Revista Direitos Humanos e Democracia

Editora Unijuí • ISSN 2317-5389 • Ano 12 • nº 23 • Jan./Jun. 2024 • 15909
Programa de Pós-Graduação *Stricto Sensu* em Direito da Unijuí
https://www.revistas.unijui.edu.br/index.php/direitoshumanosedemocracia
http://dx.doi.org/10.21527/2317-5389.2024.23.15909

FROM SOPHIA TO GENERAL ARTIFICIAL INTELLIGENCE:
Anthropomorphism and Consciousness at the Current State of Art of Technology

Mateus de Oliveira Fornasier
Universidade Regional do Noroeste do Estado do Rio Grande do Sul (Unijuí).
Programa de Pós-Graduação em Direitos Humanos. Ijuí/RS, Brasil.
http://orcid.org/0000-0002-1617-4270

ABSTRACT

The main goal of this article is to define the possibilities and obstacles of the development of an artificial intelligence (AI) capable of being as intelligent and capable as (or even more than) the human being. As a result, it was found that despite the great scientific advances in human behavior and the functioning of the brain, little is known about what consciousness is and how it works, which obliterates the development of a general artificial intelligence (GAI). Methodologically, it is a research developed according to the dialectical method of procedure, with a qualitative and transdisciplinary approach, and bibliographic review research technique.

Keywords: artificial intelligence; consciousness; anthropomorphism.

DE SOPHIA À INTELIGÊNCIA ARTIFICIAL GERAL: ANTROPOMORFISMO E CONSCIÊNCIA NO ESTADO DA ARTE ATUAL DA TECNOLOGIA

RESUMO

O objetivo geral deste artigo é definir as possibilidades e óbices do desenvolvimento de uma inteligência artificial (IA) capaz de ser tão inteligente e capaz (ou até mais) do que o ser humano. Como resultado, tem-se que, apesar dos grandes avanços científicos sobre o comportamento humano e o funcionamento do cérebro, ainda sabe-se pouco sobre o que é e como funciona aquilo que tem sido chamado de consciência, o que oblitera o desenvolvimento de uma inteligência artificial geral (IAG). Metodologicamente, trata-se de pesquisa desenvolvida conforme o método de procedimento dialético, tendo abordagem qualitativa e transdisciplinar e técnica de pesquisa de revisão bibliográfica.

Palavras-chave: inteligência artificial; consciência; antropomorfismo.

Submitted: April 2, 2024
Accepted: April 27, 2024

INTRODUCTION

A humanoid robot named Sophia has sparked controversy around the world as it has received citizenship in Saudi Arabia and has made media appearances since 2016. Although its maker Hanson Robotics has praised it as representing the future of artificial intelligence (AI), thinkers of several areas of knowledge are less optimistic about its capabilities, describing it as a sophisticated puppet or chatbot. Indeed, it is very likely that their performances are choreographed to promote specific political and economic interests, such as the interests of technology industries and their governmental promoters (Parviainen; Coeckelbergh, 2020).

Furthermore, the limits of technology have been increasingly exceeded socially, and popular culture has also been causing, through its most popular representations (films, games, etc.), issues about human values and the attributes that differentiate us from other entities – but the specific legal aspects related to personality are out of step in this regard, with only sporadic and punctual advances. Although popular culture creates useful representations, those fictions do not allow the formulation of a coherent model for entities to which legal personality may (or not) be attributed (Arnold; Gough, 2018, p. 31-33). AI is a very broad expression, encompassing a series of technologies – from physical entities, such as autonomous cars, to more abstract ones, such as parts of software – and this spectrum will be increasingly nuanced and extended by technological evolution. Thus, the regulation of any form of AI cannot take a "one-size-fits-all" approach, nor can any intelligence capable of developing knowledge be expected to be effectively regulated as well (Chen; Burgess, 2019).

AI has become daily, but in most countries its use is not yet regulated, which results in a legal vacuum. Thus, when damages due to its use occur, liability may theoretically be assumed by various parties – consumers, producers, third parties (such as trainers or designers), even the robot itself, and the definition of the liable one depend on how each country normatively considers AI: in some (such as Saudi Arabia), AI-based entities may be considered citizens, and robots may become legitimized to sue, to obtain equality before the Law, etc. Thus, thinking about AI legislation has many complexities – including those related to the involvement of many stakeholders. Various normative frameworks concerning accountability and legal personalization of such existing apparatuses can be considered – from equating AI to living beings, through analogy to common products, to creating entirely new concepts for frameworks regulating AI (Sumantri, 2019).

In this context, this research problem emerges: what are the possibilities of establishing an AI as consciously and cognitively capable as the human being? As a hypothesis, although there have been great technological advances in the field of AI due to the investment of large financial resources and significant time of research and innovation in the development of AI in the last two decades, there is still a long way off (decades, perhaps) of the development of a general artificial intelligence (GAI), as it is still not generalized, among scientists of the most varied areas, the necessary transdisciplinary thinking to promote the evolution of knowledge until the day when the understanding about how to artificially develop consciousness will be possible. In other words, despite great scientific advances in human behavior and the functioning of the brain, little is known about what consciousness is and how it works.

The main objective of the research reported in this article, which was done through the dialectical method of procedure, with a qualitative and transdisciplinary approach, and literature review research technique, is to define the possibilities and obstacles for the development of an AI which is capable of being as intelligent as (or even more) the human being. To achieve this objective, its development was divided into two parts, each corresponding to a specific objective. Thus, it starts with establishing what GAI and super-intelligence are. Afterwards, the relationships between AI, emotions and consciousness are studied, focusing particularly on the question about the need for AI to be anthropomorphic.

1 FROM AI TO SUPERINTELLIGENCE: RISKS AND PROMISES

The argument according to which artificial beings could act by emulating intelligence is called the weak AI hypothesis; and the claim that the machines that do it are actually thinking, and not just simulating reason, by its turn, is called the strong AI hypothesis (Russell; Norvig, 2016, p. 1020). Most AI researchers take weak AI for granted and do not care about strong AI – which means that as long as their software works, they do not care whether the intelligence is simulated or real. However, every researcher should be concerned about the ethical implications of their work, even in this sense.

One of the most influential and persistent critiques to AI as an enterprise is the behavioral informality argument, according to which human behavior is too complex to be captured by any simple set of rules and that, being computers, they cannot do more than follow a set of guidelines, and so, they cannot generate behavior as intelligent as that of humans. And the inability to capture everything in a set of logical rules is called a qualification problem in AI.

For Eliasmith (2015), humanity is currently at a unique point in the development of critical technologies for the realization of artificial minds – the rapid evolution of robotics, brain-like computing and new theories of large-scale functional modeling mean that there will soon be a significant increase in the abilities of artificial minds. It canals be predicted that in about five decades intelligence and physical abilities at the human level will be attained by artificial minds. Added to this are the large amounts of public and private resources directed to the construction of artificial minds. High-tech companies (IBM, Qualcomm, etc.) have invested billions of dollars in machine intelligence. And State funding agencies including Darpa (US Defense Advanced Research Projects Agency), EU-IST (European Union Information Society Technologies), Iarpa (Advanced Intelligence Research Projects Agency), ONR (US Office of Naval Research), and AFOSR (US Air Force Scientific Research Office) have contributed with major investments to a wide range of brain-inspired computing projects. And the two billion-dollar US and EU special initiatives will further deepen the understanding of biological cognition, which inspires artificial-mind builders.

The alignment of those forces will support unprecedented advances in understanding biological cognition, but there are several challenges to reaching the artificial mind in fifty years. First, robotic actuators still fall far short of the efficiency and speeds found in nature. It is clear that advances in materials science will help overcome these limitations, but it is impossible to predict how long this will take. Second, artificial touch and proprioception sensors, which are essential for fast, fluid motion control, that are similar in scale and precision to those naturally available, do not yet exist – although visual and auditory perception technologies are

already very advanced. Thirdly, current theoretical methods for integrating complex systems on a large scale will have to undergo major development yet, but this will only become clear as humanity seeks to build even more sophisticated systems. Also, perhaps as important as all these factors, there are many others, linked to the neurobiology of cognition, that may influence this fifty-year prediction: perhaps other cells, in addition to neurons (such as glial cells), still play some unknown key roles in intelligence. Furthermore, genetic transcription processes that influence learning must still be studied in detail. And finally, perhaps it is still necessary to understand issues at the quantum level to explain the development of cognition.

It is important to realize, however, that the media shapes, mediates and amplifies expectations around AI, influencing its potential for intervention in the world thus (Brennen; Howard; Nielsen, 2020). The media enables the creation of expectations about pseudo-artificial general intelligence, which would be, in fact, a collective of technologies that would be capable of solving almost any problem. In other words, rather than a single system, such an expectation offers a collective of systems capable of doing that. Thus, it supports a pseudo-artificial general intelligence. In part, the construction of the pseudo-IAG is based on the lack of specificity in the way the media talks about AI. But the pseudo-IAG builds on the fundamental optimism regarding the promise and potential of AI that infuses both academic discourse and media reporting on AI.

In *On Defining Artificial Intelligence*, Pei Wang (2019) presents the following definition: "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources". While the author considers this an adequate definition, Yampolskiy (2020) believes that there is a fundamental difference between defining intelligence in general, human intelligence in particular, and AI, as the title of Wang's article claims to do.

Typically, AI is explicitly designed to benefit its developers and users, and this is an important factor to include in the definition of AI. Another factor that should be explicitly mentioned is the definition of AI, or at least what its necessary subcomponents are, which give it controllability, explainability, understandability, predictability, and possibility of being corrected – otherwise, any definition would be dangerously incomplete. The development of the IAG, when it is finally successful, is seen as a future shift in the trajectory of human civilization – and in order to reap its benefits and avoid its pitfalls, the ability to control it is critical, and presupposes the ability to limit its performance (by setting it to a certain level of IQ equivalence, for example), as well as the ability to turn off the system, the presence of free will, the possibility of autonomous target selection, and specification of the moral code that the system might apply in its decisions. Another key capability will be the possibility to modify the system after it is deployed to fix issues discovered during use. An AI must also be able to have its decisions explained in a language understandable to humans, as well as the ability to enclose such intelligence in a constrained environment or where it operates with reduced computing resources is imperative. Finally, as little bias and as much transparency as possible should characterize the decisions and actions of an AI that is humanely friendly, secure, and protective.

Because of all these safety-related factors (practically, theoretically and ethically speaking), Yampolskiy (2020, p. 3) believes that AI should be defined as "a fully controlled

agent with the capacity of an information processing system to adapt to its environment while operating with insufficient knowledge and resources". Furthermore, the direction of AI development towards emulating (or overcoming) the human presages predictable ethical complications (Bostrom; Yodowsky, 2014). Social roles can be filled by algorithms, implying new design requirements such as transparency and predictability. Furthermore, AI algorithms may no longer run in predictable contexts, requiring new types of security assurance and engineering ethical considerations. Perhaps AIs with sufficiently advanced mental states will claim moral status, which can trigger their legal regard as persons – albeit personas which are very different from the kind that now exist, and governed by different rules. Still, the perspective of AIs with superhuman intelligence and abilities presents the extraordinary challenge of establishing algorithms for superethical behavior.

Biological minds are just one of the types of minds that can come into being when AI technology is fully mastered. However, much of morality is based on assumptions about human nature that are not necessarily valid for digital minds. It is therefore necessary to reflect on morals as humanity gets closer to the era of advanced machine intelligence. For Shulman and Bostrom (2020), digital "utility monsters" can emerge in this context – mass-produced minds with moral status and interests similar to those of humans or other morally sizable animals, so that collectively their moral claims outweigh the moral claims of the incumbent populations. It would be easier, alternatively, to create individual digital minds with individual interests and resource claims much stronger than humans. And if, on the one hand, disrespecting their moral status can produce a catastrophe of immense proportions, on the other hand, a naive way of respecting them can be disastrous for humanity. Wisdom thus demands reforms of moral norms and institutions, along with prior planning of the kinds of digital minds that will be created.

Techno-scientific progress can change people's capabilities or incentives in ways that would destabilize civilization: good examples of this are arms races, liberalization of the use of dangerous technologies by any layman, and the invention of economically advantageous processes that produce difficult disastrous negative global externalities of regular. Thus, Bostrom (2019) states that the world, in fact, must be characterized as vulnerable, as there is a level of technological development in planet Earth at which civilization will almost certainly be devastated, unless mankind leaves its "semi-anarchic standard condition". While the overall ability to stabilize a vulnerable world requires greatly expanded capabilities for preventive policing and global governance, the vulnerable world hypothesis offers a new perspective for assessing the risk-benefit balance of developments toward ubiquitous surveillance or a unipolar world order. Therefore, it is necessary to outline the technology policy in a different way – then, it should not unquestionably assume that all technological progress is beneficial, or that complete scientific openness is always better, or even that the world can manage any potential disadvantage of a technology after it is invented.

It simply cannot be believed that IAG and superintelligence will necessarily be beneficial when possible. Their widespread use on the Internet will allow the construction of profiles that are offensive to the privacy of consumers, patients, and citizens in general. When used in the achievement of services of the most varied natures (whether public or private), if there are not mechanisms that enable the explanation of the logic of their operating processes,

they will obliterate the possibility of contesting the legitimacy, legality and moral adequacy of its decisions in Courts – otherwise, no human will be able to understand the fundamentals and processes that such machines devise. They will replace human labor in various tasks and professions, so that if human education policies are not developed for symbiosis with technological entities and social well-being for those who will be replaced and incapable of reinserting themselves in productive activities – such as taxation for machines and universal basic income – human obsolescence will trigger violent social processes in the most varied senses. And the use of IAG and superintelligence in warfare could make the moral decisions about life and death of people out of human control, thus creating dystopian scenarios worthy of a *Matrix*.

Barfield (2015, p. 68 ff.) considers erroneous the view through which technology is considered merely a tool for human use, whose sole purpose is to improve humanity. While much of the technology to improve humans is just one way to help designing the next generation of AI machines, it is either: mankind is in the process of inventing the future of its own extinction, or in the eve of inventing technology to free itself from the confines of body and mind. In this context, it is interesting to note that, when acquiring DeepMind, a cutting-edge AI technology company, Google was forced to create an ethics and safety review board to ensure the safe development of such technology under its control. This type of strategy is interesting, as companies have agendas that do not always coincide with society's best interest; therefore, human-friendly machines must be designed, for when a super-intelligence is finally developed, it may come to the conclusion that perhaps existence on Earth is more sensible without humans.

Current "artificially intelligent brains" drive cars, deliver ordered packs by drones, assist in surgery, write sports and weather reports, and manage inventories – tasks that require impressive intelligence and, in some cases, complex motor skills. However, these abilities are still a long way from human levels, and it is very naive to believe that such impressive machines will continue to be mere tools at the disposal of human interests as they evolve towards superintelligence.

So, if AI threatens the existence of mankind, how might it possible, through Courts, regulatory agencies, police of all kinds (and other repressive State apparatus), and parliaments, to put an end to such a threat? Current proposals range from outright banning AI research to programming "sympathy" into the "minds" of AI systems, and also to government regulations designed to give AI certain rights (but to deny others). With regard mainly to constitutionalists, there are already Law theorists proposing the titling of rights originally designed for humans also to self-conscious machines.

As mankind moves closer to human-like AI, an "Artificially Intelligent Brains Law" will be needed to perfect legal institutions, which will provide a framework on which to discuss social and legal issues that will emerge when such AI arises. Civil liability, contractual rights and criminal culpability for artificially intelligent machines operating without any human intervention will perhaps be the areas where the main legal issues will arise. But this branch of Law will also have to focus on software, operating systems, and computational architecture of artificial brains.

Scientists have been working on reverse engineering the human cerebral cortex – the basis of human cognition – to create a computing architecture based on it. The cortex has about 22 billion neurons and trillions of synapses. A software simulation of the human brain would require computational capacity of 36.8 petaflops (or more), and memory capacity of 3.2 petabytes – which is technologically possible, and it would require perhaps 1 million lines to code all the information contained in the human genome that refers to the cerebral cortex. Furthermore, even if 100 million lines were needed, according to Barfield (2015, p. 75), the exponential acceleration of technologies that are currently being developed already indicate the future technology that will be needed, as well as the knowledge required to unravel the human brain – and in that sense, it would be a mere matter of time to simulate the human brain.

The brain architecture is different from the digital technology of computers (Mainzer, 2020, p. 221 *et seq.),* because while the technology has been optimized in a targeted, conscious way and in a short period of time, the brain's architecture has evolved more or less randomly over millions of years, under varying conditions and requirements. Biological nerve cells evolved over long periods of time from cells that first casually, then more and more frequently, generated nerve signals, to eventually specialize in generating action potentials for tasks of control and regulation. This led to highly sophisticated neurochemical signal processing, with synapses and ion channels, which enabled human intellectual abilities.

But biological neurons are very slow when compared to modern microprocessors. This slowness was compensated in human brain by an increased expansion of parallel signal processing, which led to the enormous network density. These complex networks and learning algorithms made it possible to recognize patterns in the brain, which is crucial for human survival. In the common architecture of computers, the signals are processed sequentially, and technology depends on enormous processing speeds – something possible with silicon hardware.

Mathematically, both approaches are equivalent; and with superintelligence, the advantages of one approach can be used to offset the disadvantages of the other. Thus, the material of microprocessors and transistors is more stable and resilient than biological neurons, and in case of defects, they can be replaced. Biological tissue, on the other hand, is under aging and pathologies. Therefore, artificial brain networks are technically conceivable and can process much faster and more relentlessly than nerve cells.

Short-term biological memory has the advantage of short access to it, but the disadvantage of low storage capacity, while Big Data technology already presents high speed of access to huge records. Moreover, errors and redundancies are typical of biological brains, when considering the difficulty that these organs have in learning and storing little data – but computers accurately transmit huge amounts of data through simple commands, and duplicate them however needed to other computers. On the other hand, dealing with errors, noise, and redundancies, has led to an "essential look" that does not get lost in the details: the effective evaluation of patterns and the discovery of general contexts characterize human intelligence. Thus, the advantages of human intelligence include neuronal areas, which allow to intuitively assess information and decide very quickly; but it is not technically impossible to develop algorithms that specialize in this in a superintelligence.

According to Kurzweil (2016, p. 162 ff.), the rate of technological innovation doubles every decade. The power of information technologies, on the other hand, doubles every year, which also applies to the amount of human knowledge. And with regard to information technologies, there is a second level of exponential growth – the exponential growth in the exponential growth rate – because as a technology becomes more economical, more resources are deployed to advance it, so that the exponential growth rate increases with time. While the computer industry of the 1940s was based on just a few projects that are now historically important, today the total revenue of the computer industry is more than a trillion dollars, for example.

Human brain scanning is one of the most exponentially evolving technologies, being that its temporal and spatial resolution, as well as brain mapping bandwidth, double every year. Only today is science gaining access to the means necessary to begin serious reverse engineering of the principles of operation of the human brain. Impressive models and simulations of a few dozen of the several hundred regions of the brain are possible now, but within two decades, a detailed understanding of how all regions of that organ work will be obtained – so much so that Kurzweil anticipates obtaining software models effective human intelligence for the mid-2020s.

The day technologies reach the level of development when human intelligence will have been fully emulated, computers will be able to combine the traditional strengths of human intelligence with the potentials of machine intelligence. The enormous human brain's ability to recognize patterns – which is due to the massively parallel and self-organizing nature of the organ – the ability to learn new knowledge – by applying perceptions and inferring principles of experience, including information obtained through language – the ability to create mental models of reality, and to conduct "what if" thought experiments with varying aspects of those models, will add to the artificial capabilities of instantly remembering billions of facts accurately, in high-speed, optimally repeated execution, precision and with no fatigue, from patterns learned by the machine, and with a very high capacity of knowledge sharing.

Non-biological intelligence will be able to download skills and knowledge from other machines and, perhaps, from humans, at the speed of light – which is much higher than that of electrochemical signals on which mammal biological brains are based. All knowledge of man-machine civilization will be accessible to machines via the internet. The machine's intelligence can be fully released in this way, not limited to biological barriers – slow switching speed of interneuronal connections, fixed size of the individual adult skull, etc. And once machines are able to design and develop technology like humans, but at much greater speeds and capabilities, they will have access to their own source codes and will be able to manipulate them. And biological limitations will be overcome with machine intelligence: for example, building living beings from proteins into one-dimensional chains of amino acids makes them weak and slow; but machines will enable the reengineering of all biological organs and systems, making them much more capable and resilient.

Although human intelligence is much more capable of changing its structure than it was recently imagined, the architecture of the human brain is profoundly limited. Only a hundred trillion interneuronal connections in each brain can be established, and a fundamental genetic change, which allowed for greater cognitive capacity in humans compared to other primates,

was the development of a larger cerebral cortex and a greater volume of gray matter in certain regions of the brain. But this evolution has taken place very slowly, and it has inherent limits to the brain's capacity. Machines, by reformulating their own designs and increasing their own capacities, mainly through the use of nanotechnology, will have much greater capacities.

Machines will also benefit from the use of very fast three-dimensional molecular circuits – which could be powered with devices like nanotubes, which are five hundred times smaller than today's silicon-based transistors. Because signals will need to travel shorter distances, they will also operate at speeds of terahertz (trillion operations per second), exponentially greater when compared to current speeds of a few gigahertz (billion operations per second). Thus, as the rate of technological change will not be limited to human mental speeds, machine intelligence will improve its own abilities in a feedback loop that human intelligence will not be able to keep up without artificial support. And this cycle of machine improvements in its own design will be faster.

Along with the accelerated improvement cycles of non-biological intelligence, nanotechnology will allow the manipulation of physical reality at the molecular scale, through microbots that could replace human cells. This will allow the reversal of human aging, the interaction with biological neurons – to broadly extend the human experience through creating virtual reality from the nervous system – and the increase in intelligence, with the exponential growth of the establishment of non-biological intelligence in the human brain – which has already started in computerized neural implants.

Human emotional intelligence – the ability to understand and respond appropriately to emotion – will also be understood and mastered by future machine intelligence. It is important because it enables the adjustment of emotional responses to optimize intelligence in the context of limited and fragile biological bodies. Future machine intelligence will also have "bodies" to interact with the world, but these virtual or nano-engineered bodies will be much more skillful and durable. Thus, future AI "emotional" responses will be redesigned to reflect their vastly enhanced physical capabilities.

There are real social utopias imaginable with the advent of superintelligence. The advent of GAI and Artificial Superintelligence (ASI) raises expectations of a true revolution in existence, whereby the intelligence that exceeds the greatest human minds ever known will be scaled in hardware and solve all of humanity's main problems (Atreides, 2019, p 191-193). Mediated ASI (mASI) will intelligently restructure social interactions, optimizing collectivities and corporations and, ultimately, replacing the organizational and governmental structures as they are currently known. Consequently, the structuring of society (into classes, statuses, etc.) will no longer make sense in a society where everyone will intelligently contribute to the benefit of all. With the mASI, the main contributors will perhaps be organized in classes, but in a post-scarcity social structure this distinction will not require higher salaries, as the struggle to meet basic needs will not be guided by currency. Among many of the smartest humans, the ideal distribution of individual times is obliterated by this struggle, by friction resulting from differences in intelligence among humans, and by the psychosocial negativities arising from such frictions. As mASI will be free of stigmas and prejudices, the solutions already proposed by humans would face less resistance, paving the way for more desirable futures, with more options available. Such systems will also make it possible to restructure fundamental elements

of the social structure that only evolved in disorderly and overlapping steps, through irrational and inefficient bureaucracies.

In this utopian scenario, the Western dream of achieving democracy – which currently seems to be nothing more than a hybrid between republic and oligarchy – will have new definitions. In practice, the basic problem of democracy is that it provides the option for quantity over quality. In societies where the quality of individuals' knowledge base is extremely variable, and where social predators live without effective controls, democracy is bankrupt by corruption and ignorance. But the emergence of mASI will level democracy, giving everyone better decisions, with systems that would explain exactly why they recommend certain actions. Of course, many individuals would still choose to act and think against this wisdom, but most could choose more wisely.

These are only speculations and probabilities, but based on well-founded theories and discoveries. Even if they do not materialize exactly in this way, they indeed warn for the need to change the approach of scientific and ethical knowledge that is related to the security of the GAI and superintelligence. In this sense, Yampolskiy (2016, p. 128-129) states that these concerns, not so long ago, were found exclusively in the scope of science fiction and few philosophers. Slowly, however, due to government concerns about security in some governments, the theme of superintelligent AI has started to appear in mainstream and prestigious scientific publications.

But super-intelligent AI can become a scientific field in its own right, supported by significant interdisciplinary foundations and attracting good professionals and scientists from a wide range of fields. The increased acceptance of transdisciplinary studies will enable publications in many academic circles. Increasing publication possibilities will enable scientists to replace philosophers so that practical algorithms can be developed in real experiments related to AI security engineering. This will further solidify AI security research as a major scientific topic of interest, and will yield some long-awaited answers. Lee (2019, p. 166-167) points out signs along this path, noting that the University of Illinois, Stanford University and other educational institutions started to offer a degree program of the "CS + X" type, integrating Computer and Human Sciences – a step forward in the right direction, to be taken by all schools of technology. Humanity is represented by the variable "X" in the formula CS + X.

Research on the ethical dilemmas of artificial superintelligence in coexistence with the human must change the domain of interest – from the exclusivity of concern of ethical theorists and philosophers in general, to the direct involvement of computer scientists (Yampolskiy, 2016, p. 140). It is also necessary to develop limited AI systems to experiment with non-anthropomorphic minds and improve safety protocols related to such entities. It should be added that not only active data scientists should be brought into the circle of ethical concerns, but also jurists, legislators and policy makers in general, given the wide range of socioeconomic consequences that achieving superintelligence will entail for the humanity.

However, it is always necessary to remember that technological development is not politically neutral. The main protagonists of the web – Google, Apple, Facebook, Amazon and Microsoft – promote the Singularity with great resources. And the big names that fund Singularity University with millions of dollars – Nokia, Cisco, Genentech, Autodesk, Google, Elon Musk, Bill Gates – and Google's December 2012 recruitment of Kurzweil demonstrate

corporate interest in promoting this idea of uniqueness. Other scientists and philosophers who vigorously announce the Singularity, such as Russell and Bostrom, raise funds from institutions financed by the industry itself (Ganascia, 2017, p. 108).

Those who are responsible for the massive and accelerated development of information technologies, however, warn of the great dangers that these same technologies represent for humanity. In other words: they create the problem and pretend to try to solve it. Google promises to create an ethics committee that will promulgate a universal bill of rights for technologies to prevent violations of human values, democracy and standards of good living – while the company itself violates EU regulations and turns a deaf ear to individual requests that invoke the right to be forgotten, among other illegalities. The ultimate motivations of these companies remain even more obscure because they generally do not cultivate philanthropy. As a result, three hypotheses can be formulated about what would encourage these companies to promote the idea of Singularity.

The first concerns the drunkenness of excess, the arrogance of the heads of the big web companies that managed to change society in a few years while carrying out unprecedented market capitalizations at astonishing speed. The recent successes in deep learning and processing related to Big Data encourage them and make believe that they will dominate the future, opening a new era for humanity. The myth of Technological Singularity, therefore, fits perfectly with the excitement of technology demonstrated by the web giants.

The second hypothesis for the interest of technology giants in the Singularity adds a mixed feeling of enthusiasm and fragility to the arrogance, of lack of control and loss of autonomy that Singularity itself would echo. Although this sounds strange at first sight, for such companies have conquered true empires, when considered in isolation, each of the main protagonists of the web experiences development as something random and anxiety-provoking, precisely because of the impossibility of controlling such evolution. The emphasis given to the exploration of Big Data stems from the need to identify the main trends and perceive the “weak signs” that betray them, based on the immense amount of information collected.

A third hypothesis concerns advertising: narrating past disasters and speculating on the possibility of a dystopian future always find great success. And Technological Singularity, turned into science fiction, anticipatory films, or sweeping advertisements by professors with prestigious professorships in the biggest universities, Nobel Prize winners or prodigious entrepreneurs, is something that is very successful commercially. There is an echo, in the mass media, of what these communicators report in scientific media.

But it can also be suspected that the goals of these tech giants are also political. The immeasurable success of these companies is accompanied by ambition not only for material achievements in the short term: from the beginning, these giants aspired to build a new society. In 2001, Larry Page, co-founder of Google, stated that his goal was to organize the world’s information and make it universally accessible and useful.

The economies of these large groups are often based paradoxically on free services (search engines, social networks, etc.) that drain paid activities (advertising, for example). For many of these companies, only capital raising or stock market prices measure success. Consequently, they do not plot hidden coalitions, as they are rivals in the search for funding.

But at the same time, they seem to feel that, whatever happens, none of them would be able, on their own, to dominate all the others, mainly because anti-monopoly laws would prevent this.

What feeds such a myth is also the emergence of new places of power, which distort the notions based on the old geographic territories, sometimes overlapping them, other times identifying with them. These regions, which somehow escape US influence, arouse the desire of high-tech industrialists, who see them as both very profitable and great sources of energy. It is easy to see that they place them at the center of strategies that, therefore, assume an unprecedented political dimension, capable of totally transforming planetary balances. The modern sovereign State, which was supposed to have competence for a number of functions, is now duplicated by the giants of technology, who intend to better perform its same functions and at a lower cost – security (with biometrics, cryptography, and security techniques, and management of civil records on social networks, for example), collection of taxes (with the registration of individual information in databases), definition and control of currency (creation of cryptocurrencies), services (development of education, culture, research applications and health, added to the use of wearable devices and the internet of things), environmental preservation, etc.

2 AI, EMOTIONS AND CONSCIOUSNESS

The expression "AI" was coined at a small conference at Dartmouth College in 1956, whose attendees are now seen as the fathers of that technology field. Perhaps these "founding fathers" were a little naive and incredibly optimistic – as they thought that humanity was ten years away from the machine's victory over a great chess master (which took about forty years, in fact), and that this period it would also be enough for the advent of computers capable of communicating with natural language (which has not yet been fully developed) (Wallach, 2017). Despite this, there are currently quite accurate AI-based language translators.

One of the most promising periods in the evolution of AI came, a few decades later, with the emergence of neural networks, computer platforms that use multiple processors, being that each one of them represents a neuron mathematically, emulating the thought processes of human beings. Neural networks were expected to make great strides, but they have not done that yet – although some time later this allowed for the development of deep learning.

But deep learning, perhaps the most promising advancement in AI to date, is not similar to the way human children learn – through roaming the environment learning about everything. It is a specific type of structured learning, which is only a subset of several different approaches to machine learning algorithms. Deep learning algorithms research within large amounts of data about certain subjects and find significant relationships among that data – relationships that humans would not discover or recognize without the help of excellent computing power. Deep learning can be applied to any database for a wide variety of applications.

While the AI is currently capable of comfortably beating humans in games like chess and Go, it does not mean to say that human faculties have been fully artificially recreated. Humans run on about 20W of power, and other incredibly efficient processes take place with humans, who are still able to solve all sorts of problems that computers are not. AlphaGo system, which

defeated human beings in Go, has perception as the only developed cognitive ability. But of other higher-order cognitive abilities, such as common sense, planning, analogies, reasoning, and language, no machine is yet truly capable.

Since the founding fathers of AI, there have been developers who are not interested in creating small, discrete, specialized apparatuses, but rather, in GAI, systems capable of doing everything that humans are capable of, to a comparable degree of competence. But not all AI researchers are interested in ASI – there is much disagreement about whether humanity will one day see the advent of such technology, not least because science does not even know enough about human intelligence to try to competently emulate it in artificial apparatuses.

Even with the great skepticism about superintelligence, this does not mean that there are no risks in relation to what is already being developed. There is a lot of concern about the use of AI in warfare, especially in systems capable of selecting their own targets and killing human beings. Many politicians, jurists, scientists, and philosophers have expressed themselves in the sense of banning such systems, given the superhuman efficiency with which they act. High concern has also been expressed to cyber warfare – the use of AI in social networks to hack systems and spread fake news. Building machines that make morally important decisions is a challenge that has suddenly occurred to AI researchers, and it is referred as the "value alignment problem." But humanity still lacks transdisciplinary knowledge: it is not enough, in this sense, for social scientists to be tasked with pointing out to computer scientists which ethical problems they must solve, as both fields (technical and ethical) must develop adequate technical skills to work together.

With regard to AI governance, it is a fact that policymakers and legislators generally do not understand the sciences, and work at a much slower and slower pace than technological evolution, with growing gaps between emerging technologies and their ethical oversight. cool. And regulation through soft law – which is generally thought as being more flexible, faster to develop and change – does not have the binding and coercive power of hard law.

There are those who believe that it will be possible to build artificial systems whose intelligence is comparable to that of humans in the not-too-distant future, regarding the current technological evolution of intelligent machines. Also that artificial human-like consciousnesses – that is, consciousness in machines – will be developed with such systems. However, even a distant realization of artificial consciousness gives rise to several philosophical questions of a nature – concerning thinking capacity of computers (whether they are thinking or just calculating); or the possibility that thought maybe could not a human prerogative, emulating in other beings, therefore (even if artificial); or the possibility of creating consciousness from other materials (silicon and metals, for example) that are not based on carbon, as it is the case of the human brain (Chowdhary, 2020, p. 9). These questions are currently unanswered, mainly because they require a combination of knowledge from Computer Science, Neurophysiology, and Philosophy. But the very argument about artificial consciousness – a possible product of human imagination, which would express desires and fears about future technologies – may influence evolution.

For Lee (2020), traditional value systems and conventional wisdom often fail when applied to technological innovations. It will be increasingly important to understand the limitations of human intelligence to adapting to such rapid social changes due to the ubiquity

of AI, and it is necessary to understand that intelligence is not limited to IQ. Intelligence has to be considered as the ability to make good decisions and solve all kinds of problems that lives face in ever-changing environments. The best solution in each circumstance depends on the needs and preferences of an organism. Therefore, it is nonsense to reduce the intelligence of a way of life to a single number, such as IQ – it may be convenient for some situations, such as evaluating one's adequacy to important productive tasks, but also gives the false impression that it is possible to compare biological and artificial intelligence on the same scale. IQ focuses on a single aspect of human intelligence and may be used to classify different individuals, but it does not reflect all of a person's intelligence.

As the range of AI applications increases, the unique abilities of individuals will become more important than standard measures of intelligence. As computers and AI become more sophisticated, the type of work needed to maximize production will change. In the past, huge amounts of time and effort were required to accumulate and recover the knowledge essential to the economy – which has led to high economic compensation (wages) for specialists in highly-trained areas, such as Medicine, Engineering, and Law – then, IQ and other standardized tests were commonly used to identify suitable candidates.

Human intelligence has biological and evolutionary roots – therefore, it is contiguous with animal intelligence. Humans and other primates share many common cognitive traits, even though human intelligence is different in at least two respects: social intelligence and metacognition, where human cognition overlap less with the abilities of other animals when compared to more basic learning algorithms such as classical and instrumental conditioning. It is not surprising, then, that its precise nature is not understood, nor are the biological mechanisms of intelligence.

Perhaps the most valuable impact of emerging AI on civilization depends on a better understanding of social and metacognitive abilities. Culture, science, and arts depend on social intelligence and metacognition, and as machines and AI increase their contribution to the production of various types of goods and services, freeing man from work, increasing the value of entertainment, and development staff will continue to increase. Society tends to devote more and more resources to these domains, requiring a more accurate understanding of human intelligence.

Human intelligence results from the brain evolution, and a better understanding of its functions is essential for the conception of theories about intelligence. Currently, the accuracy of instruments for measuring the activity of living brains is very limited. The development of non-invasive and accurate techniques for investigating and controlling human neural activity will accelerate the progresses in this area. Computer Science, Data Science, and AI-related studies are also closely related to Neuroscience, as they provide valuable mathematical frameworks for analyzing complex behaviors and their underlying physical mechanisms. Continued advances in digital technology will transform the industry and the understanding of human social and metacognitive intelligence, helping to find causes and cures for many devastating brain pathologies.

Although social intelligence and metacognition clearly distinguish humans from other animals, it does not mean that such forms of intelligence cannot be performed by AI. But it also does not mean that this development means approaching technological uniqueness,

or else, that AI will replace humans in all areas of intelligence. Intelligence is a function of life defined by self-replication, and life invented several solutions during evolution based on the principal-agent relationship principle to improve efficiency of self-replication. As long as computers do not reproduce physically, humans will remain the principal control behaviors of AI-beings, just as the brain is incapable of replicating itself and, as a result, continues to function as an agent for genes.

AI-gifted entities are designed to solve a relatively narrow set of specific problems in a mathematical way, which must be more efficiently performed than humans would in their applications – otherwise, there would be no economic demand for AI, and such entities would only exist for entertainment or research. Thus, maybe competition between AI and human performance is not a threat to human society, but a necessary condition for AI. Brains evolved as sophisticated learning machines, and this was a solution, not a threat, to the principal-agent relationship between brains and genes. Likewise, AI advances would not pose a threat to humans. Only if it has a set of its own independent values and utility functions antagonistic to those of humans will AI become a real threat. Otherwise, AI will be one of the many human tools to increase efficiency.

But if mankind intends to remain the main entity in relationships with AI, we should not create machines that reproduce without human intervention. A self-replicating intelligent machine could be considered a form of life. And generally Philosophy of Mind notions tend to characterize socio-cognitive abilities as if they were unique to sophisticated human beings – but if it is assumed that man is likely to share much of our everyday life with various types of artificial agents soon, a conceptual structure that explains agents other than the human is necessary (Strasser, 2018).

Although AI has achieved practical successes, many researchers focus on its scientific and philosophical potential, seeking to answer traditional questions about what the mind is, how it works, and how various types of minds can be produced by evolution – including minds in different stages of development in individual organisms. AI is still unable to faithfully replicate or model the minds conceived in such theories, even the oldest ones (Sloman, 2018). Young children and other animals make simple topological and geometric discoveries and use them to form intentions or control actions – a thing machines cannot do.

There are many implications for AI as Science, as Engineering, and as Philosophy, and also profound implications for Psychology and Neuroscience, as AI studies are not yet able to address the problem of how minds make discoveries about necessary truths and impossibilities that are not merely logical truths or falsehoods. There are also difficult biological problems to be solved about the evolutionary histories of the characteristics of human brains and minds that possess these capabilities. Only after these questions are answered will engineers be able to design artificial minds with the ancient mathematical capabilities of thinkers like Archimedes. Psychologists and neuroscientists also fail to realize that the explanation of mathematical cognition is not just about explaining numerical competences – as it depends on deduction, analogies, comparison, introspection, representation, and other conscious competences beyond statistics and probability.

Human beings are not only conscious, but also self-aware, as we are aware of ourselves as ourselves (Milliere; Metzinger, 2020). Awareness of oneself means thinking consciously

about (and like) oneself, using a concept of "self". But many philosophers understand that self-awareness is more pervasive in conscious human mental life than sophisticated cognitive states that involve the conceptual representation of oneself as oneself. Some even suggest that a more basic form of self-awareness is ubiquitous in all conscious experience.

Investigation of machine awareness began in the mid-1990s and is only gaining momentum – perhaps because such research relies on research into functional components necessary for awareness, including emotions (Scheutz, 2014). However, machine consciousness researchers do not have much contact with AI emotion researchers. And although these ones have been establishing proximities with Psychology, those of machine consciousness are more connected to philosophers interested in providing a functional and implementable view of consciousness.

AI emotions researchers work with the dimensions of communication, among others, but they ignore what emotions are and how they are implemented in humans, and many AI consciousness researchers do not research human consciousness – being more interested in so-called "weak artificial consciousness" (simulation of processes essential to consciousness), or in using principles of human consciousness to design better control systems. But there are scientists interested in conscious machines, who must address what is meant by "consciousness" and, eventually, what it would take to implement it – which is a very difficult problem, as philosophers and psychologists do not even agree about what conscience is.

Most proposals on consciousness in artificial beings made hitherto are merely conceptual, providing potentially implementable principles – for example, one might list the following as architectural requirements for a conscious system:

(I) An adequate method for representing information must be developed;

(II) Suitable elements for processing information, allowing its manipulation by chosen representation methods, must be designed;

(III) A machine architecture that accommodates sensors, effectors, processes of perception, introspection and meaning foundation, as well as the flow of speech and internal images, must be designed;

(IV) System design must accommodate functions of thinking, reasoning, emotions and language.

Five principles considered sufficient for an entity to be considered conscious in a sensorially accessible world can also be cited. This combination of sensory, imaginative, attentional and attentional representations could lead a being to have a first-person perspective – in humans, the "self". Such principles should not be motivated by a particular theory of consciousness, but by a set of individual discoveries which suggest that such principles are abstractions:

(I) Representation: entities must reach perceptual states that represent parts of the environment;

(II) Imagination: entities must have internal imaginative states resembling parts of the environment, or that build sensations similar to those of the environment;

(III) Attention: entities must select which parts of the environment to represent or what to imagine;

(IV) Planning: entities must control the state of imagination and its sequences to plan actions;

(V) Emotion: entities must have additional affective states that evaluate planned actions and determine the subsequent action.

Research on emotions has been an active interdisciplinary subfield in AI, and Machine Consciousness is about to establish a research community that pursues the design of conscious machines. Based on current trajectories, both communities are likely to grow together, as AI emotion community of researchers has been looking for more complex emotions that require several architectural features necessary for conscious machines (regret about one's behavior and disappointment with the attitude of another person towards oneself, for example) as postulated by the community of researchers on consciousness and AI – representations of one's perceptions, internal focus of attention, memories of past actions, representations of possible futures, etc.

Research in both areas promises to advance AI technology and to understanding human emotion and consciousness. It is also possible that both areas contribute to a better understanding of the trade-offs between emotional and conscious systems when compared to systems lacking one or both of these properties. However, as they are very early areas of research, satisfactory criteria for their success is not possible yet – that is, criteria that allow for identifying whether a particular machine has emotions or consciousness. This will involve arguments about the machine's functional architecture and the types of states it supports, as well as algorithms to determine whether a given system actually implements the functional architecture. Ideally, criteria for identifying whether a machine is in a particular emotional or conscious state already exist – and such criteria may well involve procedures analogous to those psychologists use to determine whether a person is in a particular emotional or conscious state.

It is also interesting to analyze Bedau's (2014) study on artificial life in order to understand the possibilities of understanding would an artificial being be endowed with conscience, emotions and, eventually, morality and legal personality. Artificial life constitutes a type of interdisciplinary study of life-like processes. In this sense, studies on artificial life have two distinct properties:

(I) They are focused on life in whatever form it may exist, concentrating on its essential characteristics, not on the contingent ones;

(II) Life is studied by synthesizing and simulating new forms of life and their fundamental processes, which allows for very flexible experimentation, allowing us to answer many general questions about the nature of life in a feasible and precise way.

Although research into artificial life is primarily a scientific activity, it raises philosophical issues, particularly with regard to the emergence, creative evolution and nature of life, and to the connection between life and mind, and social/ethical implications of creating life from "zero". Artificial life plays, in the first place, several roles in debates about the emergence of life. Emerging phenomena involve the relationship between wholes and their parts – each whole, concomitantly, depending on and autonomous from its parts. Philosophical problem of emergence involves assessing if the emergence of life is metaphysically legitimate and whether

it plays a constructive role in scientific explanations of apparent emergent phenomena. Ascendant models of artificial life generate examples of weak emergent macro-level phenomena. Thus, artificial life expands human understanding of the kinds of macro-level complexity that can have simple micro-level explanations – which provides Philosophy with new ways of thinking about the kind of emergency that seemed to be involved in life and mind to many people.

Second, studies on artificial life allow us to observe how life evolution produces increasing complexity – starting by very simple unicellular forms; then producing complex unicellular forms, with complex internal structures; passing through multicellular forms; also through large vertebrates with sophisticated sensory processing capabilities; and intelligent creatures with language and technology, finally. This growing complexity makes it possible to question whether evolution has an inherent tendency to create more adaptive complexity, or if it is just a contingent by-product. Artificial life provides a method to "play the tape of life again" (Bedau, 2014, p. 307): one can build an artificial biosphere analogous to the real one in the relevant aspects, in order to learn its typical and expected behavior by repeating the simulation. Software systems are the easiest artificial biospheres to build and "repeat the tape of life" in several different biosphere models, and they certainly shed light on the inherent creative potential of biological evolution, as long as the creative evolutionary potential of that biosphere is open enough.

Thirdly, studies of artificial life have helped to revitalize and reshape the controversial issue of life's nature. But it is only possible to simulate or synthesize essential features of living systems if there is some idea of what life is – so anyone looking to synthesize life in the laboratory is forced to face the general question of what life is. The connection between life and mind has had great philosophical interest in this regard, and all organisms have mental capacities, albeit rather rudimentary, generally speaking – they are sensitive to the environment in many ways, and this sensitivity affects their behavior. Furthermore, the sophistication of these mental abilities seems to match the complexity of these forms of life. Thus, studies on artificial life also question whether there is any important connection between life and mind, especially when one thinks that a central mind function is the capacity to adequate behavior in a complex world. Since all forms of life must deal with such a complexity, perhaps adaptive flexibility intrinsically connects life and mind.

Fourth, it is philosophically interesting to debate if a simulation of life may be considered artificial life, or whether only what is really concretized into a real entity may be. One might think that it is wrong to confuse a computer simulation of life with a real instance of it – because, however detailed and realistic the simulation is, it is a mere representation, without really exercising life, being the the ontological status intrinsic of this representation nothing more than the symbolization of certain electronic states inside the computer – no more alive than a series of sentences describing an organism – and which will appear alive only when appropriately interpreted. But many artificial life systems are not mere simulations of the known real world, but new digital worlds, which exhibit their own distinct forms of spontaneous self-organization: Conway's "Game of Life" being a perfect example of this. Thus, when run on computers, they contain new instances of self-organization, evolution, and multiplication, and can be incorporated into many different media, including the physical

media of properly programmed computers. Thus, as the essential properties of living systems involve processes such as self-organization and evolution, computers programmed according to such knowledge could be new realizations of life.

2.1 AI and anthropomorphism

The importance of conferring human traits on beings produced with AI must also be analyzed. The choice of anthropomorphic entities depends on the objectives of each technology researcher. According to Shneiderman (2020), emulation researchers prefer to use humanoid robots to better understand human perceptual, cognitive and motor skills, in order to build systems that perform tasks analogously to humans, in order to build intelligent systems that match or exceed the human performance. App developers, on the other hand, often prefer gadgets that are similar to tools or devices – then, they often apply AI methods to build widely used products and services. But while humanoid robots remain a popular emulation goal, they have had far less commercial success than tool-like devices. Even so, emulation inspires research and generates public interest. Powerful AI methods such as machine learning enable recommendation systems, speech recognition, image comprehension and natural language processing. When properly combined with users' data collection methods, design iteration, usability testing, and regulatory compliance testing, valuable products and services often emerge. Means also arise to support human effectiveness, to stimulate human creativity and to facilitate human social participation. Design compromises, which combine AI with other methods, need to be further shaped by the contextual needs of each application domain and thoroughly tested with real users. The resulting products and services have a great chance of meeting human needs in business, education, health, environmental preservation, and community safety, then.

AI has been historically conceptualized in anthropomorphic terms. There are algorithms whose design is biomimetic, seeking a certain isomorphism in relation to human brain. Others have more general learning strategies that coincide with theories from cognitive science and social epistemology. It is undeniable that several of the most innovative achievements in contemporary machine learning are somehow inspired by theories of neuroscience, cognitive psychology and social epistemology, but Watson (2019) postulates that the tendency to focus on structural affinities between biological and artificial neural networks suggests a mechanistic interpretation of intelligence that fails to explain its functional complexities.

Consequently, it would be wrong to consider AI as always being anthropomorphic, something Watson classifies as a misleading rhetoric arising from the impulse to humanize algorithms – and that would be an obstacle to properly conceptualize ethical challenges imposed by emerging technologies. The extent to which modern algorithms mimic human intelligence is in some cases exaggerated, and underestimated in others. The boundaries between machine learning methods are somewhat fluid, and more than one method is often combined to others. Moreover, while anthropomorphic analogies often help structuring learning strategies and inspire new approaches to AI research, the rhetoric built from there must be carefully thought, as the anthropomorphic bias in AI is not ethically neutral. Granting algorithms decision-making authority in socially sensitive applications may undermine the

human ability to hold powerful individuals and groups accountable for their technologically mediated actions.

Cardon (2018) has already demonstrated the possibility of designing systems that intentionally generate valid forms of information to express very high representative values – such as mental representations of the concepts of meaning, and the temporality of time as well. Such systems manipulate, within their organizational architecture of informational layers, variable aggregates of elements to control themselves the categorization of the appearance of the forms apprehended and used by the system itself. This is a specific feature of human intelligence, which thinks for itself in its psychic system. The understanding of the human psychic system and the computational model of artificial consciousness conceptually interact, and the conceptualized elements in one will serve to specify and deepen knowledge in the other – and this is an example of the transdisciplinarity necessary for the evolution of science in this area.

But the application of an artificially conscious metasystem, endowed with freewill and tendencies, and unifying many localized systems endowed with artificial consciousness, presents a significant ethical problem. Before this type of system is built, mankind must answer to problems that are ethical, not merely technical – which concern so much the social need to develop such systems, connectable to everything that is computerized, creating thus a layer of informational domination in real time, as to whether it is correct to have full human control of all material activities at all levels.

As more tasks traditionally performed by humans are entirely taken over by machines, questions about their governance become more complex in environments where the status of automation is apparent – that is, where humanity seems being present in the decision-making cycle, although not actually; and where humanity does not seem to be present, but actually is. And where the reality and appearance of decision-making systems are misaligned, such questions will become even more contingent. Jurists who study machine-driven processes have so far focused primarily on two questions: whether (and when) keeping humans involved will improve decision-making outcomes (making them more accurate and secure); and whether (and when) legal values that are not related to precision – especially those linked to ideas of legitimacy and dignity – are justified by the inclusion of humans in decision-making. To such questions, Brennan-Marquez, Levy and Susser (2020) add another, distinct but related one: does it make any difference whether humans appear to be in the decision-making cycle, regardless of actually being present?

Although the authors do not provide answers to these questions, they propose that control over the possibility of such misalignments should be democratic – both in deliberating such a possibility and in supervising their occurrence – although this ideal is not always confirmed, in practice, by several reasons: high costs, dependence on often scarce political desire, functional impossibility, etc. While democratic oversight is always important in principle, more worrisome than the occurrence of such appearance deviations is the illusion of familiarity that citizens will suffer when confronted with the passive acceptance of automation in situations where it is inadvisable. In that sense, people will not be able to consistently assess the costs and benefits of automation when its operation sounds too good to be true.

This does not mean that the apparent presence of humanity or false automation is always regrettable. Each of these appearances can have desirable characteristics that override concerns about deception in specific situations. However, weighing the damages of deception against other context-specific values requires knowing that deception is happening in the first place. While these misalignments are not always intended to sow confusion and alienation, they are also the cause of the frustration of the very cost-benefit investigation necessary to decide whether misalignments are permissible from the point of view of democratic legitimacy. The question about when such misalignments are allowed – and if not, what the appropriate remedy would be for each situation – will only have complex solutions that will require public deliberation and democratic oversight, not something imposed.

Technological advances in robotics and cognitive science have opened the way for even more sophisticated AI systems, capable of acting completely autonomously, to the point of mimicking human behavior in unprecedented ways, and making interactions between algorithms and humans very difficult to predict – racist comments and decisions from AI systems are good examples of this, as their early programmers did not intend to give them prejudiced biases, which were acquired after machine learning (Karanasiou; Pinotsis, 2017). This forces Law to inevitably develop a new concept of personhood to address the behavior of human-like agents.

Attempts to develop legal concepts for what "intelligence" is has been proving to be treacherous, given the relativity that such a concept carries. A simple question that emerges when trying to define what intelligence is for the first time is the following: is it correct to consider that an algorithm that generates human-like behavior is intelligent, or is it the human who perceives an artificial agent as intelligent? Turing (1950) already suggested that "intelligence" is related to the way human beings perceive it in a similar way to how the legal system operates. And Turing's "perception" of intelligence is similar to the principle of legal interpretation, as the ways through which Law interprets human behavior is not directed towards understanding the mechanisms (algorithms) that may have generated such behavior (Karanasiou; Pinotsis, 2017). This may be one of the reasons why automated systems are not easily perceived in Law and Humanities in general. Since intelligent processing is something opaque, it is desirable to go beyond the prima facie anthropomorphism of automated systems, and improve understanding of what machine "intelligence" can be. Deep learning, for example, can produce results that even programmers cannot predict. Therefore, the perception of what intelligence in a machine is must be improved.

CONCLUSION

Current AI learns through statistical and probabilistic mathematical strategies. However, mathematics itself was not conceived by the human mind with such procedures in the past – and this obliterates the conception of a proper general AI. Nor is there even a glimpse of the presence of emotions or consciousness in machines – not least because we still cannot understand, scientifically and in general, what consciousness is in the human mind. So, as advanced as it is, machine learning is still far from being general, intuitive and human-like learning. Despite this, research in the fields of emotions and AI, as well as consciousness and AI, has formulated a series of assumptions to be able, one day, to identify whether a machine

has any emotional or conscious state. Among such principles are, in addition to information processing and reasoning, representation, imagination, attention, planning, introspection, and emotion. However, as those fields of study are very recently emerged, there is still no great integration between them, nor the elaboration of criteria for identifying consciousness or emotions in machines – although it is already intuited that such criteria can benefit from current Psychology, to the extent that it determines human emotions and consciousness.

The possibility of one day being able to consider artificial beings as people could also draw on understandings of the field of study of artificial life, in which, through freedom of experimentation (mainly with software, hardware and laboratory tests) and synthesis of new forms of life, the essential characteristics for life in whatever form it may come to exist is researched. Philosophical questions about what life is can be transformed and answered not only by using formal logic as well, but also with scientific data and state-of-the-art computational technology. And from the Law, the abandonment of anthropomorphism will be required to define what intelligence is, being that the legal understanding of the operation of such machines should be conceived for the development of a notion of what "intelligence" is in advanced and unpredictable artificial systems, which behave like humans.

While many of the advances in contemporary machine learning have been inspired by Neuroscience, Psychology, and the Social Sciences, interpreting AI as always being anthropomorphic may be a controversial and dangerous misconception. The rhetoric of the analogy between machine and man could obliterate the ethical challenges that the machine's granting of decision-making authority over sensitive social processes poses. In other words: when comparing the machine to man in a general and simplistic way, in fact one is favoring the lack of knowledge of machine learning techniques that can cause ethical damage to humanity.

There are data and computer science researchers who has demonstrated that it is already possible to build artificial systems capable of generating, in their self-organization, mental representations of very high notions, such as temporality, and that think for themselves, such as the human psyche does. And this raises questions that are not technical, but ethical – which leads to the questioning of how necessary and how right and desirable it would be for humanity to develop such systems.

There will be situations where AI will take on human functions – such as in providing public services or administrative/judicial decisions – and it will appear that people (human intelligences) are present (but actually not). And there will also be situations in the future where people will not seem to be in control of certain decisions, but actually they will. When thinking about the cost-benefit of the appearance of humanity's presence in decision-making processes in public situations, it is estimated that AI decision-making is not always inadvisable: the problem lies in the passivity with which citizens will get used to decision automation. In this sense, although effective deliberation and democratic control are always difficult, public automation policies must not only be informed to citizens, but also consulted by authorities and legislators, so that the deceptive appearance of humanity does not cause irreversible damage to transparency public processes which are required by democratic rule of law.

There is a lot of effort and resources employed in the development of artificial minds nowadays, which leads authors to make the most diverse predictions about the time needed for the advent of IAG, usually within a few decades. However, even with so much being

invested in that, there are many technological and biological issues that still need to be resolved in order to understand the processes of cognition in nature in order to achieve such a purpose. Furthermore, due to all the ethical and practical risks that AI poses to humanity – especially the possibility of loss of human control of the moral agency, opacity, discriminatory bias in the operation and antipathy towards the human – it is essential to address, in whatever one define, the possibility of making it safe to use and deploy.

But more than economic and scientific advances are needed to achieve a IAG and a super-intelligence in a humanly acceptable way. Conceiving a vulnerable world due to technological advances means assuming that technoscientific advances can bring not only advantages, but destruction as well. Of course, this argument cannot be considered a reason for stagnating innovation, but rather pointing to a more pressing need to redirect policy towards an even more detailed and humanized weighing of the risks and benefits of the generalization of technology. Mankind is really under the blade of being replaced, relegated to the background and, ultimately, destroyed by the technologies it creates, including superintelligence and IAG.

Believing in a highly technological future may seem a politically neutral creed, arising from the pure "belief" in a technological Singularity due to the development of computers, data science, interdisciplinarity and the development of a "collaborative economy". However, the same companies and scientists that promote and fund institutions that warn about the risks and benefits of singularity and IAG are also those that mostly open up the emergence of such risks, and that act illegally – especially with regard to the privacy of the its consumers and customers. It is evident that, in addition to the arrogance resulting from the success and the fragility resulting from the uncertainty in technological development on the part of such companies, there is a great desire for global sociopolitical transformation behind this great advertising enterprise underlying the idea of singularity.

REFERENCES

ARNOLD, B. B.; GOUGH, D. Turing's People: Personhood, Artificial Intelligence and Popular Culture. *Canberra Law Review,* v. 15, n. 1, p. 4-33, 2018. Available at: http://www7.austlii.edu.au/cgi-bin/viewdoc/au/journals/CanLawRw//2017/1.html. Access in: June 22 2021.

ATREIDES, K. Choices in a Mediated Artificial Super Intelligence Assisted World: the Future Before Us. *In:* LEE, Newton (ed.). *The Transhumanism Handbook*. Cham: Springer, 2019. p. 189-214, 2019. DOI: https://doi.org/10.1007/978-3-030-16920-6

BARFIELD, W. *Cyber-Humans:* our future with machines. Cham: Springer, 2015.

BEDAU, M. A. Artificial Life. *In:* FRANKISH, K.; RAMSEY, W. M. (ed.). *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 2014. p. 296-316.

BOSTROM, N. The Vulnerable World Hypothesis. *Global Policy*, v. 10, n. 4, p. 455-476, 2019. DOI: https://doi.org/10.1111/1758-5899.12718

BOSTROM, N.; YUDKOWSKY, E. The ethics of artificial intelligence. *In:* FRANKISH, Keith; RAMSEY, William M. (ed.). *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 2014. p. 316-334.

BRENNAN-MARQUEZ, K.; LEVY, K.; SUSSER, D. Strange Loops: Apparent versus Actual Human Involvement in Automated Decision Making. *Berkeley Technology Law Journal*, v. 34, n. 3, p. 745-772, 2020. DOI: https://doi.org/10.15779/Z385X25D2W

BRENNEN, J. S.; HOWARD, P. N.; NIELSEN, R. K. What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news. *Journalism*, p. 1-17, 2020. DOI: https://doi.org/10.1177/1464884920947535

CARDON, A. *Beyond Artificial Intelligence:* from Human Consciousness to Artificial Consciousness. London; Hoboken: Iste: Wiley, 2018.

CHEN, J.; BURGESS, P. The boundaries of legal personhood: how spontaneous intelligence can problematise differences between humans, artificial intelligence, companies and animals". *Artificial Intelligence and Law*, n. 27, p. 73-92, 2019. DOI: https://doi.org/10.1007/s10506-018-9229-x

CHOWDHARY, K. R. *Fundamentals of Artificial Intelligence.* New Delhi: Springer, 2020. DOI: https://doi.org/10.1007/978-81-322-3972-7

ELIASMITH, C. On the Eve of Artificial Minds. *In:* METZINGER, T.; WINDT, J. M. (ed.). *Open Mind.* Frankfurt am Main: Mind Group, 2015. p. 1-17. DOI: https://doi.org/10.15502/9783958570252

GANASCIA, J.-G. *Le Mythe de la Singularité:* Faut-il craindre l'intelligence artificielle? Paris: Éditions Du Seuil, 2017.

KARANASIOU, A. P.; PINOTSIS, D. A. A study into the layers of automated decision-making: emergent normative and legal aspects of deep learning. *International Review of Law, Computers & Technology,* 2017. DOI: http://dx.doi.org/10.1080/13600869.2017.1298499

KURZWEIL, R. Superintelligence and Singularity. *In:* SCHNEIDER, S. (ed.). *Science Fiction and Philosophy:* from Time Travel to Superintelligence. 2. ed. Hoboken: Wiley, 2016. p. 146-170.

LEE, N. Beauty is in the A.I. of the Beholder: Artificial and Superintelligence. *In:* LEE, N. (ed.). *The Transhumanism Handbook*. Cham: Springer, 2019. p. 153-174. DOI: https://doi.org/10.1007/978-3-030-16920-6

MAINZER, K. *Artificial Intelligence:* when do machines take over? Berlin: Springer, 2020. DOI: https://doi.org/10.1007/978-3-662-59717-0

MILLIERE, R.; METZINGER, T. Radical disruptions of self-consciousness. *Philosophy and the Mind Sciences,* v. 1, n. I, p. 1-13, 2020. DOI: https://doi.org/10.33735/phimisci.2020.I.50

PARVIAINEN, J.; COECKELBERGH, M. The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. *AI & Society*, p. 1-10, 2020. DOI: https://doi.org/10.1007/s00146-020-01104-w

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence:* a Modern Approach. 3. ed. Essex: Pearson, 2016.

SCHEUTZ, M. Artificial emotions and machine consciousness. *In:* FRANKISH, K.; RAMSEY, W. M. (ed.). *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 2014. p. 247-268.

SHNEIDERMAN, B. Design Lessons from AI's Two Grand Goals: Human Emulation and Useful Applications". *IEEE Transactions on Technology and Society*, v. 1, n. 2, p. 1-17, 2020. DOI: https://doi.org/10.1109/TTS.2020.2992669

SHULMAN, C.; BOSTROM, N. Sharing the World with Digital Minds. *Nick Bostrom's Home Page*, p. 1-18, 2020. Available at: https://www.nickbostrom.com/papers/monster.pdf. Access in: June 22 2021.

SLOMAN, A. Huge, but Unnoticed, Gaps Between Current AI and Natural Intelligence. *In:* MÜLLER, V. C. (ed.). *Philosophy and Theory of Artificial Intelligence 2017*. Cham: Springer, 2018. p. 92-105.

STRASSER, A. Social Cognition and Artificial Agents. *In:* MÜLLER, V. C. (ed.). *Philosophy and Theory of Artificial Intelligence 2017*. Cham: Springer, 2018. p. 106-116.

SUMANTRI, V. K. Legal Responsibility on Errors of the Artificial Intelligence-based Robots. *Lentera Hukum*, v. 6, n. 2, p. 333-348, 2019. DOI: https://doi.org/10.19184/ejlh.v6.i2.10154

TURING, A. Computing Machinery and Intelligence". *Mind*, New Series, v. 59, n. 236, p. 433-460, 1950.

WALLACH, W. Rise of the Automatons. *Savannah Law Review*, v. 5, n. 1, p. 1-12, 2017. Available at: https://www.savannahlawschool.org/lawreview/. Access in: June 22 2021.

WANG, P. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence,* v. 10, n. 2, p. 1-37, 2019.

WATSON, D. The Rhetoric and Reality of Anthropomorfism in Artificial Intelligence. *Minds and Machines*, v. 29, n. 3, p. 417-440, 2019. DOI: https://doi.org/10.1007/s11023-019-09506-6

YAMPOLSKIY, R. V. *Artificial Superintelligence:* a futuristic approach. Boca Raton; London; New York: CRC Press, 2016.

YAMPOLSKIY, R. V. On Defining Differences between Intelligence and Artificial Intelligence. *Journal of Artificial General Intelligence,* v. 11, n. 2, p. 68-70, 2020. DOI: https://doi.org/10.2478/jagi-2020-0003

Corresponding Author:
Mateus de Oliveira Fornasier
Universidade Regional do Noroeste do Estado do Rio Grande do Sul (Unijuí).
Programa de Pós-Graduação em Direitos Humanos
Rua do Comércio, nº 3000 – Bairro Universitário. Ijuí/RS, Brasil. CEP 98700-000
E-mail: mateus.fornasier@unijui.edu.br