filosofia unisinos | Unisinos Journal of Philosophy

ISSN 1984-8234

Filosofia Unisinos
Unisinos Journal of Philosophy
25(1): 1-13, 2024 | e25113

Unisinos – doi: 10.4013/fsu.2024.251.13

Dossier

AI's black box and the supremacy of standards

Caixa-preta da IA e a supremacia dos padrões

Murilo Karasinski[1]
https://orcid.org/0000-0002-6099-6968
[1]Pontifícia Universidade Católica do Paraná - PUCPR, Educação Continuada da Escola de Educação e Humanidades, Curitiba, PR, Brasil. Email: k.murilo@pucpr.br

Kleber Bez Birolo Candiotto[2]
https://orcid.org/0000-0002-2000-4776
[2]Pontifícia Universidade Católica do Paraná - PUCPR, Programa de Pós-Graduação em Filosofia, Curitiba, PR, Brasil. Email: kleber.candiotto@pucpr.br

ABSTRACT

This article investigates the metaphor of the "black box" in artificial intelligence, a representation that often suggests that AI is an unfathomable power, politically uncontrollable and shrouded in an aura of opacity. While the concept of the "black box" is legitimate and applicable in deep neural networks due to the inherent complexity of the process, it has also become a generic pretext for the perception, which we seek to critically analyze, that AI systems are inscrutable and out of control, as well as supposedly endowed with intelligence and creativity. To challenge these ideas, we will address what we call the supremacy of patterns and the two significant phenomena that result from it: enchanted determinism and the dictatorship of the past.

Key-words: artificial intelligence, black-box, supremacy of patterns; enchanted determinism; dictatorship of the past.

RESUMO

Este artigo investiga a metáfora da "black box" na inteligência artificial, uma representação que frequentemente sugere que a IA é um poder insondável, politicamente incontrolável e envolto em uma

aura de opacidade. Embora o conceito de "black box" seja legítimo e aplicável em redes neurais profundas, devido à complexidade inerente do processo, ele também se transformou em um pretexto genérico para a percepção, que buscamos analisar de maneira crítica, de que os sistemas de IA são inescrutáveis e fora de controle, bem como dotados de suposta inteligência e criatividade. Para desafiar essas ideias, abordaremos o que chamamos de supremacia dos padrões e os dois fenômenos significativos decorrentes: o determinismo encantado e a ditadura do passado.

Palavras-chave: inteligência artificial, black-box, supremacia dos padrões; determinismo encantado; ditadura do passado.

1 Introduction

When searching for transparent governance based on justified actions, the principles of transparency, interpretability and explainability play a fundamental role in the development of artificial intelligence. These principles form the basis of a mechanism of principles that serves as a foundation for evaluating issues of justice and accountability, which are essential aspects sought in various areas of contemporary society. The challenge, however, is that the adherence to these principles is not uniform in the context of AI systems, which contributes to the "black box" problem.

In the field of computer science and engineering, the terms "black-box", "gray-box" and "white-box" are often used to describe different levels of opacity regarding the internal components of a system (Adadi, Berrada, 2018). These terms make a crucial distinction between disclosing information about the internal design, the structure and implementation of a component. A so-called "black-box" component maintains total secrecy about its internal structure, revealing nothing to the user. In contrast, a "white-box" component is completely transparent, fully exposing its internal details to the user. In addition, according to Adadi & Berrada (2018), there are intermediate levels of so-called "gray-box" components which provide partial information about their internal details, with varying degrees of disclosure.

In the context of AI, a system's difficulty in providing an adequate explanation for how it arrived at a particular answer is often referred to as "the black box problem", which has given impetus to an area of AI research called XAI, a term coined by Van Lent, Fisher & Mancuso (2004). This field of research aims to make the results of AI systems more understandable to humans. The issue of explainability is a long-standing problem in AI. Moore & Swartout (1988), for example, have been addressing the problem since the 1970s, when researchers studied explanation for expert systems. In Moore's & Swartout' article there is already a reference to related works in the field of natural language generation, suggesting that some techniques from this field may be useful in the task of explanation in expert systems. The authors also indicate that even these techniques will not provide all the capabilities needed to achieve transparency with the user. Pasquinelli and Joler (2021) disagree with this position, considering that this view of an inability to be transparent has become a common pretext for the perception that AI systems could be considered out of control.

Progress in solving this problem has slowed considerably since AI has reached an inflection point, characterized by exponential advances in machine learning. Adadi & Berrada (2018) note that there has been a shift in the focus of XAI research, mainly towards the implementation of models and algorithms that highlight predictive capacity, while the ability to explain decision-making processes has taken a back seat.

In this way, the growing dependence on data-based algorithms due to the increasingly intense insertion of machine learning, deep learning and now LLMs in areas that have direct and indirect impacts on people's lives, such as the banking sector, the justice system or job recruitment, for example, without

providing precise information on the chain of reasoning that leads to certain decisions, recommendations, predictions or actions taken by them, presents significant challenges. In such cases, the need for explainable predictive models and transparent systems becomes pressing.

AI algorithms often operate in complex ways that are difficult for non-experts to understand. This raises legitimate concerns about the justification of their decisions and the potential for algorithmic bias. To address these concerns, creating predictive models that are easily explainable and adapting transparency to pre-existing AI systems are strategies often expected by affected individuals or even imposed through legal regulation.

Various techniques and algorithms are being proposed to address the need for transparency and explainability in AI systems. However, it is important to note that research in this area is still relatively young and lacks a consensus within the AI discipline regarding a specific set of technologies that can be universally adopted to address these complex challenges.

In this context, this article seeks to undertake a critical review of the powers and promises of artificial intelligence. Our aim is to deepen the analysis of the debate that highlights the perspective that AI systems are inscrutable and prone to breaking free from human control. We also want to explore how these systems are perceived as possessing intelligence and creativity. In doing so, we aim to examine the ethical and technological implications inherent in this paradigm by debating, from the black-box metaphor, what we call the supremacy of standards and the consequent annihilation of the new.

2 The black box: an old problem with a new impact

In 1961, when writing the preface to the second edition of his work Cybernetics or Control and Communication in the Animal and the Machine, originally published in 1948, Norbert Wiener (2019, p. 285) argued that the terms "black box" and "white box" were figurative expressions whose use was not yet very well determined. For Wiener (2019, p. 285), a "black box" would correspond to a device, such as a four-terminal network with two input terminals and two output terminals, which could perform a defined operation on the present and the past of the input potential, but for which there would not necessarily be any information about the structure through which such an operation would be performed. On the other hand, a "white box" would be a network similar in the relationship between input potentials and output potentials, but according to a structural plan defined to ensure a previously determined input-output relationship.

Still in the 1960s, Mario Bunge (1963) broadened the debate to propose a general theory of the black box. In the field of science, as Cupani and Pietrocola (2002, p. 111) explain, some theories were more superficial, concentrating exclusively on analyzing the behavior of a system as a simple unit, depending solely on the relationship between a set of stimuli (input) and a set of responses (output), without delving into the intermediate mechanisms. Bunge called these theories "black box theories", referring to their lack of consideration for the "inside" of systems[1]. In contrast to these theories, others presented an internal structure model to which they referred, and were called "translucent box" theories, such as the wave theories of light, or the theories of behavior and learning, since they "modeled" a mechanism.

In the case of AI, for commercial reasons, the "black box" phenomenon has additional reinforcements. Pasquale (2015) highlights three fundamental strategies for preserving the confidentiality and integrity of information, keeping it safe under the concept of "black boxes". These are the strategies of genuine secrecy, legal secrecy and obfuscation, each representing different approaches to restricting unauthorized access to sensitive information. Genuine secrecy creates an effective barrier between the

[1] Examples of these theories would include geometric optics, which does not address the nature and structure of light, and the behaviorist theory of learning, which does not consider physiological mechanisms or mental states.

hidden content and any attempted unauthorized access, with everyday examples including protecting property by locking doors or preserving personal information with passwords in email accounts. Legal confidentiality, on the other hand, imposes legal and contractual obligations that require those with access to specific information to maintain its confidentiality, as in the case of bank employees forbidden from disclosing client balances to friends. Obfuscation, the third strategy, consists of deliberate concealment practices, often triggered when secrecy is threatened. An example is when a company responds to a request for information by flooding the requester with a vast amount of documents, forcing the researcher to search for relevant information in the midst of a considerable volume. The consequence of both forms of secrecy and obfuscation is the creation of opacity, a state which, in this philosophical context, is defined as "remediable incomprehensibility" (Pasquale, 2015, p. 06-07).

The non-transparency of algorithms goes beyond simple opacity (Esposito, 2022). Even if policies were implemented requiring full accessibility to the underlying data and procedures, most algorithmic systems would remain incomprehensible to their users. Users' inability to understand the inner workings of a technology is not a new and is a problematic issue in itself, since most ordinary users have always struggled to understand the inner workings of technology: how many drivers understand how a car engine works, for example? However, the current situation is marked by a distinctive and unprecedented aspect: algorithms not only process information, but also make autonomous decisions in a variety of contexts, from medical diagnoses to university admissions, strategy games and even judicial decisions such as granting credit or parole (Buhrmester at al, 2021).

The need to explain algorithmic decisions is particularly relevant in several areas. In medicine, it is essential that patients understand the reasons behind treatment recommendations to ensure informed participation in their own healthcare procedures (VAN DER VELDEN, 2022). In university admissions, understanding acceptance or rejection decisions is key to ensuring equity and fairness in the process (Haque at al, 2023). In teaching strategies, understanding algorithm decisions can enrich the student experience (Fiok at al, 2022). In the criminal justice system, explaining algorithmic decisions related to parole or granting credit is key to ensuring a fair and equitable system (Deeks, 2019).

Moving forward in the context of Artificial Intelligence, especially after the advent of deep learning algorithms, the most common technical limitations have come to be identified as "black box" dilemmas, which, according to Pasquinelli and Joler (2021) have become a common pretext for the perception that AI systems not only lack transparency, but are also considered out of control — even though such problems are real concerns, especially in convolutional neural networks, where excessive filtering of information would prevent the reversal of the chain of reasoning. On the other hand, for Pasquinelli and Joler (2021), it would be important to recognize that the "black box" phenomenon is inherent to any experimental machine in its early stages of development. Historically, artifacts such as the steam engine have presented similar mysteries, remaining misunderstood for some time even after successful demonstrations. However, the real problem lies in the rhetoric associated with the "black box", which in turn is closely linked to theories that portray AI as a hidden power, inaccessible to study, knowledge, or political control.

Approaching the problem from a different perspective, Mariutti (2023) argued that, based on the distribution of knowledge, any complex society would need to use machines and devices that would be beyond the complete mastery of its users. This would be evidenced by the fact that, when using complex devices, comprehension, with the exception of specialists, would be limited to their general principles, given the impossibility of fully encompassing all the equipment used. Thus, while the intrinsic opacity of the mechanisms associated with artificial intelligence would be undeniable from an operational point of view, the central challenge would lie in resisting the conception of Artificial Intelligence as an oracle. "AI is not an autonomous 'alien mind' that can operate separately from humans. But its existence creates an additional zone of opacity that we will never be able to fully master" (Mariutti, 2023).

3 Correlation and causality in AI: revisiting the Post Hoc fallacy

The concept of cause has been the subject of intense debate throughout the history of philosophy. Causality implies a chain of events, in which a previous event results in the production of a specific effect. David Hume (2007) already defined cause as a link between two objects, where the presence of the first implies the presence of the second. In other words, the existence of the second object would be conditional to the existence of the first; if the first object did not exist, the second would not exist either. However, there are reasonings which, at first glance, seem to be causal. These reasonings attribute the cause of what happens later to what happened before, in a temporal succession. This characterizes the Post Hoc fallacy (post hoc, ergo propter hoc), which means "after this, therefore, because of this".

Pinker (2021) gives the following example: imagine an investment advisor who sends half of a mailing list of 100,000 people a newsletter predicting a rise in the market, while sending the other half a version predicting a drop. Every quarter, he discards the names that received the incorrect forecast and repeats the process with the rest. After two two years, he has been hired by 1,562 recipients who have been baffled by his ability to correctly predict the market for eight consecutive quarters. "If you take note of the predictions by a psychic that are borne out by events, but do not divide by the total number of predictions, correct and incorrect, you can get any probability you want" (PINKER, 2021, p. 158-159). As Shermer (2011) ponders, at its most basic level, the Post Hoc fallacy resembles a form of superstition, since the fact that two events occur in sequence does not necessarily imply a causal link. In other words: correlation does not mean causation.

According to Pasquinelli and Joler (2021), a tragic example of this misconception can be found in the work of Frederick Hoffman, a statistician who, in 1896, published an extensive report for insurance companies suggesting a racial correlation between being African-American and having a shorter life expectancy. In the context of AI, by extracting data in a superficial way, machine learning could build arbitrary correlations that would be mistakenly perceived as causal.

Part of this debate has been going on since 2008, when American physicist and writer Chris Anderson published an article entitled The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. His aim was to argue that the deluge of data available for analysis by algorithms would make the scientific method obsolete, which in turn could suggest that human activity in terms of formulating theories would be dispensable. In other words, Anderson (2008) proposed that artificial intelligence would solve practically everything, leaving humans with a much reduced contribution to thinking. This is because, for Anderson (2008), until the middle of the 20th century, only models, from cosmological equations to theories about human behavior, were capable of explaining the world in a consistent way, even if these models were imperfect. However, this dynamic would have changed drastically with the rise of digital computers, which would have started to "read" information, as well as with the internet, which would have made it possible to track this data. As a result, according to Anderson (2008), companies like Google are establishing a vast laboratory to study the human condition based on the insights refined by machines, giving rise to the petabyte era. Furthermore, in this era, information management would require a different approach to data, since this data would have to be interpreted from a mathematical perspective before being contextualized. "[...] Google conquered the advertising world with nothing more than applied mathematics. It did not pretend to know anything about the culture and conventions of advertising" (Anderson, 2008). The underlying premise was that better quality data, combined with more advanced analytical tools, would be enough to determine the success or failure of a website. Google would base its philosophy on the fact that its algorithm did not need to intrinsically understand whether a page was better or worse than others. "[...] if the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required" (Anderson, 2008).

Furthermore, an additional implication, as highlighted by Anderson (2008), would be the idea that large volumes of data and mathematical analysis could completely replace traditional tools for understanding human behavior, making disciplines such as linguistics, sociology, taxonomy, ontology and psychology unnecessary. According to Anderson (2008), the emphasis would be on the ability to track and measure precisely, without the need for a complete understanding of the motives behind human actions. Given these scenarios, Anderson (2008) argued that, in the petabyte era, human beings would no longer need to rely on models. The proposal would be simple: insert numbers into the huge computing conglomerates already built, allowing algorithms to identify patterns unattainable by conventional science. In short, data could be analyzed without the need to formulate hypotheses. As a result, according to Anderson (2008), the exhaustion of the scientific method in favor of the influx of data became society's main modus operandi. This movement would gain prominence, especially when considering the synergy between vast amounts of information and statistical tools for analysis, providing a renewed approach to understanding the world. "[...] There's no reason to cling to our old ways. It's time to ask: What can science learn from Google??" (Anderson, 2008).

Despite the perspective presented by Anderson, we notice a significant misunderstanding in the proposal to develop scientific models, or any model, without the intervention of human thought. To reinforce this point of view, it is worth highlighting the thoughts of Luciano Floridi (2014, p. 129), who suggests that Chris Anderson's article could have been written centuries ago by Francis Bacon.

Floridi (2014, p. 130) argues that Bacon considered hypotheses "suspicious", preferring to assume that a large accumulation of facts would speak for itself. However, Floridi (2014, p. 130) points out that Bacon had underestimated a crucial point highlighted by Plato: the idea that "knowledge is more than information, because it requires explanations and understanding, not just truths or correlations" (Floridi, 2014, p. 130).

According to Floridi (2014, p. 13), although the current generation of humanity has entered an unparalleled era in terms of the amount of data produced on a daily basis, the real epistemological problem lies in the small patterns, "[...] the spot where the new patterns with real added-value lie in their immense databases, and how they can best be exploited for [...] the advancement of knowledge" (Floridi, 2014, p. 16). For Floridi (2014, p. 16), small patterns would not only redefine the limits of what could be considered predictable, influencing events and behaviors, but would also open up new frontiers of thought. These frontiers would involve a diversity of areas, from wide-ranging innovations to business competition, crossing the fields of science and government and reaching challenges in the sphere of personal security:

> *In a free and open marketplace of ideas, if someone else can exploit the small patterns earlier and more successfully than you do, you might quickly be out of business, miss a fundamental discovery and the corresponding Nobel, or put your country in serious danger. Small patterns may also be risky, because they push the limit of what events or behaviours are predictable, and therefore may by anticipated. This is an ethical problem. Target, an American retailing company, relies on the analysis of the purchasing patterns of 25 products in order to assign each shopper a 'pregnancy prediction' score, estimate her due date, and send coupons timed to specific stages of her pregnancy. In a notorious case, it caused some serious problems when it sent coupons to a family in which the teenage daughter had not informed her parents about her new status (Floridi, 2014, p. 16).*

In all cases, much more than a problem of computing power, small patterns represent problems of brainpower. Floridi (2014, p. 130) argues that victory in the "game" of knowledge would belong to those who, in a similar way to Socrates' dialogues with Craticus, can formulate and answer questions critically. This would involve understanding what information is truly useful and, in the process, discarding irrelevant data. Although it was clear that advanced technologies could facilitate the identification of small patterns, Floridi (2014, p. 130) argues that this alone would not be enough. Humans would still

need a more refined epistemology to discern and extract the truly meaningful small patterns. Ultimately, Anderson's neo-Baconian approach, according to Floridi (2014, p. 130), would be outdated: "Data do not speak by themselves, we need smart questioners" (Floridi, 2014, p. 130).

Furthermore, from James Bridle's perspective (2018), Anderson's theory exemplifies the "big data fallacy", i.e. the belief that "you do not have to know or understand anything about what you study; you just have to put all your faith in the emerging truth of digital information" (Bridle, 2018). According to Bridle (2018), one point that would increasingly highlight the damaging exclusive dependence on vast amounts of data for the scientific method would be observed in pharmacological research. According to Bridle (2018), over the last sixty years, despite the remarkable growth of the pharmaceutical industry and significant investments in drug discovery, the speed at which new drugs become available has consistently and measurably decreased in comparison to the volume of financial resources directed towards research.

> *The number of new drugs approved per billion US dollars spent on research and development has halved every nine years since 1950. The downward trend is so clear that researchers have coined a term for it: Eroom's law – that is, Moore's law backwards (Bridle, 2018)*

Another perspective that allows an analysis of the Post Hoc fallacy in the context of AI is found in the thinking of Judea Pearl and Dana Mackenzie (2018). According to Pearl and Mackenzie (2018), with Bayesian networks, machines have been taught to think in shades of gray, a crucial step towards human thinking. However, in terms of causality, there was an anomaly, since machines could not be taught to understand cause and effect. This is because there would be no way of explaining to a computer why turning the knob on a barometer would not cause rain. Nor would it be possible to teach it to anticipate what would happen if a shooter in a firing squad changed his mind and decided not to shoot. Without the ability to imagine alternative realities and contrast them with current reality, a machine would not be able to pass a Turing mini-test: "It cannot answer the most basic question that makes us human: 'Why?'" (Pearl, Mackenzie, 2018).

According to Pearl and Mackenzie (2018), a generation ago, a marine biologist could spend months conducting a census of their favorite species. Today, however, that same biologist would have immediate online access to millions of data points on fish, eggs, stomach contents or any other desired information. Instead of just taking a census, the biologist could now tell a story. However, the crux of the matter would lie in what would happen next: how to discern meaning amidst this profusion of numbers, bits and pixels? As Pearl and Mackenzie (2018) point out, although the data is vast, the questions we are asking are simple: is there a gene responsible for lung cancer? Which categories of solar systems are most likely to harbor Earth-like planets? What elements are contributing to the reduction in the population of a particular fish, and how can we intervene in this situation? For Pearl and Mackenzie (2018), there is an almost religious belief in certain circles that the answers to these questions can be found in the data itself, if we are smart enough to extract them. On the other hand, this enthusiasm is misguided:

> *The questions I have just asked are all causal, and causal questions can never be answered from data alone. They require us to formulate a model of the process that generates the data, or at least some aspects of that process. Anytime you see a paper or a study that analyzes the data in a model-free way, you can be certain that the output of the study will merely summarize, and perhaps transform, but not interpret the data (Pearl, Mackenzie, 2018).*

Furthermore, as Pearl and Mackenzie (2018) point out, in recent years, the most notable advances in the field of Artificial Intelligence have occurred in the sphere known as deep learning, which employs methods such as convolutional neural networks. These networks do not follow the conventions of probability; they do not deal with uncertainty in a strict or transparent way. What's more, they do not incorporate

any explicit representation of the environment in which they operate. When training a new network, the programmer has no understanding of the calculations it is performing or why they work. If the network fails, the solution is a mystery. According to Pearl and Mackenzie (2018), a paradigmatic example of this approach would be AlphaGo, a program based on a convolutional neural network that challenged players in the ancient Asian game Go, a game that has always been considered particularly challenging for AI. According to Pearl and Mackenzie (2018), the Go community believed that computers were still a decade or more away from offering real competition to humans. Most Go players became aware of the program at the end of 2015, when it defeated a human professional 5-0. In March 2016, AlphaGo surprised everyone by beating Lee Sedol, considered the strongest human player for years, 4-1. A few months later, AlphaGo took part in sixty online matches against the best human players, not losing once. In 2017, it officially ended its career by defeating then world champion, Ke Jie. For Pearl and Mackenzie (2018), the narrative in question would arouse great enthusiasm, and the results would be indisputable: deep learning is effective in certain tasks. However, it was also the antithesis of transparency:

> *Even AlphaGo's programmers cannot tell you why the program plays so well. They knew from experience that deep networks have been successful at tasks in computer vision and speech recognition. Nevertheless, our understanding of deep learning is completely empirical and comes with no guarantees. The AlphaGo team could not have predicted at the outset that the program would beat the best human in a year, or two, or five. They simply experimented, and it did. (Pearl, Mackenzie, 2018).*

It is important to clarify, as Pinker (2021) argues, that networks are labeled as deep learning systems because of the number of layers between input and output - that is, there is no depth in the sense of understanding something. In general, deep learning networks outperform old-fashioned classical artificial intelligence (GOFAI), which performs logic-like deductions based on hand-coded propositions and rules. However, the contrast in how these two work would be remarkable: unlike logic, the internal operations of a neural network are unfathomable. Most of the millions of hidden units would not represent any coherent concept that we can understand, and the computer scientists who train them would not be able to explain how they arrive at specific answers. For this reason, according to Pinker (2021), many critics of the technology fear that, by entrusting AI systems with decisions about people's fates, they could perpetuate biases that are difficult to identify and eliminate.

Returning to the Post Hoc fallacy, Pasquinelli and Joler (2021) defend the view that machine learning, often obsessed with "curve fitting", would record correlations without offering substantial explanations. However, this logical fallacy would have transcended the technical domain, becoming a clear political issue, especially considering the global adoption of predictive policing algorithms by police forces. "[...] when machine learning is applied to society in this way, it turns into a biopolitical apparatus of preemption, that produces subjectivities which can subsequently be criminalized" (Pasquinelli, Joler, 2021). Ultimately, machine learning's obsession with "curve-fitting" would impose, for Pasquinelli and Joler (2021), *a statistical culture, which would be replacing the traditional episteme of causality (and political responsibility) with one based on correlations*, blindly guided by the automation of decision-making.

In this sense, according to Mariutti (2023), the implications of the assimilation of AI by large corporations and its interconnection with the security apparatus of states would emerge as a central issue, exerting a significant impact on the dynamics between (a) classification and (b) pattern generation conducted by machine learning applications.

In the context of classification, automation would aim to categorize targets, which could be signs, objects, faces, among others, according to parameters integrated into the statistical model through training data. This is because taxonomies or social conventions would be linked to certain statistical distributions, assigning labels to objects that align with these distributions, identifying or recording their properties.

On the other hand, in terms of generating patterns, according to Mariutti (2023), this process would expand information based on a sample or fragment of data. One procedure, which has been increasingly used by the armed forces and police[2], would include the predictive dimension: based on previously recorded trajectories and behaviors, the algorithms would project future trends. Both the data classification phase and the pattern generation process would be excessively determined by the purpose for which they were conceived, eliminating any notion of neutrality from the outset. Thus, when captured by the rivalry between states and the competition between technological giants, artificial intelligence and its applications would, according to Mariutti (2023), tend to intensify the process of normalizing behavior, as we will discuss in the following section.

4 AI creativity? The supremacy of standards and the annihilation of the new

Pasquinelli and Joler's (2021) nooscope metaphor is relevant for critically analyzing the meaning of creativity attributed to AI, especially generative AI, which operates as an enormous statistical capacity. This is questionable creativity, but it has certain consequences for achieving this desired creativity. We will deal here with one of these main consequences, the annihilation of the new, which is caused by a problem we call here the "supremacy of patterns", resulting from an inherent limitation in the operation of generative AI.

The normative power of Artificial Intelligence in the 21st century demands an analysis from an epistemic perspective, raising questions about the nature of framing collective knowledge through patterns and the significance of constructing vector spaces and statistical distributions related to social behaviors. AI, in its capacity for data processing and analysis, substantially expands what can be termed the "normalizing power" of modern institutions[3]. This includes entities such as bureaucracy, medicine and statistics, initially associated with the numerical knowledge held by the state in relation to its population, but which are now largely under the control of AI corporations.

This is the phenomenon that Gurumurthy & Bharthur (2018) call the algorithmic turn, backed by a metanarrative associated with technomodernity, which supports the idea that all nations should adopt this transformation. Digital technologies are often presented as neutral tools, capable of driving economic progress and social advancement, thus acquiring an aura of ungovernability[4]. Large technology companies often present AI systems that learn and adapt quickly, assuming an autonomous character that is too complex to be fully understood. Gurumurthy & Bharthur (2018) also point out that the auton-

[2] An example of this use is reported by Cathy O'Neil (2016), who discusses the use of crime prediction programs, such as PredPol, in police departments in the US. These programs analyze historical data to predict locations and times prone to crime. While these models optimize resources, there are concerns about their application to less serious crimes, leading to a cycle of over-policing in certain areas, resulting in disproportionate arrests in impoverished and racially segregated communities. The text highlights the need to balance the effectiveness of these models with ethical considerations and social impacts.

[3] Paola Ricaurte (2019) argues that the institutional norm, which used to be based on public records, is now heavily influenced by algorithms and data centers. The classification of subjects, bodies and behaviors is no longer a matter restricted to traditional public records, but an activity that involves algorithms and the massive processing of information. The emergence of a data-centric rationality should be interpreted as a manifestation of the coloniality of power, in which control and influence are exercised through the appropriation and manipulation of data. This transformation not only redefines the dynamics of power in modern societies, but also raises profound questions about the ethical, political and social implications of this new paradigm.

[4] Gurumurthy & Bharthur (2018) highlight the manifestation of data colonialism and the coloniality of power in contexts beyond the northern hemisphere. In the example of Mexico, the colonization of data takes two distinct forms. Firstly, at an institutional level, the Mexican government adopts and replicates prevailing data epistemologies, incorporating them into its discourse of efficiency and modernity. However, these practices are also used as control and surveillance strategies, with the direct effect of internal colonization through data. This results in the reinforcement of domination over marginalized and vulnerable communities. The study analyzed in this article illustrates how structural violence is amplified in the Mexican context due to the absence of data on feminicides, a phenomenon that predominantly affects young women belonging to low-income groups. The analysis of these aspects demonstrates the importance of considering the dynamics of data colonialism and the coloniality of power in a broader global context, transcending the scope of the West and highlighting the need for reflexive and transformative actions.

omy of digital technologies should not be seen as an obstacle to understanding, but as an opportunity to shape their impact according to society's needs and intentions.

There remains an intrinsic gap, an essential friction, and an underlying conflict between the statistical models of Artificial Intelligence and the human subject who is the target of measurement and control. This logical gap manifested between AI statistical models and society has often been debated and identified as biased, especially in the context of the problem of facial recognition in relation to social minorities. The amplification of discrimination related to gender, race and class by AI algorithms represents, in essence, one aspect of a broader problem of discrimination and normalization that is rooted in the logical core of machine learning.

It is important to note that, according to Pasquinelli and Joler (2021), most contemporary applications of machine learning can be understood within the framework of two main modalities: classification and prediction. These modalities outline the contours of a new society characterized by statistical control and governance. The classification modality is widely recognized as pattern recognition, while the prediction modality can also be characterized as pattern generation. Both processes involve the identification or creation of supposed new patterns, and this action is done by investigating the inner core of the statistical model in question.

This approach based on statistical analysis and pattern recognition/generation is central to contemporary machine learning applications, which play an increasing role in structuring control and governance practices in today's society, creating what Campolo & Crawford (2020) call enchanted determinism. Enchanted determinism is the result of machine learning's claims of "superhuman" precision and perception, coupled with the inability to fully explain how these results are produced. Deep learning systems do not simply represent the world; they play an active role in shaping it. They tend to deepen and naturalize the classifications and hierarchies that are often the object of social contestation, while simultaneously excluding the possibility of political criticism or debate. This worrying combination, which involves reducing the capacity for action of subjugated social groups and increasing the autonomy of system builders, is the essence of the phenomenon that Campolo & Crawford (2020) call enchanted determinism. Such a phenomenon reflects a fundamental imbalance in power and agency, where those who define and implement deep learning systems often operate in a position of substantial power, while marginalized communities often face a reduction in their influence and ability to contest. This complex and asymmetrical dynamic represents a central point of concern when it comes to the ethical and social implications of deep learning technologies.

The "enchanted determinism" mentioned by Campolo & Crawford (2020) comes into play when a technological system is described in grandiose terms, often characterized by the sublime, while operating within a predetermined set of rules and results. However, close observation reveals very different dynamics in the process of how the system actually produces this kind of action, the result of which is merely a mathematical optimization on a scale far beyond human capacity in specialized spheres. This achievement is not the result of individual genius or transcendental intelligence, but of a systemic mathematical process. However, it is often discursively presented as an enchanted achievement, challenging the conventional intuitions of experts on strategy and performance. This dichotomy between dazzling presentation and disenchanted functioning is a central feature of enchanted determinism and sheds light on the way technology is often perceived and communicated in contemporary society.

One of the most pressing and fundamental limitations, both of a logical and political nature, affecting AI lies in its difficulty in recognizing and predicting truly unprecedented and unique events, as Pasquinelli and Joler (2021) point out. How machine learning deals with genuine anomalies, exceptional social behaviors and innovative acts that disrupt the status quo is particularly relevant to questioning the possible creativity attributed to generative AI. This inherent limitation in machine learning modalities is much more complex than mere preconceptions and poses significant challenges in the field of AI. It

refers to AI's ability to adapt to unexpected and unpredictable situations, highlighting the complexity underlying the development of AI systems considered autonomous.

A logical challenge intrinsic to machine learning classification, or pattern recognition, lies in its inability to identify genuinely novel anomalies, such as newly conceived poetic metaphors, humorous puns in informal communication contexts or unpredictable obstacles encountered on roads by autonomous vehicles. This limitation in detecting truly novel events, i.e. those that have never previously been observed by a model and therefore do not fit into known categories, represents a highly relevant problem, especially when considering the potentially dangerous consequences for decision-making systems such as autonomous vehicles, which have already been responsible for fatal accidents.

Similarly, the predictive capacity of machine learning, or the generation of patterns, reveals similar deficiencies when trying to anticipate future trends and behaviors. The machine learning technique, as a process of compacting information, tends to automate the imposition of past patterns and taxonomies on the present, effectively instituting what could be called a "dictatorship of the past" (Pasquinelli, Joler, 2021). Such an approach tends to impose a uniform space-time perspective, thus limiting the ability to recognize and accommodate new historical events. This leads to the phenomenon that we can characterize as "regeneration of the old".

The inclination to regenerate the old reflects machine learning's inherent difficulty in dealing with events devoid of historical precedents, which cannot be adequately understood through previous patterns. This is caused by an "epistemological flattening" (Campolo, Crawford, 2020) resulting from the increasingly intense application of deep learning in diverse social contexts. The theory underlying the application of deep learning emerges from the conviction that accurate prediction is intrinsically linked to the ability to extract correct information from chaotic or "noisy" social environments. In essence, this approach aims at "taming chance", which is present in social reality. This process of "epistemological flattening", which seeks to simplify complex social contexts into clear "signals" in order to improve predictive capacity, also leaves its mark on the social applications of machine learning. The tensions inherent in enchanted determinism become particularly acute when deep learning techniques promise to extract useful information without the need for epistemological modeling or the formulation of hypotheses from a classic probabilistic perspective.

In this scenario, effectiveness and explanatory capacity are often dissociated. Instead, we observe an approach that sets aside questions of cause and effect, with a preference for identifying complex patterns in a non-linear way in large data sets. In discourses of enchanted determinism, claims about accuracy rates tend to replace causal scientific explanations. This paradigm highlights the complex relationship between the predictive capabilities of deep learning techniques and the underlying epistemological commitments, as well as the ethical challenges associated with the quest for a deeper understanding of social systems and their dynamics.

Furthermore, in order to avoid enchanted determinism in the context of machine learning, it is necessary to have a logical definition stating that a security problem also logically sets the limit of its creative potential. This is because the challenges inherent in predicting innovative events or phenomena are intrinsically related to the problems faced in generating new content (Pasquinelli, Joler, 2021). The way a machine learning algorithm predicts a trend in a temporal context is, in essence, the same way it generates a new work of art based on previously assimilated patterns — i.e. it is stuck in the past (the data generated and fed into the system).

Pasquinelli and Joler (2021) suggest that we should change the seemingly trivial question "Does AI have the ability to be creative?" to a more technical formulation: "Is machine learning capable of generating works that are not mere imitations of the past?" "Can machine learning transcend the stylistic limits imposed by its training data?" The "creativity" attributed to machine learning is in fact delimited by its ability to detect styles in the training data and then improvise within those styles. Machine learning is therefore restricted to exploring and innovating only within the logical limits set by the data used

for training. Given these considerations and the emphasis on information compression inherent in this process, a more accurate description for the artistic output resulting from machine learning would be "statistical art" (Pasquinelli, Joler, 2021).

5 Final considerations

Thus, what we call the supremacy of patterns promotes two phenomena that must be considered in line with the exponential development of generative AI and its increasingly active participation in human decision-making processes. On the one hand, its operation with enormous statistical capacity on vast amounts of data produces an "enchanted determinism" (Campolo, Crawford, 2020), producing a false impression of creativity. On the other hand, this operation leads to a "dictatorship of the past" (Pasquinelli, Joler, 2021), which is a process of increasingly intense information compression that tends to automate the imposition of patterns from the past and taxonomies on the present. As a result, both the generation and identification of the new are annihilated in this process of enchantment, which has been a challenge for specialists in the field of knowledge known as XAI.

Describing how Artificial Intelligence works, especially in the context of future technological waves, which will not only enable the generation of content but also interactivity between AIs and between AIs and humans, represents crucial ethical and epistemological requirements. This approach aims to make a significant contribution to possibly reducing the supremacy of standards, even if a level of opacity persists, especially for experts in the field.

References

ADADI, A.; BERRADA, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, **6**: p. 52138-52160, 2018.

ANDERSON, C. 2023. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Disponível em: <https://www.wired.com/2008/06/pb-theory/> Acesso em 12 Out. 2023.

BRIDLE, J. 2018. *New Dark Age: Technology and the End of the Future*. New York: Verso Books.

BUHRMESTER, V.; MÜNCH, D.; ARENS, M. 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, **3**(4): p. 966-989. https://doi.org/10.3390/make3040048

BUNGE, M. 1963. A general black-box theory. *Philosophy of Science*, **30**(4): p. 346-358. Retrieved from https://www.jstor.org/stable/186066

CAMPOLO, A.; CRAWFORD, K. 2020. Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*.

CUPANI, A; PIETROCOLA, M. 2002. A relevância da epistemologia de Mario Bunge para o ensino de ciências. *Caderno Brasileiro de Ensino de Física*, Florianópolis, **19**: p.100-125. Número especial.

DEEKS, A. 2019. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, **119**(7): p. 1829-1850.

ESPOSITO, E. 2022. Transparency versus explanation: The role of ambiguity in legal AI. *Journal of Cross-disciplinary Research in Computational Law*, **1**(2).

FIOK, K.; FARAHANI, F. V.; KARWOWSKI, W.; AHRAM, T. 2022. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, **19**(2): p. 133-144.

FLORIDI, L. 2014. *The fourth revolution: how infosphere is reshaping human reality*. Oxford: Oxford University Press.

GURUMURTHY, A.; BHARTHUR, D. 2018. Democracy and the algorithmic turn. *SUR-Int'l J. on Hum Rts.*, **15**: p. 39.

HAQUE, AKM B.; ISLAM, AKM N.; MIKALEF, P. 2023. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. **Technological Forecasting and Social Change**, **186**: p. 122120.

HUME, D. 2007. *An Enquiry concerning Human Understanding*. Oxford: Oxford University Press.

MARIUTII, E. B. 2023. Apontamentos e digressões sobre o manifesto nooscópico: Inteligência Artificial e Complexidade. *Texto para Discussão*, (453): Unicamp, IE, Campinas.

MILLER, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, **267**: p. 1-38.

MOORE, J. D.; SWARTOUT, W. R. 1988. *Explanation in expert systems: A survey*. Marina del Rey, CA, USA: University of Southern California, Information Sciences Institute.

O'NEIL, C. 2016. *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York: Crown Publishers.

PASQUALE, F. 2015. *The black box society. The secret algorithms that control money and information*. Cambridge: Harvard University Press.

PASQUINELLI, M.; JOLER, V. 2021. The Nooscope manifested: AI as instrument of knowledge extractivism. *AI & Soc*, **36**: p. 1263–1280. Disponível em: https://doi.org/10.1007/s00146-020-01097-6

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

PINKER, S. 2021. *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. New York: Viking Penguin.

RICAURTE, P. 2019. Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, **20**(4): p. 350-365.

SHERMER, M. 2011. *Why People Believe Weird Things*. London: Souvenir Press.

VAN DER VELDEN, B. H. M.; KUIJF, H. F.; GILHUIJS, K. G. A.; VIERGEVER, M. A. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, **79**: p. 102470.

VAN LENT, M.; FISHER, W.; MANCUSO, M. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In: *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004. p. 900-907.

WIENER, N. 2019. *Cybernetics: or, Control and communication in the animal and the machine*. Reissue of the 1961 second edition. Cambridge: The MIT Press.

Submetido em 06 de novembro de 2023.

Aceito em 11 de janeiro de 2024.