filosofia unisinos | Unisinos Journal of Philosophy

ISSN 1984-8234

Filosofia Unisinos
Unisinos Journal of Philosophy
25(1): 1-13, 2024 | e25110

Unisinos – doi: 10.4013/fsu.2024.251.10

Dossier

Logic and foundations of artificial intelligence and society's reactions to maximize benefits and mitigate harm

Lógica e fundamentos da inteligência artificial e as reações sociais para maximizar benefícios e mitigar danos

Dora Kaufman
https://orcid.org/0000-0001-7060-4887
Pontifícia Universidade Católica de São Paulo - PUCSP, Programa de Pós-Graduação em Tecnologias Inteligentes e Design Digital, São Paulo, SP, Brasil. Email: kaufman1955@gmail.com

ABSTRACT

Artificial intelligence is a general-purpose technology (GPT), term given to technologies that shape an entire era and reorient innovations by reconfiguring the economy's logic and functioning and bringing in new business models. AI offers unprecedented opportunities and risks. The benefits of AI are extraordinary, as are its potential harms. Potential damage does not have the same degree of problematization, since the intensity and extent of the damage varies according to the domain and the object of application. To address the scale of this challenge, regulation is necessary but not sufficient. Standards, unwritten codes of compliance and arbitration procedures, supervision and auditability, AI governance, international agreements, compliance with current local and global standards and laws. All of this needs to be integrated. Society seems to have no alternative to facing the challenges of at least mitigating the damage already identified and trying to predict future damage in advance. The purpose of this article is to encourage reflection regarding the main initiatives that are available to society to protect its citizens and organizations from the potential harm caused by AI models, vis-à-vis the technology's own limits to act in ethical and legal compliance.

Keywords: artificial intelligence, opportunities and risks, regulation.

RESUMO

A inteligência artificial é uma tecnologia de uso geral (GPT), termo dado a tecnologias que moldam toda uma era e reorientam as inovações, reconfigurando a lógica e o funcionamento da economia e trazendo novos modelos de negócios. A IA oferece oportunidades e riscos sem precedentes. Os benefícios da IA são extraordinários, assim como os seus potenciais danos. O dano potencial não possui o mesmo grau de problematização, pois a intensidade e a extensão do dano variam conforme o domínio e o objeto de aplicação. Para enfrentar a escala deste desafio, a regulamentação é necessária, mas não suficiente. Padrões, códigos não escritos de conformidade e procedimentos de arbitragem, supervisão e auditabilidade, governança de IA, acordos internacionais, conformidade com padrões e leis locais e globais atuais. Tudo isso precisa ser integrado. A sociedade parece não ter alternativa senão enfrentar os desafios de, pelo menos, mitigar os danos já identificados e tentar prever antecipadamente os danos futuros. O objetivo deste artigo é estimular a reflexão sobre as principais iniciativas que estão à disposição da sociedade para proteger seus cidadãos e organizações dos potenciais danos causados pelos modelos de IA, vis-à-vis os próprios limites da tecnologia para atuar em conformidade ética e legal.

Palavras-chave: inteligência artificial, oportunidades e riscos, regulação.

1 Introduction

Information is at the very heart of individuals' and organizations' decision-making processes: the more important the decision, the greater the search for information. Based on this logic, over the course of the last thirty years there has been an evolution of predictive models, structured with the aim of generating insights based on data, with positive impacts on decision-making. Over time, these statistical models have become more sophisticated to the point where they can now be confused with statistical models enabled by artificial intelligence (AI). What is new about this conventional process is the hyper-connected society, in which a significant part of our communication and sociability takes place in virtual environments and/or by means of virtual devices, continuously generating extraordinary amounts of new data, known as "big data".

Traditional statistical models are limited in their ability to deal with the dimensionality of large data sets. The technique of deep neural networks – known as deep learning, a sub-field of artificial intelligence - is the statistical probability model capable of dealing with high-dimensional data, which goes some way to explaining the spread of AI over the last decade in the execution of different tasks in different sectors. For example, an AI-enabled image recognition model is well able to cope with the large set of pixels that make up an image. This technique, subdivided into predictive AI and generative AI models[1], is still in its very early stages – having only been recognized by the academic world and by the market in 2012. Predictive AI models only began to be adopted on a large scale in 2016-2017 and generative AI models are still in the experimentation phase - and there are numerous limitations (Dignum, 2019; Kaufman, 2021a).

In addition to the inconclusive evidence of the results, due to the uncertainty variable that is intrinsic to every statistical probability model - it produces probable knowledge, but which is unavoidably

[1] Generative AI models are statistical (based on the technique of deep neural networks), they look for patterns in enormous amounts of text and then use these patterns to guess what the next word should be in a sequence of words that generates the best response to the user. These models manipulate the tokens without knowing what they refer to, they simply generate different sequences of tokens based not on the actual meaning, but in accordance with statistically probable arrangements.

uncertain - the technique is subject to interference from biased databases and the human subjectivity that is present in the process of preparing, applying, visualizing and interpreting the results (Wachter et al., 2016). Moreover, the complexity of the correlations established by AI algorithms in the data goes beyond human beings' cognitive capacity, forming the "interpretability problem" (or opacity, or black-box). These characteristics and imperfections jeopardize fundamental human rights, such as the right to explainability, the right to privacy, and the right to non-discrimination (Kaufman, 2019; 2022b). The fact that the field of AI does not have a theory represents an additional limitation; the field advances by means of empirical models, which makes it impossible to accurately predict the behavior of these models, thus expanding the spectrum of risks, uncertainties and potential damage.

Artificial intelligence is a digital technology, but it differs from the set of digital technologies by its nature as a general-purpose technology (GPT), which is the term given to technologies that shape an entire era and reorient innovations in the sectors in which they are applied; by reconfiguring the economy's logic and functioning and bringing in new business models, GPTs trigger periods of reorganization which the economist Joseph Schumpter called "creative destruction" (Kaufman, 2022b). Mustafa Suleyman and Michael Bhaskar (2023) regard the emergence and spread of disruptive technologies as "waves" that transform society, and predict that the next wave that will define the 21st century will come from artificial intelligence and synthetic biology; at one and the same time, our future depends on and is threatened by these technologies.

Suleyman and Bhaskar (2023) argue that technologies can fail in a number of ways; they can fail in the common sense, such as an engine not starting or a bridge falling down, or in a broader sense, if the technology causes damage to society, harms human lives, becomes ungovernable, and they wonder as to whether in this latter sense the failure is not inherent to the technology, but rather depends on the context in which it operates, the governance structures to which it is subject, the power networks and the uses. If they cannot be controlled, "the consequences for our species are dramatic, potentially dire. Equally, without its fruits we are exposed and precarious" (Suleyman, Bhaskar, 2023, p. 7). In other words, AI offers unprecedented opportunities and risks.

The development and spread of AI are driven by incentives such as geopolitical competition, particularly between the US and China, extraordinary financial rewards and a relatively open and distributed research culture. Suleyman and Bhaskar (2023) warn that there is an urgent need for containment mechanisms to this fast-approaching wave, such as providing democratic nation-states with appropriate safeguards and affordances, a challenge exacerbated by the fact that this technology is likely to become cheaper and more accessible and as a result spread faster than any previous technology.

The potential negative consequences are manifold. For example, the advent of ChatGPT has sparked concerns in the educational ecosystem in relation to authorship, plagiarism and appropriation of areas of scientific research. Artists are disputing the originality of images generated by generative AI models, and prestigious academic publications have vetoed articles submitted in co-authorship with ChatGPT (Kaufman, 2023). AI is transforming the labor market, expanding programmed automation with "intelligent automation" with AI, encompassing cognitive tasks and functions, reducing the remuneration of functions preserved for humans, and requiring qualification and retraining in order to perform the new functions. AI, in particular generative AI models, threatens democracy by encouraging the proliferation of disinformation via the addition of synthetic, inaccurate and/or false, information to internet databases, and by enabling deep fakes[2]; similarly, it also represents a threat to democracy due

[2] Deep Fake: artificial intelligence solution based on deep neural network architecture called GAN (proposed in 2014): Training setup where two networks, generator and discriminator, are pitched against each other in a competition. G has to generate fake images with random noise as input, while D has to discern between the fake images and the real images of the target domain that we would like the model to learn. Over time, both networks progressively improve at their tasks, and a trained generator model can be obtained that replicates the target domain images very well.

to its effect on lobbying, for example, by automatically generating comments on proposals that are under consideration by legislative bodies.

On the economic front, market concentration threatens nations' sovereignty, pitting increasingly fragile states against ever more powerful conglomerates. In the West, the research, development and implementation of AI is controlled by five American big techs - Apple, Meta/Facebook, Amazon, Microsoft and Google - with a market value greater than the GDP of most countries: in 2023, Apple is worth US$ 2.8 trillion, Microsoft US$ 2.5 trillion, Alphabet/Google US$ 1.58 trillion, Amazon US$ 1.25 trillion and Meta/Facebook US$ 730 billion; on May 31, 2023, the chipmaker Nvidia saw its market value hit a figure of U$ 1 trillion for the first time. Only 18 countries have a GDP of more than a trillion dollars, only 11 countries have one of more than 2 trillion dollars and only the USA and China have a GDP of above 7 trillion dollars. Daron Acemoglu and Simon Johnson (2023) warn of the power of these organizations: "The fact that these companies are attempting to outpace each other, in the absence of externally imposed safeguards, should give the rest of us even more cause for concern, given the potential for A.I. to do great harm to jobs, privacy and cybersecurity."

The tech giants do not produce a product or service, but rather they produce ecosystems and dominate the internet infrastructure by controlling two inputs that are critical to the flow of information and innovation: data and interconnections. Ariel Ezrachi and Maurice Stucke (2022) argue that big techs do not just control competition in their ecosystems, but also determine the nature of innovation, only allowing innovations that do not threaten their business models. Unlike previous monopolies, big techs, due to the fact that they control the internet infrastructure, are able to identify market patterns, discern and anticipate emerging trends and threats.

The current antitrust legislation fails to account for the operational complexity of the tech giants because the focus is on a) what is quantifiable (price and output) rather than what the actual competitive elements are (innovation, quality and privacy); b) narrowly defined markets rather than ecosystems; and c) anti-competitive practices (Ezrachi, Stucke, 2022). On July 27, 2020, Google's, Apple's, Facebook's and Amazon's CEOs testified before the US Congress as part of an antitrust investigation that was initiated in June 2019 by the House Judiciary Committee[3]; the four CEOs acknowledged that current US laws and policies foster entrepreneurship, a precondition for the success of their respective organizations, or seen from another perspective, they explicitly acknowledged the inefficiency of the antitrust laws.

Another point of attention is the relationship between artificial intelligence and sustainability, a relationship that generates a paradox: AI enables society to learn about and monitor climate change, due to the technology's ability to cope with large data sets, but in doing so it has a negative impact on the environment due to the intensive consumption of energy in the computational processing of its models, and consequently on carbon emissions (CO2). In 2019, 23 scientists from the field of AI, including Andrew Ng and Yoshua Bengio, launched a Manifesto (Manifesto, 2019) with a bibliography containing 826 references, inviting the AI community to join the effort to use technology on behalf of the climate. The Manifesto includes 13 areas where AI can make a positive contribution, such as speeding up the development of clean energy technologies, improving demand forecasts, and optimizing the management and monitoring of these systems; in the transport sector, which accounts for roughly a quarter of all energy-related CO2 emissions, the Manifesto warns that progress towards decarbonization is negligible, and that AI has the potential to improve vehicle engineering, enable intelligent infrastructures and generate relevant information for public policies.

The purpose of this article is to encourage reflection regarding the main initiatives that are available to society to protect its citizens and organizations from the potential harm caused by AI models, vis-à-vis

[3] Available in: https://www.cnbc.com/2020/07/01/apple-google-amazon-and-facebook-ceos-to-testify-in-congress.html). Accessed on: October 12, 2023.

the technology's own limits to act in ethical and legal compliance. It should be borne in mind that the potential damage does not have the same degree of problematization, since the intensity and extent of the damage varies according to the domain and the object of application.

2 The unprecedented human-technology relationship

The human-technological relationship is the bedrock of human life from its origins to the present day; technologies are human because they are made by humans and require human intervention for their development and use, and humans are technological because they depend on technologies to increase their capabilities and are simultaneously shaped by technologies. Michel Puech (2008) defends the view that contemporary human beings are of a different nature to the "original" Homo Sapiens: "We are a species, Homo Sapiens Technological (HST), whose lifestyle has no known equivalent. This is not a new natural type, but an artificial species" (Puech, 2008, p. 5). The HST participates in a dynamic in which there is a coevolution between living beings and artifacts, whose transformations seek mutual adaptation, known as "coevolution". The priority of use places the artifact in an existential niche and, reciprocally, establishes a new life for human beings. Puech believes that the conquest of existential niches gives the theory of the evolution of artifacts its competition mechanism, which is necessary to account for selection.

We can therefore talk about a "natural" evolution of technological artifacts, and even about a "Darwinian" evolution: by mutation, sometimes almost at random, and by selection that is as ruthless in the economic and social environment as it is in the Darwinian natural environment; man's use is what defines the survival of technical objects. Changes are not always intentional, sometimes the very process of adaptation to use generates random innovations, which may or may not be incorporated by man. For Puech (2008), the concept that we need to consider is the "Technosphere", the technological environment that surrounds and constitutes man's interface with the world.

From a broader perspective, he believes that the "I" is not limited to the biological "I", but that it is integrated into the technosphere, in particular the infosphere. In this sense, for example, the information stored on a cell phone, a computer or even online is a constituent part of the human being; the ability to send digital messages is as much a part of the human being as the ability to speak out loud - without this ability being a "prosthesis". One aspect considered by Puech is the apparent lag between the evolution of technology and that of man in the contemporary world, a kind of delay within the coevolution of this relationship, a question of rhythm arising from the acceleration of technological evolution versus human beings' inherent inertia to adapt to changes, whether technological or not. The arrival of ChatGPT and other similar generative AI solutions (Google's Bard, Meta/Facebook's Llama), since they are based on large language models (LLMs), require a reframing of the relationship between humans, technology and language, "Both humans and technologies are then not absolute authors but participate in that meaning-producing process" (Coeckelbergh, Gunkel, 2023, p. 5). LLM models expand our writing capacities and, at the same time, give rise to unprecedented ways of writing and, consequently, novel ways of thinking.

> *Whereas computers were typically positioned as intelligent typewriters (with editing functions and spell check, for example), applications such as ChatGPT can be used to create a first draft. The users then no longer think as they write the text; instead, they think about what prompt to give to the application and hence generate various versions of the text they want to generate. This thinking in terms of prompts is not purely instrumental; it is likely to change the way we think and experience the writing process and ourselves as writers. In sum, humans and technologies are entangled with one another. (Coeckelbergh, Gunkel, 2023, p. 3).*

Coeckelbergh and Gunkel (2023) argue that the role of language goes beyond expressing or representing what humans think or want, but rather molds our thinking and shapes our world, acting as a kind

of author or agent. In this sense, the authors are of the opinion that the concern that LLM technologies, such as ChatGPT or Bard, will replace human authors is misplaced, because humans have never had absolute authority and agency, technologies have always been co-authors contributing to meanings, and they propose that humans, language and technology be considered as co-authors in the processes and performances of these generative AI models. "In the case of ChatGPT, for example, there is a process that has computational elements and human elements (the human user but also developer and the company) participating in the creation of text. [...] It's a hybrid human/non-human performance" (Coeckelbergh, Gunkel, 2023, p. 5). We are moving towards a scenario in which the authorship of a text or image is not only facilitated by AI, but in which there is no identifiable human author.

Although recent advances in artificial intelligence have even come as a surprise to the developer community itself, the solutions are still far from human cognitive capacity, and there is no scientific evidence that they will reach this level. Judea Pearl and Dana Mackenzie (2018) reflect on two fundamental attributes of human cognition that scientists do not yet know how to incorporate into AI systems: causality and counterfactual reasoning. Pearl and Mackenzie mull over the meaning of thinking rationally based on cause and effect, credit and regret, intention and responsibility, distinguishing correlation from causality, even acknowledging the difficulty of establishing exactly what "causality" means; using data it is possible to measure the faster recovery of patients who have used a certain drug, but there is no way to measure why, to establish causality, in other words, what factor is behind this faster recovery. For them, "A causal reasoning module will give machines the ability to reflect on their mistakes, to pinpoint weaknesses in their software, to function as moral entities and converse naturally with humans about their own choices and intentions" (Pearl, Mackenzie, 2018, p. 11).

In turn, counterfactual reasoning, a type of expression of causal reasoning, involves retrospective thinking; in the example above, assuming that the patient who took the medication died a month later, the question is whether the death was caused by the medication. As Pearl and Mackenzie (2008) explain, in order to answer this question, it is necessary to imagine a scenario in which the patient was about to take the medication, but for some reason or other changed their mind; in this case, would the patient have survived? The ability to reflect on one's previous actions and envision alternative scenarios lies at the very heart of human cognition, and is not part of the nature of current AI systems. Another key question for humans is "How?"; to answer it requires a causal model that allows one to mentally realize the options before deciding whether and how to do it in the real world (Pearl, Mackenzie, 2018).

Intentionality is also a critical element in human decision-making. The ability to conceive one's own intention and then use it as evidence in causal reasoning denotes a level of self-awareness (if not consciousness) that no machine has yet achieved. We learn about causes and effects even before we understand language, and long before any mastery of mathematics. In addition to causality, counterfactual reasoning and intentionality, current AI systems lack flexibility and neuroplasticity (the biological brain's ability to change, adapt and mold itself at a structural and functional level when subjected to new experiences from the internal and external environment).

In another assessment, Melanie Mitchell (2019) ponders that human subconscious perceptions and choices stem from a lifetime of experiences and learning such as touching hot objects and getting burnt. This type of knowledge is reflexive, does not require conscious thought and is regarded by humans as "easy knowledge", but it is exactly the opposite for machine systems. For Mitchell, human beings have a tendency to anthropomorphize non-human things, from animals to inanimate objects, using the same words we would use for acts arising from human intelligence, confusing our own understanding of AI.

Although today's AI models are far removed from the biological brain, they are mediating communication and sociability and bringing in new business models that are likely to become dominant in wealth creation. We are moving from a world of programmed machines to a world of probabilistic machines, with extraordinary benefits and potential damage that need to be weighed up.

3 Main ethical points of attention

3.1 Discriminatory bias

It is considered that there is a bias in the result when the system exhibits a systematic error, also known as statistical bias. Like any statistical model, deep neural networks are designed to generalize based on a sample (even a sample made up of large data sets); bias refers to the error that can occur in the generalization process (Cozman, Kaufman, 2022). Discriminatory bias arises when the same group of people is systematically discriminated against, for example, discriminatory gender or ethnicity bias.

Karen Hao (2019) warns that in order to detect bias it is essential to understand its origins, recognizing that the trend is to attribute bias solely to biased training data, when it can arise in the various stages of the process, particularly a) in the framing of the problem, when the developer translates the objective to be achieved into computable language; b) in the collection of the data, when the base is not representative of reality or reflects existing prejudices in society and c) in the preparation of the databases. However, even when it is detected, it is hard to correct the bias, particularly in the case of detection when systems in full use. One potential discrimination not considered by Hao is that originating in data production, present both in the predominance of users from developed countries which have greater access to technologies and social networks, which results in an image database biased by the light-skinned racial biotype, and in the failure to disaggregate data by gender and/or the treatment of men as "standard human beings " (Perez-Criado, 2021).

In the development stage of a neural network system, once the objective has been determined, it is up to the computer scientist to translate it into variables that can be computed, the so-called hyperparameters. It is the developers who define the architecture to be used, the search terms to collect the data, and who select the databases. Identifying the influence of human subjectivity is no minor task nor is it possible to eliminate it even if it is identified (Hao, 2019). Inter- and multidisciplinary teams can mitigate discriminatory effects, but their effectiveness depends on building "bridges" between fields of knowledge (Kaufman, 2021b).

In the database, discriminatory bias occurs if its composition is less demographically diverse than the target population, in other words, if the database does not reproduce the proportionality, among other things, of ethnicities and genders present in the universe that is the subject of the action (Cozman, Kaufman, 2022). In the same way, the data labeling process has the potential to produce biased results. In this case, the challenge is to represent the complexity of the world in taxonomies to label the data, a precondition for supervised learning - the type of machine learning used in image and sound/voice recognition in which the system developer defines the goal (output) and labels the input data. The level of complexity increases in the case of taxonomies of images of human beings, for example, accurately classifying a person's gender or ethnicity or even identifying their profession from a photograph (Crawford, 2021).

Society's sensitivity to the problem of discriminatory bias in data is relatively recent. For years, biased databases have been used to develop and train AI algorithms (in part, they still are).

Table 1 - Different sources of bias

Sources of bias	Explanation
When generating data	Discrimination in data production is present both in the predominance of users from developed countries who have greater access to technology and social networks, which results in a database biased by the light-skinned racial biotype, and in the failure to disaggregate data by gender and/or the treatment of men as "standard human beings".
In the choice of developers	During the development of a DLNN model, the initial task for computer scientists is to identify the problem to be solved by the system, in which situation and for what purpose the system will be used. The second step is to translate the problem to be solved into variables that can be observed and manipulated (feature engineering process). These are the things that define, for example, which research terms will be used to collect the data, the number of hidden layers and the number of nodes in each layer. Identifying the influence of human subjectivity on the design and configuration of the AI algorithm is no simple task, nor is it possible to eliminate it even if it is identified. The Alan Turing Institute points out that one of the critical problems that allows systemic biases to infiltrate data is the attitude of algorithm developers and designers, who do not prioritize actions to identify and correct potentially discriminatory imbalances in demographic and phenotypic representation. The institute attributes these biases to the complacency of technology producers, who are generally part of the dominant group and therefore exempt from the adverse effects of discriminatory results.
In the database	Bias occurs if the reference data is less demographically diverse than the target population, in other words, if the database contains few or no examples of a specific sub-population by ethnicity and/or gender. The difference between controlled environments (laboratories) and uncontrolled environments (the real world) also has the potential to produce biased results; for example, on the street, cameras can capture images in low resolution, the angle captured of the face and the lighting can make it hard to extract facial features or even distort them, causing errors in facial recognition.
In the data labeling process	Creating a training database means sampling an almost infinitely complex and varied world, and fixing it into taxonomies made up of classifications. [...] Maintaining uniformity in the manual classification of large data sets is a challenge, which becomes almost impossible when it entails classifying images of people; there are countless classification categories, including race, age, nationality, profession, economic status, behavior, character and even morality. Structuring a taxonomy in order to classify images of people with the logic used for objects generates countless distortions and, consequently, biases.
In the training data of the algorithms	It is deemed that there is a bias in the database when the system displays a systematic error in the result ("statistical bias" or "algorithmic discrimination"). Strictly speaking, any dataset could be impartial for the performance of a given task. However, potentially there is a risk that, if it is used for a different task, it will be biased towards that second task.

Source: Cozman, Kaufman, 2022, p. 200-205.

Both academic and non-academic experts are committed to finding ways to detect and remove, or at the very least mitigate, the discriminatory bias of AI systems, in particular applications in sensitive fields such as health, security and education.

3.2 Transparency or explainability

Transparency is a concept that dates back to medieval Latin, and has its origins in the term *transparentia*, which is a quality attributed to a transparent object. The term is related to the Latin *transparere*, made up of the term *trans*, which means through, to cross, and *parere*, which means to appear, to let light shine through, with objects that allow light to come or pass through being transparent (Arruda, 2021). During the period of the Enlightenment, transparency became the hope of "discovering" a singular and objective truth, capable of overcoming individual perspectives, as well as being seen as essential for establishing a just and harmonious society (Daston, 1992). A number of historians recognize the importance of practices of the Enlightenment as the first contexts in which transparency emerged in its modern form (Crary, 1990; Daston, 1992; Daston, Galison, 2007; Hood, 2006).

In the deep learning neural network technique, the way in which the algorithms correlate the parameters contained in the data is so complex that it goes beyond human cognitive capacity; an incompatibility is established between high-dimensional mathematical optimization and human reasoning and semantic interpretation, a phenomenon that is referred to by scientists in the field as the "interpretability problem" (Goodefellow et al., 2016). This limitation stems from not knowing how the so-called "input data" generated the output data.

Coeckelbergh (2020) tackles the problem of explainability and transparency of AI systems from the perspective of assigning responsibility: the responsibility of the agents (users of the technology) stems from the expectation of the recipients (the other side of the relationship) that they will be able to explain the reasons for the decision. In healthcare, for example, the assumption is that the doctor controls the procedure and is able to explain it to the patient. Responsibility is treated as accountability: to act responsibly (in this case, the agent needs to know what he or she is doing, to justify his or her action) and to explain the reasons to those affected by the action ("patients"), who can and should demand and deserve answers about what was decided and how it was decided.

Scientific efforts are underway to generate user-friendly interpretations of how these systems work, known as "Explainable AI" (XAI), but the results of these efforts are not yet effective. The difficulty increases when the systems are produced by private companies that are legally protected by trade secrets.

3.3 Privacy

The threat to privacy is less related to the nature of the deep neural network technique, and more to the origin, diversity and quality of the personal data used in the development and training of the algorithms of AI systems; this threat is more pronounced, for example, in systems designed to predict the future behavior of users, in the improvement of surveillance systems and in the Chinese personal credit system (Social Credit System). The so-called Internet of Bodies, another example, by gaining access to and control of vital functions of the human body, produces a set of sensitive data that is used to develop and train the algorithms of AI devices in the "wearable technology" categories associated with healthy living (smartwatches, fitness trackers), in microchips for biometric identification and/or granting authorization, and in health devices in general.

In July 2020, the World Economic Forum (WEF) published the report by Xiao Liu and Jeff Merritt (2020) "Shaping the Future of the Internet of Bodies (IoB): New challenges of technology governance", dealing with the implications for privacy and fairness of smart devices attached to bodies; Liu and Mer-

ritt examined IoB data governance in the US by comparison with the regulation in the European Union, pointing out that in both regions there is a gap between the anti-discrimination legislation and the unprecedented risk of discrimination, a function of inferences, profiling and clustering of data originating from smart devices.

Data protection laws - such as the European General Data Protection Regulation (GDPR) of 2016, which came into force in 2018 and inspired the California Consumer Privacy Act (CCPA, 2020), China's Personal Information Protection Act of 2021 and Brazil's General Personal Data Protection Law (Law No. 13,709/2018) - partially minimize threats to privacy.

4 Society's reaction paths

Effective social protection against the potential damage from artificial intelligence models should combine ethical principles, global guidelines, self-regulation and regulatory frameworks. The first significant initiative on this front was the "Asilomar Principles", which emerged at the Conference on Beneficial AI, held in 2017 at the initiative of the Future of Life Institute. On that occasion, a set of 23 general principles was discussed in order to ensure the development of AI technologies beneficial to society, subdivided into three categories - research; ethics and values; and long-term issues. The essence of these general principles is at the founding base of various institutes - in addition to the Future of Life Institute, the Future of Humanity Institute, which is headed up by the English philosopher Nick Bostrom; the AI Now Institute, New York University; Human-Centered Artificial Intelligence (HAI), Stanford University; and the Leverhulme Center for the Future of Intelligence, Cambridge University -, and of initiatives by multilateral and European organizations such as AI4People, organized in 2018 by the OECD (Organization for Economic Cooperation and Development). The general principles have increased awareness among society, but they have proved to have little practical applicability; in addition to their abstract and non-universal nature, the general principles are not translatable into mathematical language, which is a pre-requisite for incorporating them into machine systems.

As a result of the growing pressure from society, self-regulation has come on to the agenda of technology companies, but with dubious effectiveness. Luciano Floridi (2021) recalls countless meetings in Brussels between policymakers, legislators, politicians, civil servants and technical experts who were openly in favor of the idea of "soft law" based on codes of conduct and ethical standards of the technology industry itself, without the need for external controls or regulatory constraints; over time, this path has not become effective. The most recent initiative was the establishment of the "Frontier Model Forum" on July 26, 2023, with Google, Microsoft, Open AI and Anthropic as founders, for the purpose of a) moving forwards with AI safety research in order to promote responsible development of frontier models and minimize potential risks, b) identifying best safety practices for frontier models, c) sharing knowledge with politicians, academics, civil society and others to advance the responsible development of AI; and d) supporting efforts to harness AI to address society's greatest challenges. The limits of self-regulation are partly due to the fact that when society's interest conflicts with commercial interests, the trend is for the latter to prevail.

At the same time, regulatory initiatives are emerging based on the assumption that a) the production and application of AI is transversal, influenced by economic and social issues; b) the process is a complex one, it's not like regulating a product or service; and c) the degree of risk varies for each implementation domain. Furthermore, an AI system can initially be in compliance both from the ethical as well as the regulatory point of view, but over the course of use cease to be so, or it can be classified as low risk and due to subsequent modifications become high risk, and vice versa.

It is no simple matter to pre-identify and isolate risks and their consequences in AI systems. Understanding how AI systems' supply chains work, and how to assign different responsibilities, takes time and

training, including for the regulators themselves (Kaufman, Coelho, 2023). The European Commission's proposed regulation (AI Act) – which was presented in April 2021, with more than 3,000 amendments by November 2022, whose implementation as a law is still unknown -, a reference in the Western world, including Brazil, contains a number of inaccuracies, such as, for example, not properly defining the supply chain, which is generally made up of more than one supplier and implementer, and whether obligations are shared or there is one main actor. Generative AI solutions are equally imprecise: the obligations of those who develop the language models (LLMs) should not be the same as those who apply these solutions in use cases, and it is unclear whether open-source systems are subject to the same obligations[4].

The American approach is different to the European approach: there is no federal regulation and it does not seem likely that there will be in the near future, the authority and responsibility for AI regulation and governance are distributed among federal agencies; this approach has advantages, for example, it speeds up the process, but it also contributes to the uneven development of AI policies. Among the initiatives, highlight goes to the White House's "Blueprint for an AI Bill of Rights" (October 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights/), and the National Institute of Standards and Technology/NIST's "NIST Risk Management Framework" (https://csrc.nist.gov/projects/risk-management/about-rmf); both are voluntary guidance documents, in other words, they do not have the force of law. In May and June 2023, public hearings in the US Congress intensified. On July 21, 2023, the White House released the "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed". [5]

Unlike the AI regulatory process in Europe, which is centralized under the coordination of the European Commission, in Brazil a number of bills have been put forward by federal deputies and senators. The first bill was presented to the Federal Senate in September 2019 by Senator Styvenson Valentim (Bill 5,051/2019); the second one, in February 2020, was proposed to the Chamber of Deputies by Federal Deputy Eduardo Bismarck (Bill 21/2020) and approved by the Full House of Representatives on September 29, 2021 with 413 votes in favor and 15 votes against and, in accordance with the standard legislative process, the bill was sent to the Federal Senate for assessment; the third bill was proposed to the Federal Senate in March 2021 by Senator Veneziano Vital do Rêgo.

In February 2022, the President of the Senate, Rodrigo Pacheco, set up a Committee of 18 legal experts with the mission to draw up a replacement for the three previous bills, which was delivered on December 6, 2022, and transformed into a bill, Bill 2,338/2023, on May 3, 2023, currently being considered by the Federal Senate. This bill is a starting point, but it is not yet ready to become the Regulatory Framework for AI in Brazil (Kaufman and Coelho, 2023). There is a consensus among experts that the effectiveness of any AI regulation depends on establishing standards.

In line with the mobilization for "Responsible AI" or "AI for Good", global entities are attempting to define good practices with techniques for explainable AI, auditable AI, accountable AI, mitigation of discriminatory biases (bias mitigation)[6], such as UNESCO, the OECD[7] the IEEE (https://standards.ieee.org/ieee/7000/6781/, https://standards.ieee.org/ieee/7003/6980/). In March 2022, a consortium between the Universities of Bologna and Oxford, led by Luciano Floridi, launched "capAI", a set of procedures to guarantee the reliability of AI systems; in March 2023, the Ada Lovelace Institute - an independent research institute founded in 2018 in collaboration with various British organizations such as the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society and the

[4] Zenner's considerations, in the panel….at the UN's headquarters in Geneva, July 5, 2023 part of the AI for Good Global Summit, with the author's participation.
[5] Available in:https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/. Accessed October 9, 2023.
[6] Available in: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Accessed October 9, 2023.
[7] Available in: https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf. Accessed October 9, 2023.

Nuffield Council on Bioethics - published the document "Inclusive AI Governance: Civil Society participation in standards development".[8]

AI models, whether predictive or generative, share four main characteristics that partly explain the challenge of regulation: they are inherently general and therefore ubiquitous in use, they hyper-evolve, they have asymmetric impacts. "Regulation alone doesn't get us to containment, but any discussion that doesn't involve regulation is doomed. Regulation should focus on those incentives, better aligning individual, states, companies, and the public as a whole with safety and security" (Suleyman and Bhaskar, 2023, 260).

5 Conclusion

As a general-purpose technology, artificial intelligence will define the 21st century and, consequently, the future of humanity by becoming hegemonic in the generation of wealth, creating unprecedented economic value. We are rapidly migrating to the Data Economy, or Data Capitalism, or Datacentric Capitalism, terms that express an economic model whose strategic raw material is data and artificial intelligence as the protagonist.

Science fiction films and subsequent narratives, including from prominent experts in the AI ecosystem, blur the line between fiction and reality. Like all technology, AI is social and human, its effects depend on what human beings do with it, how they perceive it, how they experience and use it, how they insert it into technical-social environments. It is up to society to deliberate, among countless questions, on whether AI should be applied in all domains and to perform all tasks, and whether the use of AI in high-risk applications is justified.

To address the scale of this challenge, regulation is necessary but not sufficient. Standards, ownership structures, unwritten codes of compliance and arbitration procedures, supervision and auditability, AI governance, international agreements, compliance with current local and global standards and laws. All of this needs to be integrated.

The benefits of AI are extraordinary, as are its potential harms. Society seems to have no alternative to facing the challenges of at least mitigating the damage already identified and trying to predict future damage in advance.

References

ACEMOGLU, D.; JOHNSON, S. 2023. *Big Tech Is Bad. Big A.I. Will Be Worse*. New York Time, em 9 de junho, disponível em: https://www.nytimes.com/2023/06/09/opinion/ai-big-tech-microsoft-google-duopoly.html). Acesso em: 10 de outubro de 2023.

ARRUDA, C. S. L. 2021. O princípio da transparência. *Revista do Direito da Administração Pública*, (1): p. 39-111.

COECKELBERGH, M. 2020. *AI Ethics*. Cambridge, MA: MIT Press.

COECKELBERGH, M.; GUNKEL, D. J. 2023. ChatGPT: deconstructing the debate and moving it forward. *AI & Society*. Disponível em: https://doi.org/10.1007/s00146-023-01710-4. Acesso em:10 de outubro de 2023.

COZMAN, F. G.; KAUFMAN, D. 2022. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. *Revista USP*, (135): p.195-210. DOI: https://doi.org/10.11606/issn.2316-9036.i135p195-210.

[8] Available in: https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/. Accessed October 9, 2023.

CRARY, J. 1990. *Techniques of the observer*. Cambridge, MA: The MIT Press.

CRAWFORD, K. 2021. *Atlas of AI. New Haven and London***.** Yale University Press: Yale.

DASTON, L. 1992. Objectivity and the escape from perspective. *Social Studies of Science*, **22**(4): p. 597-618.

DASTON, L.; GALISON, P. 2007. *Objectivity.* Brooklyn, NY: Zone Books.

DIGNUM, V. 2019. *Responsible Artificial Intelligence:* How to Develop and Use AI in a Responsible Way. Cham: Springer Netherlands.

EZRACHI, A.; STUCKE, M. E. 2022. *How Big-Tech Barons Smash Innovation—and How to Strike*. NY: Harper Business.

FLORIDI, L.; COWLS, J.; KING, T. C.; TADDEO, M. 2020. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*. **26**(3): p. 1771-1796.

FLORIDI, L.; *et al*. 2018. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, **28**: p. 689-707. Available in: https://link.springer.com/article/10.1007/s11023-018-9482-5. Access on: 4 abr. 2021.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. 2016. *Deep Learning*. Cambridge: MIT Press.

HAO, K. 2019. Intelligent Machines: This is how AI bias really happens - and why it's so hard to fix. *MIT Technology Review*. Disponível em: https:// www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-reallyhappensand- why-its-so-hard-to-fix/. Acesso em: 7 ago. 2021.

HOOD, C. 2006. Transparency in historical perspective. *In*: HOOD, C.; HEALD, D. (eds.). *Transparency*: the key to better governance? Oxford: Oxford University Press.

KAUFMAN, D. 2019. *A inteligência artificial vai superar a inteligência humana?* SP: Estação das Letras e Cores.

KAUFMAN, D. 2021a. Inteligência Artificial e os desafios éticos: a restrita aplicabilidade dos princípios gerais para nortear o ecossistema de IA. *PAULUS: Revista de Comunicação da FAPCOM*. São Paulo, **5**(9).

KAUFMAN, D. 2022b. *Desmistificando a inteligência artificial*. BH: Autêntica.

KAUFMAN, D. 2023. ChatGPT assusta porque ameaça nossa "reserva de mercado. *Valor Econômico*. Available in: https://valor.globo.com/eu-e/noticia/2023/02/10/dora-kaufman-chatgpt-assusta-porque-ameaca-nossa-reserva-de-mercado.ghtml. Accessed October 7, 2023.

KAUFMAN, D.; COELHO, A. 2023. Regular a IA, mas sem precipitação. *Valor Econômico*. Available in: https://valor.globo.com/opiniao/coluna/regular-a-ia-mas-sem-precipitacao.ghtml. Accessed October 7, 2023.

LIU, X.; MERRIT, J. 2020. Shapping the Future of the Internet of Bodies: New challenges of technology governance. Briefing Paper. *World Economic Forum*. Available in: WEF_IoB_briefing_paper_2020 (4).pdf.

MITCHELL, M. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. NY: Farrar, Straus and Giroux.

PEARL, J.; MACKENZIE, D. 2018. *The Book of Why: The New Science of Cause and Effect*. NY: Basic Books.

PEREZ-CRIADO, C. 2021. *Invisible Women: Data Bias in a World Designed for Men*. US: Abrams Press.

PUECH, M. 2008. *Homo sapiens technologicus:* philosophie de la technologie contemporaine, philosophie de la sagesse contemporaine. Paris: Éditions Le Pommier.

SULEYMAN, M.; BHASKAR, M. 2003. *The Coming Wave*. UK: Penguin Random House.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. 2016. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law,* **7**(2): p. 76–99.

Submetido em 26 de outubro de 2023.

Aceito em 11 de janeiro de 2024.