A HUMANIST PERSPECTIVE ON ARTIFICIAL INTELLIGENCE

Jaume AGUSTÍ-CULLELL
Marco SCHORLEMMER

Abstract

We try to give an answer to the following question: What should we take account of when designing and using AI systems so they contribute to the development of the socially embodied human intelligence on which human fulfilment depends? First, we consider the current impact of AI on the development of human intelligence showing some of its ambiguities. The deterioration of human intelligence is the prospect that should frighten us, not the rise of 'super-intelligent' machines. It is our contention that computers do not understand anything: AI systems only compute. In order to avoid misunderstandings, we need to accurately distinguish between intelligence, knowledge, information, and data. We denounce the abuses of metaphors that attribute human qualities and powers such as experience, knowledge, and intelligence to machines by stretching their meaning excessively. We glance through the two main AI models of intelligence—the representational model based on 'knowledge representation' and the connectionist model based on 'artificial neural networks'—and reflect on the necessity for combining the best of each. Finally, we consider the need to govern AI in ways that will be truly beneficial for all of humanity.

Key words: Intelligence, Knowledge, Information, Data Processing, Intelligent Machines.

1. Introduction

Rather than providing an abstract and neutral presentation of Artificial Intelligence (AI), we set out to make an evaluative reflection, with the purpose of clarifying AI's cultural situation. Predominantly, the perspective on AI offered here is not techno-scientific from inside the field, but rather from the outside, in order to locate it in relation to human intelligence. A particular aim is to understand the importance of AI, the values and counter-values it generates, in order to contribute to its effective governance. We offer a critical view, but not with the intention of belittling AI's huge value. Our aim is to help it advance in what we believe is the right direction. The massive current hype around AI is not helping its healthy development.

1.1. Some key questions

Information, especially information-processing by computers, has become increasingly powerful and pervasive in all human activities. The power of computers dazzles us. The first author still remembers his amazement when, in the mid 1970s as part of his PhD in Physics, he was programming an Intel 8080 microprocessor in binary code, using a teletype and paper ribbon, to control the data acquisition in atomic collision experiments. Computer utility has fascinated us so much that we have tended to abuse it. The temptation is to use it like a magic wand that will tell us our individual and collective future. But we remain unaware of its limitations, the possible perversions of its usage, the hidden designs it has on us.

Empowered by this fascination, along with its undoubtedly useful capabilities, Artificial Intelligence sets out to capture the very heart of humanity, the intelligence that makes us human. This is a living intelligence, the foremost embodiment of universal intelligence, something that has been constantly evolving for millions of years. Seen from this perspective, AI might be said to have the most ambitious, even presumptuous goal of all the techno-sciences: to understand, describe, reproduce and even improve this amazing human intelligence by the use of computing machines. An ambition we can accurately describe as overweening.

Thanks to AI, it seems that everything will become 'intelligent:' phones, cars, homes, factories and cities are just the most obvious examples. We already rely on computing devices to navigate us through city streets, recommend movies to us, and provide us with answers to search queries. However, a question arises. What effect will all these AI products and services—mostly designed according to an economic rationale—have on humanity? From the perspective of intelligence, the primary question a wisdom-based AI needs to face is clear. Instead of asking itself how to emulate human intelligence, the wise question is this: how can AI systems be designed and deployed so that they contribute to the development of a socially embodied human intelligence–something on which all human fulfilment depends?

AI studies functional intelligent behaviour, in particular the accomplishing of goals by means of data processing. Given this fundamental understanding, it is clear that AI needs to be guided in ways that will be beneficial for all of humanity. AI only makes sense when integrated with the development of this socially embodied human intelligence.

1.2. Uses and misuses of AI

The current impact of AI on the development of human intelligence shows some ambiguities. We see very positive uses of those AI systems which are at the service of our creative intelligence, extending its power and reach. Think of the computer assis-

tant and other apps in your phone, which save you from endless tedious tasks and offer easy access to all kinds of information, freeing you up to concentrate on your creative work. Even more significantly, AI systems can be employed to deliver high-quality education to people wherever they are on Earth.[1] AI also makes possible the creation and renewal of human teams and *teams of teams*, thus facilitating the most powerful form in which intelligence can be exercised.[2]

However, there are also many clear misuses of AI which tend to work in exactly the opposite direction. The result is that each of us becomes a consumer of information, no longer fully exercising our own intelligence, now degraded into one which is merely information-programmed.[3] Rather than directly observing and controlling our life through collective intelligence, we credulously look for second-hand information to solve our problems.[4] The terrible effect on adolescents and young adults produced by the time spent online in social networks and the abuse of electronic devices is all too obvious.[5] Unconscious of the power of monopolistic information-technology corporations such as Google, Facebook and many other hugely profitable data-extraction firms, most of us all too frequently rely uncritically on what they tell us, rather than using them with a cautious and critical attitude, guarding our privacy. By merely *consuming* information, we end up leading second-hand lives, at the mercy of the powers of domination, plutocracy and imperialism. This is the way that not only information, but also the centralised power of modern states, is controlled. We become puppets in the violent hands of dominating forces.[6]

The massive degradation of human intelligence is the possibility that should alarm us, not that of a supposed super-intelligence of machines. Getting humans to conform to machine behaviour, rather than the other way around, is much easier to achieve and

[1] UNESCO, Beijing Consensus on Artificial Intelligence and Education, 2019. Available at https://unesdoc.unesco.org/ark:/48223/pf0000368303. Last accessed on February 8, 2021.

[2] E. Andrejczuk, F. Bistaffa, C. Blum, J.A. Rodríguez-Aguilar and C. Sierra, «Synergistic team composition: A computational approach to foster diversity in teams». *Kowledge-Based Systems*, 182:104799, 2019.

[3] This is a consequence of *computationalism* and its emphasis on *effective computability*, i.e., «the quest for universal knowledge and perfect self-knowledge», to «make cultural practices not just computational but programmable—susceptible to centralized editing and revision»: E. Finn, *What Algorithms Want*. The MIT Press, 2017.

[4] AI-governed social interaction in privately-owned social network platforms that follow their own economic rationale without any regard for truth is no longer a 'marketplace of ideas' in which freedom of speech can flourish. This contributes to political polarisation, breeding divisiveness and eroding social solidarity (see, e.g., A. E. Waldman, «The Marketplace of Fake News». *Journal of Constitutional Law*, 20(4), 2017, p. 845-870).

[5] See, for instance, the study carried by the International Center for Media & the Public Agenda (ICMPA), University of Maryland, USA (https://theworldunplugged.wordpress.com); and also: The Washington Post, *Generation Z. What it's like to grow up in the age of likes, lols and longing*. Diversion Books, 2016.

[6] S. Zuboff, *The Age of Surveillance Capitalism: The fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.

much more dangerous.[7] This is a new version of an understanding that human wisdom traditions have had throughout history.

To emphasise the central insight: the great power of computers to process information has bewitched us to the point where we hold the computer up as a mirror to human intelligence. We look at the computer as a creation with a privileged status. In it we seek to understand the intelligence that created it. From the perspective offered here on intelligence, the AI ideal of emulating or even superseding human intelligence is a distracting and misleading goal, simultaneously provoking optimistic fantasies and unjustified fears about the future.[8]

Underlying this belief in an artificial super-intelligence is an understanding (or *mis*understanding) of intelligence in general, and human intelligence in particular, as narrow, abstract, disembodied, individualistic and representational. The fundamental misconception is that it thinks of intelligence as a primal quality, with individuals creating representations of the world and interacting externally between each other and with the environment. Further, this is taken to be a quality which is reproducible by machines that interact between each other and with the environment, using myriads of sensors that differ profoundly from the senses associated with human intelligence.

1.3. Computers do not and need not understand anything

It is important to begin distinguishing the intelligence of AI researchers, engineers and users from the intelligence attributed to the information-processing AI systems that they themselves create and use. Human beings, by designing, creating and using AI systems, exercise—and so improve—all dimensions of human intelligence, even when not explicitly aware of them. (We define and discuss these dimensions—*functional, axiological* and *liberating*—later in this article.) Setting ourselves to become fully conscious cultivators of all dimensions of intelligence would have several important effects. It would enhance the intelligence and wellbeing of the researchers themselves as well as adding to the overall effectiveness of the research. It would also add to the quality and to the potential of AI systems to produce humanly beneficial outcomes. In the absence of characteristics that we would describe as *sense-creating and loving engagement, communication, cooperation* and *freedom,* there can be no authentically

[7] «Through brain plasticity and changing social norms, we are adapting ourselves to become more knowable for algorithmic machines. In this way we are evolving ... in conjunction with our technical systems, slowly moving toward some consummation of the algorithmic love affair»: E. Finn, 2017 (as n. 3 above).

[8] M. Schorlemmer, «La distracció d'una intel·ligència artificial sobrehumana». *Quadern de les idees, les arts i les lletres*, 217, 2019. Available at https://www.quaderndelesidees.press/la-distraccio-duna-intelligencia-artificial-sobrehumana/. Last accessed on February 8, 2021.

creative and beneficial research. AI systems only compute. They are devices programmed to perform or acquire a known skill in order to achieve a given goal.

Intelligence lies in the creators and users of AI systems, not in the systems displaying the skills that have been programmed into them. The expansion of human intelligence is one of the main positive contributions of AI. The danger, of course, is that it might degrade that intelligence. Generally speaking, 'intelligent' machines do not understand anything; they compute complex functions only understood by their programmers and users. However, it is possible to see how confusion arises. Simply through data processing, AI systems perform tasks such as language translation that at first sight seem irreducible to computation. Together with the fact that humans do these tasks through intelligence and understanding, this has led to the metaphoric attribution of intelligence and understanding to computers.[9] The obvious fact that they do not operate through sense-making and understanding has become blurred by their seemingly magical powers, helped by many people's superficial understanding of their own intelligence. (This ignorance, it is worth noting *en passant*, is one of humanity's major problems. The development of our intelligence, through continuing education, is a political must.) Even scientists of great reputation have endorsed the prediction of future 'super-intelligent' machines, and issued dark warnings about them.[10] Here, once again, are the effects of the split between the sciences and the humanities.

Given the limitations that this lack of a sense of *meaning* builds in to AI systems, we should be wary about leaving important and far-reaching decisions concerning, for instance, finance and employment, in the hands of such machines. (As of 2017, 70% of all financial transactions were performed by algorithms.[11]) Although this automatic decision-making may empirically be very useful in routine matters, for more complex and challenging issues, where human values and the dynamic of shifting contexts apply, it is simply not up to the job. The inevitable conclusion is that automatic decision-making can only be harmful to social justice.

Another important consequence of AI's rapid development and increasing ability to perform complex tasks is that people are now obliged to develop and exert their crea-

[9] G. Lakoff and M. Johnson, *Philosophy in the Flesh. The Embodied Mind and its Challenge to Western Thought*. Basic Books, 1999.

[10] See, for instance, S. Hawking, M. Tegmark, S. Russell and F. Wilczek, «Transcending Complacency on Superintelligent Machines». *Huffington Post*, 19 April 2014. This a consequence of adopting computational functionalism and its assumption that intelligence is mere information-processing, which is often taken as an unquestionable fact (see, e.g., the statement by Sam Harris: «Intelligence is a matter of information processing in physical systems. Actually, this is a little bit more than an assumption ... [W]e know that mere matter can give rise to what is called 'general intelligence'»: S. Harris, «Can we build AI without losing control over it?». *TEDSummit*, June 2016. Available at https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it. Last accessed on February 8, 2021).

[11] F. Helbig et al., «Will Democracy Survive Big Data and Artificial Intelligence?». *Scientific American*, 25 February 2017.

tive intelligence as fully as possible, particularly in the workplace. Otherwise, at the level of mere skills, they will be displaced because of the increasing number of tasks now performed quite satisfactorily by 'intelligent' machines. (You would hope that writers and translators, in love with language, forever enquiring into meaning and creating new forms of expression, would never be replaced by machines.)

A warning is also called for in connection with the recently developed field of transhumanism, in which individual brains are enhanced by AI systems implants (or vice versa), attempting to dump a human mind into a computer. These projects arise from the old (and misleading) *individualist* understanding of intelligence. As an idea and as a suggested direction of travel, it only distracts from humanity's greatest need: to develop and manifest a mature and harmonious socially embodied intelligence, based on universal lifelong education.

2. The Role of Information

2.1. The need to dispel some confusions and ignorance of fundamental principles

Information abounds in modern societies, and an understanding of its central role is vital. In order to avoid its misuse, we need to distinguish between *intelligence*, *knowledge*, *information* and *data*. Unfortunately, these words are all too frequently confused and their meanings profoundly misunderstood.

This confusion is rife in most discussions of AI. Computers, properly speaking, process data in the form of bits, zeros and ones, rather than through knowledge or information, both of which have *meaning*. Meaning is alien to computers; their operations are meaningful only to us, not to computers themselves. So, when we use the widely accepted term *information processing*, we are actually referring to *data processing*.

Our understanding of intelligence can be framed as *cultural* or *humanist*. We consider human intelligence to be a collective phenomenon. In other words, an individual, embodied intelligence depends on the *socially embodied intelligence*. Intelligence belongs to the commons; it grows because it is cultivated and held in common, not owned privately. It is our primary common good, and the source of everything else in our lives. Its essence is *communication* and *symbiosis*—the quality of being able *to live together*. Its growth or decline is dependent mainly on cultural factors.

Our focus is on five distinctly identifiable aspects—or powers—of intelligence: *interest in reality, communication, subsidiary symbiosis or cooperation, research* and *freedom* (this last being the most important of all). How these powers of intelligence are exercised—the degree of intensity and the priority of each power over the others—will determine the different uses of any particular aspect of human intelligence.

For the moment, three specific uses of this intelligence demand our attention. The first two address our material needs as living organisms and our investment in meaning and survival. They constitute the intelligence of need, and they are conditioned by our biology. We call the first *functional intelligence*. This is mainly an instrumental, abstract form of intelligence, the intelligence proper of techno-science—the entanglement of science, technology, economy, and their products and services. By way of this intelligence we humans focus on what intelligence is for: its function for adaptation, for problem-solving, for attaining goals, for success, for survival. The second, we call *axiological intelligence.* In interdependence with functional intelligence, this is the intelligence by way of which we humans imbue meaning and value to what we do. It connects us to the aesthetical dimension of life. Artists, for instance, exert very much this sort of intelligence. Through its creation of values, axiological intelligence responds to our need for meaning and direction in life. Functional and axiological intelligence constitute the intelligence of need, which depends very much on our bodily interactions with other organisms and the environment. It creates lawful models of reality, a stable world relative to our human needs and interests.

Inseparable of the intelligence of need, three is also a third, contemplative dimension to intelligence. This is the dimension that provides us with the insight that functional and axiological models are ultimately relative to our contingent needs as human beings. This is a subtle but powerful form of intelligence that frees us from getting caught within any particular reality as we conceptualise and experience. Its highest manifestation is the silent contact with the origin, of what is unique and free of law: the intelligence of the whole, of the creative freedom and unity of reality, the source of human creative freedom. Because it is centred on freedom, we call it *liberating intelligence.* Through it, the other powers of intelligence can attain their greatest potential. Interest, the basic energy of intelligence, can reach its highest degree—unconditional compassion and love—as the hallmark of full intelligence; communication can become sincere and trusting to the point of silent communion; subsidiary symbiosis becomes a union of love and service; research reaches the highest degree of collective creativity and is undertaken for the good of all humanity. It is important to keep in mind that intelligence is, at its core, as untameable as our freedom and creativity. Human intelligence, we come to see, is primarily an intelligence to be lived and developed rather than defined.[12]

[12] J. Agustí-Cullell, «Reflexions sobre intel·ligència humana i intel·ligència artificial: el repte d'una intel·ligència lliure i creativa». *Poblet*, 38, June 2019, p. 30-36. J. Agustí-Cullell, «Intel·ligència alliberadora». *Ars Brevis*, 25, 2019, p. 29-66. Available at https://www.raco.cat/index.php/ArsBrevis/article/view/371411. Last accessed on February 8, 2021. J. Agustí-Cullell, «Una inteligencia que libere la condición humana». *FronterasCTR,* January 29, 2020. Available at https://blogs.comillas.edu/FronterasCTR/?p=4819. Last accessed on February 8, 2021.

2.2. Distinguishing between knowledge, information and data

Knowledge is an essential component of the models of reality created by the intelligence of need: the co-working of functional and axiological intelligence, informed by the creativity of liberating intelligence. Thus, knowledge is not only descriptive (its information content) but fully charged with sense, emotions and values. It originates in a creative act where freedom is operative. Because of the hidden workings of this freedom, in the known always lurks the unknown, the basic awareness and stimulus behind all research.

To distinguish it from knowledge, we characterise information as the form of knowledge which is descriptive, de-contextualised, free of emotion, value-neutral and frequently conceptually abstract. It is most liable to be formally represented and then used via computation. If someone tells you that a person you hardly know has died, that is received as part of the inevitable pattern of birth and death. This news is merely information. However, if this person is your sister, then it becomes knowledge, full of meaning and deeply personal emotion.

Knowledge and information both involve the use of data. The material and conventional signs are meaningless in themselves (every language uses different signs and sounds to convey similar meaning). So, simplifying the complex evolutionary history of data (i.e., the creation of pictures and writing), we can say that data or signs support and transport meaning for us. The articulated sounds of speech, or the pure syntax of texts or pixels in physical images, are the main ways in which data is held and knowledge and information transported between us. Syntax holds semantics.

Data is the object of computing after it has been represented by lists or matrices of zeros and ones, or bits. Once data is digitised, the bits are the objects processed by computers, the atoms of computation being the electronic logic gates—or artificial neurons, in the connectionist architecture of computation that we will look at later in this article. We feed the computer with data that has informational content for us, but not for the computer. The computer processes these data, we interpret the results as new information, and we say that «information has been processed».

Information is an abstract form of knowledge that is specifically characteristic of the techno-sciences. Information is essentially logical in character and deals with the regularities of the world. Its main goal and purpose are the prediction and control of phenomena. Techno-scientific information has proved to be indispensable in helping humanity to deal with its needs and to improve the quality of life. The fundamental mistake, though, is to confuse what is *abstract* (a tool for the satisfaction of life's needs and interests) with the *concrete*—with life itself. This mistake is universally made, to the point of converting the abstract into the concrete, as when matter is taken to be the concrete foundation of the universe, or frequency is made into a colour, or a complex algorithm into an intelligence. It is because of its power to reveal and resist this very

tendency that contemplative practices such as silence, meditation, beholding, or deep listening are so significant for the human understanding of intelligence. These practices help quieting the mind and thus offer a liberating possibility, putting us in direct contact with concrete reality, with its beauty and freedom. Techno-scientific inquiry thus needs to go hand in hand with the sort of inquiry nourished by contemplative practices, in order to sustain the freedom and creativity that enables human intelligence to go beyond the forms into which it shapes reality by way of abstraction and information.[13]

Because of its abstract and de-contextualised nature, information is considered to be universally applicable, an assumption that we should be very wary of making. The case is rather different with knowledge, which is clearly inseparable from the values and the overall context of the culture within which it originates. This misuse of information is particularly dangerous in medical settings where programs make diagnoses using much more information than a doctor is able to deal with. However, they do it without fully knowing the patient in the present moment and—among other things—in ignorance of the possible harmful effects of even a correct diagnosis when informed without care. To avoid doing harm, doctors need to contextualise the diagnoses produced by these programs. The importance of contextualisation can be illustrated with the following personal experience: The first author knew a very apprehensive woman whose wise doctor, with the agreement of the woman's family, decided not to tell her immediately or all at once that she had leukaemia, opting for a gradual, sensitive and respectful communication with her over the years. His much less experienced successor told her the diagnosis right away; from that moment she felt much worse and she died in a matter of months.[14]

For us to attain full knowledge of anything, our level of interest in it has to reach its maximum degree—in other words, *love*. In the case of information, its form of interest is best described as *curiosity*. Scientists and engineers are immersed in information, moved mainly by curiosity about how things work. As such, they are liable to forget the abstract character of information, ruling out the possibility of integrating it into full evaluative knowledge.

The untamed curiosity of scientists holds the danger of impelling them into indiscriminate research, reaching after information regardless of the consequences and the

[13] M. Schorlemmer, *Indagació Contemplativa per a l'era de la intel·ligència artificial*. Quaderns 121, Fundació Joan Maragall, 2019.

[14] This is a problem of so called 'shallow medicine,' when emotional connections between patients and doctors break down, combined with a systemic problem that focusses on excessive diagnosis. These issues can worsen if AI-based medicine is not developed and deployed with care, in the context of a rich multi-dimensional understanding of intelligence, in order to foster presence, attentive listening, empathy, and human-to-human bonding, which are difficult to quantify and digitise (see, e.g., E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019).

possible harmful uses to which it might be put. In AI, this results in the tendency to automate everything that can be automated, despite the terrible consequences of creating potentially monstrous machines.

Computational power has doubled roughly every year and a half, while the cost of computers has halved. If the automobile industry had followed the same trend, you would be able to drive to the sun on a few litres of fuel. Information supported by data and processed by computers has become the most pervasive and useful commodity throughout much of the world. Such is its importance that the time we live in has been designated as the Information Age. More precisely, we are at the beginning of the so-called digital transition, the age of technological hyper-connectivity (i.e., by means of the Internet of Things, the vision of adding sensors to all kind of everyday objects—buildings, domestic appliances, furniture, clothes, vehicles, etc.—and to have all these objects interconnected through the Internet[15]). This is a social process through which technologies of information and communication are so widely deployed that they are transforming our daily lives. It is a reflection of the exaggerated importance and role currently awarded to information, much to the detriment of full evaluative knowledge.[16]

Using information when the corresponding data has been processed and applied mechanically, without appropriate contextualisation and interpretation by human intelligence, is a source of many possible errors. In general, a clear understanding of the data, their provenance and their characteristics, must be captured, so that others using the data set can understand the potential flaws.

The general paradigm of this information revolution is to reduce everything to information and then to data. The culmination of this is the installing of information as a basic component of the Universe, processed by a supposedly universal algorithm. "It from bit" is the catchphrase that summarises this myth.[17] However, the *bit* is in the machine and the *it* is in human intelligence. The flexible and wide use of the interaction *it-bit* is what brings about the enhancement of human intelligence. Taking information to be primordial is a misleading belief. The true role of information and data is that they are extremely useful *instruments* of intelligence, created by intelligence to extend its power. In truth, the *bit* comes from *it* and, once processed, returns to *it*.

The amount of data we produce doubles every year. Every minute, millions of Google searches and Tweets occur. This is testimony to the enormous usefulness of information technologies. When used with care by a creative intelligence they hugely increase its power. However, they bring with them new risks and threats. These search-

[15] S. Greengard, *The Internet of Things*. MIT Press, 2015.

[16] L. Floridi (ed.), *The Online Manifesto: Being Human in a Hyperconnected Era*. Springer, 2015.

[17] J. A. Wheeler, «Information, Physics, Quantum: The Search for Links», in: *Proc. 3rd Int. Symp. Foundations of Quantum Mechanics*. Tokyo, 1989, p. 354-368.

es and posts contain information that reveals how we think and feel. Those in control of information technologies can easily control us. Specialised tech companies that you have probably never heard of are tapping vast troves of our personal data, using extraction technologies such as Big Data. These tech companies know far more about their users' future behaviour than they themselves know. These firms sell their scoring services to major businesses and make huge profits as a result. From Big Data comes big money. People are profiled for targeted advertising by online service providers and in political campaigns. This undermines their freedom and so their humanity. We are now exposed to behavioural scrutiny, prediction, control and—eventually—alteration.[18] As a result, our tendency to become programmed intelligences, drowning in information, is progressively reinforced. And so we undergo an atrophying of the intelligence, a degradation of intelligence's truly creative powers.

To avoid being manipulated by AI systems or being overwhelmed by the current flood of information, we urgently need to develop and empower these creative powers of intelligence—above all, its power to undertake research informed by its essential quality of freedom.

3. Different Views on AI

3.1. A bit of history

The idea of thinking artificial beings arises in antiquity in the form of story-telling devices. There has long been a great fascination for imagined automatons. More significantly, the study of mechanical or formal reasoning began with early philosophers and mathematicians. In the 13th century the philosopher, theologian and mystic Ramon Llull attempted to automate reasoning in order to convert Muslims to Christianity. Much later his work was of a great influence on Leibniz's *Mathesis Universalis,* a computational view of the universe, giving rise to his celebrated dictum: «Let us calculate, without further ado, to see who is right». In 1834, Charles Babbage conceived the Analytical Engine, a universal computing device. The study of mathematical logic determined Alan Turing's theory of computation. By shuffling data or signs as simple as 0 and 1, a machine could simulate mathematical deduction. This insight, that digital computers can simulate any process of formal reasoning (or, more precisely, any computable function), is known as the Church-Turing thesis. Beyond this, discoveries in neurobiology, information theory and cybernetics led researchers to the idea of building an electronic brain.

Artificial Intelligence as an academic discipline was born in a workshop at Dartmouth College, New Hampshire in the summer of 1956. The term was coined by John

[18] S. Zuboff, 2019 (as n. 6 above).

McCarthy, then a young assistant professor of mathematics. He wanted to distinguish the new field from cybernetics and the influence of one of its leading proponents, Norbert Wiener. Initially research concentrated on the definition of general methods applicable to most problems, whereas contemporary AI research has focused on more specific methods that will be most effective in each particular type of problem.

Although only 65 years old, AI has become a highly significant branch of the techno-sciences and also one of the great scientific objectives, arousing expectations around its enormous benefits, particularly economic ones. It also provokes much speculation about risks and dangers.[19]

We can best understand it as an outpost of computer science, but one that benefits from many other disciplines, in particular the cognitive sciences. It has an enormous social and economic impact. AI techniques and systems have become an essential part of the technology industry, helping to solve many challenging problems in computer science, software engineering and operations research. Exploiting concurrent advances in computer power, large amounts of data, and theoretical understanding, it has become one of the most revolutionary of the techno-sciences.[20]

The initial hypothesis of AI is that human intelligence can in principle be described so precisely that a machine can be made to simulate it. Hence, the aim of AI is to create ever more autonomous and powerful computational systems—a true revolution in techno-science and engineering. However, as we've already established, AI has a narrow view of intelligence—one that is useful to create AI systems, but in no way comparable to human intelligence.

Within this general aim, there are two main classical approaches to AI. Firstly, the pragmatic approach called *Weak AI*. It aims at the automation of complex but well delimited tasks that are usually performed by human intelligence, such as playing chess or the recognition of objects in images. Secondly, the ambitious approach called *Strong AI,* which is oriented towards emulating or even surpassing human intelligence (i.e., machines capable of designing other machines). Within Strong AI, there has been a recent specific undertaking, with the aim of defining and measuring intelligence in general, and that of machines in particular. It proposes the testing of machine behaviour in respect to what is called Artificial General Intelligence (AGI). AGI is the foreseen ability of computational systems to learn or acquire skills and adapt to a wide range of environments, applying what is learned from one domain to another.

[19] «El objetivo último de la inteligencia artificial —lograr que una máquina tenga una inteligencia de tipo general similar a la humana— es de los más ambiciosos que se ha planteado la ciencia. Por su dificultad, es comparable a otros grandes objetivos científicos como explicar el origen de la vida, el del universo o conocer la estructura de la materia». R. López de Mántaras and P. Meseguer, *Inteligencia Artificial.* Colección ¿Qué sabemos de?, Editorial CSIC y Libros de la Catarata, 2017.

[20] K. Schwab, *The Fourth Industrial Revolution.* Crown Business, 2016.

3.2. Weak AI

Weak AI, an engineering and pragmatic view of AI, uses computers to automate complex tasks that, when done by humans, we describe as intelligent. Weak AI does not start from the supposition or hypothesis that all intelligence is computation, that is, information processing. It is essentially a more advanced informatics, extending the boundaries of automating more and increasingly complex tasks.

The Weak AI approach closely follows the techno-scientific method of abstraction. Not subject to the need to reduce reality to information, it leaves aside what is not solvable by information processing. As a result, it focuses research on the appropriate techniques to solve complex problems. Algorithms solve a fully described task in a fully described environment where all possible inputs can be explicitly enumerated or analytically defined. The goal is to accomplish the task efficiently, inspired by what humans do (and how), but not tied to it; the main intention is to make advances in computational techniques through exploiting the computational power and data available at each moment.

One of the most popular basic models is that of *autonomous agents* or *multi-agent systems*. Leading AI textbooks define the field of AI as the study of 'intelligent autonomous agents.' It refers to any device that perceives its environment, interacts with other agents and takes autonomous and rational actions that maximise its chance of successfully achieving its goals. Rational actions imply the ability to choose at every moment the action that will produce the best result following a pre-established performance measure.

However, Weak AI systems, although very efficient at doing the job for which they are designed, fail when confronted with activities that differ even slightly from those original tasks. This is one of the reasons why this approach to AI is described as 'weak' in terms of intelligence. IBM's champion chess-playing system Deep Blue, for instance, is incapable of playing checkers. The researchers who created it realised that the program, which is based on the AI methods of *minimax* and *tree search*, tells us nothing about human intelligence. The ability to play chess does not demand any specifically human abilities; the problems involved can be solved by techniques independent of human cognition. The same is true of many other human skills, as AI proves every day. In the match between Kasparov and Deep Blue, we can say that only Kasparov was really playing chess, with all its cultural connotations. Deep Blue was just computing. Deep Blue did not experience victory; but Kasparov felt the defeat.

Weak AI embraces most of the existing applications of AI in daily life. It has been deployed in a range of contexts and social domains, with mixed outcomes: insurance, finance, education, employment, marketing, governance, security, policing, etc. However, since the very beginning, many AI systems have failed to be fully reliable. They make errors when they are applied to tasks that have not been foreseen in every detail.

It's for this reason that research on robust AI systems has become increasingly important. The main aim of Weak AI is to create systems that are sufficiently safe and robust for us to trust them in all their applications. In particular, robustness and flexibility are perceived as important requirements for certain broader sub-fields of AI, such as self-driving vehicles, domestic robotics, or personal assistants.

3.3. Strong AI

Strong AI views intelligence as a program that works with the hardware of a brain. It believes in building machines that can outperform not only our muscles, but our minds as well. In other words, the assumption is that the best way to understand the mind is to create one on a computer, using the knowledge we have of the human mind. To Strong AI, the essence of intelligence is the processing of information or data. From this premise, it makes sense to assume that human intelligence efficiency is not an upper limit. Hence, Strong AI goes for systems with a much broader scope of application than Weak AI—systems with domains of application beyond human intelligence. This, however, does not make them intelligent or super-intelligent. Intelligence and machines should not be compared. Each has its own domain of efficiency. The wise attitude is to aim for cooperation, rather than creating AI systems that are fully independent or self-governing.

Despite its failure up to this point, the ideal of Strong AI still seduces a great part of the research community. It seems that the idea of emulating human intelligence is a powerful incentive for the advancement of AI. Despite the failures, its proponents are determined to carry on.[21] The danger remains that what is being promoted is a narrow view of human intelligence. One that deprives it of its central and most important feature: its freedom and full creative power. The outcome is a continued degrading of human intelligence, the opposite of what wisdom-based AI research should have as its goal.

The project of emulating human intelligence is frequently understood as mimicking the supposedly independent and self-governing intelligence of individuals. They are thought of as possessors of their own intelligence, relating and interacting externally with each other and with the environment; and reproducible by computational systems, processing and exchanging data captured from the environment through sensors. This, though, is a fundamental misconception. The embodied autonomy of an individual's intelligence is actually in a permanent process of being constituted, through sense-making interactions between individuals and with the environment. The human intelligence of individuals depends on a socially embodied intelligence. A collective or

[21] V. C. Müller and N. Bostrom, «Future Progress in Artificial Intelligence: A Survey of Expert Opinion», in: V. C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*. Synthese Library, Vol. 376. Springer, 2016, p. 555-572.

cultural intelligence that embraces all the interactions of social life, through the participatory sense-making powers of the creative intelligence. From this perspective, it starts to become clear that the project of emulating socially embodied human intelligence is, at the very least, a far-away goal. And probably a misleading and dangerous one that threatens to degrade human intelligence.

3.4. Speculations on AI

So-called *Speculative AI* anticipates a future explosion of electronic or silicon-based super-intelligences with the potential to simulate the characteristics of a human AI developer, thus having the capacity to reproduce themselves. The far-fetched goal is the creation of an AGI which will colonise the universe. According to this fantasy, 'intelligent' machines will be our successors, the new inhabitants of Earth when the evolution of the sun has rendered life on this planet impossible.

There are countless debates and predictions concerning the future, the capabilities and social impact of AI and the myriad products and services it offers. Arguments also abound about the philosophy and the ethics of AI. Throughout history the issues provoked by the idea of *autonomous intelligent machines* have been addressed not only through myth and fiction but also by philosophers. This level of discussion and scrutiny is now intensifying. Speculations proliferate about the possibilities and impact of so-called *super-intelligence*—pictured by some as a promising challenge, and as a terrible threat by others. The latter consider AI to be a danger to humanity if it progresses uncontrolled. They also believe that the technological revolution brought about by AI will risk mass unemployment for the first time in history.

Based on the view of human intelligence that has been presented in this article, we suggest that what is urgently needed is a full awakening of our *liberating intelligence*—the silent, contemplative dimension, fundamental to human freedom. This is the only way that we can avoid submission to algorithmic guidance by supposedly 'intelligent' machines on matters ranging from love to real estate. Or submission in the form of alienated work, in which humans become mere processors handling simple tasks assigned by an algorithmic apparatus. The overriding imperative is for us to break away from the dominance of an economic rationale, the force that drives all current AI development.[22]

3.5. Use and abuse of metaphors

Metaphors are constitutive of human language and so of human understanding. They have inspired much of human research—that of AI in particular. In general, AI

[22] M. Schorlemmer, 2019 (as n. 8 above).

has quite rightly taken its inspiration from human skills when designing its systems, using metaphors to describe them, even attempting itself to simulate metaphorical language. However, there are also abuses of metaphor. Attributing human qualities and powers such as experience, knowledge and intelligence to machines, stretches the metaphor to an impossible extent. This is to talk about machines as if they were human, as though they were experiencing, thinking and responding intelligently to the environment just as humans do. The outcome is much confusion and many misleading projects, such as artificial general super-intelligence and its ambition of going far beyond the human.

AI specialists in machine learning have trained an artificial neural network to detect tumours after radiography. The computers are able to undertake the task more efficiently than humans. The AI researchers say—metaphorically—that the artificial neural network learned by itself to recognise tumours. Whereas what has actually happened is that they have exploited the power of artificial neural networks to identify statistical regularities or patterns in huge amounts of data. In other words, the network was structured to approximate a computational function, to detect a data pattern that humans judge to be a tumour. The machine does not recognise tumours, it detects patterns in data; it is only a *tool*, the use of which is the responsibility of the doctor.

Given that only humans can make sense of what the artefact is doing, it follows that we are also responsible for making sure the artefact is doing something that makes sense and is beneficial to humans. Machines cannot do it themselves. Humans exercise their functional intelligence with the aid of these useful machines, but nothing is gained by metaphorically ascribing human-like intelligence to them. Metaphors provide a basic way to make sense of the world, but we need to be aware of their limits.

4. Ignoring Liberating Intelligence

Amidst current attempts to understand and characterise intelligence, the humbling, foundational experiences that liberating intelligence can open up for us are widely ignored. The origin of this ignorance lies in the confusion between models and reality, between the abstract and the concrete. This is a fundamental confusion with far-reaching consequences. In particular, it gives rise to purely materialist models of intelligence. This is to misunderstand intelligence as if it were a derived phenomenon emerging from abstract concepts such as matter, which is mistakenly believed to be primary and concrete.[23]

[23] M. Tegmark, *LIFE 3.0: Being Human in the Age of Artificial Intelligence*. Alfred A. Knopf, 2017.

This ignorance with respect to liberating intelligence is at the root of human presumptuousness and of many misconceptions about models of intelligence such as AI. Presumptuousness combined with violence renders *Homo sapiens* a species without a future.[24] In order to have one, it must transform itself into the peaceful and humble *Homo quaerens*—the inquiring human.[25] Presumptuousness is implicit in the declaration made by the founders of Artificial Intelligence itself:

«We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College ... An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer».[26]

This is evident ignorance of the primary and indefinable character of creative intelligence. In particular, they seem unaware that intelligence lies in sense-making interactions with nature, between humans, and with tools, instruments or devices. It is not—as is implicit in their rationale—a primal *possession* of individual entities, be they humans or machines.

This ignorance and presumptuousness go hand-in-hand with a world-view based on individualism and domination. It manifests in many different ways. Mirroring terminology used in software development, we can refer to Presumptuousness 1.0 as the primitive rudeness of the animal, imposing itself through sheer brute force. Presumptuousness 2.0 is that of a violent *Homo sapiens*, the one who *knows* as a way of dominating nature and other beings. Presumptuousness 3.0 would be the presumptuousness of 'singularity,' the fantasy that machine 'super-sapiens' will surpass human intelligence, start to reproduce and go on to colonise the Universe. A great deal of misunderstanding and hubris is folded into such speculations. Techno-sciences focus their attention on what is regular, predictable, quantifiable or computable. They rightly excuse themselves (or should do) from the realms of axiological and liberating intelligence, along with other complex phenomena of social life. The presumptuousness comes from the implausible pretence (and simplification), that technology—no matter how useful—can be used to explain the whole of reality and to artificially emulate its free creative agency: intelligence.

[24] One could certainly qualify this statement as presumptuous itself (from Latin *praesumptiosus,* i.e., full of boldness); but we hope the reader will indulge our confessed presumptuousness of this and the following paragraphs.

[25] See also https://www.HomoQuaerens.info.

[26] J. McCarthy, M. Minsky, N. Rochester and C.E. Shannon, «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence», 31 August 1955. Available at http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf. Last accessed on February 8, 2021.

5. Two Main AI Models of Intelligence

5.1. The representational model

In the representational view of cognition, cognitive models of reality belong to individual organisms. The process indicated by these models is understood to happen as follows: organisms receive perceptual information from the environment; based on this information they dynamically build and update internal cognitive models of the external world. Then, using these models, they reason, make decisions and act in response to the sort of entities there are in the external world, what their properties are, how those entities relate to one another and how they can be manipulated. Despite their incompleteness and possible inaccuracy, AI cognitive scientists consider these models to determine the way in which an organism views the world and acts in it. The representational cycle of an organism is conceived in this way: *perception - update models - make decisions - act*. In this view, an organism's viability depends on the efficiency of its representational cognitive models. This is a good and useful model of cognition for implementation in a machine using sensors to gather data from the environment.[27]

However, representational cognitive models are obliged—through reduction or simulation—to turn every kind of cognition into representational form. Consequently, in their attempt to be comprehensive, they become hard work. This is due to the inherent complexity involved in representing certain types of contextual intra-active cognition through formal languages. The models contain inaccuracies and are always incomplete. Formal representations of reality cannot be otherwise.

Beyond this fundamental problem there is the practical question about the type of knowledge that we humans will actually be able to model, formalise and program so that it will be finally computable in a machine (bearing in mind that not even all mathematical knowledge is computable, as Gödel's theorem reminds us). A case in point is the difficulty AI has in simulating common-sense human behaviour. This difficulty surely results mainly from the fact that human intelligence deals with contextual meaning and computers do not. It is for this reason that behaviour based on common sense resists representation as knowledge in machine-interpretable form. Any attempt at such representation cannot match the continual increase generated by the sense-creating activity of human intelligence. No wonder that the problem of relating abstract knowledge-representations in the machine to specific real-world situations, at scale and in an efficient manner, has still not been solved. Nor has the challenge of transferring what has been modelled in one context over to another. A further indication that

[27] G. Marcus, «The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence». *arXiv:2002.06177 [cs.AI]*, 14 February 2020. Available at https://arxiv.org/abs/2002.06177. Last accessed on February 8, 2021.

the role of AI is not to emulate human intelligence, but to free it from tasks which are eminently performable by computable functions.

Using these languages of logic, AI researchers and engineers create detailed representational and computational models of the external world. Through them computers perform functions corresponding to different types of human skills: handling logistics, theorem-proving, playing games such as chess or Go. Generally, these computable models are used to automate jobs which can be reduced to a complete, mathematical, formal and rational description. We also see them being applied in industrial robotics, the control of home appliances, investments in the stock market and so on.

This AI representational model is in sharp contrast with the *enactive* models of intelligence that cognitive science offers. In these, sense-creating cognition is understood to arise through a process of self-individuating interactions between a living organism and its environment. In other words, in the enactive model, organisms do not create internal representations of the world after passively receiving information from their environments. The intelligence of the life-form is not accessing its world in order to build accurate pictures of it. Rather, in each organism, intelligence is actively participating in the generation of meaning in connection with what matters to the organism, *which lives in the world it creates*.[28]

In the case of human intelligence, language is a special kind of social agent. It emerges from the intra-dependence of creative intelligence's powers, most especially those of symbiosis and research. Human language makes us conscious of our freedom, of the liberating intelligence within and between us which is our means of immediate access to reality.[29]

The contrast with representational models, which rely on mere sensors capturing environmental data, is clear. Passive sensors and their signals are not enough to make self-individuating, embodied intelligences out of machines. The same can be said about connectionist models in which training and learning is based on processing huge amounts of data. Both the symbolic and the connectionist models produce an intelligence with powers radically different from that of humanity.

5.2. The connectionist model

The connectionist model of cognition conceives of the mind as a *tabula rasa*. The term is borrowed from the philosopher John Locke who used the metaphor of the blank slate to support his view of the mind as a flexible, adaptable, highly general pro-

[28] E. Di Paolo and E. Thompson, «The Enactive Approach», in: L. Shapiro (ed.), *The Routledge Handbook of Embodied Cognition*, chapter 7. Routledge, 2014.

[29] J. Agustí-Cullell, 2019 and 2020 (as n. 12 above).

cess, adept at turning experience into behaviour, knowledge, and skills. Later, Alan Turing, one of the great precursors of AI, thought of the child's brain as being something like a notebook—not much mechanism and many blank sheets—thus, transferring his invention of a universal computing machine into the workings of the human brain.

The connectionist model tries to get away from model building through the sheer hard labour of knowledge representation. Because of the implied potential for learning, the metaphor of the 'tabula rasa' has been attractive to designers and proponents of this model. Simply put, connectionist computing systems are loosely inspired by biological neurology. An artificial neural network (ANN) simulates the physiological structure and functioning of animal brain structures. (In particular, the brain of the nematode *Caenorhabditis elegans* has been studied and simulated computationally.) An ANN is based on a collection of connected units or nodes called *artificial neurons*, loosely modelling the neurons in a biological brain. ANNs derive their skills from the huge amounts of appropriately labelled or classified training data used to program the network.

The connectionist model started to be successful when implemented by several levels of hierarchically interconnected ANNs. In such an arrangement—known as *convolutional*—each level performs what is known as an *object detection step*. At the lowest level, the straight lines of a chair might be detected; at the highest level, it is the full detection of the chair, with interaction at all levels between. This method has proved to be immensely powerful for machine learning—so called *deep learning*. From huge amounts of training emails, for instance, which have already been classified as spam or not, an ANN is programmed through a kind of algorithmic (and very long) trial-and-error process which successively upgrades its performance. The result is a computational function that can determine whether new emails are spam or not. An early outstanding success came when an ANN classified 150,000 digital images with an 85% success rate. Before long, this rate had moved up to 98%.

Deep learning has been applied mainly to image processing, games, speech recognition and natural language processing. Its success has made it the dominant cognitive framework in which AI research is taking place. This model is adequate to deal with those situations where there are vast correlative databases—more data than knowledge. However, these processes tend to be excessively data-hungry. They need ever-larger training sets and more and more computer power. 30 million training situations, for instance, is not enough for a deep learning model to learn to drive a car in a plain supervised setting. Moreover, they are brittle because of their lack of memory and reasoning power and so are limited in their ability to adapt to anything new. The main problem is that they are not fully reliable. A slight modification of the pixels in an image—imperceptible to human eyes—can result in the trained ANN making extensive errors of object detection. This is particularly so when real world

circumstances deviate from the training data, as is so often the case. When robustness and security are of the essence, they cannot be relied on. An image classifier trained only with pictures of brown horses and black cats might classify all black patches as cats, or brown ones as horses. Future research will no doubt push back these limitations. But the resulting systems will still not be comparable to human learning based on creative intelligence.[30]

The limitations of both models—the representational and the connectionist—are well known. Many researchers have proposed ways of combining them, harnessing connectionist power to certain aspects of the representational model to avoid its inherent brittleness. Such a model would, for instance, combine perceptual knowledge of what dolphins look like with a verbal characterisation that they live in water along with other features, thus producing more reasoning power. The AlphaGo programme, designed to play the game Go, merges a dynamically constructed, symbolically represented search tree with a variety of connectionist modules for estimating the value of various positions.

As we tried to explain until now, AI studies functional intelligent behaviour, in particular the accomplishing of goals by means of data processing, leaving aside considerations of value. Given this fundamental understanding, it is clear that AI needs to be governed in ways that will be beneficial for all humanity. A particular aim is to understand the importance of AI, the values and counter-values it generates, and so to contribute to its effective governance. This governance of AI must be part of the democratisation of the techno-sciences. It demands democratic, interactive initiatives within civil society.

6. The Governance of AI

In society's current state of development, algorithms work at the intersection of computing, culture and human intelligence. Their increasing presence in every aspect of life emphasises the need for proper democratic and ethical governance. This governance will become more urgently necessary as the social impact of AI increases. There is already research on the ethical regulation of the design and use of AI systems. It is a complex but necessary task. The challenge is to devise the global norms, policies and institutions which will best ensure the beneficial development and use of advanced AI. Some examples of such proposals suggest the banning of AI armaments, of AI bots in financial decisions and of any invasions of privacy by AI.[31]

[30] R. López de Mántaras, «El traje nuevo de la inteligencia artificial». *Investigación y Ciencia*, July 2020.

[31] A. Daly et al., «Artificial Intelligence Governance and Ethics: Global Perspectives». The Chinese University of Hong Kong Faculty of Law Research Paper No. 2019-15, 4 July 2019. Available at https://doi.org/10.2139/ssrn.3414805. Last accessed on February 8, 2021.

Independent review boards will be increasingly required to assess the ethical validity of AI applications. Although such boards currently assess academic research, AI applications by governments (in the military sector) or private corporations (undertaking subtle invasions of privacy) are unlikely to fall under their oversight. AI research, unlike other techno-sciences, is largely undertaken by those technological companies that currently rank top among the public corporations with greatest market capitalisation. Only ten years ago these top positions were dominated by the oil and gas industry.[32]

There is a clear need to significantly expand the purview of independent review boards and government control. However, the ethical analysis of AI is complex because it involves countless interactions—between designers, developers, users, software, and hardware. This dispersed nature of AI implies a shared responsibility for its effects and actions. In addition, any ethical analysis must not be merely general; it needs to take on particular AI technologies and systems such as self-driving cars or automatic decision-making by AI systems which are not fit for purpose.

AI has to face specific governance challenges, some of them extreme. These include labour displacement; inequality; a global market dominated by a small group of large sellers; totalitarian regimes as digital dictatorships with autonomous weapons or AI sensor technology enabling cheap, extensive and effective surveillance; shifts and volatility in civil society; strategic instability; an AI race that sacrifices safety among many other values.

6.1. Proper wording

The ancient wisdom of Confucius about proper wording being the prerequisite of good governance is more apposite than ever. The world's current crisis of democracy is closely related to the misunderstanding of language. Such misunderstanding gives rise to its misuse—it goes from superficiality to fake news (facilitated by information technologies) followed by the corruption of all the other powers that make us human. Trust—another basic condition for good governance—is no longer possible. Trust requires comprehension and comprehension requires the right use of language's creative power.

Proper wording is the first requirement in the governance of AI. Why do we use the word *intelligent* for a machine that is a stranger to both meaning and creative freedom and does not understand anything? As a general principle, the terms *intelligence*, *autonomy,* and *ethics* should not be applied to machines. This simple recognition would clarify and solve many of the issues connected with the governance of AI.

[32] List of public corporations by market capitalization. In *Wikipedia*. Retrieved February 8, 2021, from https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization.

6.2. The inadequacy of rational ethics

There have already been many different attempts to guarantee the beneficial development of AI. Governments have designed strategic plans in anticipation of its growing economic and social impact. In 2019 the EU drew up *Ethics Guidelines for Trustworthy AI*. The Institute of Electronics and Electrical Engineers (IEEE) have produced its *Global Initiative on Ethics of Autonomous and Intelligent Systems* to advance public debate about the values and principles underpinning the uses of AI. There are several declarations of principles; they include the *Asilomar AI Principles* (2017), the *Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe* (2017), and the *Montreal Declaration for the Responsible Development of Artificial Intelligence* (2018). There is also an open letter on *Research Priorities for Robust and Beneficial Artificial Intelligence* that emerged from a 2015 conference.

These debates are mainly shaped by developed nations and dominated by economic rationales. This raises obvious concerns about the neglect of other countries and their contributions; local knowledge is dismissed; cultural pluralism and demands for global fairness are overlooked. Moreover, in these documents, values are treated mainly at a purely *rational* level—as concepts—rather than as collective feelings. They are addressed more to people's heads than to their hearts. But if we consider, for instance, the value of privacy, this can only be effectively addressed by being aware of its countervalue: the surveillance to which we are all constantly subjected. Approached in this way, we are aware of what it is that we find repellent about it, and so have a clear sense of the value of privacy.

Ethical principles, established conceptual values, regulations and codes of conduct are not in themselves enough. We need to develop our creative axiological intelligence which in turn must be rooted in an awakened liberating intelligence. Rational ethical considerations are a part of axiological intelligence, but by themselves they are incapable of establishing values which will motivate and guide human activity. This impotence of modern rational ethics is what lies behind the crisis of values in modern societies.

6.3. Beneficial AI

Much current AI research is underpinned by the old ideal of designing computational systems to be autonomous agents that will adapt and self-improve until they become essentially autarchic, i.e., independent and self-governing. These computational systems are thought of as individuals in a society which in turn is conceived as being merely a set of individuals. This view has given rise to questions about ways to ensure these systems will benefit humanity. Such questions even extend to the apportioning of benefits and costs between machines and humans with conflicting desires.

Declarations have emerged: «Machines are beneficial to the extent that their actions can be expected to achieve our objectives».[33]

Working with the assumption that near-autarchic 'intelligent' machines are coming, different proposals have been made to ensure a positive outcome for humankind. The general thrust of these is that such machines should comply with the understanding of values as outlined above. Some essential conditions have been advanced: machines should be understandable; they should have transparent purposes; they should ask permission before carrying out any potentially dangerous actions; they should accept inspection and correction by humans, with the potential to be switched off when necessary. It has also been stressed that we should prevent machines from interfering with parts of the world the real value and interests of which they are ignorant. In addition, it is proposed that machines should be capable of learning more about our true, underlying preferences for how the future should unfold.[34]

Based on the metaphoric attribution of reasoning power to AI systems, there has been an attempt to create beneficial AI systems by making the machines themselves ethical—empowered to take ethical decisions without human intervention.[35] They would be able, for instance, to operate in this way in the field of automatic hiring.[36] Even more ambitious is the attempt to ensure that autonomous vehicles, and systems in other safety-critical contexts, would make the ethically preferable choice.[37]

However, if human rational ethics are often inadequate, this will be even more the case with those produced by 'rational' machines. Generally speaking, 'ethical' machines *could* functionally simulate some given ethical behaviours. However, they lack the sensibility of *life* that comes with axiological intelligence. 'Ethical' machines only address a fixed set of values; they cannot deal with the extensive background field of values that must be considered in making any ethical decision. 'Ethical' machines, making independent decisions, threaten to erode human autonomy.

The priority within AI should always be the design of beneficial machines rather than ones that are supposedly intelligent, independent and self-governing. We have a name for such machines: *tools*. Their essential characteristic is that they are aligned with human values. If we design our AI systems as tools or as services which automate specific tasks—always remaining open to our intervention in order to direct and im-

[33] S. Russell, *Human Compatible: AI and the Problem of Control*. Penguin Books, 2019.

[34] Ibid.

[35] M. Anderson and S. L. Anderson, «Machine Ethics: Creating an Ethical Intelligent Agent». *AI Magazine* 28(4), 2007, p. 15-26.

[36] P. Tambe, P. Cappelli and V. Yakubovich, «Artificial Intelligence in Human Resources Management: Challenges and a Path Forward». *California Management Review*, First Published August 2, 2019.

[37] J. C. Gerdes and S. M. Thornton, «Implementable Ethics for Autonomous Vehicles», in: M. Maurer et al. (eds.), *Autonomous Driving. Technical, Legal and Social Aspects*. Springer, 2015.

prove their activity—there is no room, and no need, for autarchic machines and the problems they bring.[38]

6.4. Autonomous AI systems

Within the framework of enactive cognitive science, the autonomy of an organism is understood to be based on a network of intra-dependent, recursively enabling processes through which the organism constitutes and sustains itself under precarious conditions. This precariousness requires it to interact with the world. These interactions have intrinsic value (positive, neutral, or negative) for the organism itself and for the continuation of its own autonomy, its freedom. This is a constitutive autonomy that living systems enjoy by virtue of their self-individuation. We can understand autonomy as a basic feature of life's sense-making intelligence.[39] For humans, precariousness demands not only interactions with the environment in order to maintain autonomy, but also continuous social-learning interactions, based on the exercise of creative intelligence. Human autonomy depends on participatory sense-creating, through language and the background field of values, out of which responsibility is central.[40]

In contrast with the autonomy of an organism, there is a limited version, as seen in intelligent autonomous agents. Their type of interaction with the environment cannot be described as a sense-making activity that is constitutive; therefore, the machine cannot be said to have actual autonomy, much less any kind of responsibility. Humans should always be the only actors ultimately responsible for technological artefacts. It is far from clear how we could align our goals with those of 'autonomous' machines by getting them to learn, adopt and retain those goals. Their very design by itself puts many obstacles in the way of their potential to be of benefit to humanity. The prospect of 'intelligent' machines with their own goals of self-preservation, self-understanding and self-replication, threatens to create monsters which it would be impossible for humanity to live with.

[38] The *Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe*, for instance, calls for 'constrained autonomy' («to have clear rules constraining the behavior of autonomous AI systems») and for not neglecting the importance of the 'human role' («All AI systems critically depend on human intelligence. ... and often real benefit comes from the synergy between human and artificial intelligence»: L. Steels and R. López de Mántaras, «The Barcelona declaration for the proper development and usage of artificial intelligence in Europe». *AI Communications*, 3(6), 2018, p. 485-494.

[39] M. Villalobos and D. Ward, «Living Systems: Autonomy, Autopoiesis and Enaction». *Philosophy & Technology*, 28, 2015, p. 225-239.

[40] H. De Jaegher and E. Di Paolo, «Participatory sense-making». *Phenomenology and the Cognitive Sciences*, 6, 2007, p. 485-507.

7. Conclusion – Towards Cooperative AI systems

We should abandon the fixation with producing autonomous or autarchic computational systems that are made to emulate or even rival humans. Instead we should address the challenge of creating tools and services that will increase the extent of human intelligence. This is the kind of AI we need in order to handle the complexity of our current world.

In designing AI, attention should be focused on its interaction with human intelligence in its social embodiment. Research in human intelligence and in AI should not only collaborate but also be alongside each other, with interdisciplinary teams in the same institutions. In such a cooperative model each partner is doing what it does best, thus avoiding the metaphorical confusions that abound in any discussion of AI.

The great benefit of AI systems is that they can free us from tedious tasks involving what we already know, liberating us to be creative—the specific power of human intelligence. The question now is how AI will help to develop each of the powers of this creative intelligence, particularly in creating *subsidiary symbiosis*, the kind of non-hierarchical ways of cooperation that will be key in any future creative democracies, i.e., democracies that are organised following the principle of subsidiarity, preventing the concentration of power and control, and in which creativity and freedom is nourished in a generalised way and at all levels.[41]

Humans and AI can actively enhance each other. For humans, leadership, teamwork, creativity, social skills and a sense of humour all come naturally. For machines, speed, scalability and analysing terabytes of data are all straightforward undertakings. An AI system might handle thousands of designs matching the designers' specifications. They then choose what they like or dislike, leading to a new round of designs. Axiological intelligence—their creativity, professional judgment and aesthetic sensibility—can thus be deployed.[42]

In the near future, many activities will be redesigned to support the partnership between human intelligence and AI. New jobs will appear to ensure that AI systems work properly and safely; data officers will guarantee that data feeding AI systems complies with consumer-protection regulations. Cooperative AI systems bring maximum transparency—the best way to safeguard the values described earlier. They reduce the prospect of a future AGI more powerful than human intelligence.

This cooperation requires a particular education of human intelligence, axiological as well as technical. It is the user rather than the designer who manages software coupling, who creates and communicates meaning. As is the case in human-computer

[41] J. Agustí, «Democràcies creatives». *El Punt Avui,* January 27, 2018. Available at http://www.elpuntavui.cat/opinio/article/8-articles/1328420-democracies-creatives.html. Last accessed on February 8, 2021.

[42] S. L. Epstein, «Wanted: Collaborative intelligence». *Artificial Intelligence* 221, 2015, p. 36-45.

interface research, AI should put psychological and sociological insights at the heart of the design process in order to create a better fit with everyday human activity, understanding and interaction.

The governance of AI must be a part of the democratisation of the techno-sciences. It demands democratic, interactive initiatives within civil society along with increased attention paid to the distribution of whatever benefits (and they are predicted to be huge) that might arise. All of this is dependent on the development of a mature, harmonious and socially embodied intelligence within the majority of the population. This is the necessary foundation of a creative democracy, the only form of society that can ensure a future for humankind.

Jaume AGUSTÍ-CULLELL
Cofundador de l'Institut d'Investigació en Intel·ligència Artificial, IIIA-CSIC
jaumeagusti@gmail.com

Marco SCHORLEMMER
Institut d'Investigació en Intel·ligència Artificial, IIIA-CSIC
marco@iiia.csic.es

Article rebut: 16 d'octubre de 2020. Article acceptat: 1 de febrer de 2021