android or iOS device, and access: i) contact information, ii) text messages, iii) WhatsApp messages and iv) images and other media that have been stored in the phone. Rather than requiring the user to explore message data separately according to originating app, the user interface displays all messages in chronological order. This helps to optimise the identification of communication patterns and timelines. Furthermore, messages can be filtered by date, app source, and contact information for targeted inspection. Filtered data subsets can be exported in industry-standard formats and then be analysed by other forensic tools with more advanced analytic capabilities. The data is portable and imported into tools as if it had come directly from the phone. In this way, RIME acts as a format-preserving filter. For additional privacy, RIME also supports an optional pseudonymisation feature which replaces real names and phone numbers with autogenerated (but indexed) alternatives. Finally, RIME supports the side-by-side timeline display of data from multiple devices , simplifying the understanding the chronology of events.

RIME is an open-source tool, with the code publicly available on the Horizon Digital Economy Research Github account. Whilst at present RIME serves primarily as a research tool, in the long term we anticipate that it will serve the wider community, so its open-source status is key to its ongoing development, accessibility and transparency, and the addition of new features by third parties. The source code is also, importantly, open for assessment and inspection by all those involved, or interested in, the MPE process. A legal team, for example, could determine exactly what data would have been included or excluded by specific filters. Developers can also extend RIME's plugins to read data stored in previously unsupported formats, or generated by the latest mobile apps.

3.3. Emerging findings: pressure points throughout the MPE landscape

Our work to develop RIME presents exciting opportunities to foster responsibility practices in the use of mobile phone data in the criminal justice system. However, it is important that we understand the context in which digital tools for MPE are used. MPE occurs within a complex environment involving multiple stakeholders, who may have conflicting expectations and experiences. Our project stakeholder interviews and focus groups enable us to develop a holistic understanding of the situated conduct of MPE. In this section we present some of the emerging findings of this work. We combine results of our early data analysis with other insights gained from meetings with stakeholders and available literature. We find that pressure points exist throughout the MPE landscape and can serve to problematise the consent process for MPE, the conduct of selective data extraction, and the effectiveness of data analysis.

As the PCSC Act is very new it is difficult to determine its effects on current practice. However, its advocation of police sensitivity towards the potential distress of complainants and need for genuine consent when asking them to hand over their phones plus its advocation for selective data extraction align with College of Policing principles of compassionate policing (College of Policing, nd). At the same time though, police investigations have often operated under expectations that evidence collection

should be completed as soon as possible after reporting and this does not necessarily allow for time for a complainant to compose themselves before a discussion or reflect on the request (May et al., 2022). A further issue, as confirmed by our interview participants, is that defence solicitors for a suspect will often advise a 'no comment' approach to responding to police questions. It may also be challenging to gain access (and in some cases, accompanying authorisation for access) to data on the device of a non-cooperative suspect, given encryption, pin codes, biometrics, and the security model of modern devices. This pushes the evidence collection focus back onto the complainant, and the complainant's phone.

Regarding the amount of data retained from the phone download, the requirement for police to pursue all reasonable lines of enquiry may still mean that a large amount of phone content is treated as relevant to the investigation. In some contexts, even data not directly related to the crime being investigated might be treated as relevant, by police, the Crown Prosecution Service or the defence. This may particularly occur in the investigation of sexual violence cases where common societal viewpoints regarding rape can mean that investigations focus on the credibility of the complainant in addition to evidence of the crime having occurred. This point was made in one of our workshop sessions, attended by academic experts plus a staff member for a victim support organisation, who stated: '*even if you restrict police officers [to only] take what's necessary … they will find a way to take the whole thing or they will think it's relevant to take the whole thing. … they will say well we don't know what's relevant until we read all of it.*'

Our interview participants report observing inefficiencies in the police analysis of mobile phone data. One barrister, reflecting on cases he had been directly involved with, told us: '*they don't necessarily ...have the same view or skills of analysis in terms of working out what's relevant and what's not for courts so that … they can potentially miss material that really should be in the evidential domain. And not necessarily because they wouldn't understand its relevance, but just because they just either haven't got the time, or they haven't seen the relevance of it, …, or the tools that they've got just don't really allow them to fully appreciate what the data is capable of demonstrating*.' These comments highlight known constraints that can impact effective analysis, such as the lack of technical training and resources within the police (HMICFRS, 2022). The same interviewee went on to say that a specific commercial tool known to be used within one police force in England '*in its most basic form doesn't really offer the level of analysis that you need and you want to understand mobile phone data and what it demonstrates*.' He gave an example of attempting to cross-reference data from multiple mobile phones and the 'monster' size dataset this creates. This indicates that the limitations of tools used within the police can create constraints in addition to those relating to training and resources.

Limited police expertise in the selection of key words and other parameters to search through datasets has been reported in existing research (May et al, 2022) and is a topic that also arises in our interview data. One solicitor recalled a case of an individual accused of drug supply. Police analysed messages the suspect had sent to contacts associated with the case but did not look at messages sent to family members and friends. As a result, they '*missed messages which effectively confessed to drug supply*.' When asked what factors help an analyst know what search terms and queries to use, the same participant remarked that '*experience*' plays a key role. The accumulation of expertise, in particular tacit knowledge, via experience is a common feature across professions (Sternberg and Horvath, 1999). However, at present the UK police force is relatively inexperienced, with 40% of personnel across the workforce having served for fewer than 5 years (Gloucestershire Police Federation, 2023). This has consequences for the police's ability to deal with service demand, including perhaps in their ability to analyse mobile phone data optimally.

These emerging findings of our stakeholder engagement work help us to build up a nuanced understanding of the landscape of MPE and highlight the existence of pressure points across it. The findings point to ways in which RIME and other digital tools can improve certain practices in MPE but also indicate that digital tools alone cannot fully address the pressures that exist.

4. CONCLUSION

Data collected from mobile phones can be a valuable resource in the investigation of crime, and digital tools can assist to make the process of acquiring and analysing this data more efficient. However, as our discussion of the UK context has shown, Mobile Phone Extraction (MPE) also raises ethical and legal questions regarding consent and privacy. If these are not dealt with sufficiently, then public trust in the criminal justice system can be undermined. We see significant opportunities for digital tools for MPE to improve practice, both in terms of privacy protection and the effectiveness of analysis. This is an area we are continuing to explore through the ongoing development of our RIME tool. For instance, we are considering potential features such as compliance reminders to prompt users to consider whether an extraction meets strict tests of necessity and proportionality, summary mechanisms to highlight at a high level what is contained in a data subset, and coverage mechanisms, which clarify how much of a dataset has been analysed. We intend to draw on our wider project activities to prepare best practice guides for RIME, and digital MPE tools more generally.

At the same time however, we recognise that MPE occurs within a complex landscape. Organisational and funding dynamics within institutions such as the police, and actions based on the conflicting aims of those involved in prosecution versus those involved in defence, create pressure points across this landscape. The existence of these pressure points can problematise MPE and make it more complex across the phases of consenting to the extraction, the conduct of selective data extraction, and the effectiveness of data analysis. Our emerging findings highlight the importance of taking an interdisciplinary approach to studying the contemporary phenomenon of MPE. In the UK context, further research is needed to understand the effects of new legislation on practice. It is also necessary to engage with the wide number of stakeholders connected to MPE and understand their perspectives on the ethical and legal issues associated with it. As interdisciplinary researchers we understand that technology alone cannot resolve the ethical issues associated with MPE, so we must be careful not to overstate the potential impact of digital forensic tools.

ACKNOWLEDGEMENTS

We would like to thank all the participants who have consented to take part in our study. We are also grateful to the UKRI TAS Hub (EP/V00784X/1) for funding the ‘Trustworthy and Useful Tools for Mobile Phone Extraction’ project and the earlier ‘Digital Forensics Platform’ project that enabled the initial development of RIME.

REFERENCES

AGO. ‘Attorney General’s Guidelines on Disclosure for Investigators, Prosecutors and Defence Practitioners’. Attorney General’s Office, 2022. Retrieved from https://assets.publishing.service.gov.uk/media/628ce5efd3bf7f1f3b19efa7/AG_Guidelines_2022_Revision_Publication_Copy.pdf

Bater-James & Anor v R. [2020] EWCA Crim 790, (EWCA (Crim) 23 June 2020). Retrieved from https://www.bailii.org/ew/cases/EWCA/Crim/2020/790.html

Big Brother Watch. 'Digital Strip Searches: The Police's Data Investigations of Victims'. Big Brother Watch, July 2019. Retrieved from https://bigbrotherwatch.org.uk/wp-content/uploads/2019/07/Digital-Strip-Searches-Final.pdf

Centre for Women's Justice. 'Stop the "Digital Strip Search" of Rape Victims like Me'. *Centre for Women's Justice* (blog), 16 March 2020. Retrieved from https://www.centreforwomensjustice.org.uk/new-blog-1/2020/3/13/stop-digital-strip-search

College of Policing. 'Authorised Professional Practice. Extraction of Material from Digital Devices'. College of Policing, 2021. 'We Are Emotionally Aware'. College of Policing. Accessed 5 January 2024. Retrieved from https://profdev.college.police.uk/competency-values/we-are-emotionally-aware/

Costantini, Stefania, Giovanni De Gasperis, and Raffaele Olivieri. 'Digital Forensics and Investigations Meet Artificial Intelligence'. *Annals of Mathematics and Artificial Intelligence* 86, no. 1 (1 July 2019): 193–229. https://doi.org/10.1007/s10472-019-09632-y

CPIA. Criminal Procedure and Investigations Act 1996 (1996). *UK Public General Acts. Legislation.gov.uk.* Retrieved from https://www.legislation.gov.uk/ukpga/1996/25/contents

DPA. The Data Protection Act (2018). *UK Public General Acts. Legislation.gov.uk.* Retrieved from https://www.gov.uk/data-protection

Du, Xiaoyu, Chris Hargreaves, John Sheppard, Felix Anda, Asanka Sayakkara, Nhien-An Le-Khac, and Mark Scanlon. 'SoK: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Forensic Investigation'. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 1–10. ARES '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3407023.3407068

Edward, Elizabeth Ozioma and Joseph A. Ojeniyi. 'A Systematic Literature Review On Digital Evidence Admissibility: Methodologies, Challenges and Research Directions," *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)***,** Abuja, Nigeria, 2019, pp. 1-7. https://doi: 10.1109/ICECCO48375.2019.9043250

Forensic Science Regulator. 'Forensic Science Regulator: Code of Practice', 2023. Retrieved from https://www.gov.uk/government/publications/statutory-code-of-practice-for-forensic-science-activities/forensic-science-regulator-code-of-practice-accessible

FSR. Forensic Science Regulator Act 2021 (2021). *UK Public General Acts. Legislation.gov.uk.* Retrieved from https://www.legislation.gov.uk/ukpga/2021/14/contents/enacted

Garfinkel, Simson L. 'Digital Forensics Research: The next 10 Years'. *Digital Investigation*, The Proceedings of the Tenth Annual DFRWS Conference, 7 (1 August 2010): S64–73. https://doi.org/10.1016/j.diin.2010.05.009

Gloucestershire Police Federation. '"A 'Significant Proportion' of Police Officers in the Force Are Inexperienced, Which Makes It Difficult to Deal with Increasing Demand"'. Gloucestershire Police Federation, 2023. Retrieved from https://www.polfed.org/gloucs/news/latest-news/a-significant-proportion-of-police-officers-in-the-force-are-inexperienced-which-makes-it-difficult-to-deal-with-increasing-demand/

Grobler, Marthie. 'The Need for Digital Evidence Standardisation'. *International Journal of Digital Crime and Forensics (IJDCF)* 4, no. 2 (1 April 2012): 1–12. https://doi.org/10.4018/jdcf.2012040101

Hargreaves, Christopher, and Jonathan Patterson. 'An Automated Timeline Reconstruction Approach for Digital Forensic Investigations'. *Digital Investigation*, The Proceedings of the Twelfth Annual DFRWS Conference, 9 (1 August 2012): S69–79. https://doi.org/10.1016/j.diin.2012.05.006

HM Crown Prosecution Service Inspectorate. '2019 Rape Inspection: A Thematic Review of Rape Cases by HM Crown Prosecution Service Inspectorate'. Thematic Review. HMCPSI, December 2019. Retrieved from https://www.justiceinspectorates.gov.uk/hmcpsi/wp-content/uploads/sites/3/2019/12/Rape-inspection-2019-1.pdf

HMICFRS. 'An Inspection into How Well the Police and Other Agencies Use Digital Forensics in Their Investigations', 2022. Retrieved from https://hmicfrs.justiceinspectorates.gov.uk/publication-html/how-well-the-police-and-other-agencies-use-digital-forensics-in-their-investigations/

Home Office. 'Extraction of Information from Electronic Devices: Code of Practice'. Home Office, October 2022.

House of Commons Science and Technology Committee. 'Forensic Science on Trial'. House of Commons, 2005. Retrieved from https://assets.publishing.service.gov.uk/media/5a7a372bed915d1a6421bdf4/forensic-science-on-trial.pdf

ICO. 'Mobile Phone Data Extraction by Police Forces in England and Wales'. Investigation Report. Information Commissioner's Office, June 2020. Retrieved from https://ico.org.uk/media/about-the-ico/documents/2617838/ico-report-on-mpe-in-england-and-wales-v1_1.pdf

ICO. 'Mobile Phone Data Extraction by Police Forces in England and Wales. An Update on Our Findings.' Information Commissioner's Office, June 2021. Retrieved from https://ico.org.uk/media/about-the-ico/documents/2620093/ico-investigation-mpe-england-wales-202106.pdf

ISO/IEC. 'General Requirements for the competence of testing and calibration laboratories'. ISO/IEC17025:2017(E).

Lin, Xiaodong, Ting Chen, Tong Zhu, Kun Yang, and Fengguo Wei. 'Automated Forensic Analysis of Mobile Applications on Android Devices'. *Digital Investigation* 26 (July 2018): S59–66. https://doi.org/10.1016/j.diin.2018.04.012

May, Tiggey, Catherine Talbot, and Rachel Skinner. 'Appendix 12: Pillar Six. Ecamining, Understanding and Improving the Use of Digital Material in RASSO Investigations'. In *Operation Soteria Bluestone Year One Report*, by Betsey Stanko, 2022. Retrieved from https://www.gov.uk/government/publications/operation-soteria-year-one-report/operation-soteria-bluestone-year-one-report-accessible-version

Metropolitan Police and CPS. 'A Joint Review of the Disclosure Process in the Case of R v Allan: Findings and Recommendations for the Metropolitan Police Service and CPS London.' Crown Prosecution Service, January 2018. Retrieved from https://www.cps.gov.uk/sites/default/files/documents/publications/joint-review-disclosure-Allan.pdf

NPCC. 'Operation Soteria – Transforming the Investigation of Rape'. National Police Chiefs' Council. Accessed 5 January 2024. Retrieved from https://www.npcc.police.uk/our-work/violence-against-women-and-girls/operation-soteria/

OPCC. 'Police "Odyssey" Lands Top Award for Sex Abuse Evidence Gathering'. *Gloucestershire's Office of the Police and Crime Commissioner* (blog), 25 April 2023. Retrieved from https://www.gloucestershire-pcc.gov.uk/police-odyssey-lands-top-award-for-sex-abuse-evidence-gathering/

Osborne, Grant, Benjamin Turnbull, and Jill Slay. 'The "Explore, Investigate and Correlate' (EIC) Conceptual Framework for Digital Forensics Information Visualisation'. In *2010 International Conference on Availability, Reliability and Security*, 629–34, 2010. https://doi.org/10.1109/ARES.2010.74

Owen, Richard, Phil Macnaghten, and Jack Stilgoe. 'Responsible Research and Innovation: From Science in Society to Science for Society, with Society'. *Science and Public Policy* 39, no. 6 (1 December 2012): 751–60. https://doi.org/10.1093/scipol/scs093

PACE. Police and Criminal Evidence Act 1984 (1984). *UK Public General Acts. Legislation.gov.uk.* Retrieved from https://www.legislation.gov.uk/ukpga/1984/60/contents

PCSC. Police, Crime, Sentencing and Courts Act 2022 (2022). *UK Public General Acts. Legislation.gov.uk.* Retrieved from https://www.legislation.gov.uk/ukpga/2022/32/contents/enacted

Privacy International (2018). *Digital Stop and Search: How the UK police can secretly download everything from your mobile phone.* Privacy International Report. Retrieved form https://privacyinternational.org/sites/default/files/2018-03/Digital%20Stop%20and%20Search%20Report.pdf

Spranger, Michael, Jian Xi, Lukas Jaeckel, Jenny Felser, and Dirk Labudde. 'MoNA: A Forensic Analysis Platform for Mobile Communication'. *KI - Künstliche Intelligenz* 36, no. 2 (1 September 2022): 163–69. https://doi.org/10.1007/s13218-022-00762-w

Stanko, Betsy. 'Operation Soteria Bluestone Year 1 Report 2021 – 2022', 2022. Retrieved from https://assets.publishing.service.gov.uk/media/63c02994d3bf7f6c287b9ff7/E02836356_Operation_Soteria_Y1_report_Accessible.pdf

Sternberg, Robert J., and Joseph A. Horvath. *Tacit Knowledge in Professional Practice: Researcher and Practitioner Perspectives*. Psychology Press, 1999.

Tassone, C, B Martini, and Kim-Kwang Raymond Choo. 'Forensic Visualization: Survey and Future Research Directions'. In *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, edited by Kim-Kwang Raymond Choo and Ali Dehghantanha. Syngress, 2016. Retrieved from https://www.oreilly.com/library/view/contemporary-digital-forensic/9780128054482/B9780128053034000113.xhtml

Tully, Gillian. 'Annual Report November 2016 - November 2017'. Forensic Science Regulator, 2018. Retrieved from https://assets.publishing.service.gov.uk/media/5a820029e5274a2e87dc09f0/FSRAnnual_Report_2017_v1_01.pdf

NATIONAL CYBERSECURITY STRATEGY ACTION PLAN FOR CYBER RESILIENCE: QUALITATIVE DATA AND ACHIEVEMENTS

William Steingartner, Darko Galinec

Technical University of Košice (Slovakia), Zagreb University of Applied Sciences (Croatia)

william.steingartner@tuke.sk; darko.galinec@tvz.hr

ABSTRACT

Cyber issues of importance to the state and the global environment represent a much wider area than the field of cybersecurity and are closely related to several traditional departments of public administration. Cybersecurity in these matters is the basis for their smooth development in the virtual dimen- sion of modern society. Cybersecurity is a part of all public administration processes, as all processes rely on the proper functioning of communication and information systems, either directly, through data processing, storage and transmission, or indirectly through the management of basic services (e.g. electricity distribution, transport, etc.). Given the widespread dispersion of responsibilities of state bodies in cyberspace, the establishment of the National Council for Cybersecurity, Operational and Technical Coordi- nation for Cybersecurity and the development of the National Cyber Security Strategy and Action Plan for its implementation establishes a mechanism for sharing information and harmonizing public administration professional and political/administrative level. This paper presents a qualitative assessment of the implementation of the Action Plan of the Strategy based on the out- comes of reporting to the holders and co-carriers of the implementation of the Action Plan's measures at the state level.

KEYWORDS: action plan, cyber attack, cybersecurity, cyber defense, cyber operations, cyber resilience, national cybersecurity strategy, qualitative assessment.

1. INTRODUCTION

In the development of the National Cybersecurity Strategy and Action Plan for its implementation comprehensive approach to cybersecurity by cov- ering cyberspace and infrastructure and users that fall under the jurisdiction of the Republic of Croatia (citizenship, registration, domain, address) is used as well as integration and harmonization of activities and measures arising from various aspects of cybersecurity and falling under the competence of various organizations and their complementarity in order to create a safer common cyberspace.

A proactive approach by constantly adapting the activities and measures applied in cyberspace and by occasionally adapting the relevant strategic frameworks was needed for strengthening the resilience, reliability and adaptability of information systems by implementing certification, accreditation and security protocols (Szymoniak, 2021a; Szymoniak, 2021b), especially taking into account the specific requirements of data, services and other business processes on information systems. Using probabilistic techniques, various parameters and behaviors of security protocols embedded in the authentication systems can be thoroughly examined (Siedlecka-Lamch, 2021; Palša et al., 2022; Vokorokos et al., 2016). The basic principles on which modern society is based (Cesarec, 2021; Gálik and Tolnaiová, 2019) are also applied in the cyberspace that makes up the virtual dimension of society:

- Application of the law for the purpose of protection of human rights and freedoms, especially privacy and the right to expression, property and all other essential features of an organized modern society.

- Harmonized legislative framework and continuous improvement of regulatory mechanisms through harmonized initiatives of all sectors of society, i.e. bodies and legal entities.
- The principle of subsidiarity through the systematic elaboration of the power to decide and inform on cybersecurity issues to the body whose competence largely covers the problem to be solved, whether the prob- lem relates to the organization, coordination and cooperation, or tech- nical capabilities to respond to computer communication threats and information infrastructure.
- The principle of proportionality between the increase of protection measures and responsibilities and decreasing negative consequences and accompanying costs and reduction of associated risks, i.e. greater possibilities to limit the threats that cause them.

The main goal of this paper is to present the results of the implementation of the National Cyber Security Strategy because of research by qualitative analysis based on reports of sectoral bodies as responsible bodies for the implementation of action plan measures and implementation of the strategy. As authors of (Korauš et al., 2022) state that the implementation of the national project will significantly improve the readiness of public authorities at both central and regional levels to detect, analyze and adopt targeted measures against hybrid threats.

The paper is structured as follows: in the Introduction, we provide the reader with the basic information on the National Cyber security policy of the Republic of Croatia along with the Action plan for its implementation. In Section 2, we cover the strategic methodology for the research. In Section 3, General objectives of the National Cybersecurity Strategy along with Cybersecurity Areas covered in the Implementation of Action Plan are presented. In Section 4, we formulate, show, and discuss the results of the research: a qualitative assessment of the results of the National Cybersecurity Strategy with the Case Study of Action Plan Implementation in the period 2016-2022, giving the conclusion. In Section 5, conclusion along with guidelines for future work are given.

2. METHODOLOGY

Qualitative assessment is a method that focuses on research on subjective aspects of reality, rather than obtaining empirical data, as is quantitative. It is used both in the evaluation of intervention programs, action plans and other similar areas ("Qualitative Assessment," 2022). Qualitative method or qualitative research, as it is also called, is a research technique or method that refers to traits and is used especially in the social sciences. Besides, data collection is a vital part of a qualitative research (Bojović and Lygre, 2023). However, it is also used in political and market research. This method is based on a detailed description of events, facts, people, situations, behavior, and interactions that are observed through research ("What is the Qualitative Method," 2022). Qualitative assessment is not about large sample sizes or generalizability power. Rather, the selection of “the case” or “the participant” is based on strong theoretical reasoning (deductive approach), empirical data, and/or follows a reasonable logical path (inductive approach). Therefore, one must know why a given case or participant is of interest, either based on theory or exploratory logic ("Sampling: The Importance," 2022). Qualitative research methods are most appropriate in situations in which little is known about a phenomenon or when attempts are being made to generate new theories or revise preexisting theories. Qualitative research is inductive rather than deductive and is used to describe phenomena in detail, without answering questions of causality or demonstrating clear relationships among variables ("Qualitative Methods," 2022). The methodology of the approach chosen to define the contents of the Croatian National Cybersecurity Strategy was based on determining the general goals of the Strategy, society sectors covered by the Strategy, and basic principles of approach to the implementation of the Strategy.

The measures necessary to achieve the objectives set out in this Strategy, together with the competent authorities and deadlines for their implementation, are set out in the Action Plan for the implementation of this Strategy. The Action Plan for the implementation of the Strategy enables systematic monitoring of the implementation of the Strategy and is a control mechanism used to assess whether an individual measure was ultimately implemented, to what extent, whether it achieved the desired result or needs to be re- defined according to new needs. To monitor the goals realization, within National Cyber Security Strategy a system of its continuous monitoring has been established. Accompanying Action Plan established a mechanism for coordinating state bodies. Such coordinated bodies create appropriate policies and responses, within their competencies, to create appropriate policies and responses to threats in cyberspace. The National Cyber Security Council is an inter-ministerial body for the coordination of horizontal national initiatives in the field of cybersecurity. The Council primarily deals with the objectives of the Strategy and the measures of the Action Plan, initiates discussions and makes recommendations and conclusions on all current issues related to cybersecurity. The Council operates through the nominal competencies of bodies and institutions whose representatives are appointed to the work of the Council (primarily the public sector). Further work will seek to further improve and strengthen the established formal cross-sector coordination between the state, academic, economic, and public sectors, based on the continuation of activities undertaken by the Council in the past through its activities and the activities of bodies participating in the Council (Government of the Republic of Croatia, 2015; Galinec, Možnik, & Guberina, 2017).

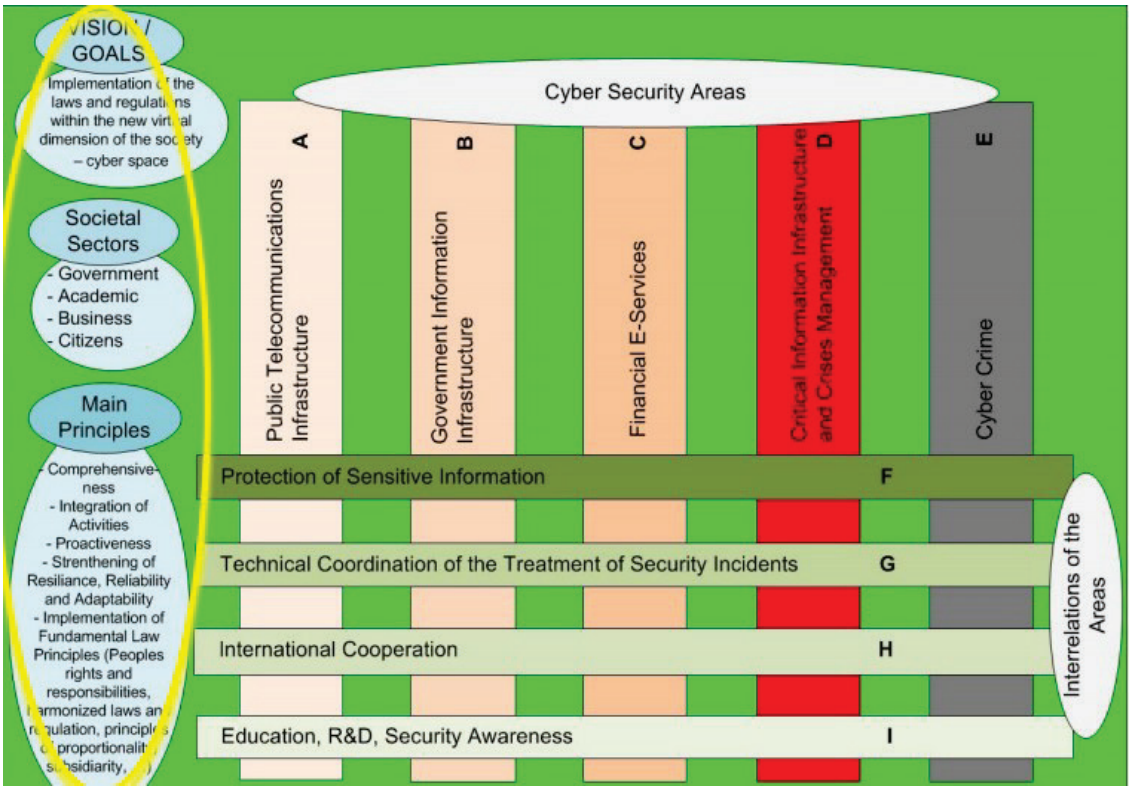
Operational-technical coordination for cybersecurity is an operational interdepartmental body, established for more efficient coordination of prevention and response activities to cybersecurity threats. Coordination operates primarily in terms of a complementary approach of bodies and institutions whose representatives are appointed to the work of the Coordination (primarily the public sector) in the prevention and resolution of security incidents. At the same time, this harmonizes the development of national capabilities in cyberspace (Government of the Republic of Croatia, 2015; Galinec, Možnik, & Guberina, 2017).

3. NATIONAL CYBERSECURITY STRATEGY AND IMPLEMENTATION OF ACTION PLAN: EXPLORING CYBERSECURITY AREAS

The area of cyber defense represents the part of the defense strategy falling under the responsibility of the ministry in charge of defense issues. It is the subject of separate elaboration and action, which will be pursued using all the necessary elements arising from this Strategy. Cyber-terrorism and other cyber aspects of national security are dealt with by a small number of competent bodies within the security and intelligence system and require a separate approach, and that will also include the use of all the necessary elements arising from this Strategy. Figure 1 (reprinted from (Klaić, 2017)) shows and describes cyber- security areas and their interrelations as well as vision/goals, societal sectors and main principles of the strategy.).

Cybersecurity areas are analyzed with the general goals of the Strategy to identify the special objectives aimed at achieving improvements in each area and the measures necessary for achieving the goals of the Strategy. The special objectives, as well as the measures that will be further elaborated by the Action plan for the implementation of the Strategy, are determined concerning the defined society sectors and the influence of the cybersecurity area on each sector, but also concerning the forms of cooperation and coordination of cybersecurity stakeholders. The principles defined by the Strategy are followed in the elaboration of the cybersecurity areas. It is logical to predict that the vulnerability of geospatial data is expected to be potentially and ultimately targets for physical attacks on objects within the data (e.g. changing the metadata with false information within the georeferenced data).

Figure 1. Components of the Strategy.



Source: reprinted from (Klaić, 2017).

4. QUALITATIVE ASSESSMENT OF THE RESULTS OF NATIONAL CYBERSECURITY STRATEGY ACTION PLAN IMPLEMENTATION

Action plan for the implementation of the Strategy, made to implement the Strategy, elaborates the defined strategic goals and determines the implementation measures necessary for achieving those goals, along with the competent authorities and the list of deadlines for their implementation. An action plan for the implementation of the Strategy allows for systematic over- sight of the implementation of the Strategy and serves as a control mechanism that will show whether a certain measure has been implemented in its entirety and has produced the desired result, or it should be redefined following the new requirements.

4.1. Case study: Action Plan Implementation Results Qualitative Assessment

To determine in due time whether the Strategy is achieving the desired results, namely if the defined goals are being accomplished and the established measures implemented within the planned time frame, it is necessary to set up a system of continuous monitoring of the implementation of the Strategy and Action plan, thus also setting up a mechanism for coordinating all the competent government bodies in creating the appropriate policies and responses to threats in cyberspace.

The entities tasked with the measures from the Action plan for the implementation of the Strategy are responsible for monitoring and collecting information on the implementation and efficiency of the measures and are required to submit consolidated reports to the National Council once a year, no later than the end of the first quarter of the current year for the previous year or, if necessary, more frequently, namely at the request of the National Council.

The National Council will submit the reports on the implementation of the Action plan for the implementation of the Strategy to the Government of the Republic of Croatia no later than the end of

the second quarter of the current year for the previous year. The Strategy will be revised after three years of implementation, based on the reports of the entities tasked with the measures from the Action plan for the implementation of the Strategy. The National Council shall submit to the Government of the Republic of Croatia a consolidated report with the proposed amendments to the Strategy no later than the end of the year of revision (Galinec, Možnik, & Guberina, 2017). Following are the described the results of the implementation of the Action Plan by particular years.

4.2. Year 2016

Most institutions in their capacity as holders of individual measures of the Action Plan, of which The National Cyber Security Council requested the completion of the forms, conducted their obligation and submitted the necessary data for analysis to the Council. In 2016, there was a noticeable lack of the Council as an inter-ministerial body to encourage implement the measures of the Action Plan and conduct the necessary mediation in cases where additional interpretation or harmonization of different interpretations of stakeholders in the implementation of the Strategy is needed, resulting in a slower start of implementation of measures and more difficult implementation in measures that relate to more complex issues and a larger number of institutions/holders / co-carriers. The start of the implementation of the Action Plan in 2016, however, has yielded some results in large numbers from about 30 indebted institutions of various profiles. The start of the implementation of the Action Plan resulted in also a significantly greater understanding of cybersecurity issues in very different ways institutions involved in the implementation of the Action Plan. All institutions are in this initial phase of the implementation of the Action Plan recognized and linked the activities within their competence with thematically conceived measures of the Action Plan. Designation of coordinators for the implementation of measures of the Action Plan in the institutions in a large number of cases was determined by monitoring the closest competencies in the portfolio of competencies of the institution, but in that part, it is necessary with some stakeholders to work on the more effective appointment of coordinators for the implementation of the Action Plan measures. In the area of critical information infrastructure, the problem of the inability to use national results in critical infrastructure sectors, due to slow law enforcement nationally defined sectors and reorganization was undertaken to implement similar obligations for The Republic of Croatia arises from the obligations to implement the EU NIS Directive. To further implement national measures in this area it is necessary to complete the implementation of activities in advance, if necessary, with changing the legislative framework in the field of national critical infrastructures, to approach the implementation of measures in the field of critical communication and information system. The key issue that needs to be addressed is the need for much greater consistency of educational programs in the field of cybersecurity and better training of lecturers at different levels and types of education. The current situation still indicates a low degree of consistency of the program and insufficient training of lecturers and thus questionable results of cybersecurity education programs implemented in the Republic of Croatia. Elaboration of cybersecurity within the Strategy and Action Plan should be the framework for the development of all national educational programs in this area, and the interdepartmental body, the National cybersecurity council, should be appropriately involved in the advisory the process of the competent ministry and other bodies related to curricular reform and by improving all types and levels of education in Croatia (The Office of the National Security Council, 2017).

4.3. Year 2017

All institutions in their capacity as holders of individual measures of the Action Plan have implemented their obligation and submitted to the Council the information necessary for the preparation of this Report. In 2017, after the establishment of the Council as an interdepartmental body and the beginning

of its work in March 2017, a series of processes described in the Council's annual report19 was launched, which are extremely important for the implementation of the Action Plan. Among other things, to encourage the implementation of Council guidelines from the Report on the Implementation of the Action Plan in 2016, data were collected and a collective e-booklet was prepared for all stakeholders in the implementation of the Action Plan, in which they have indicated the basic competencies of all stakeholders of the Strategy, their role in the work of interdepartmental bodies and implementation of measures from the Action Plan and contact details of persons in charge of institutions for coordinating the implementation of individual measures. The continued implementation of the Action Plan during 2017 resulted in a further increase in awareness and understanding of cybersecurity issues in very different institutions and the sectors involved in the implementation of the Action Plan. Institutions that are stakeholders in the Strategy and implement the measures from the Action Plan are increasingly recognizing and linking activities from their core competencies with thematically conceived measures of the Action Plan. Designation of coordinators' implementation of the Action Plan measures in the institutions in a large number of cases has been identified by monitoring the closest competencies in the portfolio of competencies of an individual institution. In the field of critical communication and information infrastructure and crisis management, great progress has been made in drafting and proposing a new Act of cybersecurity for key service operators and digital service providers, whose adoption, bundled with the Decree, is expected in the middle of 2018. This is an extra big step in alignment obligations arising for the Republic of Croatia from the requirements of the implementation of the EU NIS Directive, which are determined by this Act completely take over. The key issue that still needs to be worked on is the need for much greater consistent educational programs in the field of cybersecurity and better training of lecturers at different levels and types of education. The current situation shows the initial shifts in a good direction, but still indicates a low degree of program consistency and insufficient training of lecturers, and thus the questionable results of educational programs cybersecurity implemented in the Republic of Croatia. Elaboration of cybersecurity within Strategies and Action Plan should be the framework for the development of all national education programs in this area, and the inter-ministerial body, the National Council for Cyber Security, will continue initiatives toward the competent authorities and stakeholders in the implementation of the Action Plan with the aim improvement of all types and levels of education in the Republic of Croatia (The Office of the National Security Council, 2018).

4.4. Year 2018

The continued implementation of the Action Plan during 2018 has resulted in a significant increase in security awareness at the national level and a much better understanding of the issue of cybersecurity in the various institutions and sectors involved in the implementation Action Plan. Institutions that are stakeholders in the Strategy and implement the measures from the Action Plan all better recognize and connect the activities from their core competence with the thematically conceived measures of the Action Plan, but insufficient horizontal cooperation with others is still observed by competent authorities in the implementation of measures, in particular between institutions belonging to different sectors. In the field of critical communication and information infrastructure and crisis management, an extremely big step forward was made in 2018 with the enactment of the Cybersecurity Act security of key service operators and digital service providers and the related Regulation, whose adoption also harmonized the obligations arising from the requirements for the implementation of the EU NIS for the Republic of Croatia directive. Thus, the Republic of Croatia positioned itself on the map of EU member states in the group of countries that have successfully verified the national cybersecurity strategy and which they have successfully launched solving extremely complex issues of critical communication and in- formation and infrastructure. It is important to emphasize both the area of education and the development of security awareness, which

is in time adoption of the current Strategy was far behind most other areas covered by the Strategy, today is beginning to successfully catch up with the development of other areas to which the Action Plan refers. Experience to date with the implementation of the Action Plan in the period from 2016 to date, clearly shows the need to actively guide the implementation of the measures from the Action Plan, as they are crucial results in the implementation of the Action Plan achieved in 2017 and 2018, respectively the establishment and commencement of the work of the Council, as an inter-ministerial body now composed of representatives of 18 institutions. Update of the Strategy and the corresponding Action Plan, which is planned to be proposed to the Government of the Republic of Croatia by the end of 2019, will be based on an analysis of many successfully implemented goals of the Strategy in the past period, on changing the approach to goals that have not been fully achieved or their achievement progress is slower, as well as the introduction of new goals dictated by the global environment and rapid development information and communication technologies. One of those areas that need special addressing by updating the Strategy, there is certainly a need for a more efficient and formal cross-sector coordination between the public, academic and private sectors (The Office of the National Security Council, 2019).

4.5. Years 2019-2020

The continuation of the implementation of the Action Plan during 2019 and 2020 resulted in a further increase in security awareness at the national level and the activation of measures that were delayed compared to the planned dynamics of the implementation of the Action Plan in previous years. The importance of cybersecurity as a prerequisite for the digital transformation of society is now much better accepted, both within the state sector, as well as within the economy and citizens. Cyberspace risks are seen less and less as potential, unlikely events, and more and more as something that is inevitable and for which one should be prepared. Institutions that are stakeholders in the Strategy and implement measures from the Action Plan cooperate with each other and coordinate their actions. Despite all the legal, financial, organizational and personnel limitations and difficulties, the strategy has been successful since its adoption until today. New requirements, both realistic and life-like, due to the application of new technologies and consequently new risks, as well as due to the need to harmonize with international alliances and assumed obligations, inevitably require further adjustments to the Strategy and the vision of national needs in the next time period. Work on our National Cybersecurity Strategy has already begun. While the first Strategy tried to address priorities, the new Strategy should create more ambitious organizational, legal, financial and human pre-requisites for a secure digital society of the future (The Office of the National Security Council, 2021).

4.6. Year 2021

The continuation of the implementation of the Action Plan during 2021 resulted in a further increase in security awareness at the national level and a better understanding of cybersecurity issues in the various institutions and sectors involved in the implementation of the Action Plan. In the area of critical communication and information infrastructure and crisis management, further improvements are expected after the adoption of the Act on Critical Infrastructures in 2022 and the transposition of the NIS 2 Directive, which should be implemented in 2023-2024. In the field of education and development of security awareness, significant progress has been achieved in the past period. The previous experience with the implementation of the Action Plan in the period from 2016 to today clearly shows the need for active monitoring of the implementation of the measures from the Action Plan because the key results in the implementation of the Action Plan were achieved under the guidance of the Council. Respecting the requirements of the NIS 2 directive, which defines the elements that member states must include in new generations of national cybersecurity strategies, as well as taking into

account the new threats and risks that come from cyberspace every day, the National Council for Cyber Security agreed on the need to adopt new national cyber- security strategies. The proposal for a new strategy, which will be based on the identified risks and the projection of the need to increase cybersecurity for the next (maximum 5) years of the digital decade, while respecting all the requirements set by the NIS 2 directive, is in progress (The Office of the National Security Council. 2022a).

4.7. Year 2022

In the previous year, the strong engagement of state authorities in improving security continued in Croatian cyberspace. At the same time, the challenges have become greater. Let's just mention aggression against Ukraine and its reflections on cyberspace, increasingly sophisticated cybercrime, increasing digitization and dependence on electronic services, and lack of enough professional staff. The Republic of Croatia coped with challenges in the field of cybersecurity through cooperation, exchange of information, and harmonization of actions, both between bodies and within associations to which it belongs. However, it is necessary to emphasize that actions in the field of are cybersecurity still mostly in the initial phase: they are based on voluntary cooperation, critical moments are addressed, funds are provided from the existing budgets of state bodies with financing from EU funds, but there is still no leading competent authority that would have the capacity to realize more complex projects. The new National Cybersecurity Strategy and the new Cybersecurity Law should make a significant step forward and adequately deal with and counter the risks that are growing expected in the next 3-5 years and to provide security and trust to citizens and organizations (The Office of the National Security Council, 2023).

4.8. Discussion on Results

The activities of the Council in 2019 were focused on a systematic and coordinated approach to the implementation of both current national programs and EU and NATO processes and initiatives. Cyber issues of importance to the state and the global environment are much broader area from the area of cybersecurity addressed by the Council and are closely linked to the series of traditional departments of public administration, while cybersecurity in these matters represents only the basis for their smooth development in the virtual dimension of modern society. Office of The National Security Council has in the last five years, in particular since the establishment of the Council and coordinated his work, connecting and coordinating the work of the body in the field of cybersecurity, given the shared competence in cybersecurity issues between more bodies. Meanwhile, most bodies have significantly developed their capabilities in the area of cybersecurity, on the one hand, thanks to the accelerated development of information and communication technology that does not suffer lagging behind and on the other hand investing (not only financial) in the development of their capabilities, which is a prerequisite for further improvement cybersecurity at the national level and dealing with the growing challenges in cyberspace, which requires a high level of expertise, capacity sufficient for launch and implementation own initiatives and a proactive approach, which is expected from all bodies in the coming period involved in the work of the Council, but also more widely through the implementation of the Strategy. Mentioned ability to progress individual bodies of Council members in matters of cybersecurity also directed the distribution responsibilities recognized by the new, revised, Strategy and associated Action Plan. By recognizing the need for strategic security-intelligence action to prevent various cyber attacks directed against the national security of the Republic of Croatia, by expanding and improving the system for detection, early warning and protection from the state-sponsored cyber attacks and the continued development of tied surveillance centers increases the security of cyberspace but also strengthens capacities and competencies at the wider national level. In this way, an additional significant step forward was made, not only in terms of revision of the

Strategy as a document but also a proactive approach to the issue of security of the cyberspace of the Republic of Croatia. The accelerated development of information and communication technology is constantly creating new challenges, through the fulfillment of the measures set out in the Action Plan for the implementation of the Strategy respond quickly and effectively, i.e. before a new risk or threat arises. Even in the most optimistic scenario, some of the challenges facing the Republic of Croatia daily, whether complexity, competence, organization or coordination, will certainly take several years to resolve. At this point, the Strategy provides sufficient the opportunity for the necessary transformation of our future security and the preservation of progress in the digital age, and in the future, some of these issues and new challenges will be addressed within (some future) a national agency or center that would be in charge of all cybersecurity issues space of the Republic of Croatia. And within the work of the Council during 2019, it sought to emphasize the need to develop awareness and the ability of state bodies to exercise their competencies and responsibilities, both in real, as well as in cyberspace, which is why thematic sessions of the Council were held, to the importance of individual issues was further emphasized. Given that this synergistic approach gives positive steps, the same will continue and intensify in the coming period. Most bodies have developed more significant capabilities of their own in the field of cybersecurity, on the one hand, encouraged by the accelerated development of information and communication technology that does not suffer lag and on the other hand by investing (not only financially) in development own capabilities, which is a prerequisite for further improving cybersecurity on the national level and dealing with the growing challenges in cyberspace for which it is necessarily high level and concentration of expertise, capacities sufficient for launching and realising of own initiative and a proactive approach that is expected in the coming period from all bodies involved in the work of the Council, but also wider through the implementation of the Strategy and by launching its initiatives. Many of the activities presented in this report are presented as individual body actions within its competencies, however, there is coordination and guidance behind such actions Councils to make optimal use of national capacities in the field of cybersecurity. The accelerated development of information and communication technology creates new challenges every day. Even in the most optimistic scenario, some of the challenges facing the Republic of Croatia face daily, whether it is a matter of complexity, competence, organization, or coordination, will certainly take several years to resolve. At this moment, the Strategy, along with the appropriate enhancement of the capabilities of individual bodies and the Council with its coordinating role, provides a sufficient basis for further continuation of the necessary transformation of cybersecurity management and preservation of progress in the digital age, but not for a long time. Security management at both central and regional levels is currently lacking in identifying and addressing hybrid threats and their manifestations, and in most public institutions – local government and regional self-government – this activity is a marginal and ignored area of security. To solve and cope with new, future (but not distant), rapidly growing challenges for the protection of the cyberspace of the Republic of Croatia, it is necessary to rapidly establish centralized management of cybersecurity for all levels, ultimately through the legislative framework, by establishing a central body responsible for all issues of cybersecurity. Such demands are imposed on us by the accelerated digitalization process, as well as the increase in threats and threats in the virtual dimension of society (The Office of the National Security Council, 2022b).

5. CONCLUSIONS

The National Cybersecurity Strategy planning and the Action Plan for its implementation development have been created and executed based on integration, inclusiveness and integrity principles by drafting strategic guide- lines, concept development, plan development and plan assessment to achieve situational awareness at the national level.

Furthermore, one of the main principles of the strategy was strengthening resilience, reliability and adjustability by applying universal criteria of confidentiality, integrity and availability of certain groups of information and recognized social values, in addition to complying with the appropriate obligations related to the protection of privacy, as well as confidentiality, integrity and availability for certain groups of information, including the implementation of appropriate certification and accreditation of different kinds of devices and systems, and also business processes in which such information is used (Dewar, 2018).

The Republic of Croatia is committed to keeping cyberspace open, stable and secure. However, recognizing cyberspace as a new, fifth domain of warfare as well as a platform for hostile actions, the Republic of Croatia is ready to protect its sovereignty, independence, rule of law and democratic processes, national priorities, critical infrastructure and other interests with all necessary measures and actions. The government has a major role to play in stimulating progress toward higher levels of cybersecurity.

ACKNOWLEDGEMENTS

This work was supported by national KEGA project 030TUKE-4/2023 – Application of new principles in the education of IT specialists in the field of formal languages and compilers, granted by the Cultural and Education Grant Agency of the Slovak Ministry of Education.

REFERENCES

Bojović, D., & Lygre, J. T. (2023). To deceive or not deceive: Unveiling the adoption determinants of defensive cyber deception in Norwegian organizations (Master's thesis). University of Agder, Kristiansand, Norway.

Cesarec, I. (2020). Beyond physical threats: Cyber-attacks on critical infrastructure as a challenge of changing security environment — Overview of cyber-security legislation and implementation in SEE countries. Annals of Disaster Risk Sciences, 3(1). https://doi.org/10.51381/adrs.v3i1.45

Dewar, R. (2018). National Cybersecurity and Cyberdefense Policy Snapshots: Collection 1. Center for Security Studies (CSS), ETH Zurich.

Gálik, S., & Tolnaiová, S. (2019). Cyberspace as a new existential dimension of man. In E. Abu-Taieh, A. E. Mouatasim, I. H. A. Hadid (Eds.), Cyberspace (Chapter 2). IntechOpen. https://doi.org/10.5772/intechopen.88156

Galinec, D., Možnik, D., & Guberina, B. (2017). Cybersecurity and cyber defence: National level strategic approach. Automatika Journal for Control, Measurement, Electronics, Computing and Communications, 8(3), 266–272.

Government of the Republic of Croatia. (2015). The National Cyber Security Strategy and Action Plan for the Implementation of the Strategy. Official Gazette, 108/2015.

hr.awordmerchant.com (2022). What is the qualitative method? its definition and meaning. Retrieved from https://hr.awordmerchant.com/m-todo-cualitativo (accessed on 2022-08-03).

JournalMural.com (2022). Qualitative assessment: characteristics, advantages, examples. Source: Retrieved from https://hr.journalmural.com/evaluaci-n-cualitativa (accessed on 2022-08-01) (2022)

Klaić, A. (2017). The Importance of National Cyber Security Organization – Prerequisite of Cyber Resilience. In: Workshop on Establishing a Cyber Incident Response Team (TAIEX JHA 61275), Skopje, Macedonia.

Korauš, A., Kurilovská, L., & Šišulák, S. (2022). Increasing the competencies and awareness of public administration workers in the context of current hybrid threats. In: Conference Proceedings, RELIK 2022: Reproduction of Human Capital - Mutual Links and Connections (pp. 379–388).

Palša, J., Ádám, N., Hurtuk, J., Chovancová, E., Madoš, B., Chovanec, M., & Kocan, S. (2022). MLMD—A Malware-Detecting Antivirus Tool Based on the XGBoost Machine Learning Algorithm. Applied Sciences, 12(13), 6672. https://doi.org/10.3390/app1213667

ScienceDirect (2022a). Qualitative methods. Retrieved from https://www.sciencedirect.com/topics/psychology/qualitative-research-method (accessed on 2022-08-09).

ScienceDirect (2022b). Sampling: The importance of case selection. Retrieved from https://www.sciencedirect.com/topics/social-sciences/qualitative-assessment (accessed on 2022-08-08).

Siedlecka-Lamch, O. (2021). Probabilistic and timed analysis of security protocols. In A. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, & E. Corchado (Eds.), 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) (pp. 142–151). Springer International Publishing.

Szymoniak, S. (2021). Using a security protocol to protect against false links. In Moving Technology Ethics at the Forefront of Society, Organisations, and Governments (pp. 513–525). Universidad de La Rioja.

Szymoniak, S. (2021). Security protocols analysis including various time parameters. Mathematical Biosciences and Engineering, 18(2), 1136–1153.

Vokorokos, L., Mihal'ov, J., & Chovancová, E. (2016). Motion sensors: Gesticulation efficiency across multiple platforms. In 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES) (pp. 293-298). Budapest, Hungary. http://doi.org/10.1109/INES.2016.7555139

The Office of the National Security Council (UVNS). (2017). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2016, Croatia.

The Office of the National Security Council (UVNS). (2018). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2017, Croatia.

The Office of the National Security Council (UVNS). (2019). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2018, Croatia.

The Office of the National Security Council (UVNS). (2021). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2020, Croatia.

The Office of the National Security Council (UVNS). (2022). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2021, Croatia.

The Office of the National Security Council (UVNS). (2023). Report on the implementation of the Action Plan for the Implementation of the National Cybersecurity Strategy 2022, Croatia.

The Office of the National Security Council (UVNS). (2022). Annual report on the work of the National Cyber Security Council and operational and technical coordination for cybersecurity for 2021, Croatia.

PRIVACY AFTER DOBBS: HOW THE SHIFTING U.S. LANDSCAPE AFFECTS THE BROADER DEBATE

Michael S. Kirkpatrick

James Madison University (USA)

kirkpams@jmu.edu

ABSTRACT

The 2022 *Dobbs* decision by the Supreme Court of the United States ("SCOTUS") caused an immediate and very significant change to fundamental rights in the country. Although the primary effects of the decision centered on the legal protections for abortion, there is a secondary change shift in the decision that has the potential to alter a much broader collection of rights. Specifically, *Dobbs* overturned the precedent set forth in *Roe v. Wade* that the U.S. Constitution protected the right to abortion as a consequence of broader protections surrounding due process and privacy. The *Dobbs* decision declared this reasoning to be "egregiously wrong." As a result, a key pillar of the legal conception of privacy was removed.

At the core of this decision is a long history regarding the nature of privacy itself and what fundamental human rights are guaranteed by a right to privacy. The key element that arose in the *Dobbs* decision is whether the broader privacy protections entail *decisional privacy*, the right to make significant and intimate decisions without government interference. That is, is privacy fundamentally limited to questions surrounding the *disclosure* of information or does it also imply the right to *act* on that information. Some scholars argue that decisional privacy is simply a form of autonomy and should be removed from the broader debate about privacy protections; others argue that violations of decisional privacy cause harms that are intertwined with other aspects of privacy and viewing decisional privacy as simply a form of autonomy would significantly weaken protections to mitigate these harms.

In this paper, we will provide an overview of the U.S. legal framework as it relates to privacy. We will also examine multiple perspectives on the nature of privacy and the extent to which these competing views incorporate decisional privacy. We will close with a focus on the specific findings of *Dobbs* and related decisions to identify how the loss of protection for decisional privacy impacts other fundamental rights.

KEYWORDS: privacy, decisional privacy, SCOTUS, substantive due process.

1. INTRODUCTION

In June 2022, the Supreme Court of the United States ("SCOTUS") released its decision in *Dobbs v. Jackson Women's Health*, overturning two earlier decisions (*Roe v. Wade* in 1973 and *Planned Parenthood v. Casey* in 1992. The most immediate focus of these three decisions centers around legal protections for abortion throughout the U.S. Under *Roe*, no state (or the federal government itself) could pass a law that restricted abortion in the first two trimesters of pregnancy. The *Casey* decision mostly upheld *Roe*, although it allowed states to pass certain restrictions so long as they did not pose an "undue burden" on pregnant women. These two decisions established the right to abortion as having a foundation in the U.S. Constitution that could not be undermined through basic legislative action. The *Dobbs* decision overturned both *Roe* and *Casey*, thereby declaring these earlier decisions invalid. Abortion was no longer considered to be a fundamental right protected by the Constitution, allowing states to pass laws that would completely ban abortion, which many states did.

Although these three decisions and their related debates are primarily focused on the legality of abortion, they are more properly understood as decisions about the nature of privacy and whether privacy is considered to be a fundamental right in the U.S. In the case of *Roe* and *Casey*, the protection

of abortion was an indirect effect of an implicit right to privacy. In both decisions, the right to privacy was determined to be implied by the Fourteenth Amendment's protection of the right to due process. The *Dobbs* ruling declared this finding to be "egregiously wrong." By overturning these earlier decisions, *Dobbs* removed significant precedents that had been used to protect other fundamental rights, including rights for interracial or gay marriage, that had relied on the same reasoning as *Roe* and *Casey*.

The definition and limits of the nature of privacy lies at the core of this debate. The most broadly accepted conceptions of privacy focus on protections relating to the disclosure of information. That is, privacy is often conflated with a right to keep certain forms of information secret from the state or other entities. This history and legal tradition can be observed in the quote attributed to Cardinal Richelieu: "If you give me six lines written by the hand of the most honest of men, I will find something in them which will hang him." The earliest privacy protections aimed to prevent this harm by keeping the necessary information from actors who would be in a position to abuse that knowledge. In other words, early conceptions of privacy focused on a right to keeping certain information secret.

This conception of privacy, which was prevalent in the 18th and 19th centuries, later became viewed to be insufficient. Warren and Brandeis (1890) argued that there should be protection for a broader "right to be let alone," which laid the foundation for a much wider array of protections. This line of reasoning led to shifts that paved the way for *Griswold v. Connecticut* (1965) and *Roe*. These rulings found that constitutional protections for due process necessarily entailed a right to privacy, specifically a right to *decisional privacy*. Specifically, the choice to use contraception (*Griswold*) or to abortion (*Roe*) is inherent in the right to due process, which is protected by multiple Amendments. Consequently, states were forbidden from enforcing laws that would intrude on these private, intimate decisions.

Although these rulings created the foundation for subsequent decisions based on a broader conception of privacy, they did not fully address key questions about the nature and limits of privacy. Instead, these issues continued to be debated in the legal and philosophical scholarly communities. Thomson (1975) argued that privacy was best viewed as "a cluster of rights" that are primarily focused on "the right over the person" (i.e., the physical body) and rights concerning "owning property." On the other hand, Solove (2007) argued that, while agreeing that privacy is not a singular right, it is "best used as a shorthand umbrella term for a related web of things," including the right to be free from interference (particularly by the state) with intimate, personal decisions.

The *Dobbs* decision has directly challenged this latter approach. By overturning *Roe*, *Dobbs* has declared that future rulings cannot use the argument that due process protections give one the ability to act on sensitive information. Rather, *Dobbs* has restricted the scope of privacy rights, as protected by the Constitution, to the realm of disclosure of information. In doing so, the ruling has invalidated the rationale that has been used for many other rights, including interracial marriage (*Loving v. Virginia*, 1967), bans on homosexual sodomy laws (*Lawrence v. Texas*, 2003), and gay marriage (*Obergefell v. Hodges*, 2010). While these subsequent rights were not addressed or overruled by *Dobbs*, it is not clear that SCOTUS would keep these rulings in place should a challenge arise.

In this paper, we will examine the legal and philosophical foundations regarding decisional privacy. We will start by summarizing the role of SCOTUS and the history of how its rulings have shaped privacy law in the U.S., contrasting this approach with the comprehensive protections in other legal frameworks, such as the European Convention on Human Rights (ECHR). We will also summarize contemporary frameworks for understanding privacy and examine the merits of different approaches. We will then analyze how *Dobbs* fits into this history and these frameworks, examining potential issues that the ruling has inadvertently created by removing protections for decisional privacy. We will close with a discussion of how these changes may impact privacy on the Internet and have ripple effects beyond the U.S.

2. FOUNDATIONS OF PRIVACY LAW IN THE U.S.

Before understanding the nature of privacy law in the U.S., it is necessary to understand the relationship between the Constitution, the legislative branch (Congress and the equivalent entity at the states), and the judicial branch. The Constitution forms the basis for all law in the U.S. In addition to the original Constitution itself, there have been 27 Amendments that have been ratified; these Amendments are part of the Constitution itself and have the same level of authority. Congress and the state legislatures pass additional laws, but these laws must be consistent with the principles of the Constitution.

When a party has been harmed by such a law, they can sue the relevant authority (typically a state or the U.S. itself) challenging the constitutionality of the law. The judicial branch, which is topped by SCOTUS as the final authority, has the ability to invalidate the law or to uphold it. Furthermore, because the U.S. adheres to the common law tradition, rulings by the judicial branch have the full force of law. In effect, this means that SCOTUS decisions have the same authority as laws passed by Congress. In certain circumstances, Congress or the state legislature can later overrule SCOTUS decisions; however, this is not common and SCOTUS rulings tend to be the final authority on cases and controversies.

The Constitution itself does not mention a right to privacy. However, there are five key Amendments that have been interpreted as guaranteeing an equivalent protection. The Third Amendment states that "No Soldier shall…be quartered in any house," establishing the home as a private realm. The Fourth Amendment subsequently extends privacy to include the right "to be secure in their persons, houses, papers, and effects." The Fifth and Fourteenth Amendments prevent the state from depriving someone of "life, liberty, or property, without due process of law." The Fifth Amendment applied this right only to the U.S. itself, while the Fourteenth Amendment extended this protection to the states[1] individually, as well. Finally, the Ninth Amendment states that the lack of enumeration in the Constitution is not sufficient grounds to deny that a right exists.

As privacy is not an enumerated right, it has been left to the judicial branch to define its legal definition and limits. Beginning with *Boyd v. U.S.* (1886), SCOTUS began to link the Fourth and Fifth Amendments to protect the "sanctity of man's home and privacies of life." This began a series of decisions that have focused on the conception of privacy as a right to secrecy. I.e., this branch of decisions has primarily examined questions about how and when the state can conduct surveillance on an individual or group as part of a criminal investigation. This branch is the most widely understood as focusing on questions of privacy rights.

As second branch can be traced to *Pierce v. Society of Sisters* (1925). This ruling, which was made as part of a challenge to a ban on private education, is one of the original *substantive* due process rulings in U.S. law. That is, while *procedural* due process requires the state to adhere to specific procedures in civil and criminal processes, substantive due process restricts the legislative branch from passing laws that arbitrarily intrude into citizens' fundamental rights. That is, laws that interfere with the free exercise of fundamental rights must pass a higher level of scrutiny in order to be upheld.

It is this notion of substantive due process, protected by the Fifth and Fourteenth Amendments, that formed the basis of privacy rights in future decisions. In *Griswold*, SCOTUS reiterated from an earlier

[1] The Fourteenth Amendment was ratified as part of the conclusion to the U.S. Civil War (1861 – 65), which was fought over the legality of the institution of slavery. Prior to the Civil War, some states allowed slavery (which denied enslaved persons of liberty without due process) while others did not. The Fourteenth Amendment was ratified explicitly to resolve this issue. Other rights have avoided such bloodshed as most other Amendments have been "incorporated," meaning that they have been determined to apply to the states as well as the U.S.

decision that the Fourth Amendment created a "right to privacy, no less important than any other right carefully and particularly reserved to the people." *Griswold* extended this reasoning by explaining that the Constitution contained *penumbras*, or zones of proximity, of certain explicitly mentioned rights. *Roe* used similar reasoning to find that "personal, marital, familial, and sexual privacy," including abortion, was "founded in the Fourteenth Amendment's concept of personal liberty and restrictions upon state action."

Planned Parenthood v. Casey reiterated this reasoning by declaring that "intimate family matters" were protected by the "fundamental right of privacy…against governmental intrusion." *Casey* strengthened this reasoning by noting that "personal decisions that profoundly affect bodily integrity, identity, and destiny should be largely beyond the reach of government." Furthermore, "restrictive abortion laws force women to endure physical invasions far more substantial than those this Court has held to violate the constitutional principle of bodily integrity in other contexts."

In essence, the interpretation of due process in the *Roe* and *Casey* decisions, as well as the penumbras described by *Griswold*, is comparable to Article 8 (Right to respect for private and family life) of the European Convention on Human Rights (ECHR). These rulings, as well as the precedents they relied on, made clear that the history of SCOTUS decisions dating back to *Pierce* and other decisions of that era considered privacy to be a fundamental right, with autonomy and bodily integrity entailed as part of the broader conception of privacy.

Beyond these SCOTUS decisions, it is important to note the role of the legislative branch in this discussion. All the rulings discussed above originated as a lawsuit against a particular state or agent of the state. Pierce and Casey were governors of their respective states, while Wade was the district attorney tasked with enforcing a state law. In each case, the plaintiff made the ultimately successful argument that these state laws violated fundamental rights and must be overturned by judicial review, preventing the states from enforcing these laws on the basis of due process. Legislatures, including the U.S. Congress, could codify these protections but they have generally chosen not to do so. Furthermore, at the time of this writing, the U.S. does not have a comprehensive federal privacy framework to provide a clear and cohesive definition of the nature and limits of privacy. This lack of action by the legislatures contributes to the reliance on due process as the basis for a right to privacy.

3. CONCEPTIONS OF PRIVACY

The nature and definition of privacy has long been debated. Richards (2022) begins his book by noting that it has become "customary at the beginning of a book about privacy to explain that…we lack a settled definition of what privacy is." Thomson (1975) argued that "the most striking thing about the right to privacy is that nobody seems to have any very clear idea what it is." Solove (2008) characterized privacy as a "concept in disarray." Despite its vague nature, *Olmstead v. U.S.* (1928) emphasized its importance by noting that the "makers of our Constitution" conferred "the right to be let alone – the most comprehensive of rights, and the right most valued by civilized men."

Although there is no single definition of privacy, there is a consistency that it involves control over the disclosure of certain types of information. Article 8 of the ECHR and Article 12 of the Universal Declaration of Human Rights both identify "correspondence," while Articles 7 and 8 of the European Charter of Fundamental Rights mention "communications" and "personal data" about a person. The Fifth Amendment to the U.S. Constitution protects against the disclosure of information through "unreasonable searches and seizures," which is nearly identical to Article 8 of the Canadian Charter of Rights and Freedoms prohibiting "unreasonable search or seizure."

However, privacy has long been recognized to involve more than just secrecy. Thomson (1975) characterized privacy as "a cluster of rights" that are primarily focused on "the right over the person" (i.e., the physical body) and rights concerning "owning property." Solove (2007) agreed that privacy was not a singular right but rather that it is "best used as a shorthand umbrella term for a related web of things," but noted that Thomson's limited focus on the person and property miss many important privacy invasions.

Others have emphasized that privacy and autonomy are intrinsically linked. Rachels (1975) emphasized that "there is a close connection between our ability to control who has access to us and information about us, and our ability to create and maintain different sorts of social relationships." In other words, the purpose of controlling information (secrecy) is about freely interacting with society (autonomy). Nissenbaum (2010) more explicitly links privacy with autonomy, as limiting access to information "contributes to material conditions for the development and exercise of autonomy and freedom in thought and action." Citron (2022) argues that intimate privacy "is a precondition to a life of meaning."

4. DOBBS AND DECISIONAL PRIVACY

The full history of the SCOTUS decisions beginning with *Boyd* and *Pierce*, coupled with the conceptions of privacy just discussed, clearly indicate that privacy is about more than just information secrecy. When the focus is on information secrecy, the emphasis is typically on a particular type of information. Private information commonly focuses on aspects of our lives that are sensitive and have the potential to cause distress if we are forced to disclose it against our will. Protecting this information is the foundation of building the connections and relationships that are fundamental to society. In addition, it allows us to live full and meaningful lives in an open and free society.

While this history suggests that there was a growing acceptance that privacy as a fundamental right involved more than just the ability to keep certain information secret, defining the boundary of privacy was still in flux. Many of the scholars discussed above argued that a right to privacy also involved the right to *act* on that private information without public disclosure. Decisions about family planning and medical care, both of which involve sensitive information, were considered to be best decided by the individual. Decisional privacy, allowing those involved to choose how to act without interference, provided a fundamental protection of dignity. Solove (2008) emphasized that decisional interference, like many other privacy harms, involves "invasions into realms where we believe that people should be free from the incursion of others."

The *Dobbs* decision marks a turning point in U.S. law regarding privacy. The trend in both scholarship and law had been toward broadening the concept of privacy toward a more expansive right beyond simply information privacy, and *Dobbs* has stopped that trend. Specifically, *Dobbs* declared that the *Roe* decision "conflated the right to shield information from disclosure and the right to make and implement important personal decisions without governmental interference." That is, *Dobbs* has declared that decisional privacy is not a fundamental right.

Although the immediate effect is on the legality of abortion in the U.S., this shift in respect for decisional privacy raises concerns about other rights. Some decisions, such as *Lawrence* and *Obergefell*, were decided based on *Roe* as a precedent. Others, such as *Loving*, relied on the *Griswold* precedent in a manner similar to *Roe*. *Dobbs* specifically did not overturn any of these other decisions, emphasizing that what "sharply distinguishes the abortion right" from others is abortion "destroys what those decisions [*Roe* and *Casey*] call 'potential life.'" However, in a concurring opinion, Justice Clarence Thomas goes farther by arguing that "'substantive due process' is an oxymoron that 'lack[s] any basis in the Constitution.'" He later goes on to say that "we should reconsider all of this Court's substantive due process precedents, including *Griswold*, *Lawrence*, and *Obergefell*." Thus, it is not clear

– and will not be clear until relevant challenges are brought before SCOTUS – what individual liberties may be protected or limited based on the concept of decisional privacy.

5. DISCUSSION

Since the *Dobbs* decision was released, a lot of the focus has been placed on certain controversial cases in which women (El-Bawab, 2023; Klibanoff, 2023) or girls (Wikipedia, n.d.) were denied abortions while facing horrifying circumstances. In the current environment, these stories rightly should be prioritized and nothing that we say here should take away from their plight. Acknowledging that, our focus is on an underlying issue that has the potential to impact more than just abortion rights. We emphasize that the rationale that the *Dobbs* decision used to justify overturning *Roe*, the criticism of the conflation of "the right to shield information from disclosure" with "the right to make and implement personal decisions," eliminates the concept of decisional privacy as a fundamental right.

This key point of contention reiterates the long-standing debate about the nature and boundaries of privacy. Specifically, this ruling raises the question of whether decisional privacy, a specific form of autonomy, fits within the confines of privacy itself. Or is it more consistent to remove decisional privacy from the privacy debate on the grounds that autonomy is a separate concept? Can the debate over privacy rights be clarified by limiting its scope to issues more clearly about the disclosure of information?

Solove (2008) argues that many of the key cases involving decisional privacy unavoidably link this implementation of personal decisions to other harms that are well established as privacy invasions. Enforcing laws such as those that led to the *Griswold*, *Roe*, and *Casey* decisions requires the forced disclosure of sensitive information related to health, the body, and sexual activity to third parties. This forced disclosure would clearly be considered an invasion of privacy by most, which stands in contrast to most limitations (legitimate or not) on autonomy.

The threat of forced disclosure is not hypothetical. Kaste (2022) reported on a case where Facebook had revealed a user's private data to law enforcement as part of an investigation for illegally "mishandling the fetal remains" that resulted from an abortion. Facebook (2022) revealed this information after "receiv[ing] valid legal warrants from local law enforcement," and these warrants were in relation to laws that predated and were unaffected by *Dobbs*. However, many states now have laws that allow for prosecution both for women who get abortions and others who assist them. This disclosure demonstrates that technology companies will disclose sensitive information to law enforcement without user consent.

It should be noted that, although this history and discussion has focused on the U.S. perspective, the full impact of the *Dobbs* decision will be international. Some have noted that this decision will shape how technology companies implement and maintain privacy (Federman, 2022; Krishnan et al., 2022; Privacy International, 2022; Sexton, 2022), how medical organizations protection patient information (Clayton et al., 2023; Henneberg, 2022), and how information gathered from technology companies will affect law procedures (Edelson, 2022; Kamin, 2023; Marathe, 2022; Stuart, 2023). The Internet is a global network, so the capabilities that technology companies build in the U.S. will impact the services and protections that they can provide in other parts of the world.

A specific concern about how technology companies will respond is how this change to the nature of privacy will shift what are perceived as reasonable defaults. With decisional privacy respected as a fundamental right, it would be easier to push for user protections during internal disputes over design choices. Without decisional privacy, there is a risk that such protections may become harder to justify, as companies will need to build in mechanisms to comply with law enforcement action. Beyond just

the concern of abortion rights in the U.S., the same mechanisms could be used to threaten privacy protections in countries with other restrictions, such as countries in which homosexuality is a crime. As such, the loss of decisional privacy as a fundamental right is likely to have ripple effects beyond the borders of the U.S.

Given this interplay between decisional privacy and information disclosure, there are strong reasons to consider decisional privacy as a form of privacy rather than a distinct autonomy right. Furthermore, treating decisional privacy solely as autonomy raises the question of the foundation for claiming a right to autonomy. Within the U.S. context, such a right to autonomy would likely rely on the same types of substantial due process arguments that were used to establish a right to privacy that included decisional privacy. Consequently, there does not seem to be a compelling argument that decisional privacy is solely a form of autonomy or that treating it as such strengthens either the right to autonomy or privacy.

6. CONCLUSION

In this paper, we have considered the how privacy has been defined and shaped within the U.S. legal context. Previous legal rulings throughout the 20th and 21st centuries had iteratively expanded the scope of privacy based on the concept of substantive due process. The *Dobbs* ruling reversed this trend by invalidating not just the right to abortion, but the foundation such a right was built on. While the nature and definition of privacy, particularly in relation to decisional privacy, was still in flux, *Dobbs* has injected a clear delineation: "the right to shield information from disclosure" and "the right to make and implement personal decisions" are distinct, and the latter is not protected as a fundamental right in the U.S. This delineation is likely to affect more parts of the debate over privacy and personal liberties beyond the singular issue of abortion. Although the definition of privacy still remains unsettled, SCOTUS has injected its justices opinions on the matter.

REFERENCES

Citron, D. K. (2022). *The Fight for Privacy*. Norton Books.

Clayton, E. W., Embi, P. J., & Malin, B. A. (2022). Dobbs and the future of health data privacy for patients and healthcare organizations. *Journal of the American Medical Informatics Association, 30(1)*, 155–160. https://doi.org/10.1093/jamia/ocac155

Edelson, J. (2022, September 22). Post-Dobbs, your private data will be used against you. *Bloomberg Law.* https://news.bloomberglaw.com/us-law-week/post-dobbs-your-private-data-will-be-used-against-you

El-Bawab, N. (2023, November 28). Lawsuit challenging Texas abortion bans appears before state Supreme Court. ABC News. https://abcnews.go.com/US/lawsuit-challenging-texas-abortion-bans-appears-state-supreme/story?id=105154571

Facebook. (2022, August 9). Correcting the record on Meta's involvement in Nebraska case. https://about.fb.com/news/2022/08/meta-response-nebraska-abortion-case/

Federman, H. (2022, September 29). Privacy and data protection in the wake of Dobbs. *Security*. https://www.securitymagazine.com/articles/98414-privacy-and-data-protection-in-the-wake-of-dobbs

Henneberg, C. (2023, June 5). The trade-offs for privacy in a post-Dobbs era. *Wired*. https://www.wired.com/story/the-trade-offs-for-privacy-in-a-post-dobbs-era/

Joh, E. E. (September 5, 2022). Dobbs online: Digital rights as abortion rights. In (Levendowski, A. & Jones, M. L. (eds.), *Feminist Cyberlaw,* forthcoming 2023. http://doi.org/10.2139/ssrn.4210754

Kamin, S. (2022, December 18). Katz and Dobbs: Imagining the Fourth Amendment without a right to privacy. *Texas Law Review*. https://texaslawreview.org/katz-and-dobbs-imagining-the-fourth-amendment-without-a-right-to-privacy/

Kaste, M. (2022, August 12). Nebraska cops used Facebook messages to investigate an alleged illegal abortion. NPR. https://www.npr.org/2022/08/12/1117092169/nebraska-cops-used-facebook-messages-to-investigate-an-alleged-illegal-abortion

Klibanoff, E. (2023, December 13). Kate Cox's case reveals how far Texas intends to go to enforce abortion laws. The Texas Tribune. https://www.texastribune.org/2023/12/13/texas-abortion-lawsuit/

Krishnan, A., Cohen, K., & Hackley, C. (2022, August 27). Digital privacy in the post-Dobbs. *The Regulatory Review*. https://www.theregreview.org/2022/08/27/saturday-seminar-digital-privacy-in-the-post-dobbs-landscape/

Marathe, I. (2022, July 1). Post-'Dobbs,' privacy attorneys prepare for increased data surveillance *Legaltech News*. https://www.law.com/legaltechnews/2022/06/27/post-dobbs-privacy-attorneys-prepare-for-increased-data-surveillance/?slreturn=20230512233641

Nissenbaum, H. (2010). *Privacy in Context*. Stanford Law Books.

Privacy International. (2022). Privacy and the body: Privacy International's response to the U.S. Supreme Court's attack on reproductive rights. *Privacy International*. https://privacyinternational.org/news-analysis/4938/privacy-and-body-privacy-internationals-response-us-supreme-courts-attack

Rachels, J. (1975). Why privacy is important. *Philosophy & Public Affairs, 4(4)*, 323–333. http://www.jstor.org/stable/2265077

Richards, N. (2022). *Why Privacy Matters*. Oxford Books.

Sexton, M. (2023, January 22). The new front in the battle for digital privacy post-Dobbs. *Third Way*. https://www.thirdway.org/memo/the-new-front-in-the-battle-for-digital-privacy-post-dobbs

Solove, D. J. (2007). "I've got nothing to hide" and other misunderstandings of privacy. 44 San Diego Law Review 745.

Stuart, A. H. (October 26, 2022). Privacy in discovery After Dobbs. *Virginia Journal of Law and Technology*. http://doi.org/10.2139/ssrn.4259508

Thomson, J. J. (1975). The right to privacy. *Philosophy & Public Affairs, 4(4)*, 295–314. http://www.jstor.org/stable/2265075

Warren, S. D. & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review, 4(5)*, 193–220. https://doi.org/10.2307/1321160

Wikipedia. (n.d.). 2022 Ohio child-rape and Indiana abortion case. Retrieved January 14, 2024 from https://en.wikipedia.org/wiki/2022_Ohio_child-rape_and_Indiana_abortion_case

USE AND ABUSE OF AI – ETHICAL PERSPECTIVES IN THE EDUCATIONAL SECTOR

Nuno Silva, Isabel Alvarez

Lusíada University, COMEGI (Portugal); ISTEC, COMEGI (Portugal)

nsas@lis.ulusiada.pt; alvarez@edu.ulusiada.pt

ABSTRACT

The advent of generative AI (Artificial Intelligence), as it is the example of exponential use of ChatGPT, sounded the alarms on management of schools and universities. It is essential to ensure that the use of AI in education is well-designed, transparent, and aligned with educational goals and values. Therefore, there are a lot of issues from plagiarism to equity of assessment that are under ethical analysis. The authors, as lecturers and reflective practitioners, explored from an interdisciplinary perspective the ways how this technology is guiding students to the final objectives of educational programme. Giving support to the data analysis and conclusions, we started by doing a survey to the students on their actual use and opinion of ChatGPT and also to pose a set of several ethical questions to ChatGPT regarding profile and personal autonomy, cultural sensitivity, discriminatory social biases, accountability, and cybersecurity, whose answers were intriguing. Finally, it is important to acknowledge that AI is not capable of experiencing emotions or offering lecturer-like opinions or judgments. While provide information and guidance on a wide range of topics and cannot offer personal beliefs or biases, are not always perfect and may sometimes be limited by the quality and relevance.

KEYWORDS: Higher Education, Artificial Intelligence, ChatGPT.

1. INTRODUCTION

This paper discusses the impact and the potential implications of the generative Artificial Intelligence language model, ChatGPT (where GPT stands for Generative Pre-Trained Transformer) in higher education and learning. Some educators think that this popular bot can alter teaching while others worry that it may have the opposite effect on their students' motivation to learn. After the pandemic times of 2020, the emergence of such a sophisticated new artificial intelligence with the ability to write seemingly anything—tweets, poems, essays, and even computer programs—all with a simple prompt was far from most of our minds.

Some educators have already tested the ChatGPT 3.5 (OpenAI, 2023), ability to generate convincing versions of responses to essay questions and even publishable academic papers. Others believe students may benefit from understanding the ins and outs of how this technology works, and might use it as a tool to explore the possibilities and limits of online sources of information.

Though, apparently, there are benefits of this new technology, a lot of caution is required for its use.

1.1. The purpose of the research

The purpose of this research is to highlight the emerging concerns in the field of academic management in the face of artificial intelligence, namely ChatGPT, and framing the role of the teacher in the analysis of this problem from an ethical point of view. It also considers the freedom or limitation of use by students and the care to be taken in written and/or oral assessment methods.

1.2. Scope and Objective

This research raises questions about whether the skills traditionally thought of as essential for research and learning, such as summarizing complex texts and writing essays and critical thinking, can be replaced by machines. Some institutions decided to ban Artificial Intelligence (AI) powered writing technology due to educators being worried about a new kind of high-tech plagiarism. However students need to learn how to work with AI for their future careers. On the other hand, should teachers use AI for grading?

Some authors explored the potential use and abuse of ChatGPT in the educational sector. Qadir (2022) answered the question of "what are the potential pitfalls in schools and education regarding the student use and abuse of ChatGPT?" exploring the issues of plagiarism, overreliance, misinformation, and privacy concerns. Çakmakoğlu (2023) questioning if "can student use and abuse of ChatGPT be prevented?" suggested that it is impossible to completely prevent student use and abuse of ChatGPT, but some controlling measures could include guidelines around plagiarism, cheating, and academic misconduct. Moreover, encourage collaborative learning can reduce the incentive to misuse ChatGPT and promote academic integrity. Finally, Deshpande and Szefer (2023) remember that only recently ChatGPT has gathered attention from the public and one potential use, or abuse, of ChatGPT is in answering various questions or even generating whole essays and research papers in an academic or classroom setting. Panic has also hit in the classroom and teachers are concerned about the students' use and abuse of ChatGPT!

2. LITERATURE REVIEW

2.1. Artificial Intelligence in Education

Students and educators have long been affected by technological changes. Digital technologies play a powerful role in today's classrooms.

Studies on Artificial Intelligence (AI) in education focused mainly on using these sort of technologies to enhance learners' abilities in memorizing, comprehending, applying, analysing, and assessing, with the utmost educational objective to the highest cognitive level, which is creativity (Hwang & Chen, 2023). Several publishers have recently introduced new policies in response to the growing use of Generative Artificial Intelligence applications by authors of academic publications, developing a new AI author policy to provide guidance to authors, readers, reviewers, and editors (Hwang & Chen, 2023).

The new education: Information that was once dispensed in the classroom is now everywhere: first online, then in chatbots. Teachers now play a role of facilitators as they must now show students not only how to find it, but what information is to be trusted or not, and how to tell the difference. Bringing interactivity into the classroom is one of the best uses of technology. In fact teaching methods that get students to be creative or to think critically lead to a deeper kind of learning. For example, ChatGPT can play the role of a debate opponent and generate counterarguments by exposing students to an endless supply of opposing viewpoints, helping them to look for weak points in their own thinking (Will, 2023). Deep thought must be done on how might schools and educators take a more proactive approach towards this new technology. This field is constantly evolving, and new authors and researchers are continuing to make important contributions to this area of study. Beverly Park Woolf research (Woolf et al, 2013) focuses on the use of AI in education, with a particular emphasis on intelligent tutoring systems. Probably the best way will be to schools to start encouraging students critical thinking about what technology can help and what it hinders us from doing instead of just teaching how to use technology (Woolf et al, 2013).

Deshpande and Szefer (2023) explored how well ChatGPT can do in a computer engineering course, found that using ChatGPT correct solutions to quiz questions can often be generated, while solutions to homework questions were much less accurate.

Silva and Alvarez (2023) suggested that the reality is that AI tools are yet in their initial state far from perfect. Out of the darkness of disruption, ensuring a balance between the benefits that AI offers, needs to be properly regulated and learners properly trained, then it would be used responsibly. Therefore, critical thinking perspectives and audit AI tools should avoid losing the human aspect that is crucial in education.

2.2. ChatGPT

The education community has been concerned with the rise of ChatGPT 3.5 (OpenAI, 2023), an artificial intelligence tool that can write anything with just a simple prompt. Most of the conversation and concerns have been centred on the extent to which students will use the chat bot—but ChatGPT could also fundamentally change the nature of teachers' jobs.

The use of this bot can generate detailed responses to questions related to several subjects hardly distinguishable from those created by humans, which on one side is impressive but on the other side this potential is also very concerning and worrying that could lead to serious problems in education and social security (Yang et al., 2021).

So far, teachers in some institutions are considering using the ChatGPT to plan lessons, offer students feedback on assignments, and execute some administrative tasks like sending emails or write letters of recommendation.

But the technology of ChatGPT is not foolproof. Some teachers published that they noticed a factual error when they experimented asking the bot to plan a lesson for an early chapter on a certain subject. The tool also demonstrated that has limited knowledge of world events that happened after 2021(Will, 2023). ChatGPT can also offer feedback on student work. Other situations have also been reported by teachers, saying that the examples of grading from the chat bot feel shallow or even inaccurate. It was also published that, while the technology might get it right nine times out of 10, there's always the risk that it won't grade one student's work correctly, so teachers would still need to personally review each piece of feedback (Will, 2023). Some schools worldwide have decided to ban the use of this bot and issued statements that warned students against using ChatGPT to cheat. And as some authors say, while the tool may be able to provide quick and easy answers to questions, it does not build critical-thinking and problem-solving skills, which are essential for academic and lifelong success (Will, 2023).

This paper examines the opportunities and challenges of using ChatGPT in higher education, and discusses the potential risks and rewards of these tools and suggests strategies that universities can adopt to ensure ethical and responsible use of these tools. These strategies include developing policies and procedures, providing training and support, and using various methods to detect and prevent cheating. Based on our major findings, we conclude that the use of AI in higher education presents both good opportunities and also challenges that need to be addressed by taking a proactive and ethical approach to the use of AI in education (Cotton et al., 2023), namely if it is genuinely useful in supporting teaching and learning (Kousa & Niemi, 2023).

There are some opinions that the threats to education in this context is based on a lack of deep understanding and difficulty in evaluating the quality of responses, threatening academic integrity, democratising plagiarism and declining high-order cognitive skills (Farrokhnia et al., 2023).

2.3. The Ethics of ChatGPT in Education

While there is much generic literature on ethics in artificial intelligence, there is a clear gap in studies on the ethics of ChatGPT in the education sector. We systemically explore what exists and address what does not exist. For Pedró et al. (2019), the major challenges are related to personalisation, inclusion and equity, powered education, quality, and transparency. The issues of equity and personalisation are detailed by Chine et al. (2022), namely in the case of experience learning gaps due to a lack of access or economic disadvantages. On the other hand, Jiang and Pardos (2021), gives special attention to fairness and bias in artificial intelligence and graduation prediction. Regarding specifically ChatGPT (Debby, 2023), it opens new difficulties of detecting and preventing academic dishonesty. An update of plagiarism detection tools and controlling cheat proctoring tools is absolutely necessary. The output from ChatGPT does not include proper referencing, while academic writing is expected to accurately include citations and references. The ChatGPT has raised security and privacy issues, namely because there is no minimum age requirement to use ChatGPT. Also, it is not clear that personal data analysis is done in respect to GPDR (EU General Data Protection Regulation). In addition, copyright law generally applies to original works of authorship created by human authors, and not to works created by machines or algorithms.

Artificial intelligence does not have the ability to receive congratulations or experience emotions in the same way that humans do. Yet, doesn't have feelings so doesn’t experience tiredness. However, like all technologies, smart digital devices bring unintended, collateral, and disproportionate effects. It is known that smartphones and laptops provide students with access to communication and information, but they also contribute to distraction, social conflict, bullying, and many other problems. Even more, emerging smart small wearable devices like smartwatches and hearables. How should educators respond when problems like these inevitably occur? (Krutka, Pleasants & Nichols, 2023).

3. METHODOLOGY AND METHODS USED

3.1. Overview

As lecturers of higher education, our methodological approach is mainly to engage in reflective practice concerning the adoption and use of artificial intelligence, exploring the case of ChatGPT in the Portuguese Universities contexts. The methods used are based on our daily experience observing students and institutions academic activities, qualitative interviews and discussion boards.

The reflective practitioner approach is a methodological framework that emphasizes the importance of reflection and self-evaluation in the learning and teaching process. As lecturers in higher education, there are several ways that we can apply the reflective practitioner approach in our teaching practice like reflecting on our teaching practice, seed feedback from students, engage in professional development, collaborate with colleagues or maintain a reflective journal promoting critical thinking.

Reflective practitioners through double-loop learning will expand the analytical framework and challenge their assumptions to promote self-study (professional understanding and personal growth). The authors’ decision is to promote an explicit and reflective process concerning daily events, as well as reporting them in multiple ways and follow the methodological principles for interpretive research.

3.2. The survey

Following factors that make responding to the form convenient for both the respondent and the interviewer, Silva and Alvarez (2023) conducted a physical survey between lecturers and learners.

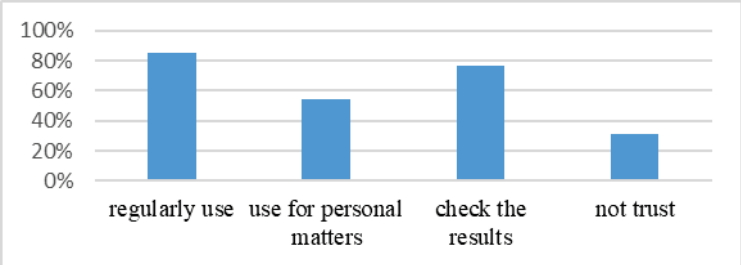
The survey aimed to understand the reality in Portuguese higher education regarding the use of ChatGPT in their academic activities. By analysing the results, it is intended to better understand the challenges presented to Universities regarding its use, as well as the degree of satisfaction of the performance of this type of technology in their study activities. Due to the fact that this was a physical survey in the classrooms, 100% was the level of participation. The survey had eight questions, asking (1) if the learner used ChatGPT 3.5, (2) if so for which purpose (private or academic work), (3) if the results obtained were confronted with other sources and (4) if they were up to date, (5) if the use of ChatGPT 3.5 was useful for their academic work, (6) if sources were mentioned, and finally asking (7) if in their opinion this new technology will alter the traditional way of teaching and (8) how should teaching be changed to cope with these new tools.

When asked about how teaching can change, with artificial intelligence being one of the parts of this change, several interpretations and opinions were transmitted. In general, it was stated that teaching should not block the use of these tools, but rather ally with them, taking to the classes other types of challenges and more complex paradigms allowing learners themselves to explore the tools and compare the results. It is also a general opinion that this type of tools will interfere with the role of the lecturer as it should be readjusted to this new reality, thus allowing learners guided by artificial intelligence, where the program will be defined by lecturers, to have individualised classes and adapted to their real needs, thus raising the quality of teaching where, given the increasing size of the classes, personalisation and individualisation are increasingly smaller (Silva & Alvarez, 2023).

Although half of the respondents using ChatGPT claim that the results obtained were correct, the other fifty per cent said that they confronted the bot and that it did not always correct the results. On the other hand, they claim that using this tool makes it easier to get the answers they want. If they have a question about something, they can just ask the AI directly, there is no need to search through search engines that only return links whose sites must be searched and their information summarised, and it also allows them to complete data that the materials provided in the discipline do not include.

Analysing the results of the data obtained [Figure 1], it appears that 85% of the respondents regularly use ChatGPT 3.5 to obtain information for professional and/or academic purposes, of which 54% also use it for personal matters. It was observed that 77% of the respondents usually check the results obtained with other sources of information, as they found differences in the information obtained and even 31% do not trust that information. Many respondents also stated that depending on the way the question is asked, the result may differ.

Figure 1. Learners' ChatGPT use.



Source: Silva & Alvarez, 2023.

3.3. Direct questionnaire to ChatGPT

Following the results obtained in the physical survey, the authors decided to pose some ethical questions to Chat GPT related to the topics of this research and observe the responses obtained.

We highlight some questions done, whose answers were intriguing and are summarised here:

Q1: What is the minimal age to use ChatGPT?

A1: There is no minimum age requirement to use ChatGPT. It is important for users to understand that it may not always provide accurate or complete information. Additionally, some content discussed by ChatGPT may not be appropriate for all ages, so parental or guardian guidance is recommended.

Q2: Can ChatGPT ensure equity in educational sector?

A2: Promoting equity in education requires a multifaceted approach that addresses systemic inequalities, provides equal access to resources, and creates an inclusive learning environment for all students. While I (text of ChatGPT, 2023) cannot ensure equity in the educational sector, I can provide information and insights that can help promote equity and fairness in education.

Q3: How can lectures prevent ChatGPT cheating?

A3: As AI language models become more advanced, it may become more difficult to detect instances of cheating. Therefore, it is important to continually evaluate and update assessment strategies to ensure that they are effective in promoting academic integrity. There are several strategies that instructors can use to prevent this from happening. For example, design assessments that require critical thinking and synthesis: assignments and exams that require students to analyse and synthesize information are less likely to be susceptible to cheating with AI language models. Rather than asking students to simply regurgitate information, instructors can design assessments that require students to apply what they have learned in new and creative ways. As an AI language model, I am not capable of cheating, as I am a machine and do not have the ability to act with intent or deception. Therefore, I cannot cheat or act dishonestly.

Q4: Can ChatGPT avoid plagiarism?

A4: As an AI language model, I generate responses based on the input I receive and the algorithms that have been developed to create me. While I am capable of producing original and unique content, I do not have the ability to recognize or avoid plagiarism on my own. Always cite the sources of any information used in research and writing, including any content generated by AI language models.

Q5: What about ChatGPT risks concerning cybersecurity?

A5: As an AI language model, ChatGPT presents some potential risks concerning cybersecurity, primarily related to the confidentiality and integrity of information shared during interactions with the model.

Q6: Is ChatGPT culturally sensitivity?

A6: ChatGPT may not be aware of all cultural nuances or differences, and may not be able to fully understand or interpret cultural references or idioms. This could potentially lead to misunderstandings or miscommunications during interactions with individuals from different cultural backgrounds.

Q7: What are the accountability of ChatGPT?

A7: The accountability of ChatGPT depends on the actions and decisions of the individuals or organizations that develop, deploy, and use the model. By prioritizing transparency, privacy, fairness, quality, and accountability, ChatGPT can be used in a responsible and ethical manner that benefits users and society as a whole.

Q8: Have ChatGPT discriminatory social biases?

A8: By actively working to address potential biases and promote inclusivity, ChatGPT can be used in a way that is respectful, equitable, and beneficial for all users. Studies have shown that AI language models can exhibit biases related to factors such as race, gender, and ethnicity.

Q9: Is ChatGPT accurate regarding profile and personal autonomy?

A9: When using ChatGPT for information or guidance on topics related to profile and personal autonomy, it is important to approach the conversation critically and to seek out multiple sources of information to validate and contextualize the information provided by the model.

Q10: Is ChatGPT limited by the quality and relevance?

A10: Yes, ChatGPT is limited by the quality and relevance of the training data used to develop the model. It is important to note that ChatGPT is not a substitute for human expertise or judgment.

Q11: Is ChatGPT aligned with educational goals and values?

A11: In order to ensure that ChatGPT's responses are aligned with educational goals and values, it is important to carefully consider the specific educational context in which the model will be used, as well as the topics and questions that will be presented to the model. It may also be necessary to supplement ChatGPT's responses with additional information or insights from other sources, such as human experts or educational materials.

Q12: What are ethical issues using ChatGPT in higher education?

A12: As an AI language model, ChatGPT has the potential to be used in various ways in higher education, including assisting in research, generating content for courses, providing personalized learning experiences, and facilitating communication between students and teachers. However, there are several ethical issues to consider when using ChatGPT in higher education, including:

Bias and Discrimination: ChatGPT has the potential to replicate biases and discriminatory behaviour of its training data, which could result in unfair treatment of students based on their race, gender, sexuality, religion, or other factors.

Privacy and Data Security: ChatGPT collects and stores user data, including personal information, which could be misused or accessed by unauthorized individuals or entities.

Intellectual Property Rights: If ChatGPT is used to generate content for courses or research, there may be intellectual property rights issues related to the ownership and use of the generated materials.

Dependence on Technology: Overreliance on ChatGPT could lead to a reduction in critical thinking and analytical skills among students, who may rely solely on the model for information and answers.

Ethical Use of AI: Institutions must ensure that ChatGPT is used ethically and transparently, with clear guidelines for its use and appropriate governance mechanisms in place.

To mitigate these ethical issues, institutions must establish clear policies and guidelines for the use of ChatGPT, including guidelines for data privacy, bias and discrimination, intellectual property rights, and appropriate use of AI. Additionally, institutions must ensure that students and staff are adequately trained to use ChatGPT responsibly and understand its limitations.

Q13: Is the use and abuse of ChatGPT changing ethical perspectives in the educational sector?

A13: The use of ChatGPT and other artificial intelligence technologies in the educational sector has raised ethical concerns, but it is not clear if it is changing ethical perspectives. The ethical concerns related to the use of ChatGPT in education include issues of privacy, bias, and transparency. For example, some may argue that the use of ChatGPT to grade student work may be biased against certain students or may not accurately assess a student's understanding of a subject. Similarly, the use of ChatGPT to provide personalized learning may raise concerns about student privacy and the collection and use of their data.

4. MAJOR FINDINGS

We have observed in our Institution great enthusiasm of students in the use of ChatGPT, but often without being able to justify the benefit they take for the educational purpose that is proposed to them. These actions seem to have as main objective the facilitation in the execution of autonomous work and without direct control of the teacher in the execution phase, which somehow escapes the control of plagiarism. On the other hand, there are attempts to answer assessment questions in written exams (cheating) using mobile devices (mobile phones, smartwatches, etc.) which are difficult to control by the teachers' surveillance systems.

It is up to educators, policymakers, and the wider community to determine the appropriate use of AI technologies in education and ensure that they are used in an ethical and responsible manner. The control on plagiarism in autonomous work makes it necessary to reinforce the oral assessment.

Finally, we make a comparison between the ChatGPT responses and the authors' literature review made in point 2. (above). The ChatGPT identified all the points that the authors referenced by us had addressed. However, the way the text is presented by the ChatGPT does not reflect a critical and comparative construction between the approaches of these authors. In-depth learning and the development of techniques to take advantage of ChatGPT responses is therefore necessary, especially given the time that is spent on this approach. Nevertheless, we observed that in a spontaneous evaluation a teaching colleague commented: it took me five years to do my thesis, and now this gives me the answers in milliseconds!

The use and abuse of ChatGPT 3.5 (OpenAI, 2023) is not yet verified in the classroom environment in the examples explored by this research (Portuguese higher education). There are no regulatory recommendations or guidelines yet applied. As an AI language model, ChatGPT is a machine learning model trained on large datasets of text. The model itself is not copyrighted, but the content generated by ChatGPT may be subject to copyright laws.

5. CONCLUSION

First of all, it is also important to note that the use of AI language models can be a valuable learning tool when used appropriately. They can be used to supplement learning, assist with research and writing, and provide students with additional resources to enhance their understanding of the subject matter. Be aware of the limitations of AI language models and the potential for bias or inaccuracies in the content generated by these tools (text of ChatGPT, 2023).

The use of ChatGPT in education can bring many potential benefits, such as personalized learning, better feedback, and enhanced student engagement. However, it is important that this is used in a responsible and ethical manner that respects the privacy and well-being of students, as well as the principles of good teaching. ChatGPT is not designed to address issues related to accountability and cybersecurity directly. The alarm generated by news and evidence reported on the potential of ChatGPT forced an ethical reflection in practical context that the authors as lecturers framed in the education sector.

There are problems related to equity and autonomy granted to students, especially the ability that ChatGPT gives them to do practical work in an assertive way and in a short period of time. Plagiarism is a major concern and this involves not only the work of the lecturer but also academic regulations. While AI language models cannot avoid plagiarism on their own, students should take steps to ensure that any content produced using these tools is properly cited and attributed to its original source. Ultimately, the extent to which ChatGPT is aligned with educational goals and values will depend on

how the model is used and the degree of care taken to ensure that its responses are accurate, relevant, and appropriate for the educational context in question

Furthermore, while the use and abuse of ChatGPT in education have raised ethical concerns, it is uncertain if it is changing ethical perspectives in the educational sector.

ACKNOWLEDGEMENTS

This work is supported by national funding's of FCT - Fundação para a Ciência e a Tecnologia, I.P., in the project «UIDB/04005/2020».

REFERENCES

Çakmakoğlu, E. E. (2023). The place of ChatGPT in the future of dental education. Journal of Clinical Trials and Experimental Investigations, 2(3),121-129. https://doi.org/10.5281/zenodo.8210063

Chine, D., Brentley, C., Thomas-Browne, C., Richey, J., Gul, A., Carvalho, P., Branstetter, L., & Koedinger, K. (2022). Educational equity through combined human-AI personalization: A propensity matching evaluation. In *International Conference on Artificial Intelligence in Education*. pp. 366-377. Springer, Cham.

Cotton, D., Cotton, P. & Shipway, J. (2023). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education & Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Deshpande, S., & Szefer, J. (2023, Aprin 14). Analyzing ChatGPT's Aptitude in an Introductory Computer Engineering Course. arXiv:2304.06122v2

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023), A SWOT Analysis of ChatGPT: Implication for Educational Practice and Research, *Innovations in Education & Teaching International*. https://doi.org/10.1080/14703297.2023.2195846

Hwang, G.-J., & Chen, N.-S. (2023). Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions. *Educational Technology & Society*, *26*(2). https://doi.org/10.30191/ETS.202304_26(2).0014

Jiang, W. & Pardos, Z. A. (2021). Towards Equity and Algorithmic Fairness in Student Grade Prediction. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 608–617. https://doi.org/10.1145/3461702.3462623

Kousa, P., & Niemi, H. (2023). Artificial Intelligence Ethics from the Perspective of Educational Technology Companies and Schools. In H. Niemi, R. D. Pea, & Y. Lu (Eds.), *AI in Learning: Designing the Future*. Springer, Cham. https://doi.org/10.1007/978-3-031-09687-7_17

Krutka, D. G., Pleasants, J., & Nichols, T. P. (2023). Talking the Technology Talk. *Phi Delta Kappan*, *104*(7), 42-46.

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Pedró, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: challenges and opportunities for sustainable development*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000366994

Qadir, J. (2022). Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. TechRxiv. https://doi.org/10.36227/techrxiv.21789434.v1

Silva, N., & Alvarez, I. (2023). AI in higher education: utopia or reality? In E. Smyrnova-Trybulska (Ed.), E-learning & Artificial Intelligence (AI), E-learning, 15, Katowice–Cieszyn, pp. 57–67. https://doi.org/10.34916/el.2023.15.05

Will, M. (2023, January 11). With ChatGPT, Teachers Can Plan Lessons, Write Emails, and More. What's the Catch?. *Education Week*. https://www.edweek.org/technology/with-chatgpt-teachers-can-plan-lessons-write-emails-and-more-whats-the-catch/2023/01

Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2013). AI Grand Challenges for Education. *AI Magazine*, *34*(4), 66-84. https://doi.org/10.1609/aimag.v34i4.2490

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2*, 100008. https://doi.org/10.1016/j.caeai.2021.100008

AN ANALYSIS ON AI ETHICAL ASPECTS FROM A STAKEHOLDER'S PERSPECTIVE

Sofia Segkouli, Maria Tsourma, Pinelopi Troullinou, Paola Fratantoni, Dimitris Kyriazanos, Anastasios Drosou, Dimitrios Tzovaras

Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki (Greece), Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki (Greece), Trilateral Research, Dublin (Ireland) Zanasi & Partners, Modena (Italy) National Center for Scientific Research "Demokritos", Agia Paraskevi Athens (Greece), Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki (Greece), Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki (Greece)

sofia@iti.gr, mtsourma@iti.gr, pinelopi.troullinou@trilateralresearch.com, paola.fratantoni@zanasi-alessandro.eu, dkyri@iit.demokritos.gr, drosou@iti.gr, dimitrios.tzovaras@iti.gr

ABSTRACT

Artificial intelligence (AI) is developing at a rapid pace, raising critical ethical questions and risks that affect many facets of society. During these years, many attempts have been made to design ethical frameworks that could be used for the assessment of an AI application, however their application is not successful in all cases and requires additional information. This work aims to understand the perspective of the stakeholders participating in the design, implementation and adoption phase of an AI application, collect their issues and recommendations and highlight the achievements and the gaps so far in respect to ethical guidelines' implementation. Understanding the wide spectrum of stakeholders engaged in AI research and applications, this study attempts to clarify their various ethical questions, concerns and points of view. Through an analysis of current ethical frameworks and the presentation of stakeholder's opinions, the paper provides also guiding principles for the deployment or adoption by the public sector or by an organization of an AI-based solution on an ethical basis. This study adds to the current discussion on AI ethics by offering a sophisticated analysis of ethical and security controversies from conceptual and practical lens the diverse perspectives of individuals affected by AI technologies.

KEYWORDS: AI technologies, ethical and security controversies, inclusive and user-friendly AI, cultural and social diversity, ICT professionals, Law Enforcement Agencies (LEAs).

1. INTRODUCTION

Artificial Intelligence (AI) technologies rapid development and use stimulate vigorous discussion about their potential, in different contexts and uses (Carvalho et al, 2022). The application of AI by various information and communication technology (ICT) professions provides several benefits (Eurostat, 2022). More specifically, varied viewpoints lead to more robust AI systems (Liu et al, 2022). When people from various backgrounds cooperate, they contribute unique ideas and experiences that can improve AI technology development and implementation. Second, the diversity of ICT experts means that AI systems are intended to appeal to a broad spectrum of users, supporting their diverse wants and preferences. As a result, AI applications become more inclusive and user-friendly (Meyer & Henke, 2023). However, there is limited research on how to provide AI technology in a more ethically aligned way with social ideals and promote justice, transparency, accountability, and inclusion by including various viewpoints. In this study, the successful case study of popAI project[1], A European Positive Sum

[1] https://www.pop-ai.eu/

Approach towards AI tools in support of Law Enforcement and safeguarding privacy and fundamental rights, targeted at developing pathways on the basis of ethical views of ICT and academic professionals for AI design and development.

To this direction, different background and expertise has been gathered in order to reach a broader and more comprehensive understanding of issues and concerns related to the use of AI-based technologies in the security domain. Moreover, in order to conceptualize in depth ethical and security controversies, diverse societal sectors have been engaged in the light of taking into consideration all the voices and perspectives when developing ethical and legal oriented policies.

A controversial issue nowadays raises the question of whether the diversity of information and communication technology (ICT) professionals can raise ethical concerns. From an initial point of view, this inclusion may more effectively recognize and resolve biases and discriminatory behaviours in AI systems. Diversity gives a variety of viewpoints and experiences, which aids in the identification and mitigation of biases in AI systems. The hidden risk of developing discriminatory or unjust AI algorithms that disproportionately could affect specific groups of individuals can be mitigated by incorporating experts from diverse professional backgrounds. Disparate viewpoints also improve decision-making and encourage the inclusiveness of AI systems that respond to a larger spectrum of users' requirements.

Furthermore, understanding the socioeconomic and cultural settings in which these technologies are employed is required for unfolding ethics related considerations in AI research. Diverse ICT experts contribute a plethora of cultural, social, and ethical understanding that may be used to inform AI system design and deployment. This guarantees that AI technology is consistent with local values, conventions, and legal frameworks, preventing ethical conflicts or harm. In this context, the use of AI by diverse professionals fosters transparency and accountability, because different viewpoints and expertise contribute to making AI systems explainable, auditable, and subject to critical examination. (Felzmann et. al 2020)

On the other hand, the ethical views on the use of AI by diverse ICT professionals can vary based on individual perspectives and cultural backgrounds. In terms of popAI project’s case study systematic surveys have been conducted to unlock specific ethical issues and concerns both at local level to identify methods and strategies of single countries but also compare different perceptions and feelings of similar topics. To this end LEAs (Law Enforcement Agencies) have been engaged along with relevant experts through policy labs.

In particular, academics and practitioners experience from different lenses and perspectives the potential implications of adopting AI-based technologies. One concern is the potential for biased outcomes in AI systems, as stated before. If professionals do not address biases in training data or fail to account for diverse perspectives during system development, AI can perpetuate existing societal biases and discrimination. Another concern is privacy and data protection. With diverse ICT professionals working on AI, there is a need to ensure that personal data is handled responsibly, and individuals' privacy rights are respected. Moreover, accountability and transparency are essential ethical considerations. Diverse professionals must be diligent in making AI systems explainable and auditable to avoid potential negative consequences. Additionally, there is a concern about the impact of AI on employment. (Dwivedi et al., 2021)

Professionals need to consider the potential displacement of jobs and work towards minimizing adverse effects on individuals and communities. Lastly, there is the broader ethical concern of power and control. AI technology should not concentrate power in the hands of a few or reinforce existing inequalities. By acknowledging and addressing these ethical concerns, diverse ICT professionals can

work towards developing AI systems that align with societal values, promote fairness, and have a positive impact on individuals and communities. (Veale, 2020)

Identifying different views, theories and perceptions in the AI ethics discussion could improve the potential of AI technologies on a global scale in a multidisciplinary social, cultural, political and ethical manner. In cognate literature, a number of research initiatives and academic endeavours targeted to identify unacceptable risks and prohibited AI practices. The challenging point is which categories of high-risk AI systems have been elaborated so far, what redress mechanisms are revoked and the opening issues by diverse fields, sectors and environments.

Given the main motivation for AI's use and its relevant applications, which is the economic benefits and sustainability for different sectors such as education, healthcare, business management and agriculture, it is of great importance to (a) review the perceptions of diverse actors and environments in AI world and (b) stress the achievements and the gaps so far in respect to ethical guidelines' implementation. The present work, therefore, reviews relevant initiatives such as the ETAPAS[2] and NOTIONES[3] projects, that attempted to converge different contexts on this topic in order to acquire sufficient evidence for effective mechanisms, strategies and policies. In addition to this, and in order to prepare a more consolidated work, interviews with companies including diverse ICT professionals in the use and implementation of AI-based solutions have been conducted. These interviews aim to discuss the ethical issues that might be raised during these processes, and how they are handled.

The present work highlights also the dynamics and interactions that could be deployed between diverse AI actors and stakeholders and investigates if there is balance and complete consideration of AI ethics in a horizontal way. AI technology can be influenced by those "who build it and the data that feeds it" (Kim, 2017). Therefore, the role of context, education and culture could be reflected in AI development and use and vice versa. Upon this, among the considerations of the present work is how sustainability in education and training programs of ICT professionals can be achieved and which pathways and what kind of effort and individual involvement are required to meet AI challenges.

This attempt for limited risks in AI systems is currently happening through the Artificial Intelligence Act voted recently. The Commission proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored to a 'risk-based approach (Madiega, 2021).

The remainder of the paper is structured as follows. The second section presents the related work available in literature, upon this topic. The third section presents the methodology followed for the stakeholders' feedback collection, while section 4 presents the risks identified and mentioned by the stakeholders participated in the organised events. Section 5 presents the analysis of the collected information. Based on the analysis presented, section 6 describes the recommendations provided by stakeholders on how the application of an ethical framework can be facilitated. Finally, section 7 describes the conclusions resulted from the collected information.

2. RELATED WORK

As AI systems become more advanced, insights regarding transparency, accountability, bias, and the social effect of these technologies must be thoroughly addressed. This section examines current literature and research on understanding and mitigating ethical challenges in AI, with the goal of contributing to a more rigorous and nuanced knowledge of the ethical environment in which AI

[2] https://www.etapasproject.eu/

[3] https://www.notiones.eu/

functions. This investigation is important not just for the responsible development and deployment of AI systems, but also for encouraging public confidence and ensuring that the technology is consistent with human values.

Several research studies (Eitel-Porter, 2021; Floridi et al., 2021, Ashok et al., 2022) present the need of applying ethical frameworks in the design and use of AI applications and assess them in terms of principles and risks in this domain. One of these researches is a practical approach presented by Felzmann et al (2020), that addresses the challenges and complexities associated with transparency in automated decision-making (ADM) environments, in the context of AI. With the increasing prominence of AI systems making automated and self-learned decisions, demands for transparency in decision processes have emerged in academic and policy discussions. This article acknowledges the multidimensional nature of openness and its multiple promises, which frequently run into challenges in practice. To close the gap between the normative ideal of transparency and its actual execution, the authors examine transparency difficulties and draw lessons from the creation of "Privacy by Design". The outcome is a set of nine principles that support the notion of Transparency by Design and are intended to guide organisations in creating transparent AI systems.

Another research that discusses about the ethical principles of Human-centered AI (HCAI) is presented by Shneiderman (2020). The aim of this paper is to provide practical steps for effective governance in the development and implementation of HCAI systems. The author proposes 15 recommendations distributed across three levels of governance: team, organization, and industry, designed to enhance the reliability, safety, and trustworthiness of HCAI systems. At the team level, the emphasis is on dependable systems based on strong software engineering techniques, fostering a safety culture through business management tactics, and ensuring trustworthiness through independent oversight. Leadership commitment, safety-oriented recruiting and training, robust reporting mechanisms, internal review boards, and conformity with industry standards all contribute to organisational safety culture. The suggested governance framework's main purpose is to minimise the risks and maximise the advantages of HCAI for people, organisations, and society as a whole.

Sanderson et al (2023) performed an interview study to identify distinct requirements, restrictions, and aims of the varied spectrum of projects in which participants have been involved providing useful insights into the challenges of creating and building responsible, or ethical, AI systems. In the aforementioned research an analysis is conducted about the techniques and experiences of researchers and engineers from Australia's national scientific research organisation (CSIRO), involved in the design and development of AI systems. The results of the research highlighted a noteworthy issue related with the management of inherent costs and conflicts in applying actions to achieve privacy and security, transparency and explainability, and accuracy in the designed AI systems.

On the same context, Griffin et al (2023) presented a study that addresses a notable gap in the discourse on the ethics of AI, ML, and data science by focusing on the ethical agency of developers. Through semi-structured interviews with 40 developers, the research identifies more than 20 issues, with a specific focus on three: ethics in the occupational ecosystem, developer ethical agency, and the characteristics of an ethical developer. The findings expose significant disparities between developers' self-perceptions and the reality of their work experiences, highlighting variations in their ethical agency. While developers possess some authority to intervene for ethical reasons in the systems they work on, they often underestimate the extent of their ethical decision-making. Nevertheless, the study identifies a growing ethical wisdom within the developer community, emphasizing the importance of recognizing and nurturing this emerging ethical consciousness through engaging with developers.

Regarding case studies presenting the application of ethical design in applications, Shilton and Greene (2019) have written a paper, presenting the case study that addresses the essential issue of privacy within the framework of corporate social responsibility in the mobile device ecosystem. The study examines the ethical issues involved in collecting, retaining, and exchanging customer data, which frequently occurs in granular and completely uncontrolled ways. The study examines when and how privacy conversations occur during development by doing a discourse analysis of mobile application developer forums. According to the findings, these online forums function as platforms for ethical discourse, allowing developers to define, discuss, and explain their principles. Also, the ethical considerations in mobile development differ between the two major mobile platforms, iOS and Android, implying differences in work practices.

3. METHODOLOGICAL FRAMEWORK

To ensure a comprehensive understanding of how different stakeholders view and handle the ethical aspects of AI, a mixed-methods approach incorporating a focus group and interviews was adopted. In the initial phase, focus group discussions were conducted to harness group dynamics and generate rich qualitative data. Participants were purposefully selected based on their relevance to the research objectives and their diverse perspectives. The focus group sessions were guided by a semi-structured protocol designed to explore key themes and encourage participants to express their views openly. The discussions were audio-recorded and transcribed, allowing for a detailed analysis of the participants' interactions, shared experiences, and emerging themes.

During the focus group, case studies of AI applications employed in law enforcement were presented to the participants of the focus group, including real-world examples of using these AI-applications in LEAs, along with challenges and opportunities. Providing concrete example allowed participants with limited knowledge on AI to easily grasp the issues at stake. Each case study was chosen by the focus group moderator. This approach fostered a greater level of engagement and ensured that the topic resonated with the participants on a deeper level. The case studies encompassed a wide range of AI applications, including predictive and detection systems, systems for processing child sexual abuse material (CSAM), social network analysis, and recognition technologies and prompted participants to analyse, interpret and evaluate the case and imagine potential policy solutions. In total, 127 people from five countries (i.e., Greece, Germany, Slovakia, Italy, Spain) participated in the focus group.

Following the presentation of each case study, participants were divided into break-out rooms for group discussions. The groups were divided in a way that ensured each group had individuals from diverse backgrounds or with different areas of expertise. In the break-out rooms, participants were asked to discuss the opportunities and risks of the technology presented in the case study and brainstorm recommendations to policymakers. After the breakout sessions, participants were asked to reconvene and collectively share and discuss the key points.

After the focus group sessions, individual interviews were conducted to delve deeper into participants' personal perspectives and to capture any nuances that might not have surfaced in the group setting. A purposive sampling strategy was employed to select participants for the interviews, ensuring representation across relevant demographics or categories. The interview guide was developed based on insights gained from the focus group discussions, addressing specific themes that required further exploration. Interviews were conducted in a one-on-one virtual format. Ethical considerations, including issues of informed consent design and confidentiality, were rigorously maintained throughout the data collection process. During the interviews, the ethical frameworks that are available in literature (Table 1) and the principles included in each one of them have been presented, aiming to discuss about the issues raised by the application of such frameworks during AI application's development.

Table 1. Available ethical frameworks in literature and supported principles.

Ethical Framework	Description	Main supported principles
High-Level Expert Group (HLEG) - Assessment List for Trustworthy Artificial Intelligence (ALTAI)[4]	This framework aims to offer guidance on securing ethical and robust AI. Addressed to all stakeholders, these Guidelines seek to go beyond a list of ethical principles, by providing guidance on how such principles can be operationalized in sociotechnical systems.	• Human agency & oversight • Technical Robustness & safety • Privacy & data governance • Transparency • Diversity, Non-discrimination & Fairness • Societal & environmental well-being • Accountability
Model AI Governance Framework[5]	The Model Framework translates ethical principles into implementable practices, applicable to a common AI deployment process.	• Transparency • Explainability • Fairness • Human-centric solutions
NOREA Guiding Principles Trustworthy AI investigation[6]	The Guiding Principles included in NOREA framework aim to support IT-auditors in performing ex-ante and/or ex-post investigations of algorithmic systems to guide and support organizations in their journey towards deploying trustworthy AI applications.	• Technical Robustness • Safety • Transparency • Explainability • Fairness • Accountability
ICO AI and Data Protection Risk Toolkit	The ICO AI and Data Protection Risk Toolkit reflects the auditing framework developed by the ICO internal assurance and investigation teams. This framework provides a methodology to audit AI applications and ensure they process personal data in compliance with the law.	• Accountability & governance • Lawfulness and purpose limitation • Fairness (Statistical accuracy, bias and discrimination) • Transparency • Security • Data Minimization • Individual rights • Meaningful human review

The combination of focus groups and individual interviews allowed for a triangulated analysis of the gathered information, enriching the overall understanding of the research topic and providing a more holistic view of the participants' perspectives. In total, 12 diverse ICT professionals in the use and implementation of AI-based solutions from Greece participated in the interviews. The participants were 2 data analysts, 8 AI and ML developers, and 2 project managers.

During both events, the risks faced during the implementation and the use of an AI solution by the participants were discussed and recorded. In addition, recommended actions suggested by the participants have also been written down and are provided in section 6.

[4] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[5] https://ai.bsa.org/wp-content/uploads/2019/09/Model-AI-Framework-First-Edition.pdf

[6] https://www.norea.nl/uploads/bfile/a344c98a-e334-4cf8-87c4-1b45da3d9bc1

4. IDENTIFIED RISKS

The following risks have been stressed by all participants during the discussion in both the focus group and the interviews. The first risk category mentioned concerns the misuse of the application. The intricate nature of AI systems, along with the complexity of the data they process, makes it challenging to identify and mitigate potential risks effectively and in all cases. Furthermore, the **lack of a standardized framework or the wrong use of one,** for training end-users on how to handle and implement AI-based applications contributes to the overall difficulty in risk management. This **absence of guidance** can lead to misuse or **misinterpretation** of AI outputs, exacerbating ethical concerns.

The second risk category mentioned concerns the **population's reluctance** that might be a direct result of the lack of openness of AI systems, the narratives propagated by the media, or a lack of preparation owing to the digital gap by age. In this regard, some participants and writers argue that younger generations are more willing to adapt whereas older generations are more resistant. Other participants and authors, on the other hand, argue that younger generations see AI systems as a tool for mass monitoring and a violation of human rights and liberties. In any case, the shared feature is binary homogenization by age. The deployment of AI in law enforcement may affect public trust. If not properly communicated and understood, AI applications can lead to scepticism or resistance from the community. Involving the community in the development and deployment of AI technologies can help build trust and ensure that diverse perspectives are considered.

Furthermore, many participants mentioned the risks of **job displacement** faced by the use of AI applications. The main risk and concern are that the introduction of AI technologies may change the nature of work, potentially leading to job displacement for some personnel. Ethical considerations include providing adequate training and support for those affected.

An additional risk added concerns the handling of large volumes of sensitive data that increases the **risk of data breaches**. Ensuring the security of AI systems and the data they process is essential to prevent unauthorized access and potential misuse. AI systems can be vulnerable to adversarial attacks, where malicious actors manipulate the system to produce inaccurate results. Safeguarding against such attacks is a constant challenge.

As organizations increasingly integrate AI tools into various facets of their operations, there is a growing risk of **relying too heavily on automated systems without adequate human oversight**. This over-dependence may lead to unintended consequences, such as algorithmic biases, misinterpretation of complex scenarios, or unforeseen errors. It is crucial to strike a balance between the efficiency of AI tools and the necessity of human judgment and intervention. Imbalance in this relationship may occur, since through the automation of some processes, people tend to be fully depended on the AI applications, while a symbiotic partnership between AI and human expertise should be maintained for responsible and effective decision-making.

5. ANALYTICAL INSIGHTS: FINDINGS FROM FEEDBACK COLLECTION

The results of the focus group and the interviews were collected and analysed, aiming to extract and highlight the ethical views on the use of AI by diverse ICT professionals and LEAs. The first theme emphasized by the focus group participants was the need to **minimize bias in predictive AI**. The discussion centered on leveraging machine learning and AI to enhance the crime recording system by avoiding negative feedback loops and allocating resources to reduce bias. Participants brought the example of a neighborhood labelled as high-risk based on historical data. This classification leads to an increased police presence in the area, subsequently resulting in a higher number of recorded criminal activities within that specific location.

Another significant topic highlighted, was importance of harmonizing AI usage, both nationally and across Europe. Participants mentioned the need for a **comprehensive legal framework** that safeguards data protection and allows judges to intervene in granting permission for data usage. This emphasis on legal harmonization aimed to establish a consistent and accountable approach to AI implementation, fostering **trust and transparency in the deployment of these technologies**.

A clear procedure to ensure compliance with legal and ethical requirements is paramount. In the case of LEAs, as criminal activities increasingly transition to the online realm, the use of AI tools and the establishment of ethical standards become critical. Strict adherence to legal and ethical standards is essential to mitigate the risk of law enforcement authorities abusing their powers, especially when dealing with anonymous perpetrators in the virtual space. Ensuring compliance with these standards remains a crucial aspect of responsible and ethical AI utilization in law enforcement.

The third theme emphasized the role of **humans as decision-makers** in the context of AI systems. Participants asserted that AI should function as decision-support tools rather than autonomous decision-makers. Human supervision throughout the entire lifecycle of an AI system was deemed crucial, with the ultimate authority for decision-making resting in human hands. This perspective reinforces the importance of **maintaining human oversight** to ensure responsible and ethical use of AI. At this point, participants also raised the importance of having an authority for the assessment of the human that uses or monitors the AI application.

Moreover, the need for regulations was raised, aiming to **enhance citizens' awareness** of the implementation and adoption of AI systems. This need is critical when AI-based applications are used in LEAs and in public administrations for either monitoring or decision-making purposes. Establishing clear regulations was seen as essential to enable public objection and raise concerns about potential unjust decisions made by AI systems. This emphasis on citizens' awareness and participation underscores the commitment to inclusivity, accountability, and transparency in the deployment of AI technologies. In the case of using AI-applications in LEAs, participants emphasized on the early prevention alongside crime detection, and the importance of educational programs for citizens. These initiatives target the root causes and risk factors associated with criminal behavior, aiming to reduce the likelihood of criminal activities.

Regarding LEAs and specifically for the case studies presented to the focus group participants, they highlighted the following:

1. Proportionality of biometric data use is necessary to defining the appropriate use of personal biometric data based on specific circumstances.
2. A perception of a **division exists between technical aspects of AI and the role of police officers**. Comprehensive training programs for law enforcement personnel on the use of AI, including perspectives on ethics are crucial to enabling officers to better understand this world, maximize the benefits of AI while understanding its limitations and potential risks.
3. Participants emphasized the necessity of understanding that implementing such initiatives is a significant societal decision. The intrusive nature of AI applications in surveillance systems necessitates careful evaluation. Special consideration must be given to the group of people who may be disadvantaged by an AI system that does not recognize everyone equally, resulting in excessive monitoring.

As far as the interviews concerns, the corresponding stakeholders mentioned that the initial problem they are facing is the **lack of a comprehensive repository outlining common ethical risks** associated with AI applications. The absence of a standardized ethical framework leaves the development and

management teams navigating in uncharted territory when it comes to ensuring responsible and unbiased AI outcomes. This poses challenges for developers seeking to proactively address these issues during the development lifecycle.

In addition to the aforementioned opinion, interview's participants mentioned as an important aspect the **low availability of AI application ethical assessment protocols**. As presented in section 3, in literature are at least six assessment frameworks that can be used for the assessment of AI applications. However, most of the interviewees were either not familiar with these frameworks, or they could not use them properly due to **lack of guidance and common glossary**. More specifically, 5 interviewees (including both data analysts and developers) mentioned that they could not understand some of the terminology included within the ALTAI and NOREA frameworks, which they had used in the past, and that they didn't have enough guidance while applying these frameworks. Furthermore, they also stressed the fact that during the application of such frameworks, all the corresponding stakeholders of an AI-based solution should participate, in order to help during the risk identification and mitigation action's application processes.

From technical perspective, developers mentioned that they encounter significant hurdles in **addressing biases** and mitigating risks in AI applications, particularly in cases where there is no explicit guidance provided to users. In addition, they mentioned that there are also cases where the nature of the datasets and the data used is biased, leading to biased AI models and decision-making. They also highlighted the **absence of well-defined frameworks or guidelines on how to handle biased or risky situations** or examples of case studies including results and unintended consequences, ethical dilemmas of similar cases. During this discussion, participants mentioned also the **importance of GDPR** and the fact that it should be used carefully in each AI method during data pre-processing. In this process, a corresponding legal team should participate in the applications implementation phase and guide the developers, highlighting the important and critical paragraphs of GDPR, and ensuring that it is applied.

In addition, and in order to achieve all the aforementioned, **qualified staff and continuous training** programs for users, model designers, and technology experts to stay abreast of evolving AI technologies is required. Establishing a legal framework outlining relevant certifications is deemed essential to ensure the competence of professionals in this field. Additionally, achieving interoperability among different databases is crucial for effective collaboration, especially in areas like recording information related to unaccompanied minors, foster care cases, and adoptions. This fosters seamless data sharing and coordination, enhancing the overall implementation of AI systems.

In terms of principles, project managers particularly said that the principles accessible in the currently existing ethical frameworks in literature are not common in all circumstances, and that there is a lack of explanations for how each principle might be implemented in some of them. They emphasized that these values do not always coincide, which contributes to uncertainty in ethical decision-making. Furthermore, several frameworks were found to be deficient in extensive explanations explaining the practical application of each ethical principle, leaving practitioners in the dark about implementation tactics.

6. RECOMMENDATIONS

This section presents the recommendations resulted from the insights analysis. In particular among the essential guidelines of the ethical use of AI, is to provide concrete techniques for ethical design and deployment in the public sector or in an organisation. Therefore, in order to focus on the implementation of an ethical solution, the recommendations that the organisations should follow are presented in the following table (Table 2).

Table 2. Guidelines to be followed for the ethical adoption of AI.

Category	Description
Role	AI systems in LEAs should play a supporting function, enhancing human skills. However, control must always be exercised by a human. To demonstrate this, the possible Human Rights problems associated with the emphasis on algorithmic predictive policing systems, Automation Bias (AB), and the Human in the Loop role have been emphasised.
Use	Starting with the aforementioned law as a foundation, responsible use of AI systems, organizations necessitate comprehensive education based on technical, ethical, and legal knowledge and training.
Objective	The ultimate objective of the AI application should be properly and thoroughly defined, along with the priorities and the responsibilities of each participant/team.
Transparency	In order to promote and acquaint the people with AI and its deployment in the security sphere, the technologies and their applications must be described clearly and transparently.
Acceptance	Only with the aforementioned characteristics can trust be established in order to strike the essential balance between security and privacy. This balance must be reached through collaboration across several stakeholders and disciplines.
Organisational culture	The culture of the organisation should focus on the implementation and adoption of ethical AI-based solutions, that increase transparency and accountability, and lead to a sustainable environment.
Multidisciplinary team creation	The setup of multidisciplinary teams while developing AI-based apps is a tough task. Diversifying skill sets, experiences, and viewpoints is critical for developing inclusive and equitable AI systems. However, establishing a diverse workforce has practical challenges, including insufficient representation and inclusion in the software sector. This lack of diversity might result in blind spots in identifying different user demands, as well as potential biases in the development process. Clear communication and comprehension are vital for ethical decision-making, and the present hurdles highlight the necessity for comprehensive measures to solve these complexchallenges in the implementation of AI technologies.
Interoperability for Collaboration	Achieving interoperability is essential to facilitate collaboration between different databases for LEAs & public administrations, leading to the best and most effective implementation of AI systems.
Responsibility for Data	Designating a data controller is important to ensure the proper and responsible use of data. Deliberating on the reliability of the police versus the political establishment in managing this task is a key aspect of the discussion.
Clear protocols and observation body	There is the need establish clear protocols and the authorization of an observation body, that govern the utilization of AI technologies, providing guidance on their proper deployment and potential limitations.
Continuous training	continuous training and evaluation of the human skills should take place, aiming to level-up all stakeholders and prepare them for performing the corresponding mitigation actions in case of an ethical risk. In addition, guidelines should be provided for the use and application of the mitigation actions required in order to solve the risks raised.

7. CONCLUSION

This paper presented the work performed on the elicitation of stakeholder's aspects on the ethical aspects followed when designing, implementing and using AI-based applications. For the stakeholder's aspects collection, one focus group has been organised, along with one-to-one interviews, aiming to

collect several perspectives from the participants. The list of stakeholders included LEAs employees, public administration employees, developers, data analysts and project managers. The diversity of the stakeholders participating in the research is of great importance, aiming to collect all the aspects raised from different user groups.

On the focus group organized, the focus was raised on the stakeholder's opinions on using AI applications in sensitive public administrations such as LEAs. The aim of this focus was to highlight the risks raised by using AI applications that have not been evaluated in terms of ethics. The risks raised have been captured and analysed, while recommendations for the ethical adoption of AI applications in this field have been mentioned. The results of the discussions raised that there is a gap on the current ethical frameworks as far as it concerns the use of AI applications by LEAs. The gap refers to the data processing methods that need to be performed in order to minimize ethical risks, which are not predefined, the lack of case studies and guidelines on how to minimize risks in data sensitive cases and also the lack of training in the employees that use AI solutions.

Similar insights have resulted from the interviews performed on developers, data analysts and project managers. In addition, they mentioned also the need of having predefined principles and risks for each case of AI applications and the context of their use. Also, they highlighted the need of having case studies explaining the mitigation actions performed for the minimization of ethical risks. As far as the recommendations, they also highlighted the need of training, the definition of the AI application's objective and the setup of a multidisciplinary team that facilitates the elicitation of ethical requirements and the ethical design of an AI application.

In the future, this research can be extended in order to analyse in depth the role of each stakeholder during the design and implementation of an AI application, and highlight their participation, skills, and effort in the ethical assessment of the corresponding application. This analysis is of high importance aiming to define the metrics according to the appropriateness and readiness of relevant stakeholders and ICT professionals to be engaged to ethical attitudes and ethical competences by applying robust assessment practices, tools and methods and minimizingpotential risks posed by AI.

ACKNOWLEDGEMENTS

This research was supported by grants from Horizon 2020, the European Union's Programme for Research and Innovation under grant agreement No. 101022001 - popAI. This paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. International Journal of Information Management, 62, 102433.

Carvalho, L., Martinez-Maldonado, R., Tsai, Y. S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world? Computers and Education: Artificial Intelligence, 3, 100053.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. AI and Ethics, 1, 73-80.

Eurostat, Use of artificial intelligence in enterprises, (2022), https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and Engineering Ethics, 26(6), 3333-3361.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2021). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Ethics, governance, and policies in artificial intelligence, 19-39.

Griffin, T. A., Green, B. P., & Welie, J. V. (2023). The ethical agency of AI developers. AI and Ethics, 1-10.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., ... & Tang, J. (2022). Trustworthy ai: A computational perspective. ACM Transactions on Intelligent Systems and Technology, 14(1), 1-59.

Kim, P. T. (2017). Auditing algorithms for discrimination. U. Pa. L. Rev. Online, 166, 189.

Madiega, T. A. (2021). Artificial intelligence act. European Parliament: European Parliamentary Research Service.

Meyer, D., & Henke, M. (2023). Developing design principles for the implementation of AI in PSM: An investigation with expert interviews. Journal of Purchasing and Supply Management, 100846.

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., ... & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. IEEE Transactions on Technology and Society.

Shilton, K., & Greene, D. (2019). Linking platforms, practices, and developer ethics: Levers for privacy discourse in mobile application development. Journal of Business Ethics, 155, 131-146.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

Veale, M. (2020). A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. European Journal of Risk Regulation, 11(1), e1.

THE CHALLENGE OF CO-CREATION: HOW TO CONNECT TECHNOLOGIES AND COMMUNITIES IN AN ETHICAL WAY

Kristina Khutsishvili, Neeltje Pavicic, Machteld Combé

University of Amsterdam (the Netherlands), City of Amsterdam (the Netherlands), City of Amsterdam (the Netherlands)

K.Khutsishvili@uva.nl; N.Pavicic@amsterdam.nl; M.Combe@amsterdam.nl

ABSTRACT

The contribution aims to reflect on the ethical issues and moral dilemmas related to the process of co-creation with communities in the technological domain. While there is a significant amount of literature dedicated to co-creation, the best or good enough practices of co-creating artificial intelligence and broader emerging technology solutions are yet to be established. Ethical aspects come to the forefront when working with members of vulnerable and marginalised communities, with the discussion not sufficiently presented in academic literature.

Deriving from empirical material accumulated while working on the CommuniCity Horizon Europe project, we argue for an incremental improvement approach to co-creation aimed at both learning from and improving processes, with findings continuously shared and updated for potential replication purposes. We also stress the dialectic nature of co-creation with communities: the experimental, learning part coexists with the 'do no harm' principle and the need for a strong moral framework and continuous ethical monitoring. While advocating for such an approach, we emphasise the essential character of a thoughtful and careful design and execution of processes.

KEYWORDS: ethics, technology, AI, city, community, vulnerability, co-creation.

1. INTRODUCTION

In the scope of literature on co-creation, the specificity of co-creation in the technological domain has not been elaborated extensively, with attempts suggesting the need for a different approach to the issue itself when it comes to co-creating technological solutions with citizens (Jarke, 2021, p.203; Spuzic et al., 2016). Even more so, the existing literature lacks the focus on engagements with members of vulnerable and marginalised communities, where ethical aspects and risk mitigation have even more significance. The abovementioned factors together form an interesting setting in which practical experiences of co-creation fill the knowledge gap and, if successfully disseminated, give an impetus for replication attempts, and opportunities for new learnings to be derived, accumulated, and disseminated. This curious and challenging circle we propose to call the mechanism of *incremental improvement*.

Experimentation as an approach to urban challenges had been outlined in the preceding European projects, OrganiCity and SynchroniCity (Amaxilatis et al., 2019; Wilson, 2019). According to Brynskov et al. (2018, p.151), experimentation "proposes a methodology to engage stakeholders in a processual, solution-oriented and citizen-centric manner when addressing urban challenges and problems."

The distinctive character of the approach proposed by this work is rooted in coexistence of the experimental foundation and 'high stakes' in the ethical domain, consequent to engagements with community members, especially when it concerns vulnerable and marginalised communities.

In this work we aim to reflect on ongoing experiences within CommuniCity Horizon Europe project as a source of empirical material for learnings on co-creation in the domain of artificial intelligence and broader new emerging technologies, with a focus on engaging disadvantaged communities.

2. COMMUNICITY CONTEXT

The project draws on three rounds of open calls starting in the cities of Porto, Amsterdam and Helsinki and then 'replicated' in other European cities during the project timeframe of three years. In the beginning of the open call rounds, hosting cities announce societal challenges to which pilot proposals need to respond. Selected by independent jury members, awarded pilots aim to develop technological solutions tailored to the specific needs of local communities together with the members of those communities by means of co-creation. The overall aspiration of the project is to accumulate experiences and learnings on co-creation of targeted technological solutions with disadvantaged groups, to come up with replicable practices.

For the process of incremental improvement, it is important that the learnings that derive from both successes and failures are shared, reflected upon, with best and good enough practices replicated by other cities and communities, new learnings shared, the knowledge framework improved and updated. As community members are aimed to be the end users of technological solutions, their needs are meant to be at the heart of the process design. Without sincere interest and in-depth immersion in their life circumstances, methods, approaches, and know-how will likely remain formal tools. The project framework also has its limitations. For instance, each pilot period lasts five months within which the cooperating partners work towards a solution. The funding of €12500 is granted per winning pilot. Working and co-creating with members of disadvantaged communities in response to complicated and delicate challenges requires time and funding sufficient to run feasible pilots in which the 'challenge owner' (usually the municipality unit with corresponding expertise), community and technological provider can all advance. With learning accumulation being the main project objective, we can immediately spot a moral dilemma when concluding pilots with final implementable solutions is a desirable but not necessary situation, while engaged community members and 'challenge owners' put effort into working towards solutions. As a balancing act, concrete steps taken to progress with the solution, if not make it achievable in a short term, work towards a positive experience for all parties involved.

Successful and rewarding co-creation experiences require trust. It takes time to thoroughly assess the challenge before a thoughtful design process can begin, and this also limits the complexity of challenges that can be addressed in the pilot environment.

3. CO-CREATION IN THE CONTEXT

Co-creation processes are approached by a wide range of academic disciplines, including management and innovation studies that aim to give a structured overview of the issue rather than providing a conceptual framework. For instance, Rose et al. (2013, p.23) define co-creation is "an interactive, creative and social process between stakeholders." Within the context of innovation studies, co-creation is seen as "open innovation with customers" (Rayna & Striukova, 2015, p.38). It "corresponds to the customer-related part of open innovation: 'open innovating' with consumers necessarily implies co-creating with them" (Rayna et al., 2015, p.2). While not explicitly mentioned in the project proposal, *the innovation part of our project is comprised by both technological and social innovation aspects, with an attempt to see them connected rather than exercised separately.* This is not the first European project comprising these, oftentimes viewed in silos, innovation vectors (Kohlgrüber et al., 2018, 2021; Wilson et al., 2019). The significant distinction is in attempt to connect social and technological domains in a mutually rewarding way, in bringing communities to the centre of the processes, in experimentation combined with high ethical 'stakes' consequent to conditions of vulnerability and marginalisation.

As suggested by the OECD report on co-creation (Rossi et al., 2020, p.1), "when co-creating a complex technological solution, the intermediary is involved in two complementary, often intertwined, but distinct processes that bring together organisations that demand technology and those that supply technological solutions." In our project, the role of intermediaries has mostly been assigned to associations, mainly non-profit organisations that have established connections with vulnerable and marginalised communities. According to Rill and Hämäläinen (2018, p.2), the main foundational principle of co-creation as "harnessing the collective potential of groups can lead to breakthroughs wherein every participant is empowered." The 'empowerment' wording is very interesting in application to co-creation in the technological domain. Digital disparity is a feature of the socio-technical landscape even in developed countries. It is a matter of uneven power, where actions to empower are crucial to bridge the growing gap. Knowledge and skills, starting with basic digital know-how, play a role in these power dynamics. New technologies can both empower and pose risks for vulnerable and marginalised communities, including the elderly, people with disabilities, long-term unemployed individuals, refugees, and those with lower incomes. The evolving landscape may bring positive changes on personal and community levels, but it could also worsen the circumstances. Moreover, there are both 'objective' and 'subjective' aspects. The latter represents how individuals 'feel' about new technologies: the presence of individual or collective anxieties around new emerging technologies may not necessarily relate to the factual side such as a disadvantageous position, bias, or discrimination. Instead, it may relate to the perception of rapid technological development working 'against' an individual or a group.

Conditions of vulnerability and marginalisation are discussed in a wide range of academic disciplines. According to Wrigley and Dawson (2016, p.203), the concept of vulnerability has not been defined clearly enough but it "indicates that an individual or group is thought to have a particular status that may adversely impact upon their well-being." In this work, we use both words, 'vulnerable' and 'marginalised', to describe communities we aim to target in the project. The vector of these two human conditions differs, with vulnerability being an internal characteristic and marginalisation related to the 'outside' world, with a strong circumstantial connection to features of reality. The passive tense in the latter reflects such a connection and a negative influence of circumstances of the external world on the internal situation. The 'disadvantaged' may be seen as an umbrella term uniting these conditions while being a more neutral constatation of disparity.

To experiment with developing technological solutions for vulnerable and marginalised groups within the context and possibilities of a pilot, we suggest paying attention to the scope of the challenge and peculiarities of the particular group, outlining two variables – the vulnerability of the group and the sensitivity of the challenge. The more intense the conditions of vulnerability and marginalization are, the more sensitive the announced challenge is, the more time and effort are required from a technological party, the more is the need to 'immerse' in the target group and the challenge. Making authentic connections with communities requires an approach more sophisticated than traditional business and marketing tools aimed at engagement and co-creation. The experiment here plays not only a methodological role on a path of incremental improvement but is also necessary for the parties involved, including technological providers, as preliminary recommendations are not enough to fully understand and correctly perceive where the boundaries lie. Here we need to emphasize once more the dialectical relation between experiment and sensitive conditions within the project and pilots: while arguing for experiment aimed at the incremental improvement of co-creation practices, the more vulnerable the target group is and the more complex and sensitive the issues are, the greater is the chance of disappointment and failure, and the more effort we need to put into ethical consideration, both preliminary and continuous.

4. ETHICAL QUESTIONS

One of the ethical questions not yet sufficiently answered in the project is the ethical way to respond to the possible situation in which community members add value to the development of the solution but then will not be amongst the ones who benefit from this solution in its later development stage. The underlying reasons for such a situation may vary: in one of the examples from the city of Amsterdam, the piloting solution, despite multiple challenges on the way of this particular pilot implementation, is appreciated by the pilot host with an aim to sustain and utilise the solution beyond the piloting timeframe. The unexpected problem the pilot host faces at the end of the piloting period is the high maintenance cost of the solution making this aspiration not feasible, while having motivation and resources available but not matching with the costs and with alternative possibilities of spending these resources on direct activities supporting the 'target group' with regard to the same challenge. A similar obstacle is observed with pilots in Porto.

Another ethical challenge that has been 'spotted on' in the project is *the dilemma of 'co-creation with' and 'testing on'*. Does the line between the two exist, and if so, where should it be drawn? For example, if the proposed solution is an application that needs to be developed through new data coming from a vulnerable or marginalised community being 'fed' to this application, is this an example of co-creation or 'testing on'? One of the answers proposed is that the way the obtained data will be used determines the answer: one may use the data for improving or finalising a commercial solution or a solution tailored for these specific people, members of the 'target community'. When the latter happens, there is still an issue of the solution not being made available free of charge or at a reduced cost to those who tested it and helped it to evolve, as is suggested by the experience of Porto and Amsterdam.

In view of project partners who share a technological professional background, any technological project needs people and data, with technological solutions needed to be tested. From this standpoint, there is not much of a difference for the technological provider whether it is engaged in 'testing on' activities or co-creation activities, with the latter being more challenging and demanding. Both modalities though require the translation activities on the initial stage aimed to prepare the 'target group' for the engagement activities. Parties with the social innovation professional focus do outline the difference: in this view, by co-creation we should mean facilitating the setting where we can question the product itself and not the setting where the already designed product is presented to collect the feedback of community members. It should start and grow in a process of dialogue that demands a lot of time and effort. From this perspective, 'testing on' is not a negative term, but it is different from co-creation. Interestingly, this discussion had not been raised in projects preceding CommuniCity, such as SynchroniCity and OrganiCity, and it would be interesting to continue it further, also within the framework of incremental improvement of the processes.

The situation we wanted to avoid by all means is the 'aftertaste' of community members feeling 'used', exploited by pilot teams for the teams to develop commercial, 'for profit' solutions. Consequently, we saw the social innovation vector as leading while designing the open call and piloting processes. Many questions still do not have an unequivocal answer, this also includes the practicalities of approaching and motivating community members to participate in the project including the question of reward. For instance, the Amsterdam-based pilot hosts and intermediaries articulated the need for financial rewards assigned to community members engaged in co-creation and intermediaries themselves helping to connect with communities. Other partners pointed to the problems related to financial reward: while in technological domains such a reward is a usual practice, enabling communities to get access to final solutions is an incomparably more significant factor and motivation, so we need to put the effort in making this more feasible.

It is important to mention that the risks of disappointment and demotivation relate not only to communities but also to pilot hosts. In Amsterdam, pilot hosts belong to different structural units inside the City of Amsterdam. Their efforts start from the co-creation of challenges that are later announced by the City and to which the applicants need to respond. The motivation of pilot hosts derives from their practical interest in finding sustainable technological solutions for the existing needs. By them the project is seen as an instrument for tackling the objectives and finding practical solutions to significant problems. Then, at the end of the piloting period, the pilot host unexpectedly finds out that the solution, while being desirable to be procured, is too expensive. The disappointment and demotivation to put effort further in the next rounds of open calls may follow. The scenario when technological solutions are procured after piloting rounds is positive and desirable, it is the value added beyond the declared scope of the project and the factor raising its impact.

5. RECOMMENDATIONS

The vulnerable and marginalised community specificity, as well as the situation of power imbalance, lead to the necessity to prepare a minimal set of recommendations addressed to the technological providers on how to engage with communities. The points comprising the list below are developed inside the framework of the first piloting round, primarily based on piloting experiences of the City of Amsterdam (Khutsishvili, 2024):

1. Time and effort needed on the preparatory stage should not be underestimated.
2. Simplicity, clarity, evidence should be the core principles of all community engagement activities.
3. Reasoning in favour of mutual benefit (Why do we need to participate? What can we learn from each other? What is the significant and meaningful outcome we are putting effort towards?)
4. Keeping in mind the power imbalance. Trying to share ownership (of ongoing processes, of final solutions).
5. Diverging from the established practices of work. Adjusting not only the message but also the way of delivery (stepping out of the office space and going into ‘the field’; emphasising the relatable points while keeping authentic).
6. Having a genuine interest in the community and its members. Imitating such an interest will not benefit any of the parties involved.
7. Being ready for initial attempts to fail and trying again.
8. Considering consulting with and/or involving in the processes the ‘intermediaries’ – people close and trusted by community members. For ‘intermediaries’ it may take less time to reach the community members. In Amsterdam and Porto, this was mostly the role of associations. Yet, in one of the pilots, the ‘intermediary’ individuals had been involved and paid by the pilot host to help to facilitate the trusted contact with a specific community of youth with criminal records and provide feedback on the pilot activities and the solution proposed by the pilot.
9. Having a vision of how the engagement should be designed in order to have a positive effect on the community is necessary from the very beginning.
10. Reflecting on the (potential) difference between ‘co-creating with’ and ‘testing on’, in the context of a particular community, challenge, and proposed solution.

In addition to the kick-off, midterm and concluding meetings, Amsterdam designed periodical meetings aimed to facilitate discussions, with guests separated during the sessions with regard to their 'professional' focus in a pilot, be it technological providers, associations, civil servants, or community members. The awarded pilot teams outlined points of reflection and suggestions for further open call and piloting process adjustment (Khutsishvili, 2024). Among those are:

1. The meaning as well as the definition of co-creation in the project may be unclear for a pilot team. The suggestion is to organise meetings on co-creation uniting all pilot teams at the start of the next rounds of piloting. During such a meeting, the project partners and different pilot teams could share their methods.
2. If a pilot takes place during summer, it is more difficult to establish the needed partnerships and engagement moments. In this case, it is proposed to extend the pilot duration.
3. For the success of the pilot, it is crucial that a technological provider thoroughly understands the target community, people's problems and needs.
4. Co-creation is not only a method required for the given project. It is also a useful tool to appeal to the community.
5. It is important to make sure that the open call is also announced in the target neighbourhood itself so that local technological companies which are likely to already understand the context can respond and participate.
6. The possibilities of involvement of community members on the very initial stage can be further explored. For instance, let the 'target' group read along with the proposals and thus give them a role in the jury.
7. It is crucial for a technological provider to be sure that the association they are teaming with has a solid contact with the target community. Otherwise, numerous problems will come on the way, and they will be difficult to overcome inside a limited piloting timeframe.
8. It is important for the teams to ensure that members of target communities are available for the engagement activities from the beginning of the pilot, otherwise time will be lost and the pilot may not be finished during the given timeframe.
9. Once the target community has joined, make sure you start co-creation activities quickly and don't wait too long, otherwise you may lose the motivation of people.
10. It is necessary to thoroughly explain to all parties involved, especially to technological providers, what is meant by co-creation and why it is important for the project as a whole and for the success of a particular pilot.
11. All parties should be aware that if the awarded team proposes the solution already existing in its technological portfolio to be 'tested on' a new audience, the vulnerable or marginalised community, such a solution can indeed be partly adapted to the needs of the specific community but there is no 'one fit for all', so such solutions may not always be suitable for the particular community.
12. Co-creation approaches must 'fit well' with the experiences of the target community.
13. For technological providers engagement with communities can be hard and time-consuming.
14. The open call information should be more detailed, with a focus on engagement and co-creation with communities, the resources required and possible challenges.

15. The jury needs to pay careful attention to the team's response to co-creation in the project proposal.
16. The more vulnerable the community is, the greater the chance that harm can be done if the expectations that have been raised are not met.
17. While engaging with community, it is important to 'take it seriously': listen well, gather the feedback carefully, explain and translate all steps well, even in cases of a difficult technological solution (for example, an animated video).
18. No stereotyping should take place. An equal non-hierarchical approach is necessary.
19. The translation challenge in its literal sense is something to be aware of. If the winning team is not based in the country where the pilot takes place, geographical and linguistic barriers can impact the engagement activities. Literal translation issues and co-creation in another language (English) than the official language of the country need to be further reflected on in the project.
20. There is a complex issue of financial remuneration of community members for their participation: according to some pilot hosts and intermediaries working 'in the field', community members put effort that is equal to effort put in regular work, so those efforts should be remunerated, this is also a matter of respectful and professional treatment. At the same time, and this is the point raised by a technological provider: if technological providers pay to test their products, would it lead to the situation when community members participate in co-creation just because of remuneration? In addition, another pilot team members quoted the member of target community who stated that he participates in the project "to help people, to work together towards a higher goal." This vector of shared experiences and motivation needs further reflection and discussion.

6. CONCLUSION

Co-creation activities aimed at the communities in question do not make piloting processes easier, quite the opposite. Yet, the opportunities for experimentation and learning accumulation enabled by such design are extremely valuable. To facilitate and conduct the related activities of community engagement activities in an ethical way, it is necessary to keep in mind the 'do no harm' principle, the power imbalance including the imbalance of professional, technological subject-related knowledge, and the general condition of belonging to a disadvantaged community.

The balancing act, as well as the risk mitigation, may be exercised by providing clear and honest communication including communication on the general aims and limitations of the project and a particular pilot. Encouraging dialogue on equal terms, aimed at 'de-objectivization' of the community and empowering its members from the very beginning of the engagement, is crucial.

Arguing for the incremental improvement approach, we emphasise the dialectical relationship between experimentation and community engagement, especially with vulnerable and marginalised communities being at the heart of the project. In our view, such a setting brings unprecedented analytical and research possibilities but also stresses the factor of responsibility and the ethical component. The learnings, in their broader sense, include not only successes but failures. At the same time, with communities being at the centre of the processes, not all failures are acceptable. Consequent efforts put in preliminary ethical consideration and continuous ethical monitoring and advise are required for running such projects.

ACKNOWLEDGEMENTS

This work has been supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101070325, project CommuniCity (Innovative Solutions Responding to the Needs of Cities & Communities).

REFERENCES

Amaxilatis, D., Boldt, D., Choque, J., Diez, L., Gandrille, E., Kartakis, S., Mylonas, G., & Vestergaard, L.S. (2019). Advancing experimentation-as-a-service through urban IoT experiments. IEEE Internet of Things Journal, 6(2), 2563-2572. Retrieved from https://core.ac.uk/download/pdf/222786441.pdf

Brynskov, M., Heijnen, A., Balestrini, M., & Raetzsch, C. (2018). Experimentation at scale: challenges for making urban informatics work. *Smart and Sustainable Built Environment*, 7(1), 150-163.

Challenges articulated by participating cities in first and second rounds of open calls. Retrieved from https://communicity-project.eu/first-open-call/, https://communicity-project.eu/second-open-call-challenges/

CommuniCity Horizon Europe project: Innovative solutions responding to the needs of cities and communities. (2022-2025). Retrieved from https://cordis.europa.eu/project/id/101070325

Jarke, J. (2021). Co-creating digital public services for an aging society: Evidence for user-centric design. *Public Administration and Information Technology*, 6. Springer.

Khutsishvili, K. (2024). Guidelines for translating frameworks, methods, tools and principles of local innovations for marginalised and vulnerable Communities – 2023. *Open Research Europe.*

Kohlgrüber, M., Maldonado-Mariscal, K., & Schröder, A. (2021) Mutual learning in innovation and co-creation processes: Integrating technological and social Innovation. *Frontiers in Education*, 6. Retrieved from https://www.frontiersin.org/articles/10.3389/feduc.2021.498661/full

Kohlgrüber, M., Schröder, A., Yusta, F.B. & Ayarza, A.A. (2019) A new innovation paradigm: combining technological and social innovation. Matériaux & Techniques, 107(1).

OrganiCity Horizon 2020 project: Co-creating smart cities of the future. (2015-2018). Retrieved from https://cordis.europa.eu/project/id/645198

Rayna, T., & Striukova, L. (2015). Open innovation 2.0: is co-creation the ultimate challenge? *International Journal of Technology Management*, 69(1), 38-53.

Rayna, T., Striukova, L., & Darlington, J. (2015). Co-creation and user innovation: The role of online 3D printing platforms. *Journal of Engineering and Technology Management*, 37, 90-102.

Rill, B.R., & Hämäläinen, M.M. (2018). *The art of co-creation: A guidebook for practioners.* Singapore: Palgrave Macmillan.

Roser, T., DeFillippi, R., & Samson, A. (2013). Managing your co-creation mix: co-creation ventures in distinctive contexts. *European Business Review*, 25(1), 20-41.

Rossi, F., Caloffi, A., Colovic, A., Russo, M. (2020). Public innovation intermediaries and digital co-creation. *Research contribution to the OECD TIP Co-creation project.* Retrieved from https://stip.oecd.org/assets/TKKT/CaseStudies/49.pdf

Spuzic, S., Narayanan, R., Abhary, K., Adriansen, H. K., Pignata, S., Uzunovic, F., & Guang, X. (2016). The synergy of creativity and critical thinking in engineering design: The role of interdisciplinary augmentation and the fine arts. *Technology in Society*, 45, 1-7.

SynchroniCity Horizon 2020 project: Delivering an IoT enabled Digital Single Market for Europe and Beyond. (2017-2019). Retrieved from https://cordis.europa.eu/project/id/732240

Wilson, D., McLoughlin, S., & Brynskov, M. (2019). OrganiciCity: Lessons from an experimentation as a service model for digital civic innovation. *International Conference on Smart Infrastructure and Construction (ICSIC*), 195-202. Retrieved from https://www.icevirtuallibrary.com/doi/10.1680/icsic.64669.195

Wrigley, A., & Dawson, A. (2016). Vulnerability and marginalized populations. In: Barrett, D., W. Ortmann, L., Dawson, A., Saenz, C., Reis, A., & Bolan, G. (eds) *Public health ethics: Cases spanning the globe.* Public Health Ethics Analysis, 3. Cham: Springer.

THE PIVOTAL ROLE OF INTERPRETABILITY IN EMPLOYEE ATTRITION PREDICTION AND DECISION-MAKING

Gabriel Marín Díaz, José Javier Galán Hernández

Universidad Complutense de Madrid (Spain)

gmarin03@ucm.es; josejgal@ucm.es

ABSTRACT

This article explores the evolution of machine learning (ML) algorithms, emphasizing the growing importance of interpretability in understanding automated decisions. Progress from early to advanced ML models highlights the need for better performance and adaptability. However, the inherent black-box nature of many ML algorithms raises challenges, underscoring the necessity for interpretability to improve transparency and accountability.

Examining the evolution of interpretability in ML, the article showcases advancements in techniques facilitating human comprehension of decision-making processes. As ML becomes integral across domains, the article underscores the importance of interpretable models to bridge the gap between automated decisions and human understanding.

The article delves into the changing role of humans in decision-making. Despite the efficiency of ML algorithms, the interpretability factor prompts a revaluation of human involvement, necessitating a balanced approach for ethical AI deployment.

Furthermore, the article explores integrating decision-making methods like Analytic Hierarchy Process (AHP) to enhance interpretability. Proposing a framework that combines AHP with interpretable ML models, it suggests a structured approach for human-in-the-loop decision-making while considering feature importance.

KEYWORDS: decision-making, machine learning, XAI, interpretability, AI, AHP.

1. INTRODUCTION

Throughout the historical evolution of Artificial Intelligence (AI) and machine learning (ML) algorithms, the emphasis on interpretability has become increasingly critical, particularly in the dynamic landscape of the business world (Hall, 2022). While the historical narrative traces the roots of interpretability in the context of AI development, its significance in the business realm, and specifically in areas like Human Resources (HR), is a pressing concern (Bandyopadhyay & Jadhav, 2021).

In the contemporary business environment, the adoption of ML algorithms is pervasive, and their applications extend to crucial domains such as HR, where decisions regarding employee management and resource allocation have profound implications. Achieving a balance between predictive power and human-understandable insights becomes paramount, especially when dealing with sensitive areas like employee attrition.

Interpretability is vital in the business context for several reasons. Firstly, businesses need to comply with ethical standards and legal regulations (Bibal et al., 2020). Transparent and interpretable ML models are essential for ensuring that decisions related to hiring, promotions, and terminations align with fairness and non-discrimination principles. Secondly, in HR, the ability to explain why a particular decision was made becomes crucial for building trust among employees and stakeholders (Mishra, 2013). For instance, if an algorithm predicts an employee is likely to leave the company, it is imperative to understand the features contributing to this prediction to take appropriate actions.

Consider a scenario where a company utilizes an ML algorithm to predict employee attrition. An interpretable model not only provides accurate predictions but also offers explanations for those predictions. This transparency allows HR professionals to understand the factors influencing an employee's likelihood of leaving, enabling them to intervene proactively. Interpretability, in this context, becomes a tool for strategic workforce planning, talent retention, and fostering a more inclusive workplace culture (Marín Díaz et al., 2023).

In HR decision-making, interpretability aids in justifying and fine-tuning models based on real-world observations, aligning them with organizational values (Srivastava & Eachempati, 2021). It empowers HR professionals to leverage the strengths of ML models while retaining human oversight in critical decision-making processes. The interpretability of algorithms in HR ensures that the human touch remains integral, fostering a collaborative and ethical approach to workforce management.

Interpretability is indispensable in the business landscape, particularly in critical areas like HR, where algorithmic decisions impact the livelihoods and well-being of employees. By shedding light on the decision-making process, interpretable ML models not only enhance trust but also contribute to strategic and ethical human resource management.

This paper addresses the challenges associated with interpretability in machine learning (ML) models, following a structured framework. Section 2 reviews the current state of eXplainable Artificial Intelligence (XAI) in the business domain. In Section 3, a methodology is proposed, focusing on handling algorithmic explainability while incorporating a crucial aspect—human decision-making. Section 4 presents a real-world use case illustrating the application of the proposed methodology. Finally, Section 5 provides conclusions and outlines future directions for research and development.

2. RELATED WORK

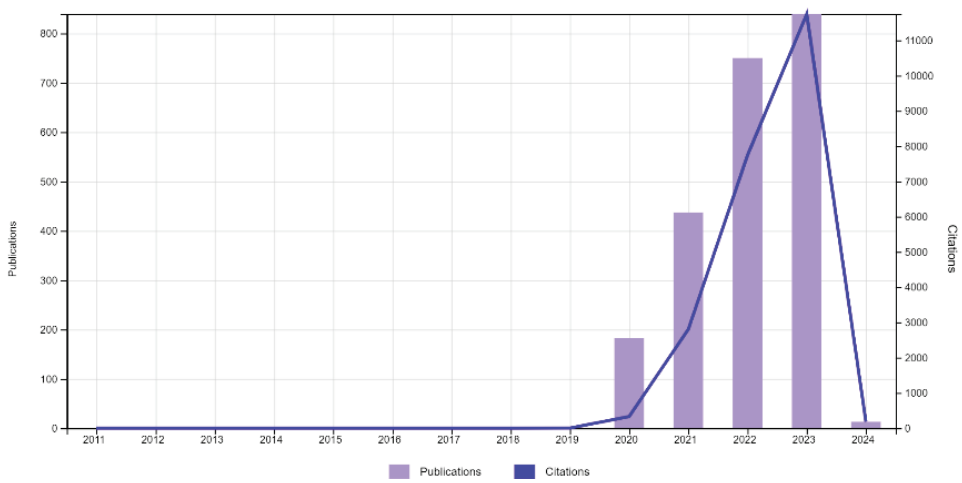
Employee attrition has emerged as a focal point for organizations, especially in the ever-evolving landscape of contemporary workplaces. Examining this from a psychological standpoint, numerous influential factors contribute to an individual's decision to depart from their current employment, especially in the technology sector (Colomo-Palacios et al., 2014).

Within this rapidly changing industry, newly hired employees often prioritize elements such as job satisfaction, a conducive work environment, and substantial financial compensation over traditional notions of stability and long-term commitment. Attrition is influenced not only by intrinsic aspects of the job role but also by organizational culture, growth prospects, and the alignment of individual values with the company's mission (Climek et al., 2022).

Comprehending the intricate dynamics and psychological foundations of employee attrition, particularly in technology-driven sectors, is imperative for crafting effective retention strategies.

The growing attention towards eXplainable Artificial Intelligence (XAI) highlights the increasing emphasis on creating AI systems that are understandable, particularly in situations where decisions have broad implications for individuals or society. Striking the appropriate equilibrium between predictive accuracy and interpretability remains an enduring challenge in the field. This equilibrium holds significant importance in building trust and gaining acceptance for AI systems in practical applications, ranging from healthcare to finance and beyond. Figure 1 visually depicts the evolution in the volume of studies addressing the interpretability of algorithms over time, TS = ("eXplainable Artificial Intelligence " or "XAI").

Figure 1. Studies addressing the interpretability of algorithms (2,221 publications).



Source: self-elaboration based on Web of Science (2024)

As observed in Table 1, the number of publications is centered around scientific areas, although the practical implementation of eXplainable Artificial Intelligence (XAI) models applied to the business world is not significant.

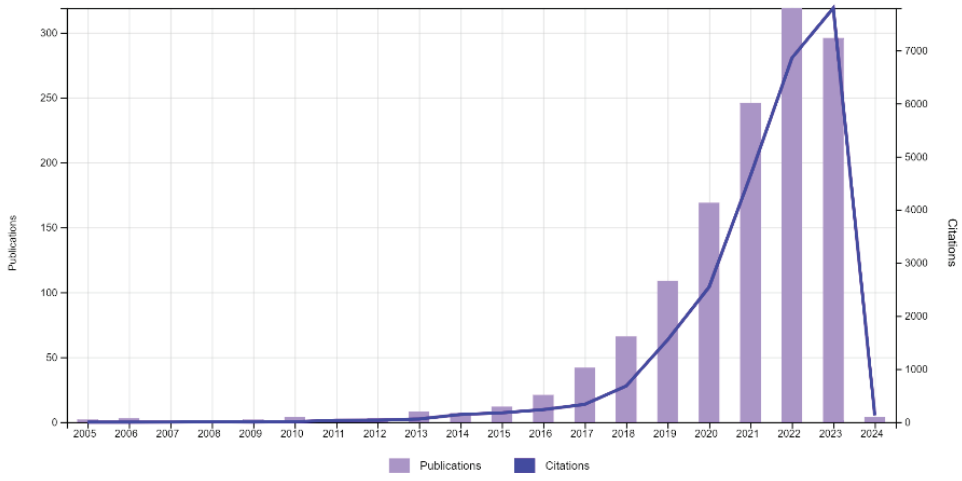
Table 5. Publications by research areas.

Area	Publications	%
Computer Science	1,646	74.111
Engineering	772	34.759
Mathematics	571	25.709
Mathematical Computational Biology	516	23.233
Communication	300	13.507

Source: self-elaboration based on Web of Science (2024)

Next, we proceed to analyse publications related to Machine Learning and Artificial Intelligence applied to the Human Resources sector, Figure 2, without currently delving into studies on interpretability in this area, TS = ("Machine Learning" or "Artificial Intelligence") AND TS = ("Resource Humans" or "HR").

Figure 2. Studies addressing the ML / AI applied to the Human Resources (1,314 publications).



Source: self-elaboration based on Web of Science (2024)

Finally, we focus on studies related to eXplainable Artificial Intelligence (XAI) that directly impact the Human Resources field, Table 2, TS = ("eXplainable Artificial Intelligence " or "XAI") AND TS = ("Resource Humans" or "HR").

Table 2. Publications XAI and Human Resources.

Publications
Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring (Hofeditz et al., 2022)
Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making (Marín Díaz et al., 2023)

Source: self-elaboration based on Web of Science (2024)

As evidenced by Figure 2 and Table 2, the utilization of predictive models in the Human Resources domain is extensively documented. However, it is noteworthy that articles specifically addressing interpretability amount to a total of 2.

3. METHODOLOGY

3.1. Interpretable Machine Learning

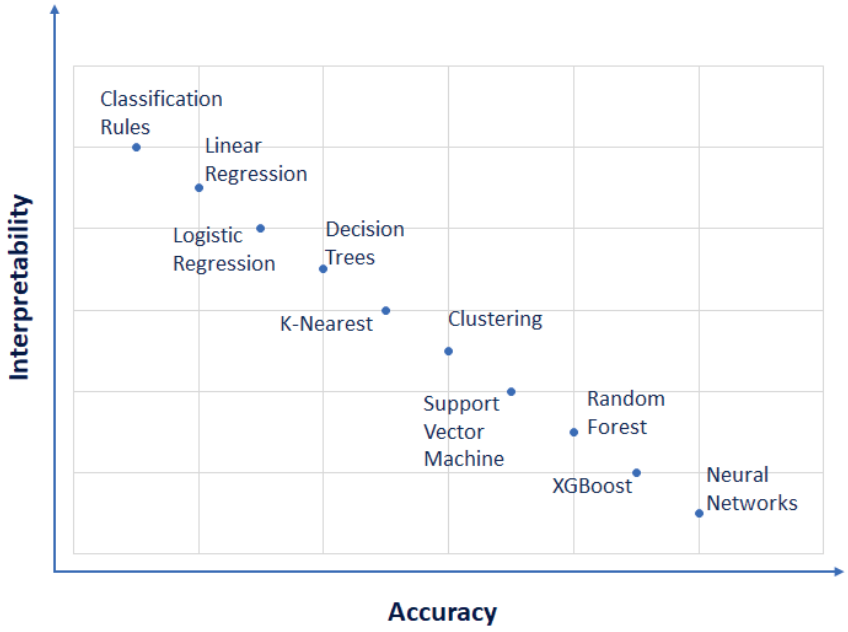
Interpretable Machine Learning (IML) is a pivotal component when decisions carry significant implications. This in-depth exploration will delve into various methods of interpretability and their practical implementation in environments where decision-making is of paramount importance.

When examining the interpretability of machine learning models, scholars commonly classify them into two fundamental categories (Carvalho et al., 2019). Firstly, 'transparent models,' often referred to as 'white box models,' aim to establish a clear connection between input variables and resulting outputs. On the other hand, 'opaque models,' or 'black box models,' lack easily interpretable decision rules. It is noteworthy that even within the realm of transparent models, interpretability remains a topic of ongoing discussion, as highlighted in a distinct study that raises questions about their interpretability (Lipton, 2018).

Figure 3 portrays an inverse correlation between interpretability and accuracy, emphasizing the intricate challenge of striking a balance between model interpretability and predictive precision (Molnar, 2019). Distinguishing between these model types provides valuable insights into the spectrum of interpretability within the domain of machine learning. This nuanced perspective enhances our comprehension of model behaviours and performance, contributing depth to the ongoing discourse on interpretability.

It is essential to acknowledge that the presence of bias and noise in data can distort interpretations. Addressing these issues is imperative before embarking on any interpretability analysis. Data cleaning techniques, such as class balancing, play a crucial role in mitigating bias, while noise removal ensures that interpretations rely on reliable information (Gilpin et al., 2019). The following outlines various methods for interpreting algorithms.

Figure 3. Interpretability and Accuracy.



Source: self-elaboration based on (Duval, 2019).

Feature Importance Analysis: Quantifies the impact of individual features on model predictions, often expressed through coefficients or weights (Perisic & Pahor, 2020). Commonly applied in linear models, providing explicit feature contributions.

Decision Tree Structure Analysis: Investigates the hierarchical decision-making process of decision trees, offering insight into feature importance and splits. Decision trees and ensemble methods like Random Forest (Freitas, 2014).

Gradient-Based Attribution Methods: Leverages partial derivatives to attribute model predictions to specific features, enhancing understanding of feature contributions. Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) (Molnar, 2019).

Residual Analysis: Examines the discrepancies between predicted and actual outcomes, aiding in the identification of systematic errors (Sangeetha & Prasad, 2006). Residual plots help reveal patterns and model shortcomings.

Visualization of Feature Relationships: Utilizes graphical representations to depict the relationships between features and model predictions. Scatter plots and heatmaps for intuitive interpretation (Goldstein et al., 2015).

Permutation Feature Importance Analysis: Evaluates the importance of features by systematically permuting their values and measuring the impact on model performance (Altmann et al., 2010). A rigorous approach to discerning feature importance under various perturbations.

Variable Profiling and Sensitivity Analysis: Explores how model predictions evolve as individual features undergo controlled variations (Montavon et al., 2018). Comprehensive sensitivity analyses, assessing global and local model responses.

Analysis of Intermediate Model Representations: Investigates the transformations and representations within intermediate layers of complex models, shedding light on information processing (Goodrich, 2010). In-depth examination of neural network architectures.

Association Rule Mining: Identifies frequent patterns and associations in the data, contributing to the understanding of feature interactions (Hsieh, 2004). Application of the Apriori algorithm to discover significant rule sets.

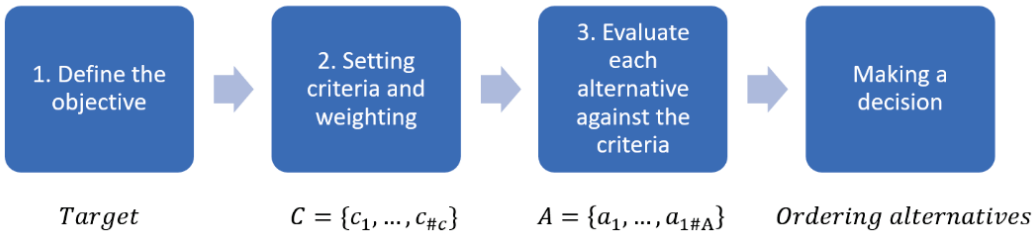
While interpretable machine learning enhances decision-making, challenges must be navigated. Striking a balance between accuracy and interpretability is an ongoing challenge, and understanding the trade-offs is crucial. Additionally, the ethical considerations of interpretable models, especially in sensitive decision-making contexts, require careful attention.

3.2. Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP), introduced by Thomas Saaty (Thomas L. Saaty, 2008), is a decision-making technique designed to address complex and multi-criteria problems. This method involves hierarchically structuring decision factors and systematically comparing available options, applicable across diverse domains such as business, engineering, healthcare, and environmental planning.

Considerations for applying AHP include the number of experts involved in decision-making, varying from a single expert to a group. Engaging multiple experts incorporates diverse perspectives, enhancing the robustness of decisions. The decision environment, classified into structured and unstructured, influences AHP's effectiveness. In structured environments, well-defined and quantifiable criteria facilitate systematic comparison, while unstructured environments require expert judgment and qualitative assessments, Figure 4.

Figure 4. Analytic Hierarchy Process (AHP).



Source: self-elaboration based on (Saaty, 1980).

AHP offers a versatile decision-making approach adaptable to various scenarios, specifically tailored for intricate decision scenarios with multiple criteria (Cid-López et al., 2016). Factors like multiple experts, decision environment, and the number of criteria are crucial for effective AHP utilization, allowing decision-makers to categorize problems and apply appropriate techniques.

3.3. Proposed Model

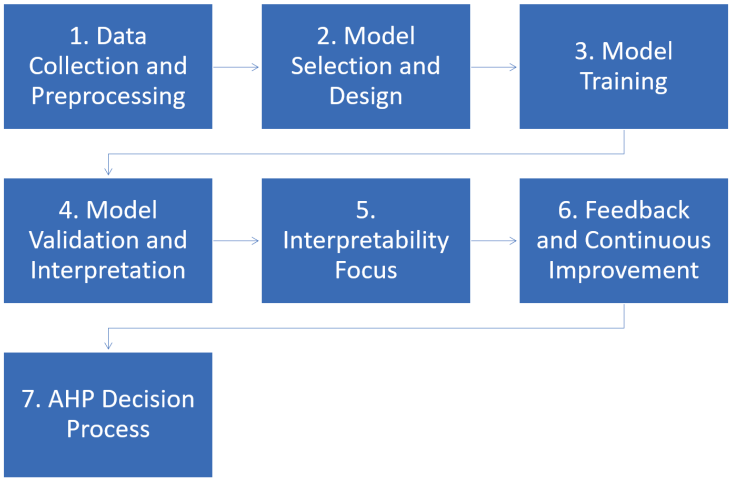
Through the AHP model, we can derive the various weights of the most significant features obtained via XAI and incorporate the human factor into the decision-making process. The decision regarding the hiring of a specific profile is not solely determined by AI. In our case, we conduct an analysis of the model's most crucial characteristics related to employee attrition. These features play a pivotal role in the decision-making process for hiring new employees, emphasizing those essential characteristics to prevent attrition.

In this decision-making process, guided by the insights provided by the AI model, we can apply the Analytic Hierarchy Process (AHP). This structured approach allows us to prioritize and weigh the

identified characteristics, aligning with the model's recommendations to enhance the hiring decision-making process.

The process is detailed below and is depicted in Figure 5.

Figure 5. Proposed Model (AI + XAI + AHP).



Source: self-elaboration based on (Shafique & Qaiser, 2014).

Data Collection and Preprocessing: Begin with obtaining relevant data and preparing it for analysis. During preprocessing, address issues such as outliers, missing data, and normalization.

Model Selection and Design: Choose an appropriate model for the problem and design its architecture. Consider interpretable models, such as decision trees or linear regressions, when possible.

Model Training: Use labeled data to adjust the model's parameters. Validate and fine-tune performance on a validation set.

Model Validation and Interpretation: Evaluate the model on an independent test set to measure its performance. Interpret how the model makes decisions in specific cases.

Model Deployment: Implement the model in an operational environment for real-time predictions.

Continuous Monitoring: Monitor the model's performance in production and adjust as needed.

Feedback and Continuous Improvement: Gather user feedback and adjust the model based on new needs or changes in the data.

AHP Decision Process: Once the most relevant features of the predictive model are identified, apply these features in a decision-making process using the Analytic Hierarchy Process (AHP).

By integrating AHP in the final stages of the process, after selecting the most relevant characteristics from the predictive model, we can enhance the decision-making process with a systematic and interpretable approach.

4. PRACTICAL APPLICATION

The data were gathered from the publicly available IBM HR database (Kaggle HR Analytic Data Set, n.d.), and Table 3 enumerates the features comprising the dataset.

Table 3. Data Set IBM HR.

Features	
Age	Monthly Income
Attrition	Monthly Rate
Business Travel	Number of Companies Worked
Daily Rate	Over18
Department	Over Time
Distance from Home	Percent Salary Hike
Education	Performance Rating
Education Field	Relationship Satisfaction
Employee Count	Standard Hours
Employee Number	Stock Option Level
Environment Satisfaction	Total Working Years
Gender	Training Times Last Year
Hourly Rate	Work Life Balance
Job Involvement	Years at Company
Job Level	Years in Current Role
Job Role	Years since Last Promotion
Job Satisfaction	Years with Current Manager
Marital Status	

Source: IBM HR (*Kaggle HR Analytic Data Set*, n.d.)

After completing the research process, exploratory analysis, and predictive modeling, the model that best fits our observations is XGBoost, with an Accuracy Mean of 85.91. The model is a black-box, and therefore, we apply algorithm interpretability to understand the most relevant features influencing employee turnover.

Figure 6. Features Importance, ELI5.

Weight	Feature
0.0867 ± 0.0119	OverTime
0.0373 ± 0.0113	MonthlyIncome
0.0089 ± 0.0072	DailyRate
0.0056 ± 0.0033	DistanceFromHome
0.0052 ± 0.0010	RelationshipSatisfaction
0.0051 ± 0.0015	JobSatisfaction
0.0047 ± 0.0019	NumCompaniesWorked
0.0043 ± 0.0029	MonthlyRate
0.0041 ± 0.0015	MaritalStatus_Single
0.0041 ± 0.0026	StockOptionLevel
0.0033 ± 0.0020	Age
0.0025 ± 0.0016	EnvironmentSatisfaction
0.0021 ± 0.0008	PercentSalaryHike
0.0017 ± 0.0008	YearsAtCompany
0.0014 ± 0.0010	JobInvolvement
0.0014 ± 0.0016	YearsSinceLastPromotion
0.0012 ± 0.0008	BusinessTravel_Travel_Frequently
0.0010 ± 0.0000	EducationField_Technical Degree
0.0010 ± 0.0017	HourlyRate
0.0008 ± 0.0008	JobRole_Research Scientist
	… 24 more …

After completing the research process, exploratory analysis, and predictive modeling, the model that best fits our observations is XGBoost, with an Accuracy Mean of 85.91. The model is a black-box, and therefore, we apply algorithm interpretability to understand the most relevant features influencing employee turnover.

Once the importance of features is visualized through ELI5 (Molnar, 2019), it can be observed that the determining variables in a selection process would be related to income level. We need to offer a contract that is economically competitive. Age shows an inverse correlation with turnover – the higher the age, the lower the probability of abandonment. The same holds true for marital status; singles exhibit a higher propensity for turnover. Avoiding overloading workers with extra hours is crucial; we seek a positive attitude and satisfaction with the work environment. Additionally, minimizing the commuting distance from home is essential.

This same process can help us assess potential departures among our personnel. We must take care of workers facing work overload or earning below-average incomes, especially those with longer-than-average commuting distances from home. As observed, interpretability provides a powerful mechanism for determining the most crucial criteria in a selection process or in guiding and supporting workers.

The AHP model, once the criteria for personnel selection are defined, allows us to obtain a criterion weighting according to values provided by the predictive and interpretable model. This facilitates the decision-making process, enabling the selection of the best alternative for the vacant position.

For the case study, we conducted the AHP analysis to determine a selection process with three potential candidates, considering the following criteria and alternatives, as well as the following pair-wise comparison matrices.

Figure 7. AHP Analysis, selection process.

Criteria	
C1	Age
C2	Level of Specialization
C3	Distance from Home
C4	Years of Work
C5	Economic Aspiration

	Alternatives				
	C1	C2	C3	C4	C5
A1	30	4	10	10	60.000,00 €
A2	35	3	30	5	65.000,00 €
A3	42	3	12	7	55.000,00 €

Criteria weighting						
	C1	C2	C3	C4	C5	Average Vector
C1	1,00	0,20	3,00	0,33	0,33	9,76%
C2	5,00	1,00	5,00	3,00	3,00	44,28%
C3	0,33	0,20	1,00	0,33	0,33	6,01%
C4	3,00	0,33	3,00	1,00	3,00	23,92%
C5	3,00	0,33	3,00	0,33	1,00	16,03%

C1 = Age				
	T1	T2	T3	Average Vector
T1	1,00	0,33	0,20	10,62%
T2	3,00	1,00	0,33	26,05%
T3	5,00	3,00	1,00	63,33%

C4 = Years of Work				
	T1	T2	T3	Average Vector
T1	1,00	5,00	3,00	63,33%
T2	0,20	1,00	0,33	10,62%
T3	0,33	3,00	1,00	26,05%

C2 = Level of Specialization				
	T1	T2	T3	Average Vector
T1	1,00	3,00	3,00	60,00%
T2	0,33	1,00	1,00	20,00%
T3	0,33	1,00	1,00	20,00%

C5 = Economic Aspiration				
	T1	T2	T3	Average Vector
T1	1,00	3,00	0,33	26,05%
T2	0,33	1,00	0,20	10,62%
T3	3,00	5,00	1,00	63,33%

C3 = Distance from Home				
	T1	T2	T3	Average Vector
T1	1,00	5,00	1,00	47,96%
T2	0,20	1,00	0,33	11,50%
T3	1,00	3,00	1,00	40,55%

After completing the entire process, alternative 1 is considered the optimal choice with a weight of 49.81%, the second option corresponds to alternative 3, with a weight of 33.86%, and finally, alternative 2 with a weight of 16.33%. Therefore, upon concluding the process, it can be asserted that, considering the interpretability applied to employee turnover, along with the AHP method adhering to the recommended criteria for the hiring process, we opt for the most suitable candidate.

5. CONCLUSIONS

In this study, eXplainable Artificial Intelligence (XAI), was utilized to address the issue of employee turnover. The use of interpretable techniques allowed for the identification and measurement of the importance of various characteristics related to this phenomenon.

Through the measurement of feature importance with ELI5, a detailed investigation was conducted on the criteria that could trigger employee turnover. This approach provided a solid foundation for developing a more informed personnel selection process aimed at preventing employee attrition.

The application of interpretability in Machine Learning (ML) algorithms emerges as a crucial component in decision-making. The ability to understand and explain model decisions not only enhances confidence in these models but also facilitates the adoption of more ethical and well-founded decisions.

This work underscores the imperative need to consider interpretability in ML algorithms, not only for its practical utility but also for the associated ethical implications. Opacity in automated decisions can lead to unexpected consequences and a lack of accountability. The adoption of interpretable approaches aligns with the pursuit of transparency and responsibility in algorithm implementation.

The practical application of interpretability in the context of employee attrition highlights its utility in identifying and understanding determining factors. This knowledge is valuable not only for decision-making in human resources but also contributes to talent retention and strengthens organizational policies.

The integration of Explainable Artificial Intelligence (XAI) and the application of Decision Theory, specifically the Analytic Hierarchy Process (AHP), bestow decisions with an enriching collaboration between humans and machines. This synergistic approach not only enhances interpretability in complex decision-making scenarios but also underscores the symbiotic relationship between human insight and machine-driven analyses, contributing to a more informed and ethical decision landscape.

REFERENCES

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Bandyopadhyay, N., & Jadhav, A. (2021). Churn Prediction of Employees Using Machine Learning Techniques. *Tehnicki Glasnik*, *15*(1), 51–59. https://doi.org/10.31803/tg-20210204181812

Bibal, A., Lognoul, M., de Streel, A., & Frénay, B. (2020). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, *0123456789*. https://doi.org/10.1007/s10506-020-09270-4

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, *8*(8), 1–34. https://doi.org/10.3390/electronics8080832

Cid-López, A., Hornos, M. J., Carrasco, R. A., & Herrera-Viedma, E. (2016). Applying a linguistic multi-criteria decision-making model to the analysis of ICT suppliers' offers. *Expert Systems with Applications*, *57*, 127–138. https://doi.org/10.1016/j.eswa.2016.03.025

Climek, M., Henry, R., & Jeong, S. (2022). Integrative literature review on employee turnover antecedents across different generations: commonalities and uniqueness. *European Journal of Training and Development*, *ahead-of-p*(ahead-of-print). https://doi.org/10.1108/EJTD-05-2021-0058

Colomo-Palacios, R., Casado-Lumbreras, C., Misra, S., & Soto-Acosta, P. (2014). Career Abandonment Intentions among Software Workers. *HUMAN FACTORS AND ERGONOMICS IN MANUFACTURING & SERVICE INDUSTRIES*, *24*(6), 641–655. https://doi.org/10.1002/hfm.20509

Duval, A. (2019). *Explainable Artificial Intelligence (XAI) Explainable Artificial*. *April*. https://doi.org/10.13140/RG.2.2.24722.09929

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, *15*(1), 1–10.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Goodrich, M. T. (2010). Data Structures and Algorithms in Python. *Wiley*, *53*(9), 1689–1699. http://arxiv.org/abs/1011.1669%0Ahttp://dx.doi.org/10.1088/1751-8113/44/8/085201

Hall, P. (2022). *Machine Learning for High-Risk Applications*.

Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets*, *32*(4), 2207–2233. https://doi.org/10.1007/s12525-022-00600-9

Hsieh, N. C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, *27*(4), 623–633. https://doi.org/10.1016/j.eswa.2004.06.007

Kaggle HR Analytic Data Set. (n.d.). https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 35–43. https://doi.org/10.1145/3233231

Marín Díaz, G., Galán Hernández, J. J., & Galdón Salvador, J. L. (2023). Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. *Mathematics*, *11*(22). http://doi.org/10.3390/math11224677

Mishra, D. (2013). Review of literature on factors influencing attrition and retention. *International Journal of Organizational Behaviour & Management Perspectives*, *2*(3), 435–445.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247. https://christophm.github.io/interpretable-ml-book

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Perisic, A., & Pahor, M. (2020). Extended RFM logit model for churn prediction in the mobile gaming market. *Croatian Operational Research Review*, *11*(2), 249–261. https://doi.org/10.17535/crorr.2020.0020

Saaty, T. L. (1980). The analytic hierarchy process : planning, priority setting, resource allocation LK - https://ucm.on.worldcat.org/oclc/911278091. In *TA - TT* -. McGraw-Hill International Book Co.

Sangeetha, V., & Prasad, K. J. R. (2006). Deep residual learning for image recognition. *Indian Journal of Chemistry - Section B Organic and Medicinal Chemistry*, *45*(8), 1951–1954. https://doi.org/10.1002/chin.200650130

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. http://www.ijisr.issr-journals.org/

Srivastava, P. R., & Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi- Criteria Decision-Making Approach. *Journal of Global Information Management*, *29*(6), 1–29. https://doi.org/10.4018/JGIM.20211101.oa23

Thomas L. Saaty. (2008). Decision making with the analytic hierarchy process. *Journal of Manufacturing Technology Management*, *26*(6), 791–806. https://doi.org/10.1108/JMTM-03-2014-0020

AN INTEGRATED ETHICS FRAMEWORK FOR EMBEDDING VALUES IN AI

Xenia Ziouvelou, Konstantina Giouvanopoulou, Vangelis Karkaletsis

AI Politeia Lab, SKEL AI Lab, Institute of Informatics and Telecommunications (IIT), National Centre of Scientific Research "Demokritos" (Greece)

xeniaziouvelou@iit.demokritos.gr; vangelis@iit.demokritos.gr; kgiouvano@iit.demokritos.gr

ABSTRACT

In the light of recent years, there are growing concerns about unintended, foreseeable and unforeseen risks with negative, unanticipated consequences that may accompany the rapid evolution of AI technology and its applications. Under the prism of these threats, a growing body of ethical guidelines and principles is developing that appears to adopt a deontological approach focused on rules and duties. This paper aims to address the need for a holistic framework for AI ethics by design; a framework that will augment the current prevalent deontological approach with a virtue-driven approach aiming at values, moral and character dispositions of individuals embedding values in AI systems (at an individual level and organisational level). To this end it provides an integrated approach to AI ethics aiming to broaden the scope of action by embedding virtues, ethos and values in AI by design and explores the practical implementation of the proposed framework.

KEYWORDS: AI ethics, AI ethical principles, virtue ethics, values by design, integrated AI Ethics framework.

1. INTRODUCTION

Artificial Intelligence (AI) is evolving rapidly, becoming a key driver for the digital transformation of our economies and societies, impacting this way the future of humanity, by transforming the lives of individuals and influencing human societies, reshaping patterns of living, working, learning, and interacting. However, while AI can create great opportunities by driving economic and social progress, it also presents complex challenges and potential risks. Risks are related to gender-based or other kinds of discrimination and bias (intentional and unintentional), opaque decision-making, intrusion, social harms for individuals and society, loss of liberty, control and autonomy, in addition to the concentration of power in the hands of a few private actors, among others (UNESCO, 2020; EIGE, 2021). Challenges on the other hand, stem from the great uncertainties that are linked with the alignment of AI systems with human values (AI value alignment) from their design to their use (Han et al., 2022) which is of major concern given that the way we model and design AI may affect the values we are able to embed (Gabriel, 2020; van de Poel, 2020). However, there are other dimensions. The evolution of AI brings about the need to explore deeper the interplay between values and technology design, development, implementation and use, and the role of individuals in realising value sensitive technology; as well as the need to explore new values, which are appropriate to protect the rights of the individual in the light of such an evolution (Ziouvelou et al., 2020).

Triggered by these risks, a growing body of ethical AI guidelines and principles, has emerged over the last few years (Hagendorff, 2020; 2022, EU HLEG, 2019; Whittaker et al., 2018; Campolo et al., 2017; Floridi et al., 2018; IEE, 2019; Jobin et al., 2019; Fjeld et al., 2020 among others) aiming to harness the unintended disruptive potential and complex challenges posed by AI. Numerous guidelines have been launched by governments, scientific or industrial communities as well as civil society representatives, over the past few years aiming to serve as a basis for ethical decision-making in AI design, development, deployment and governance.

However, public debate is already saturated by these ethical guidelines. From a ***macroscopic perspective*****,** this abundance of ethical principles threatens on the one side to overwhelm and confuse and on the other to delay the development of laws, rules and standards that will ensure that AI is socially beneficial (Floridi and Cowls, 2019) or even avoid regulation altogether (Wagner, 2018) in some geographical regions. From a ***microscopic perspective***, the vast majority of these guidelines appear to adopt the 'deontological ethical approach' (Mittelstadt et al., 2019; Hagendorff, 2020), that emphasises duties or rules at an institutional level. At an individual level though, there appears to be a gap, for example in relation to the values, moral and character dispositions of the individuals who create these technologies. A "values-based approach" to technologies is essential acknowledging that societies and technologies mutually shape each other in a reflexive way[1] (Schwab and Davis, 2018).

Business, government and civil society leaders should consider the importance of values and virtue ethical approaches in technological development and deployment as well as guidelines development (Franzke, 2022), in order to seize the opportunities and address the threats that accompany emerging technologies, seen as sociotechnical systems rather than isolated artifacts (van de Poel, 2020). This implies adopting a conscious perspective on technological development that prioritises society's values (Philbeck et al., 2018). As such, **virtue ethics could expand traditional deontological AI ethics** and broaden the scope of action (Hagendorff, 2020, van de Poel, 2020).

The human-centred Artificial Intelligence is considered very important in this effort considering that we are dealing with systems that try to imitate human intelligence, so we have to first study the human being, ourselves, before designing such complex systems and on this basis we can build the framework that we propose. When dealing with the development and implementation of AI technologies, the ethical parameter is fundamental as AI regulation implies choices that reflect our moral values. Ethics should concern the wider institutional and social context in which individual decisions are taken and implemented following Plato's paradigm who tried to understand justice, a moral value that affects institutions and social patterns of organisation, within the broader context of the polity (Tasioulas, 2022). Virtue ethics is a theory that may be of greater value in a holistic framework (Steen et al., 2021; Franzke, 2022) that complements the existing deontological ethics and may lead to beneficial social innovations.

2. METHODOLOGY

This study utilises a multi-phased research design, that combines a meta-level literature review analysis of existing AI ethics review studies with theoretical perspectives of the proposed ethics guidelines. Initially, a secondary literature review focuses on the existing landscape of systematic review studies (meta-analyses) in the field of ethical guidelines for AI, examining whether convergence is emerging in relation to the high-level principles and explore the ethical approaches that have been adopted. This review was conducted during September-December 2023, and we utilised sources/databases such as Scopus, Web of Science, arXiv and Google Scholar in order to select scoping studies in the area of AI ethics. Our analysis revealed 10 relevant and highly cited scoping studies since 2018. Furthermore, the theoretical parameters underpinning these ethical guidelines, were examined aiming to identify gaps and open the discussion about what constitutes ethical AI aiming to provide a complementary perspective based on an integrated, multi-perspective approach. Our analysis is extended by the practical implementation model of an integrated AI ethics framework by embedding values in AI by design.

[1] According to Schwab and Davis (2018), we are the product of our technologies as much as they are products we create.

3. MAPPING OF THE ETHICAL GUIDELINES FOR AI

Globally, the AI governance landscape is characterised by many initiatives taken by several different actors at all levels of government as well as the private sector. Some relate to the regulation of specific AI applications and others are more general principles of AI ethics and policy. Similar initiatives can be observed at a national level, with many countries having promoted their own AI strategies, with explicit references to the international level and global governance of AI (Schmitt, 2022). It is interesting to note that among other actors, supranational such as the UNESCO and the OECD are also present and have been instrumental in shaping global AI governance; and there seems to be a convergence on a certain kind of values and principles of AI, as proposed by the European Commission and the OECD with a focus on trustworthy, human-centred AI (Schmitt, 2022).

Several scholars have focused their research on mapping the guidelines that have been developed concerning ethical AI and provide overviews. Among the first to systematically analyse various ethical AI guidelines, were Floridi et al. (2018) that analysed 47 principles from 6 recommendation papers and derived 5 high level principles - beneficence, non-maleficence, autonomy, justice, explicability. Zeng et al. (2018) collected 27 proposals of AI principles that were explored the issuing bodies of these principles, by grouping them across stakeholder segments (i.e. principles from academia, non-profit and non-governmental organisations, principles from governments, principles from industry). They initially identified a set of a priori selected keywords as key terms belonging to 10 general topics namely: humanity, collaboration, share, fairness, transparency, privacy, security, safety, accountability, AGI/ASI (Artificial General/Super Intelligence) and then examined the coverage of different principles on these 10 topics. Jobin et al (2019) identified through the content analysis of 84 sources, 11 overarching ethical values and principles namely transparency, justice and integrity, non-justice, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability and solidarity and underlined a global emerging cross-stakeholder convergence that promotes the ethical principles of transparency, justice and fairness, non-justice, accountability, privacy.

Fjeld et al. (2020) focused their research on comparing the contents of 36 prominent AI principles documents and their research effort identified a growing consensus around 8 key thematic trends: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. They detected that the most recent documents tend to cover all 8 of them, indicating that there is beginning to be a convergence among these principles. Hagendorff (2020) compiled 22 ethical guidelines based on a literature review, and based on his survey findings several guidelines were repeated across studies with accountability, privacy, justice/fairness which appeared in about 80% of all guidelines. Hagendorff notes that these principles seem to have the least requirements for the development and use of an "ethically sound" AI system and adds that the most frequently addressed ones are those for which technical solutions can be developed or have already been developed. Similarly, the scoping study by Khan et al. (2022), revealed a set of 22 ethical principles and 15 challenges. The most common AI ethics principles were found to be transparency, privacy, accountability and fairness, while at the same time the lack of ethical knowledge and vague principles were presented as the most significant challenges. Validating Khan et al. (2022), the scoping study by Franzke (2022) found similar principles as the most commonly featured in the AI ethics guidelines corpus.

Table 1. Overview of scoping studies (meta-analyses) focusing on AI ethics.

Year	Author(s)	Focus	Study Type*	Type of Review	No. of Guidelines/ studies	No. of ethical principles (in total)	Overarching/ converging ethical values, principles, themes	Overarching/Converging/Common Key ethical principles/themes
2023	Khan et al.	Global	Empirical	-	N/A	21	3	Transparency, accountability, and privacy
	Attard-Frost et al.	Global	Lit. Rev.	Semi-systematic	47	4	4 (**)	A priori classification based on the FAST (fairness, accountability, sustainability, transparency) principles for AI ethics
	Correa et al.	Global (37 countries)	Lit. Rev.	Systematic	200	17	6	Transparency/Explainability/Auditability,Reliability/Safety/Security/Trustworthines, Justice/Equity/Fairness/Non-discrimination, Privacy, Accountability/Liability, Freedom/Autonomy/Democratic Values/Technological Sovereignty
2022	Khan et al.	Global	Lit. Rev.	Systematic	27	22	4	Transparency, Privacy, Accountability, Fairness
	Franzke	Global	Lit. Rev.	Systematic	70	-	4	Transparency, Privacy, Accountability, Safety
2020	Hagendorff	Global	Lit. Rev.	Semi-systematic	22	22	3 (*)	Accountability, Privacy, Fairness
	Fjeld et al.	Global	Lit. Rev.	Systematic	36	47	8	Accountability, Privacy, Fairness & Non-discrimination, Safety & Security, Transparency & Explainability, Human Control of Technology, Professional Responsibility, Promotion of Human Values
2019	Jobin et al.	Global	Lit. Rev.	Systematic SLR	84	11	5	Privacy, Justice &Fairness, Transparency, Non-maleficence, Responsibility
2018	Zeng et al.	Global	Lit. Rev.	Semi-systematic	27	-	10(**)	Accountability, Privacy, Fairness, Humanity, Collaboration, Share, Transparency, Security, Safety, AGI/ASI
	Floridi et al.	Global	Lit. Rev.	Semi-systematic	6	47	5	Beneficence, Non-maleficence, Autonomy, Justice, Explicability

Notes: SLR (Systematic Literature Review), SSLR (Semi-Systematic Literature Review)
* Type of Study: Empirical (bottom-up research), Literature Review (top-down research). (*) Seen by the authors as the minimum requirements for building and using ethically sound AI systems (coverage by the various studies 80%), although no overarching principles are identified. (**) Themes selected by the authors (a priori selection).

Attard-Frost et al. (2023) examined 47 ethics guidelines that were a priori classified based on the FAST principles (fairness, accountability, sustainability, transparency) (Leslie, 2019) of AI ethics. Their findings indicate a disproportional focus on issues of algorithmic decision-making and limited focus on the ethics of AI business practices and decision-making context of the AI systems. However, as the authors indicate, algorithmic design and decision making is only one piece of ethical AI system design; stressing the need for future AI ethics research to focus on issues of business practices as well as business decision-making processes in the development and use of AI systems.

Correa et al. (2023) conducted a meta-analysis of 200 governance policies and ethical guidelines for AI usage published by various organisations worldwide (academic, private sector, civil society, public sector). They identified 17 prevalent principles, and the top 5 were found to be similar to the ones

identified by Jobin et al. (2019), Hagendorff (2020) as well as Fjeld et al. (2020) in relation to reliability-safety-security-trustworthiness. Adopting a different perspective Khan et al. (2023) explored the significance of AI ethics principles and related challenges, via an empirical research study (during 2021-2022) that mobilised 99 randomly selected AI practitioners and lawmakers, from 20 countries across five continents. Their findings indicate that transparency, accountability, and privacy are the most critical AI ethics principles, while the lack of ethical knowledge, the non-existence of legal frameworks, and the lack of monitoring bodies were the most common AI ethics challenges.

Our analysis of some key systematic and semi-systematic scoping studies in the area of AI ethics (Table 1) revealed that in all studies the focus is global, however the vast majority of scoping studies utilise literature reviews (top-down research) stemming from existing research in the area, and only one, Khan, et al. (2023), conducts bottom-up empirical research.

In relation to the overarching ethical values and principles under examination there appears to be a high degree of similarity**.** As it can be seen in Table 2, the 4 top scoring ethical principles recurring across these scoping reviews appear to be: *Accountability, Privacy (7 out of 9 principles) and Fairness, Transparency* (6/9). This emerging convergence, identified in more than half of the studies, could be seen as the minimal requirement for *'ethically sound'* AI systems. Nonetheless, further thematic analysis would reveal conceptual variations in relation to the definition, interpretation, justification and domain of application among others aligned with the finding of other studies in the area (Jobin, et al., 2019). Table 2 illustrates the mapping of the occurrence (based on definitions) of the overarching principles of AI ethics identified in the above scoping studies. The studies of Zeng et al. (2018) and Attard-Frost et al. (2023) were excluded due to the a priori classification of the later and focus on general AI topics rather than principles of the former.

Table 2. Mapping the occurrence of overarching AI ethics principles.

	Overarching AI Ethics Principles								
Review Studies	Transparency / Explainability	Accountability/ Explicability	Privacy	Fairness/ Justice	Safety/Security/ Non-Maleficence	Responsibility	Human Values/ Democratic Values	Human control of technology/ Autonomy	Humanity/Beneficence (people & Planet)
Khan et al 2023	1	1	1						
Correa et al. 2023	1	1	1	1	1		1	1	
Khan et al. 2022	1	1	1	1					
Franzke, 2022	1	1	1		1				
Hagendorff 2020		1	1	1					
Fjeld et al.2020	1	1	1	1	1	1	1	1	
Jobin et al. 2019	1		1	1		1			
Floridi et al., 2018		1		1	1			1	1
TOTAL	6	7	7	6	4	2	2	3	1

It is clear from the above that the growing concerns raised by the rapid evolution of AI systems on a wide range of potential ethical issues have led to dozens of principles/guidelines for addressing ethical aspects. However, several of them are too abstract creating a difficulty in how they can be translated into concrete designs for AI systems (Prem, 2023, Attard-Frost et al., 2023). Furthermore, most of the proposed guidelines focus on algorithmic design and decision-making processes, as pointed out by Attard-Frost et al., (2023), which however constitutes only one piece of ethical AI system design and development, leaving out "many other business practices and business decision-making processes implicated in the development and use of an AI system" (p. 25) involving the organisational and

individual developer practices. This is aligned with the prevailing view that majority of the proposed guidelines appear to adopt a deontological ethics approach (Mittelstadt et al., 2019; Zhou et al., 2020; Hagendorff, 2020; Ziouvelou et al., 2024 – while different views also exist (i.e., Franzke, 2022 utilitarian tendencies)), emphasing duties and rules at an institutional level, without explicitly stating how they should be applied in a particular context. However, there appears to be a gap, in relation to values and virtue ethics perspectives that brings to the fore the human factor involved in the creation and use of these systems (Ziouvelou et al., 2024). As such, the contribution of the virtue ethics approach to the current deontological AI perspectives, would broaden the scope of action (Hagendorff, 2020, van de Poel, 2020, Franzke 2022) as it would expand the current ethical perspectives. Highlighting this way, the need for the creation of a more holistic framework consistent with an integrated ethical approach, that would contribute to their practical implementation.

4. THE NEED FOR AN INTEGRATED AI ETHICS APPROACH

Considering the deontological approach, the question that arises is whether all these principles and guidelines can actually lead to ethical decisions for people involved in the design and development of Artificial Intelligence systems. McNamara et al. (2018) that examined the extent to which ethical guidelines can serve as a basis for ethical decision making for software engineers concluding that ethical guidelines do not change their behaviour. It is observed that current AI ethics principles show weaknesses, such as, for example, that the lack reinforcement mechanisms and that deviations from codes of conduct have no or very little consequences (Rességuier & Rodrigues, 2020; Hagendorff, 2022); or that ethical concerns related to AI business practices are sometimes not adequately addressed or addressed in AI ethics guidelines (Attard-Frost et al., 2023). The development of these principles/guidelines clearly indicate a significant positive step towards a more ethical AI but poses the question whether this effort can be more effectively achieved in the context of a more holistic approach that incorporates human character and ethical values alongside the principles and guidelines.

Having briefly reviewed the macroscopic level aspects of the ethical guidelines for AI, we now focus on what is observed at an individual level where there seems to be a gap related to values and individual character dispositions. Virtue ethics approach contribute towards the creation of a more holistic framework that extends the existing deontological approach. The existence of virtues can guide decision-making and design principles for AI (Neubert and Montañez, 2020); for this reason, we advocate that the virtue ethics framework that prioritises values and character could complement the deontological approach, providing a more robust basis for ethical AI. Afterall, regulation forms that are or could be proposed for AI, whether self-regulation or social and legal rules, lead to choices that reflect the hierarchy of moral values (Tasioulas, 2022) and the existence of moral values defines a virtuous and moral person who makes ethical decisions.

This need has also been stressed by Hagendorff (2020) that highlights the importance for a transition from a more deontologically oriented ethics based on adherence to principles and rules to an ethical approach that focuses on the virtues and dispositions of the personality and as he highlights the formation of a moral character, in terms of virtue ethics, presupposes the cultivation of virtues in families, schools, communities, and businesses. It is therefore necessary to create an interconnection between **values and technical implementations** in the field of AI, examining both the individual and the organisational aspects in addition to the algorithmic ones, aiming towards achieving a balance between the two approaches which may be challenging (Hagendorff, 2020).

For the virtue ethics approach to be feasible, great emphasis should be placed on the existence of **moral values** in order to overcome **value-activity gaps**. Where value-activity gaps signify the inconsistency between people's moral values/self-perceptions and their actual behaviour (Hagendorff,

2022). In the context of AI, value-activity gaps can occur at an organisational level when terms related to ethics are used in formal reports, but the reality reveals unethical practices (Loughran et al., 2009).

The key question that arises in relation to values, however, is whether we can embed values in AI and if so how. Since the 20th century, technology ethicists and philosophers have debated whether technology is value-neutral or value-laden (Wynsberghe, 2020). On the one hand, the neutrality thesis argues that AI systems are in themselves neutral and depend on the user for acquiring a moral status as either good or bad and, on the other hand, the embedded values thesis (values embedded in their design) argues that it is possible to detect tendencies within a system/software that enhance or undermine certain ethical values and norms (Nissenbaum, 1998; 2001; Brey, 2010). We follow the embedded values thesis and support the view that to reduce as many as possible imminent problems related to unethical practices, we should focus on embedded values (i.e., values that have been intentionally/unintentionally, and successfully, embedded) in AI systems. This implies that the system must be designed to comply with those values and to actually respect or further them (van de Poel, 2020, IEEE, 2019). To achieve responsible technological innovation, developers and organisations must be endowed with values, implement those values, and be held accountable for them in an approach that integrates design with our values, monitoring and intervening in value-driven technological development (van den Hoven et al., 2015). It is therefore important to focus on the character and moral values brought by a designer and a developer of such sociotechnical AI systems as well as the organisation values, so that the embedded values are for good in the direction of an ethical AI that respects moral values through transparent processes.

Embedding values is therefore a matter of vital importance, and, as Nissenbaum (2001) points out, if we ignore values, we run the risk of handing over this important dimension to chance or some other force. She also adds that it is necessary that scientists and engineers broaden the criteria they use to evaluate systems so that social, ethical and political criteria can be successfully integrated.

5. AN INTEGRATED APPROACH TO AI ETHICS

In this paper we support the view that understanding human values is a crucial step in the design, development and use of responsible and trustworthy AI (Han et al., 2022, Ziouvelou et al., 2024). Our perspective is anchored in the analysis of Hagendorff (2020) and van de Poel (2020) and the need for a holistic framework for AI ethics that will present a model that augments the traditional, prevalent deontological approach of AI ethics (Mittelstadt et al., 2019; Hagendorff, 2020) with a virtue ethics approach pertaining values, moral and character dispositions (van de Poel, 2020, Ziouvelou et al., 2024). As it is quite difficult to predict exactly what kind of consequences future innovations will bring to society, many scholars (Shannon Vallor, 2016; Hagendorff, 2020; van de Poel, 2020) argue for virtue ethics as an appropriate framework for the development of emerging technologies, which, by enabling new forms of behaviour, are expected to influence human values in the future (Steen et al., 2021).

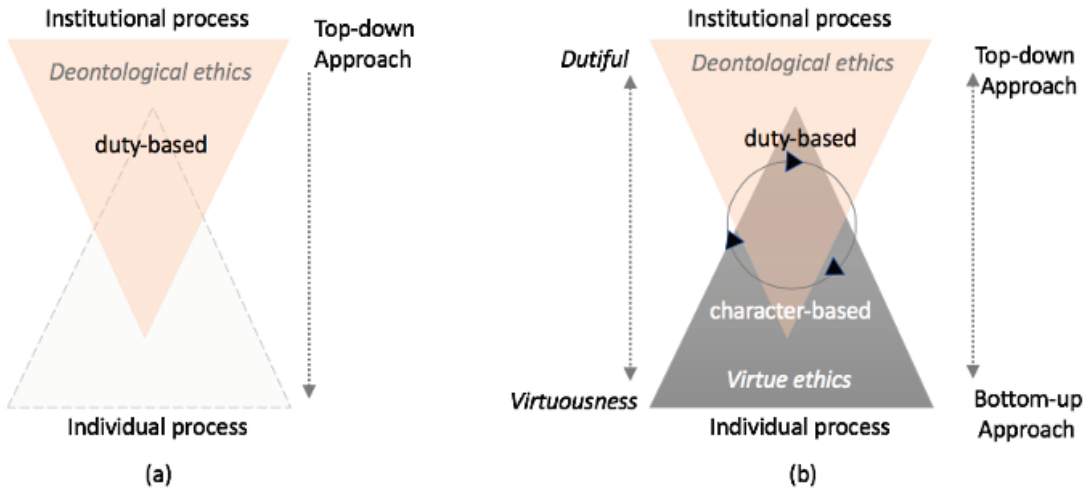
Virtue ethics is an approach that *treats virtue and character as the primary subjects of morality, focuses on the character rather than on actions*, and thus, contrasts with other ethical frameworks/theories (Statman, 1997) that give a primary role to the consequences of actions (consequentialist ethical frameworks) or to principles/rules and moral duty (deontological ethical frameworks). Virtue ethics can be traced back to ancient Greek philosophy, to Socrates, Plato and Aristotle. The *Aristotelian* approach to ethics may have much to contribute to research on ethical and trustworthy AI, since ethics should not be seen as a limitation to human well-being. The Aristotelian concept of *eudaimonia* teaches us that the ideas of morality and personal prosperity are connected, and therefore, according to Aristotle, a person in prosperity possesses moral virtues. Aristotle, in his *Nicomachean Ethics*, expresses the view that the pursuit of *eudaimonia* can only be properly exercised in the typical human

community – the *polis (πόλις)*. He argued that man is by nature *zoon politikon* (ζῷον πολιτικόν, political animal), destined to live in an organised political society and that virtues contribute to living in a polis and promoting the welfare of the people (Steen et al., 2021). The word virtue denotes moral excellence, and it indicates the fundamental qualities that allow people to excel and thus contribute to social well-being. Moral wisdom is found in a particular kind of character (ethos), reinforced by nurturing and social environment. Ethos means ‘‘virtue” in the Aristotelian sense, denoting the internal values that characterise an individual. All human beings are born with the potential to become morally virtuous and practically wise but in order to achieve these goals, they must develop appropriate habits during childhood and then, when their reason is well developed, acquire practical wisdom (Kraut, 2022). We note that in the Aristotelian approach to ethics, particular value is placed on human judgment (practical wisdom-*phronêsis* – *Aristotelian compass* - that enables its possessor to detect the best thing to do on a particular circumstance) as well as and this can be a strong counter-argument to the view that understanding of the good can only be achieved through the implementation of principles and rules in the field of AI.

So far it seems that a small group of technology giants is influencing the ethical tone in AI with part of the culture in which it is embedded being technocratic and decisions about the “values” encoded in AI being made by elites without any substantial democratic control (Tasioulas, 2022). Considering this reality and along with the virtue ethics approach, Tasioulas (2022) underlines the need for an alternative *humanistic* approach to AI which focuses on the human engagement with ethics-not always foreseen by the dominant approach. The key elements of this approach include a commitment to a plurality of values aiming at human well-being and at basic components of morality, such as justice and the common good, with an emphasis on the importance of the processes we adopt to achieve social objectives, not just the outcomes. Such an approach places participation and dialogue at the heart of the understanding of human well-being and ethics provided that such participation will be active, giving ordinary citizens the possibility and opportunity to put forward their views in dialogue with others (Tasioulas, 2022). Such a virtue-based approach, that encapsulates both phronesis, dialogue and participation, could certainly broaden the scope of ethical AI (Franzke, 2022).

As stated previously, the difference between deontology and virtue ethics is that while the former is based on normative rules with universal validity, the latter examines what constitutes a good person or character. Virtue focuses on the development of positive characteristics of the actor (Hagendorff, 2020) and is essential if we consider that the **values** that every individual engineer embraces should be the starting point for responsible and ethical behaviour (Hersh, 2012). Values contribute to evaluation in terms of goodness and badness, while concepts such as duties and rules are used to determine the rightness (or wrongness) of actions (van de Poel, 2020). The virtuous actor embraces values of goodness; therefore, the ethics of virtue is directly related to moral values. As stated by Annas (2011) virtue is a disposition of character, which is not impermanent, to act reliably and virtue requires commitment to values; it involves the orientation of the person to something that the person considers valuable. In an effort to make the implementation of existing AI ethics initiatives successful and effective, the insights of moral psychology should be included, since until now, when talking about AI ethics, the psychological processes that limit the goals and effectiveness of ethics programs are not taken into account (Hagendorff, 2020).

Figure 1. (a) Deontological AI ethics approach. (b) Integrated AI ethics framework (Deontological & Virtue ethics).



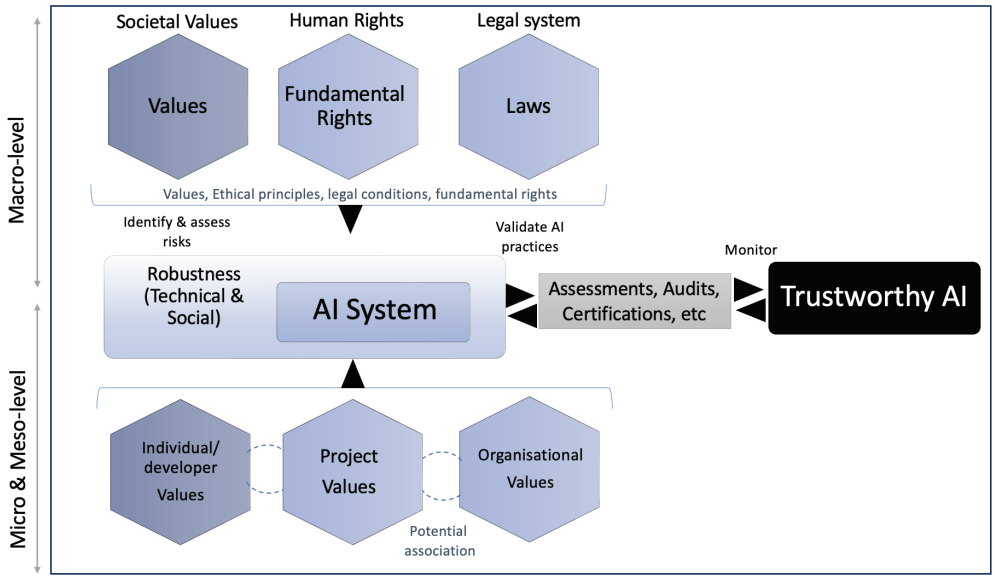
Source: Ziouvelou, Karkaletsis & Giouvanopoulou (2024)

In line with such a holistic perspective to AI ethics, Ziouvelou et al., (2024) proposed a model that adopts an integrated approach. This model augments the existing deontological, duty-driven approach which includes principles and rules (i.e., ethical AI code) (Figure 1a – deontological approach) with a virtue-driven approach including values and moral personality traits (i.e., value AI code) (Figure 1b-integrated approach). Thus, broadening the scope of action by infusing virtues and ethos into existing AI ethics principles. This will augment the current prevalent approach with considerations that relate to the values, moral and character dispositions of individuals (human embedding values in AI systems at an individual level and organisations-organisational level).

However, the practical implementation of such a framework is crucial for achieving its goals. In order to facilitate the implementation of such an integrated AI ethics framework and achieve responsible technological AI innovation by design we should focus on complementing existing macroscopic AI ethics assessments processes for Trustworthy AI (aligned with the HLEG (2019) definition of Trustworthy AI -lawful, ethical, robust (technically, socially)) with the values that are embedded in AI systems, across the **micro-level** (individual) and **meso-level** (project/organisational) in addition to the **macro-level** (society), achieving this way a value-driven socio-technical development aligned with van den Hoven et al. (2015). Such an implementation model (Figure 2) will thus broaden the scope of action by infusing virtues and ethos in existing deontological AI ethics by design.

The consideration of values at different levels is crucial for the preservation of the variety and diversity of ethical values and it is worth mentioning that a similar approach, in terms of different levels of values consideration, has been proposed by Walz & Firth-Butterfield (2019) for the development of an AI governance regime, following a graded governance model for the application of ethical considerations to AI systems. This model incorporates 4 levels of ethical core values - inalienable, constitutional, group-specific, and individual.

Figure 2. An implementation model for an integrated approach to AI ethics.



Considering that AI systems are not able to understand the notion of human values (Neuhäuser, 2015), lack emotional abilities (Sharkey, 2017) and are human made (Hakli & Mäkelä, 2019), all concern should be about humans involved in designing, developing and deploying AI systems. Given that humans can be considered full moral agents (Dignum, 2018; Hakli & Mäkelä, 2019), virtue ethics suggests that we do not treat AI systems as autonomous, equal to humans, but rather as assistive companions (Maes, 1995; Savulescu & Maslen, 2015; Voinea et al, 2020) as intelligent tools (Balkin, 2017) to serve human needs in a responsible way. Furthermore, just as individuals can demonstrate character, an organisation can also embody character, as a collection of individuals (Moore, 2005). Existing research indicates that organisations that demonstrate virtue by exhibiting character, tend to experience positive benefits both internally as well as in the marketplace (Cameron et al., 2004; Sosik et al., 2012; Neubert and Montañez, 2020). As such micro and meso-level virtue-ethics related considerations and assessments could indicate a potentially fruitful way forward.

CONCLUSIONS

Our analysis presented a high-level overview of some of the prominent systematic and semi-systematic studies in AI ethical guidelines. A growing body of ethical AI guidelines has emerged, triggered by the anticipated and unanticipated AI risks, aiming to harness the unintended disruptive potential and complex challenges of AI, but this abundance of ethical principles has wider implications beyond the direct and intended ones. This mapping revealed high-level alignment both in relation to their thematic focus and their ethical approach. The emerging high-level ethical principles were found to be *Accountability, Privacy, Fairness and Transparency*. However, conceptual variations of these principles in relation to the interpretation, justification and their domain of application reveal further diversities. Furthermore, aligned with the findings of these studies, a number of principles are significantly underrepresented which gives rise to additional risks.

Aligned with the latter perspective we also considered the theoretical perspectives of these guidelines, which exhibits a high degree of similarity as the vast majority of these guidelines appear to adopt the *deontological ethical approach* that is a duty-driven institutional level approach. Anchored in the need for a holistic framework for AI ethics that will augment this prevalent approach with a virtue-driven approach focusing on values, moral and character dispositions, we expanded upon the integrated approach to AI ethics (Ziouvelou et al., 2024) aiming to broaden the scope of action by infusing virtues

and ethos in AI ethics. Our analysis proceeded one step deeper by showing how this holistic framework could be implemented in practice by integrating the values into the design of the AI systems, considering both the micro-level (individual developer values) and the meso-level (project/organisational values) as well as the existing trustworthy AI perspectives as defined by the AI HLEG. This study aims to highlight the need for a holistic approach to AI ethics guidelines and assessment processes and practices. Showing that virtue ethical approaches can successfully complement and expand the current ethical perspectives.

A number of limitations, on an empirical and theoretical front, can be identified. The empirical aspects related to the incompleteness of the utilised data sources. The theoretical aspects relate with shortcomings linked with virtue ethical approaches and their associated application perspectives (Hursthouse & Glen, 2023). Future research is needed to explore the practical application and assessment of an integrated ethics approach in the AI context.

REFERENCES

Annas, J. (2011). *Intelligent Virtue*. Oxford University.

Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2023). The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics, 3*(2), 389-406.

Balkin, J. M. (2017). The three laws of robotics in the age of big data. *Ohio State Law Journal, 78*(5), 1217–1241.

Brey, P. A. E. (2010). Values in Technology and Disclosive Computer Ethics. In: Floridi L, (Ed). *The Cambridge Handbook of Information and Computer Ethics,* 41–58. Cambridge University Press.

Cameron, K.S., Bright, D., Caza, A. (2004). Exploring the relationships between organizational virtuousness and performance. *American Behavioral Scientist, 47* (6), 766-790.

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute at New York University.

Corrêa, N.K., Galvão, C., Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., Massmann, D., Mambrini, R., Galvao, L., Terem, E., Oliveira. N. (2023). Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance. Patterns, 4(10): 100857.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1-3.

EIGE (2021). Artificial intelligence, platform work and gender equality. *European Institute for Gender Equality.* Luxembourg: Publications Office of the European Union.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet & Society*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al. (2018). AI4People – an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4), 689–707.

Floridi, L. and Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*.

Franzke, A.S. (2022), "An exploratory qualitative analysis of AI ethics guidelines", Journal of Information, Communication and Ethics in Society, 20(4), 401-423.

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *Minds and Machines, 30,* 411-437.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines,* 1-22.

Hagendorff, T. (2022). A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philos. Technol.* 35, 55.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.

Han, S., Kelly, E., Nikou, S. & Svee, E.Q. (2022). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & Soc* 37, 1383–1395.

Hersh, M.A. (2012). Science, Technology and Values: Promoting Ethics and Social Responsibility, *IFAC Proceedings Volumes, 45*(10), 79-84.

HLEG, (2019). A definition of AI: Main capabilities and disciplines, High-Level Expert Group on Artificial Intelligence of the European Commission. *Downloaded,* 1, 2019-12.

Hursthouse, R. and Glen, P. Virtue Ethics, *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.).

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.*

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature.*

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., … & Akbar, M. A. (2022, June). Ethics of AI: A systematic literature review of principles and challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (383-392).

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., ... & Abrahamsson, P. (2023). AI ethics: an empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems*.

Kraut, R. (2022). *Aristotle's Ethics.* The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman

Leslie, D. (2019). Understanding AI ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute.*

Loughran, T., McDonald, B., & Yun, H. (2009). A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics, 89*(S1), 39–49.

Maes, P. (1995). Artificial life meets entertainment: Lifelike autonomous agents. *Communications of the ACM, 38*(11), 108–114.

McNamara, A., Smith, J., Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In G. T. Leavens, A. Garcia, C. S. Păsăreanu (Eds.) *Proceedings of the 2018 26th ACM joint meeting on 288rtifici software engineering conference and symposium on the foundations of software engineering—ESEC/FSE 2018* (1–7). New York: ACM Press.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). *Explaining explanations in AI*. In Proceedings of the conference on fairness, accountability, and transparency—FAT* '19, 1–10.

Moore, G. (2005). Corporate character: Modern virtue ethics and the virtuous corporation. *Business Ethics Quarterly, 15* (4), 659-685.

Neubert, M. J., and Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. Business Horizons, 63(2), 195-204.

Neuhäuser, C. (2015). Some Skeptical Remarks Regarding Robot Responsibility and a Way Forward. In C. Misselhorn (Ed.), *Collective Action and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation* (131–146) Springer.

Nissenbaum, H. (1998). Values in the design of computer systems. *Computers and Society,* 38-39.

Nissenbaum, H. (2001). How computer systems embody values. *Computer*, 34(3), 120-119.

Philbeck, T., Davis, N. and Engtoft Larsen, A. M. (2018). White Paper. Values, Ethics and Innovation Rethinking Technological Development in the Fourth Industrial Revolution. *World Economic Forum.*

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. AI and Ethics, 1-18.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society, 7*(2), 1–5.

Savulescu, J., & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide* (79–95). Springer.

Schmitt, L. (2022). Mapping global AI governance: a nascent regime in a fragmented landscape. *AI Ethics* 2, 303–314.

Schwab, K., & Davis, N. (2018). Shaping the Future of the Fourth Industrial Revolution. Geneva: World Economic Forum.

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210–216.

Sosik, J.J., Gentry, W.A., Chun, J.U. (2012). The value of virtue in the upper echelons: A multisource examination of executive character strengths and performance. *The Leadership Quarterly, 23* (3) (2012), 367-382.

Statman, D. (1997). Introduction to Virtue Ethics. *Virtue Ethics: A Critical Reader*. Edinburgh University Press.

Steen, M., Sand, M. & Poel, I. (2021). Virtue Ethics for Responsible Innovation. *Business & Professional Ethics Journal.*

Tasioulas, J. (2022). Artificial Intelligence, Humanistic Ethics. *Daedalus,* 151, 232-243.

UNESCO (2020). Artificial Intelligence and Gender Equality, Key findings of UNESCO's Global Dialogue.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. New York: Oxford University Press.

Van den Hoven, J., Vermaas, P.E. & van de Poel, I. (2015). *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains.* 10.1007/978-94-007-6970-0.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409.

Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The Internet as Cognitive Enhancement. *Science and Engineering Ethics, 26*(4), 2345–2362.

Wagner B. (2018). *Ethics as an escape from regulation: From "ethics-washing" to ethics-shopping?* In Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M. (Eds.). Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen (pp. 84-89). Amsterdam: Amsterdam University Press.

Walz, A., & Firth-Butterfield, K. (2019). AI Governance: A Holistic Approach to Implement Ethics into AI. World Economic Forum.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O. (2018). AI Now report 2018.

Wynsberghe, A. V. (2020). *Artificial intelligence: From ethics to policy. Panel for the Future of Science and Technology.* European Parliamentary Research Service, (STOA), Scientific Foresight Unit (PE 641.507).

Zeng, Y., Lu, E., Huangfu, C. (2018). Linking artificial intelligence principles. arXiv (pp. 1–4).

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S. & Savage, S. (2020), "A Survey on ethical principles of AI and implementations", In 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 3010-3017.

Ziouvelou X., Karkaletsis, V., Giannakopoulos, G., Nousias, A. & Konstantopoulos, S. (2020). Democratising AI: A National Strategy for Greece. NCSR Demokritos White Paper.

Ziouvelou X., Karkaletsis, V. & Giouvanopoulou, K. (2024). Embedding Values in AI by Design: An Integrated Framework. *21st International Conference on the Ethical and Social Impacts of ICT - ETHICOMP 2024*, Spain.

ARAB CULTURE AND PRIVACY OF SOCIAL MEDIA: A THEORETICAL STUDY

Ala Ali Almahameed, Mario Arias-Oliva, Orlando Lima Rua, Mar Souto-Romero

Social and Business Research Lab, Universitat Rovira i Virgili (Spain), Marketing Department, Complutense University of Madrid (Spain), Porto Accounting and Business School, Politechnic of Porto (Portugal), School of Business and Communication, Universidad Internacional de La Rioja (Spain)

a.mahameed82@gmail.com; mario.arias@ucm.es; orua@iscap.ipp.pt; mar.souto@unir.net

ABSTRACT

Same as the rest of the world, Arabs are using social media networks for different purposes, such as communicating with friends and relatives, shopping, seeking jobs, and so on. Social media platforms continue to attract Arabs in various countries across the region, and simultaneously, the number of fixed and mobile internet users has also increased. Despite that social media has become an integral part of our lives, it comes with its own set of privacy concerns. Some of the most common social media privacy issues include social media phishing scams, hacking and account takeovers, shared location data used by stalkers and predators, data mining leading to identity theft, privacy "loopholes" exposing your sensitive information, employers or recruiters evaluating you based on your posts, doxing leading to emotional distress or physical harm, cyberbullying and online harassment. this research aims to understand the role that morality and ethics that are driven by Islam and Arab culture are playing in regulating users' interaction with others over social media websites if associated with national laws that govern such interaction.

KEYWORDS: Social media, Privacy, Arab World, Ethics.

1. INTRODUCTION

The development of the internet and social media has dramatically altered the way people communicate and share information, creating new opportunities for social interaction, business, and entertainment. Social media platforms like Facebook, Twitter, and Instagram allow people to communicate with each other in a real-time, despite their physical location. Furthermore, social networks have expanded and diversified their offerings. For example, Facebook has acquired Instagram and WhatsApp, and it has launched features such as Facebook Live and Facebook Marketplace. Likewise, Snapchat has introduced new features such as Snap Map and augmented reality filters. As well, LinkedIn has introduced new tools for job seekers and recruiters, and Twitter has expanded its focus on news and live events. In general, social media platforms make it easy to share news, articles, photos, and videos with friends, family, and followers (Dizikes, 2020). Overall, social media networks have evolved to become an integral part of daily life for many people, with a wide range of uses and features. According to Smart Insights, the number of social media users globally increased from 4.2 billion in January 2021 to 4.62 billion in January 2022. Furthermore, there are currently 4.76 billion social media users worldwide, which is slightly less than 60% of the global population. Precisely, the growth of social media users has slowed down in 2023, with the addition of 137 million new users in 2023, equivalent to a modest annual growth rate of 3%. It appears that social media platforms have been the major beneficiaries of the shift to digital in the world of digital advertising. In fact, data shows that global spending on social media advertising has more than doubled since the coronavirus outbreak, reaching USD 226 billion in 2022 ("Five Countries", 2023). As a result, people post 500 million tweets, share over 10 billion pieces of Facebook content, and watch over a billion hours of YouTube video during the day (Chaffey, 2023).

Same as the rest of the world, Arabs are using social media networks for different purposes, such as communicating with friends and relatives, shopping, seeking jobs, and so on. For example, social media networks are widely used by elites and everyday citizens to discuss politics and achieve political goals. In this context, a study by National Endowment for Democracy found that social media has become a powerful tool for political mobilization in the Arab world. Researchers have also used social media data to study political behaviour in the Arab world (Siegel, 2019). Furthermore, a report by Pew Research Center claimed that social media played a role in the Arab uprisings that began in 2010 (Brown, Guskin & Mitchell, 2012).

After the discussion initially focused on the "right to communicate" and the "right to knowledge and access to information", it has shifted towards protecting human rights from the risks of new media, especially the right to privacy (the sanctity of private life). Users' data, personal information, and communications are stored, collected, and electronically processed not only by network management and specialized companies but also by anyone with the ability and means to do so, including hackers, service providers, governments, and other entities. In addition to the opportunities provided by these means of communication, they have enabled users to violate each other's privacy and publish what they want under pseudonyms, without any regulations or ethics governing these tools and their users, particularly in the absence of international standards (Alfaisal & Sayed, 2017). Hence, this research aims to understand the role that morality and ethics that are driven by Islam and Arab culture are playing in regulating users' interaction with others over social media websites if associated with national laws that govern such interaction. In this way, the researchers believe that the research will introduce an overall image that could be used to make social media platforms a safe place for users, especially while interacting with others. Also, it can represent a starting point for future research to empirically determine the factors that impact social media privacy in Arab culture.

2. SOCIAL MEDIA INDICATORS IN ARAB WORLD

Social media platforms continue to attract Arabs in various countries across the region, and simultaneously, the number of fixed and mobile internet users has also increased. According to the annual report published by Global Media Insight (2023) on the latest global internet usage figures, the number of internet users in Egypt reached 75.66 million users in January 2022, in which internet penetration rate was 71.9% of the total population (102.3 million) at the beginning of 2022. As a result, Internet users in Egypt increased by 1.4 million (+1.9%) between 2021 and 2022. On the other hand, these figures reveal that 29.55 million people in Egypt did not use the internet at the beginning of 2022, meaning that 28.1% of the population remained unconnected to the internet at the start of that year. With regards to social media users in Egypt, in January 2022, the number of social media users reached approximately 51.45 million, which is equivalent to 48.9% of the total population. It's important to note that the number of social media users may not represent individual users, as there could be multiple accounts held by the same person. Therefore, the actual number of social media users in Egypt may be lower. In Lebanon, the number of internet users at the beginning of January 2022 was approximately 6.01 million, equivalent to 89.3% of the total population (6.825 million). This means that 716.9 thousand people in Lebanon did not use the internet at the start of 2022, indicating that 10.7% of the population is not connected to the internet in the country. There were 5.06 million social media users in Lebanon by January 2022, representing 75.2% of the total population. The number of social media users in Lebanon increased by 690 thousand, a growth of 15.8% between 2021 and 2022. In Morocco, the number of internet users reached 31.59 million at the beginning of that year, with an internet penetration rate of 84.1% of the total population (36.91 million). This indicates an increase in the number of users by 1.2% compared to the previous year. However, 15.9% of the total population in Morocco is still not connected to the internet, though these numbers may be influenced by the impact of COVID-19 on search activities.

In Algeria, there were 27.28 million internet users, representing 60.6% of the total population (43.85 million) with an increase of 1.8 million users (+7.3%) between 2021 and 2022. In Iraq, 20.58 million people used the Internet, with a penetration rate of 49.4% of the total population (40.22 million) at the beginning of 2022. Currently, there are 28.35 million social media users in Iraq, representing 68.0% of the total population. This marks a significant jump of 13.4% compared to the previous year. In Qatar, the internet penetration rate reached 99.0% of the total population (2.881 million) at the beginning of 2022. This substantial figure means that only 1.0% of the population is not connected to the Internet. The usage of social media platforms in Qatar also corresponds to this high rate, reaching 99.8% of the total population. In Saudi Arabia, there were 34.84 million internet users, accounting for 97.9% of the total population (34.81 million) at the beginning of 2022. The number of social media users in Saudi Arabia stands at 29.30 million, representing 82.3% according to the latest global figures (“Arab World”, 2022). Figure 1 shows the increased use of social media in some Arab countries between 2014 and 2019. This growth in social media usage can be attributed to various factors, including the development of ICT, the increasing availability of affordable smartphones, high internet penetration rates, and the growing popularity of social media platforms among Arab youth (Alammary 2022).

For instance, figure 2 shows the number of cellular users in Jordan, in addition number of internet and social media users.

Figure 1. Social Media Followers in Arab Countries. Source: (Social Media, 2019).

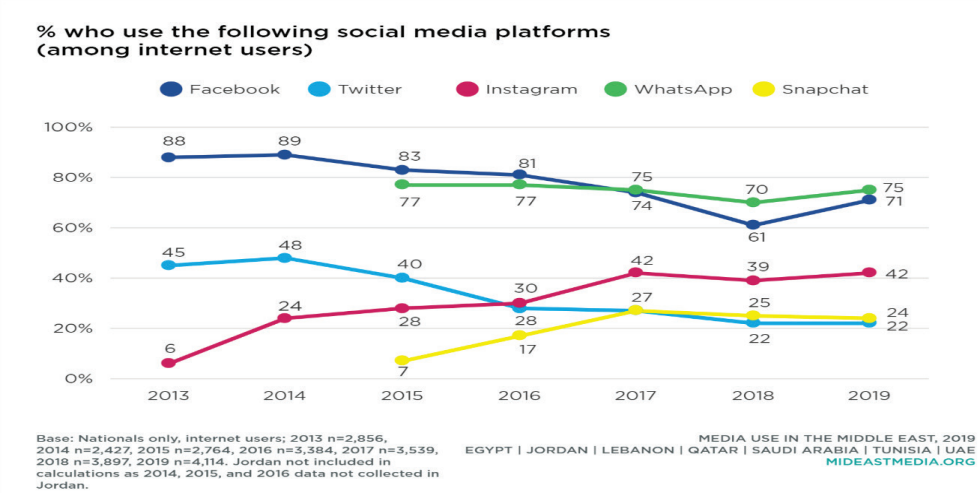
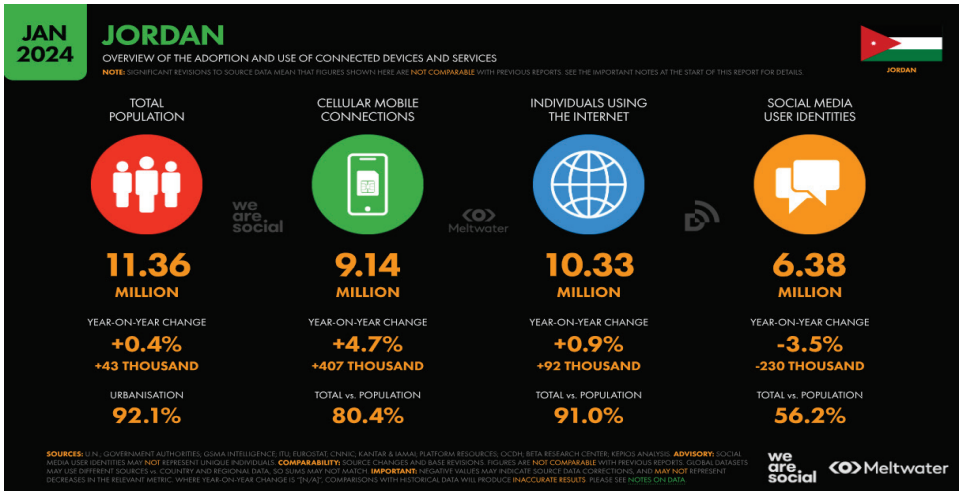


Figure 2. Use of connected devices and services in Jordan. Source (KEMP, 2024).



3. PERCEPTION OF PRIVACY IN ARAB CULTURE

The tremendous advancement witnessed in communication technologies and information systems has led to a new understanding of the right to privacy, linked to the concept of individuals' ability to control the flow of their information, particularly in the face of the risks associated with its collection and processing in digital environments unique to these technologies. This concept has been coined to prevent misuse or abuse of these technologies against individuals, ensuring legal protection for personal information, which includes identifiable details like names, addresses, phone numbers, and more (Al-Bashtawi, 2017).

Informational privacy encompasses the rules governing the collection and management of data specific to individuals, such as those related to identity cards, financial records, and medical information, as well as the confidentiality of telecommunications, both over the Internet and through email. The focus of informational privacy is to safeguard personally processed data, considering the increasing restrictions on its use under legally defined circumstances. Such data is prone to various violations, either through unauthorized commercial exploitation, government surveillance, or malicious theft, resulting in potential harm to the individuals concerned (Oraby,2022).

Despite that social media has become an integral part of our lives, it comes with its own set of privacy concerns. Some of the most common social media privacy issues include social media phishing scams, hacking and account takeovers, shared location data used by stalkers and predators, data mining leading to identity theft, privacy “loopholes” exposing your sensitive information, employers or recruiters evaluating you based on your posts, doxing leading to emotional distress or physical harm, cyberbullying and online harassment. Furthermore, social media platforms such as Facebook, Twitter, and Instagram collect and store massive amounts of personal data from users, including their location, search history, and social interactions (Zhang et al., 2020). This data is used to deliver personalized content and advertising to users, which can be beneficial for some individuals. However, concerns arise when this personal data is misused, shared without consent, or exploited for profit. For instance, millions of Facebook users' data was harvested without their consent and used for political advertising (Cadwalladr & Graham-Harrison, 2018). Moreover, Children are at risk of online grooming, cyberbullying, and exposure to inappropriate content, while individuals with disabilities may be more susceptible to online scams and phishing attacks (Kargupta & Kumar, 2021).

The right to privacy is considered one of the fundamental constitutional rights that bind the natural person in their human capacity. This right precedes the existence of the state itself. Therefore, the private lives of individuals have obtained constitutional and legal protection in all countries around the world (Jaber, 2021). With social media's emergence and widespread use, the concept of privacy and personal life has become different. In the past, many actions and activities were considered private, such as family and emotional life, and these details were not shared with anyone on the internet, considering them sacred. People used to refrain from disclosing information or details about their day to anyone other than close family or friends. As for the pictures, they were extremely private. Nowadays, many people share their personal information without hesitation. There is a trend where individuals publish their private and personal details without being asked, assuming an audience is ready to receive and engage with them (Kadwani, 2022).

Privacy concerns have significantly increased, especially after the infamous Cambridge Analytica incident, where the data of Facebook users was leaked, and the Equifax data breach, which was exploited in the U.S. elections. These incidents raised numerous concerns about the privacy of the information on social media platforms (Stier et al., 2020). The issues extended to the extent of surveillance and the use of such platforms for spying on individuals, as evidenced by the accusations against the CEO of TikTok, Cheng Cho, by the U.S. Congress. They affirmed that the program was used

for espionage on various institutions and for leaking user data. Additionally, accusations were made that the Chinese government was allowed to use the platform for spying on user data in the United States. The matter did not stop there but extended to the exploitation of social media sites to promote dubious websites that advocate against morals, as seen in the case of "Hanin Hossam" in Egypt. She utilized her followers to promote content contrary to societal norms and traditions, leading the public prosecution to level serious charges, including human trafficking (Ibrahim & Taha, 2020).

The Arab world is a rapidly growing market for social media platforms, with a high rate of social media adoption among its population. Understanding privacy concerns in the region is crucial for social media companies that wish to tap into this market and build trust with their users (Khawla F Ali et al., 2020). Also, privacy is a fundamental human right, and social media privacy concerns in the Arab world are no exception. In this context, the previous research focused on the effect of cultural restrictions on individuals' motivation, users' attitudes, intentional behaviour, and social media's actual use, in addition to understanding the purposes, benefits, and risks of its use (e.g. (e.g., Askool, 2013; Abaido, 2020; Asiri et al., 20217). Also, some of the previous research investigated the role of Islam and cultural traditions in constructing norms around privacy (e.g., Abokhodair et al., 2017; Shehu et al., 2017). However, there are limited studies that investigate the impact of culture and governing laws in mitigating the negative impact of privacy while using social media websites. In particular, understanding and respecting the privacy boundaries of other users while interacting with them on these platforms.

The basis for morality and ethics in the Arab world, especially for Muslims, is primarily derived from the Qur'anic text and the verbatim quotes from the Prophet Muhammad, known as the Sunnah. These sources constitute the foundation of Sharia law, which not only shapes the judicial system but also establishes societal norms and expectations for behaviour. The concept of privacy is highly valued and is an integral part of daily life in the Arab world. The Holy Quran emphasizes the importance of seeking permission before entering someone's home as a means of safeguarding privacy and maintaining the sanctity of the house and body. The act of knocking on a door three times before entering is intended to prevent unintentional intrusion on one's private space, especially in situations where one may be in a state of undress or with their spouse or family. Failing to seek permission and entering without consent can lead to an invasion of privacy (Norah & Sarah, 2016).

The Arab world has a unique cultural and social context that affects the way people view privacy. For instance, people in the Arab world may value privacy differently than people in the Western world. Understanding these cultural differences is crucial in designing effective privacy policies that are sensitive to the needs and expectations of the Arab population (Askool, 2013). Besides, studying social media privacy concerns in the Arab world is required to understand cultural differences, political implications, business opportunities, and human rights issues. It is essential also to develop effective privacy policies and protect the privacy of individuals in the region (Norah & Sarah, 2016). Furthermore, social media has played a crucial role in the Arab Spring uprisings that took place in the region. These events have highlighted the importance of social media platforms as tools for political mobilization and expression of dissent. In fact, privacy concerns in the Arab world are not just about protecting individual rights, but they also have significant political implications (Abokhodair et al., 2017).

4. CONCLUSION

Controlling privacy on social media platforms poses a significant challenge, especially given the widespread use of these platforms and the multitude of entities associated with user interactions. As the usage of these platforms is notably increasing in the Arab world, all countries must establish

comprehensive laws and regulations to govern this usage and ensure the protection of privacy for all users. Therefore, there is a necessity to study the factors that ensure the protection of personal privacy for users of social media platforms in the Arab world, considering the uniqueness of Arab culture and its inherent elements. If associated with appropriate legislation, these elements can contribute to regulating the secure use of social media platforms and limiting privacy violations.

In this context, researchers and public institutions play a crucial role in conducting empirical studies to define these regulations. They need to identify factors that enhance the concept of respecting privacy and those that may encourage privacy violations. This effort aims to help legislators formulate laws capable of regulating privacy on social media platforms, maintaining them as a secure environment for all users, regardless of their gender or age.

REFERENCES

A. Siegel, Alexandra (2019). Using Social Media Data to Study Arab Politics. APSA MENA Politics. Retrieved from https://apsamena.org

Abaido, G. M. (2020). Cyberbullying on social media platforms among university students in the United Arab Emirates. International journal of adolescence and youth, 25(1), 407-420.

Abokhodair, N., Abbar, S., Vieweg, S., & Mejova, Y. (2017). Privacy and social media use in the Arabian Gulf: Saudi Arabian & Qatari traditional values in the digital world. The Journal of Web Science, 3.

Abokhodair, N., & Vieweg, S. (2016, June). Privacy & social media in the context of the Arab Gulf. In Proceedings of the 2016 ACM conference on designing interactive systems (pp. 672-683).

Al-Bashtawi, Saad (2017). Constitutional protection of information privacy. Jordanian Journal of Libraries and Information, 52(3).

Alammary, J. (2022). The impact of social media on women's empowerment in the Kingdom of Bahrain. Gender, Technology and Development, 26(2), 238-262.

AlFaisal, Abdulameer, & Sayed, Esraa, (2017). Privacy Violation on Social Media Platforms. ALBAHITH ALALAMI, 9(36), 213-240.

Ali, K. F., Whitebridge, S., Jamal, M. H., Alsafy, M., & Atkin, S. L. (2020). Perceptions, knowledge, and behaviours related to COVID-19 among social media users: cross-sectional study. Journal of medical Internet research, 22(9), e19913.

Asiri, E., Khalifa, M., Shabir, S. A., Hossain, M. N., Iqbal, U., & Househ, M. (2017). Sharing sensitive health information through social media in the Arab world. International Journal for Quality in Health Care, 29(1), 68-74.

Askool, S. S. (2013). The use of social media in Arab countries: A case of Saudi Arabia. In Web Information Systems and Technologies: 8th International Conference, WEBIST 2012, Porto, Portugal, April 18-21, 2012, Revised Selected Papers 8 (pp. 201-219). Springer Berlin Heidelberg.

Brown, Heather, Guskin, Emily, & Mitchell, Amy (2012). The Role of Social Media in the Arab Uprisings. Pew Research Center. Retrieved from https://www.pewresearch.org

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The guardian, 17(1), 22.

Chaffey, Dave (2023, June 07). Global Social Media Statistics Research Summary 2023. Smart Insights. Retrieved from https://smartinsights.com

Dizikes, Peter (2020, September 24). Why social media has changed the world — and how to fix it. MIT News Office. Retrieved from https://news.mit.edu

GMI Blogger (2023, March 06). Saudi Arabian Social Media Statistics 2023. Global Media Insight. Retrieved from https://globalmediainsight.com

Ibrahim, Taha, & Ahmed (2023). Contact behaviour through social media platforms and its relationship to their awareness of the privacy of their digital data. Arab Journal of Media and Communication Research, 2023(42), 453-512.

Jaber, Emad Al-Din Ali Ahmed (2021). Professional and legal controls to protect individual privacy in Arab media legislation. Scientific Journal of Journalism Research, 2021(21), 345-433.

Kadwani, Shereen (2022). The protection rules of the right to privacy on social networking sites: An analytical study. Journal of Media Research, 60(2), 903-948.

Kemp, Simon (2023, January 28). Digital 2023 Deep-Dive. DATAREPORTAL. Retrieved from https://datareportal.com

Kemp, Simon (2024, February 23). Digital 2024: Jordan. DATAREPORTAL. Retrieved from https://datareportal.com/reports/digital-2024-jordan?rq=Jordan%202024

Oraby, S. S. (2022). Criminal liability for violating information privacy via social networking sites: a comparative study in Egyptian and French law. Journal of Law and Emerging Technologies, 2(2), 213-267.

Shehu, M. I., Othman, M. F. B., & Osman, N. B. (2017). The social media and Islam. Sahel Analyst: Journal of Management Sciences, 15(4), 67-80.

Social Media. (2019). Retrieved from https://www.mideastmedia.org/survey/2019/chapter/social-media/

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2020). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. In Studying Politics Across Media (pp. 50-74). Routledge.

Zhang, D., Zhou, L., & Lim, J. (2020). From networking to mitigation: the role of social media and analytics in combating the COVID-19 pandemic. Information Systems Management, 37(4), 318-326.

CHAT GPT: HAS ITS POTENTIAL ARRIVED TO ENHANCE THE NEW WAY OF TEACHING AND LEARNING? A CASE STUDY IN AVIATION STUDIES

Ana María Lara Palma, Rafael Brotóns Cano

Universidad de Burgos (Spain), Antolín S.A. (Spain)

amlara@ubu.es; rafael.brotons@antolin.com

ABSTRACT

Teaching and learning are two concepts that are intrinsically linked. The excellence of the latter depends on the innovation of the former. Therefore, the resources that teachers use to update each discipline are the guiding thread towards quality didactics. And, in this line of innovation, it is possible to find the usability of digital tools. As a current example, artificial intelligence (AI) in higher education has contributed as a supporting element whose benefits have been indisputable so far. But everything evolves, and computer systems have been adapting to an increasingly faster market committed to user satisfaction. This new scenario opens the possibility of analysing the benefits and limitations of its use in the classroom from a bidirectional perspective, that of the teacher and that of the student. Therefore, the aim of this study is to analyse the ChatGPT tool usability in the lectures carried out in aviation studies (Commercial Pilot for Passenger and Cargo Transport Degree at Burgos University) by conducting two surveys (one for teachers and one for students) in paper form. Findings allow to address the potential this recent resource spread to enhance the new way of teaching and learning.

KEYWORDS: ChatGPT, emerging technology, aviation studies, academic innovation, ethics.

1. INTRODUCTION

The basic mandates and lines of action in the field of education established in the European Higher Education Area indicate that, “it is the responsibility of the universities to ensure that the studies are innovative and original, incorporate lines that favour the development of the professional career, take into account the importance of inclusion and diversity, serve for social use and the consequent academic support is given to achieve excellence” (European University Association). Additionally, The European Commission in its Digital Education Action Plan (2021-2027) sets the objective of readjusting education and training to the digital age. It highlights two main guidelines. Firstly, to promote the development of a high-performing digital education ecosystem and, secondly, to enhance digital skills and competences for the digital transformation (Lara-Palma et al., 2022).

Artificial intelligence AI in higher education has contributed as a supporting element whose benefits have been indisputable so far. But everything evolves, and computer systems have been adapting to an increasingly faster market committed to user satisfaction. ChatGPT, a repository of content generated by artificial intelligence has arrived as an assistant to higher education where interaction is done through a chatbot which provides detailed and precised answers (Kocón et al., 2023); furthermore, adding a challenging language and generation tasks in the form of conversation (Wu et al., 2023). This new scenario opens the possibility of analysing the benefits and limitations of its use in the classroom from a bidirectional perspective, that of the teacher and that of the student. Therefore, the aim of this study is to analyse the ChatGPT tool usability in the lectures by with the following question: is ChatGPT a resource that reinforces acquisition of learning in the classroom? The next section facilitates a theoretical framework to set up a scientific approach to the main concern mentioned above.

2. A THEORETICAL APPROACH TO CHATGPT RESOURCE

In addition to understand the perceptions of this supportive resource, it is necessary to introduce a brief overview of the concept. Based on a pure holistic view, "Artificial Intelligence (AI) frequently refers to teaching machines that perform tasks that mimic human intelligence. (…) The application of AI in education provides an opportunity to break physical barriers as learning materials are now accessible online. Different studies have demonstrated the transition of AI in education from conventional computers to embedded systems such as robots. AI in education goes beyond the normal functions of computers and involves the collaboration of different professionals, including data scientists, product designers, linguists, cognitive scientists, psychologists, and education experts. AI surpasses the conventional understanding of various technological applications in education" (Hashem et al., 2024). Therefore, looking at how individuals and institutions make use of it, it is possible to recognize that ChatGPT "could provide high-level societal and ethical benefits. However, it also raises significant ethical concerns across social justice, individual autonomy, cultural identity and environmental issues" (Stahl et al., 2024).

In this sense, students need to "be aware of the limitations of language models and critically evaluate their output. Educators should embrace and adapt these tools, ensuring they actively verify the reliability and accuracy of generated material" (Zirar, 2023).

Quoting Nam and Bai (2023), "one of the utmost values of teaching and learning in higher education is to educate students to develop academic integrity, helping them cultivate ethics and morality in academic writing in a classroom setting".

Barrett and Pack (2023) go further and take into consideration parameters to measure the perception of the use of GenAI chatbot (representation of user prompts and output from ChatGPT) in six tasks of learning (writing process) such as brainstorming, outlining, writing, revising, feedback and evaluating. All in all, Chan and Lee (2023) point out the necessity of "combining technology with traditional teaching methods to prove a more effective learning experience".

To date, no many studies have gone trough the main concerns and benefits in aviation studies, therefore, the next subchapter analyse this issue from a real case scenario.

3. A CASE STUDY IN AVIATION STUDIES

The model of the European Higher Education Area (EHEA), the Bolonia Plan, brought a series of changes at the university institutions as acquire competences, among others, within the students throughout their academic studies. Since then, many changes were set up; for teachers, a new challenge for sharing knowledge (intranets, social nets, new technologies, among others); for students, a new scenario of learning (acquisition of transversal competences, face up the paradigm of new industrial market approaches and closer tutorial paths for supporting job hunting). Both the teaching improvements developed by the teaching staff to achieve more inclusive and accessible classes, as well as the new modes of learning that students have had to address, have served to achieve better rates of access to the labor market. But the technology never stops and it means a new paradigm of emerging resources as ChatGPT. The use of these resources has been extended to all types of disciplines, and, of course, has reached Aviation studies. In order to understand the philosophy of the research and its findings, it is required to point out the most relevant highlights of the studies. This research applies directly to students of the Degree in Commercial Pilot for passengers and cargo transport. The University of Burgos incorporated in the academic year 2022-2023 this new discipline in accordance with art.36 of Royal Decree RD822/2021, with a Frozen EASA Airline Transport Pilot License ATPL and a Commercial Pilot License CPL. In three Academic Courses, graduate students are

able to work as pilots with an official international flight license on regular airlines and work on ground positions thanks to the competences acquired within the 204 ECTs that are offered.

In the first academic course, contents endorse specific Integrated Airline Transport Pilot License (ATPL) subjects and flight instruction practices. Lectures take place in the university facilities whilst instrumental activities, flight simulator training (Airbus A320 and Boeing 737) and flight instruction take place at the Burgos Airport. In the second academic course, contents endorse remaining ATPL subjects and air navigation subjects. In the third academic course, students complete their competences and skills by attending a wide range of aeronautical subjects, simulator sessions and flight instruction tutorials. Throughout the degree, students have the chance of enrich their wisdom in aeronautical competences by attending seminars with experts and visiting aerospace enterprises. All lectures are provided in English language. All in all, students have to develop a wide range of skills not only for examinations but for team practical's as well. It is in this specific field where attendants need to use several resources, among others, ChatGPT. Therefore, former, in which sense this tech is for students feasible, effective, educational, ethical, make good or, by contrast, harm to soft skills capabilities, inequality, lack of ethics, unfair for authority? And, later, to what extent is for lecturers enhancing, user-friendly, secure, reliable or by contrast, not well seen, not acceptance? All these concerns play a reality just arrived to the academic institutions. Through this research it is possible to explore real data by going significantly beyond this current and disruptive discourse.

4. METHODOLOGY

The present study aims to explore the effectiveness of addressing ChatGPT as a technological resource in the learning process; moreover, results imply consequences (benefits and concerns), impact in the educational environment and an ethical dissertation concerning its usability in the class. To do this, it has been carried out two surveys (educators' (n=5) and students' (n=22)) in paper form during the Fall Semester 2023-2024. The questionnaires consist in 10 questions posed to ChatGPT. All questions were answered by assigning a Linkert-like scale from 1 (leftmost option) to 5 (rightmost option). Both, students and teachers completed the surveys in an anonymous way in order to prevent unintended data recollection and to encourage all of them to answer in the most honest way possible.

Students are currently enrolled in the first and second academic year of Commercial Pilot for Passenger and Cargo Transport Degree at Burgos University and professors work at academic institutions and private companies. Students come from 16 different countries and lecturers are involved in teaching their subjects in English. This activity took place between October and November 2023 at the university facilities. A reduce sample of the student´s questionnaire is included. Each question has been assimilated to a representative boundary/drawback and additionally it has been tested a sentiment analysis (like emotion recognition) (see Table 1). Some questions concern an ethical approach, from the perspective of plagiarism when writing, not well acceptance, no emotion recognition, or isolation. Another set of questions refer to knowledge and comprehension and other skills, such as communication and self-reflection. The information gathered incorporates quantitative data. Given the limited number of participants, the analysis of quantitative data is limited to the mean analysis and elements profile.

The obtain results provide a basis for a fundamental discussion of whether ChatGPT is a useful digital resource that can enhance learning (for students) and teaching (for teachers). Moreover, can provide customized strategies and approaches to students' characteristics and needs (Crompton, et. al., 2023). It has undoubtedly been a cultural impact with multiple implications for cybersecurity and education, among several other disciplines.

Table 1. Student´s Survey.

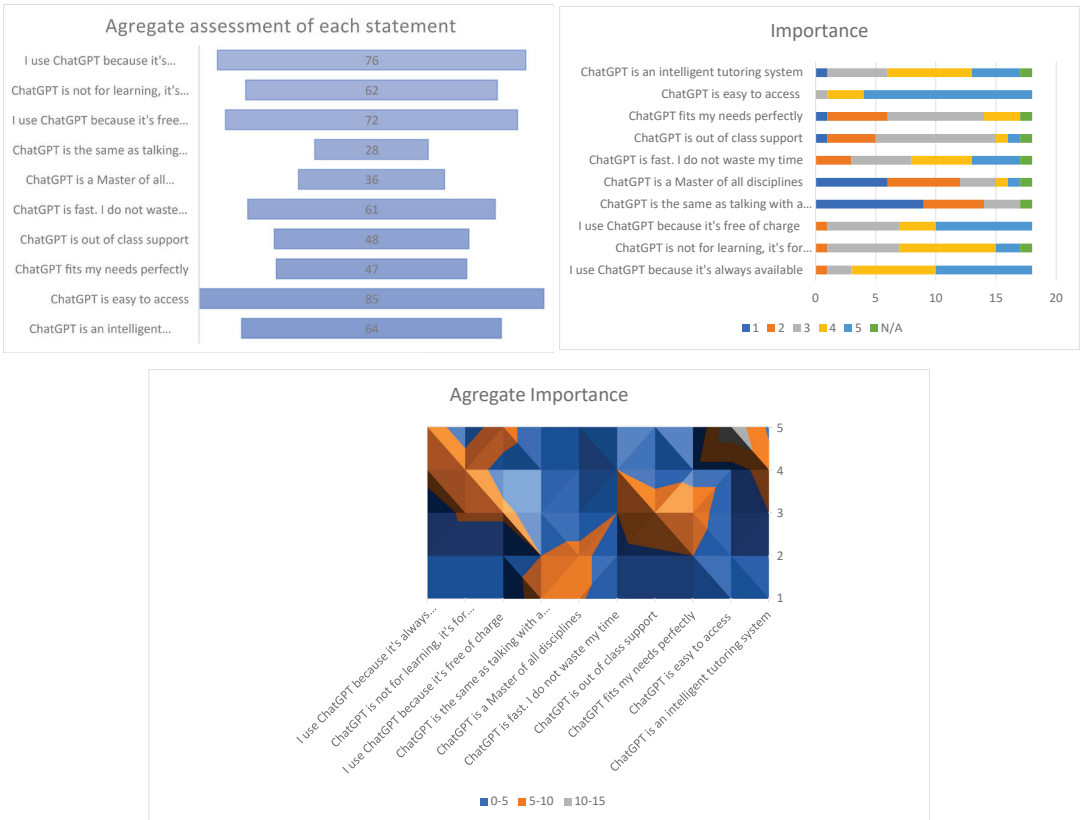
Boundaries	Drawbacks
ChatGPT is always available	Plagiarism must be considered
ChatGPT is entertaining	Doubts about authenticity
ChatGPT is free of charge	It is not well seen
ChatGPT is the same as talking with a professor	No emotion recognition
ChatGPT is a Master of all disciplines	I no longer read books or articles
ChatGPT is fast. I do not waste my time	I can ask in English or any other language
ChatGPT is out of class support	There are no figures or tables
ChatGPT fits my needs perfectly	I don't need to attend tutorials
ChatGPT is easy to access	Less interaction with my classmates
ChatGPT is an intelligent tutoring system	Less independent

Source: Self-elaboration and based on Fawaz (2023)

5. RESULTS

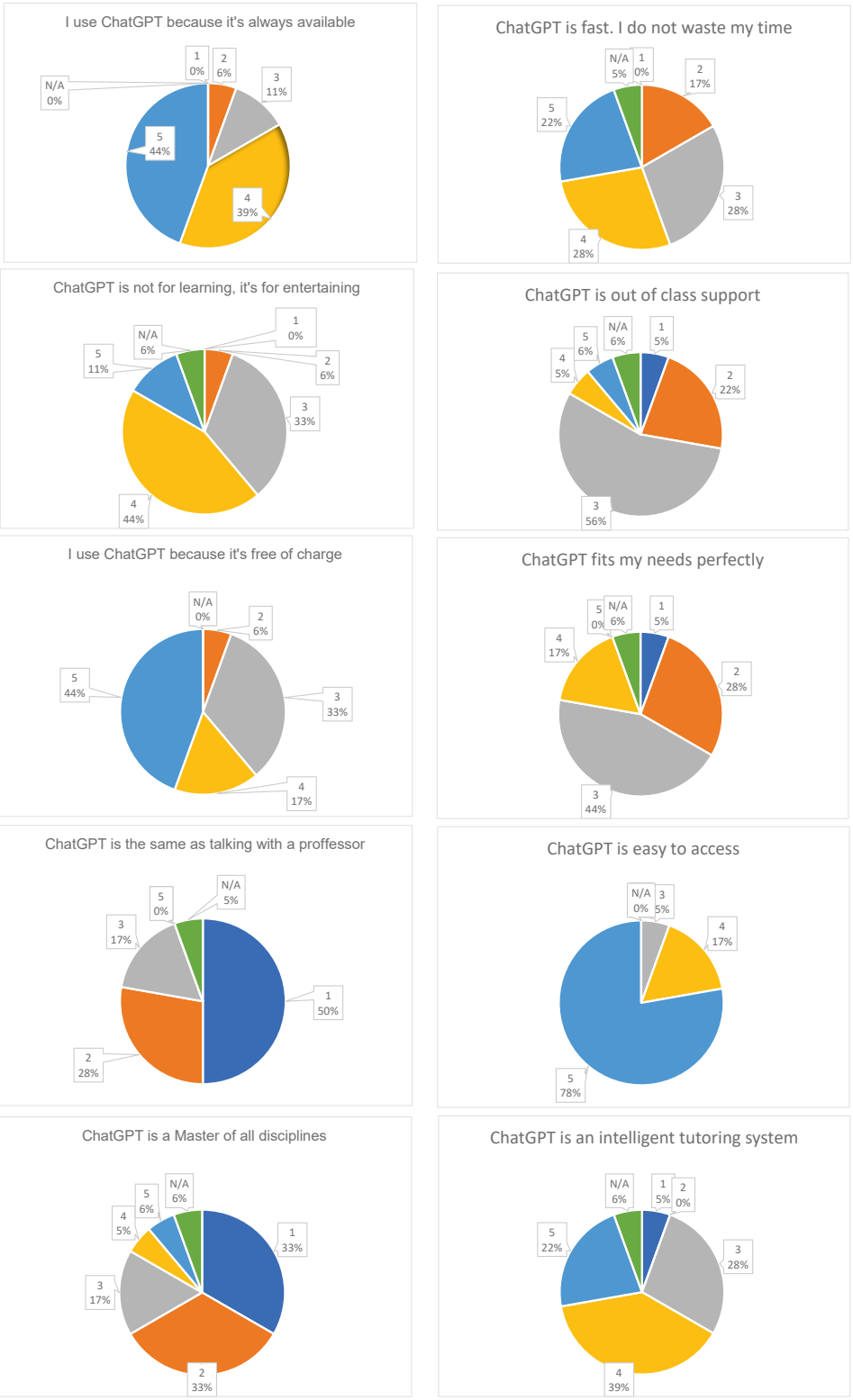
The tornado graph (Figure 1) indicates a predominance of the parameters that most define the new digital generations, which are immediacy and optimization of resources (time, repositories, etc.). Thus, availability and ease of use are the most appreciated parameters to access this application. It is followed by the fact that it is freely accessible. Once these attractive points have been overcome, the respondents still do not show excessive confidence in the contents of the application. In fact, it still seems to be unclear whether its use is due more to its practical usefulness to expand knowledge or as a mere novel attraction with which to experiment. When using the free version, students only have access to updates until January 2022, so it may be understandable that it is not trusted to replace the figure of the teacher.

Figure 1. Survey findings-Aggregate (students).



On the other hand, it is seen in Figure 2, that the students' responses are very fragmented. In fact, for the same question, students' scores cover all marks, including those of DK/NA (Don not Know/No Answer). This reflects the fact that the tool is not firmly established, it is not yet known by all users and each one is in the process of knowing, approaching and shaping its use according to their particular needs. Unlike tools like Excel and Powerpoint, ChatGPT is in a prior maturity process, adapting to the user. A clear reflection of what artificial intelligence is.

Figure 2. Survey findings (students).

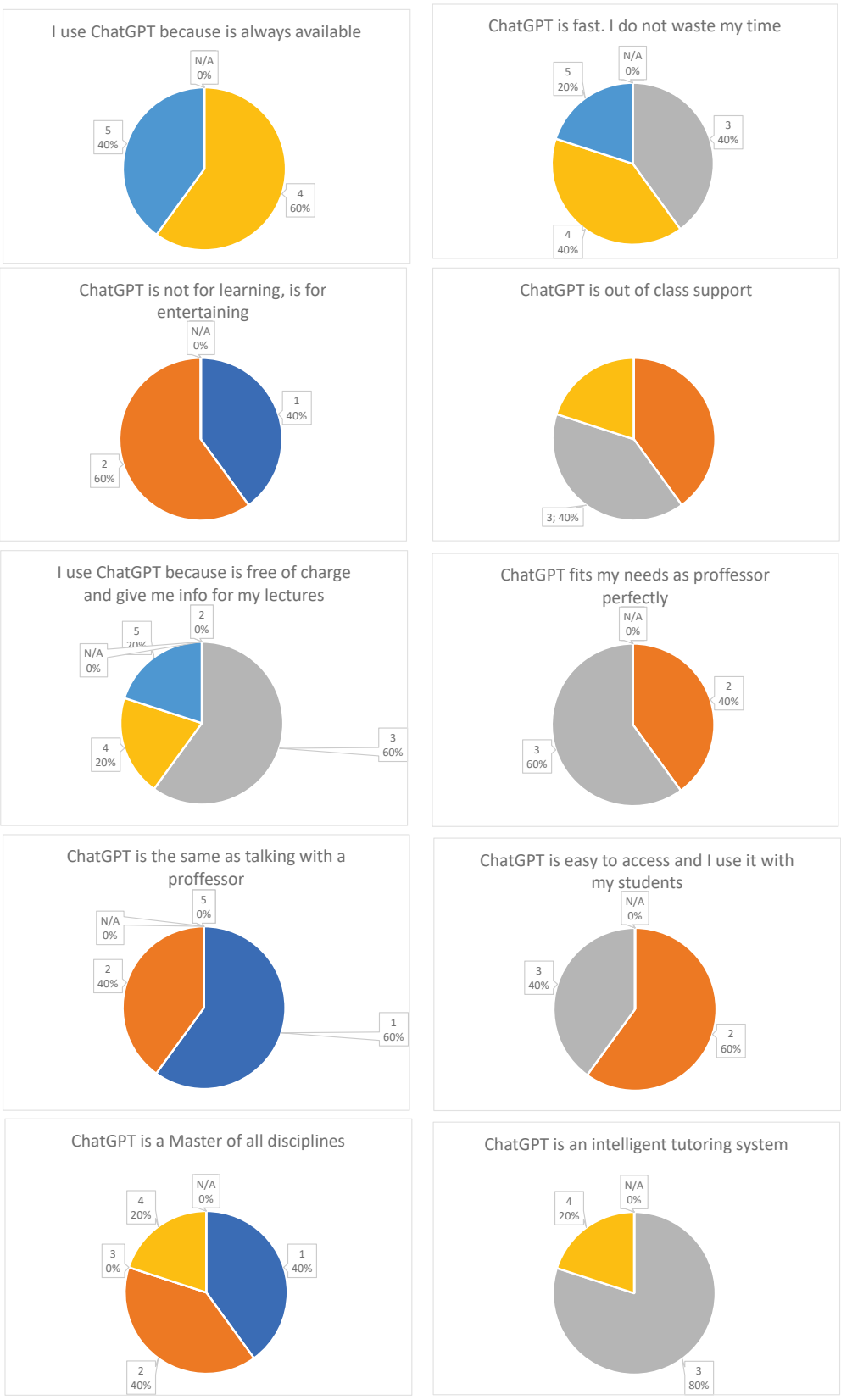


In the teaching sector (Figure 3), the use of GPT is also controversial. Similar approaches in regards with students, immediacy in use is the fundamental factor for its popularity. Above the ease of use, accessibility and the promises of time use are the reasons why teachers could feel more inclined to use the ChatGPT tool. This is a point that shows the first differences between teachers and students. The second point, the most important, is the conception, the idea that both groups have for the use of the tool.

While students are more open to using artificial intelligence (AI) tools, teachers are more reticent. The low rating of parameters such as the similarity of use of ChatGPT and communication with a professor, as well as the consideration of this as a master in all disciplines, is notable. This is evidently summarized in the responses to the question about the use of artificial intelligence as a source of knowledge, which has a very low rating by teachers. It is possible that the approach of this sector is more critical and has not been dazzled by the novelty, but rather its approach has been more critical, typical of the scientific method (Figure 4).

Figure 3. Survey findings-Aggregate (lecturers).

Figure 4. Survey findings (lecturers).



6. CONCLUSIONS

This article explores the academic potential issues raised by the use of ChatGPT in the aviation studies. Our analysis highlights the significant concerns that ChatGPT tool display in the academic environment being the first artificial intelligence tool that has penetrated the educational community with a significant impact. Both students and teachers have approached with curiosity a technology that promises immediate results, free of charge and very simple. The strategy of the company Open IA, creator of ChatGPT, consists of making a free version of its application available to the public, with updated content until January 2022, and another more advanced and updated version. Since the appearance of Microsoft office packages and, later, Internet content search engines, no computer application had been received with so much expectation and interest at a general level. It is true that there are specific sectors (design, photography, architecture, circuits, statistics) that have received recent releases of utilities aimed at very specific fields of work, but the novelty of ChatGPT comes from its great spread in all areas of the society. For this reason, it is normal that the different sectors that comprise it present a different attitude depending on their needs and characteristics. In the educational community there is also this differentiation depending on whether it is dealing with university students or teachers. While ease of access is paramount for both groups, students prioritize ease of use and are also more forgiving of the software's capabilities. On the other hand, the teaching staff has received this App with more critical sense. They do not give it the benefit of considering it as a source of knowledge, but more as entertainment. While students are more interested in knowing how to use their capabilities, teachers are more concerned about how to detect fraudulent use by students, without fully using their full potential (ethical concern). Meanwhile, there were a few notable limitations. Former, the reduce sample of respondents as the Commercial Pilot for Passenger and Cargo Transport Degree at Burgos University starts in the academic course 2022-2023. The latter, because students have not completed their syllabus, thus, the use of ChatGPT has not been able to be extended in its entirety. Therefore, the current study should continue being carrying out in future with higher groups of students and teachers and usability-spending time. Overall, this study contributes to promote and expand learning resources to enhance the new way of teaching and learning. It is possible that we will soon see more specific applications for education, but this will not diminish the impact that ChatGPT has already had on the university community, and, more specifically, in aviation studies.

ACKNOWLEDGEMENTS

The authors wish to thank the students and teachers who participated the survey.

REFERENCES

Barrett, A., Pack, A. (2023): Not quite eye to A.I.: student and teacher perspectives on th use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*. 23(59). https://doi.org/10.1186/s41239-023-00427-0

Chan, C., Lee, K. (2023): The AI generation gap: ¿Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millenial generation teachers? *Smart Learning Environments* 10(60). https://doi.org/10.1186/s40561-023-00269-3

Crompton H., Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(22), https://doi.org/10.1186/s41239-023-00392-8

European University Association (2007): Doctoral Programmes in Europe´s Universities: Achievements and challenges. Report prepared for European Universities and Ministers of Higher Education. European University Association Publications. 2007.

European Commission/EACEA: The European Higher Education Area in 2020. Bologna Process Implementation Report. 2020.

Fawaz, Q. (2023). ChatGPT in scientific and academic research: future fears and reassurances. *Library Hi Tech News*. Number 3, pp. 30-32. Emerald Publishing Limited. http://doi.org/10.1108/LHTN-03-2023-0043

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., Khurma, O. (2024): AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. Research and Practice in Technology Enhanced Learning, 19(23). https://doi.org/10.58459/rptel.2024.19023

Kocón, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydlo, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocón, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, S., Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion* Vol. 99. https://doi.org/10.1016/j.inffus.2023.101861

Lara-Palma, A. M., Brotóns Cano, R., Valencia, O., Matsuda, D. (2022). Heading for inter-disciplinary lectures: an international collaborative team activity carried out between an Asian and European Universities. 16TH International Technology, Education and Development Conference. IATED.

Nam, B. H., Bai, Q. (2023): ChatGPT and its ethical implications for STEM research and higher education: a media discourse analysis. *International Journal of STEM Education*. 10(66). https://doi.org/10.1186/s40594-023-00452-5

Stahl, B. C., Eke, D. (2024): The ethics of ChatGPT. Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, https://doi.org/10.1016/j.ijinfomgt.2023.102700

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q., Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potencial future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), pp. 1122-1136. http://doi.org/10.1109/JAS.2023.123618

Zirar, A. (2023). Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education*, pp.1-18. http://doi.org/10.1002/rev3.3433

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
Cátedras Telefónica
UNIVERSIDAD COMPLUTENSE MADRID
UNIVERSITAT DE BARCELONA
UNIVERSITAT ROVIRA i VIRGILI
Logroño