REVISTA THESIS JURIS

Member since 2020 JM14465

Recebido: 23 abr. 2024 – **Aprovado:** 25 maio 2024
Editor-Chefe: Prof. Dr. Alexandre Reis Siqueira Freire
Processo de avaliação: *Double Blind Review*
e-ISSN: 2317-3580
http://doi.org/10.5585/13.2024.26510

Check for updates

From “black box” to “glass box”: using Explainable Artificial Intelligence (XAI) to reduce opacity and address bias in algorithmic models[1]

Da “caixa-preta” à “caixa de vidro”: o uso da Explainable Artificial Intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos

Otávio Morato de Andrade
Universidade Federal de Minas Gerais / Université libre de Bruxelles
Belo Horizonte, MG - Brasil
Doutorando em Direito na Universidade Federal de Minas Gerais (UFMG), com período sanduíche na Université libre de Bruxelles – Bélgica
otaviomorato@gmail.com

Marco Antônio Sousa Alves
Universidade Federal de Minas Gerais (UFMG)
Doutor em Filosofia
Belo Horizonte, Minas Gerais – Brasil

Abstract: Artificial intelligence (AI) has been extensively employed across various domains, with increasing social, ethical, and privacy implications. As their potential and applications expand, concerns arise about the reliability of AI systems, particularly those that use deep learning techniques that can make them true “black boxes”. Explainable artificial intelligence (XAI) aims to offer information that helps explain the predictive process of a given algorithmic model. This article examines the potential of XAI in elucidating algorithmic decisions and mitigating bias in AI systems. In the first stage of the work, the issue of AI fallibility and bias is discussed, emphasizing how opacity exacerbates these issues. The second part explores how XAI can enhance transparency, helping to combat algorithmic errors and biases. The article concludes that XAI can contribute to the identification of biases in algorithmic models, then it is suggested that the ability to “explain” should be a requirement for adopting AI systems in sensitive areas such as court decisions.

Keywords: XAI; explainable artificial intelligence; algorithmic opacity; transparency.

Resumo: A inteligência artificial (IA) tem sido utilizada em larga escala em variados domínios, com cada vez mais implicações sociais, éticas e de privacidade. À medida que suas potencialidades e aplicações são expandidas, surgem dúvidas sobre a confiabilidade dos sistemas equipados com IA, particularmente aqueles que empregam técnicas de deep learning que podem torná-los verdadeiras “caixas-pretas”. A XAI (explainable artificial intelligence), ou inteligência artificial explicável, objetiva oferecer informações que ajudam a explicar o

[1] Tradução do artigo publicado - ALVES, M. A. S.; ANDRADE, O. M. de. Da “caixa-preta” à “caixa de vidro”: o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, v. 18, n. 100, 2022. DOI: http://doi.org/10.11117/rdp.v18i100.5973. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 28 jun. 2024.

processo preditivo de determinado modelo algorítmico. Este artigo se volta especificamente para o estudo da XAI, investigando seu potencial para explicar decisões de modelos algorítmicos e combater o enviesamento dos sistemas de IA. Na primeira etapa do trabalho, é discutida a questão da falibilidade e enviesamento da IA, e como a opacidade agrava esses problemas. Na segunda parte, apresenta-se a inteligência artificial explicável e suas potenciais contribuições para tornar os sistemas mais transparentes, auxiliando no combate aos erros e vieses algorítmicos. Conclui-se que a XAI pode colaborar para a identificação de vieses em modelos algorítmicos, razão pela qual se sugere que a capacidade de “se explicar” – ou seja, a explicabilidade – seja um requisito para a adoção de sistemas de IA em searas mais sensíveis, como, por exemplo, o auxílio à tomada de decisão judicial.

Palavras-chave: XAI; inteligência artificial explicável; opacidade algorítmica; transparência.

Para citar este artigo (ABNT NBR 6023:2018)

ANDRADE, Otávio Morato; ALVES, Marco Antônio Sousa. From “black box” to “glass box”: using Explainable Artificial Intelligence (XAI) to reduce opacity and address bias in algorithmic models. **Revista Thesis Juris – RTJ**, São Paulo, v. 13, n. 1, p. 3-21, jan./jun. 2024. http://doi.org/10.5585/13.2024.26510

Introduction

In recent years, systems equipped with artificial intelligence (AI) have invaded our lives and optimized processes in various fields, such as search recommendations, products, and user preferences, automated customer service through chatbots, medical diagnostics, and "smart" devices, such as autonomous cars. The advancement and popularization of AI raise a wide range of concerns, from privacy issues – vulnerable to the predictive power of big techs – to the dangers of compromising the capacity for subjectivation (Rouvroy; Berns, 2015), and ethical implications regarding discriminatory biases. There is also a lively debate on how new AI-related phenomena could affect democracy and political pluralism, through the spread of fake news, ideological radicalization, or even the sophistication of censorship and mass surveillance techniques (Sunstein, 2009; Bruno, 2013).

These concerns are intensified by the fact that the inner workings of certain algorithms – particularly those equipped with deep learning – can be a complete mystery to the average technology user and, not uncommonly, even to those with advanced competencies in the field. While in machine learning there is a more streamlined statistical learning structure between data input and output, in deep learning, there are multiple layers of neural networks that overlap each other, making the understanding of their reasoning more complex. Thus, when it comes to systems involving multiple artificial neural networks, tools and techniques capable of facilitating the understanding of an algorithm's decisions are still scarce (Cortiz, 2021).

It must be considered that AI is not a technology in itself, but rather an area of knowledge, formed by different strands, including machine learning, which should not be seen as synonymous with a "black box." After all, not every machine learning-based technique suffers from algorithmic opacity. Supervised machine learning techniques, for example, are easily explainable. In this article, our focus will be directed towards unsupervised techniques, especially deep learning or deep learning, which can be understood as a sub-area of statistical learning or machine learning. In this domain, the issue of explanation and regulation assumes a much more delicate and complex contour.

Faced with this "algorithmic opacity," i.e., the inability to see beyond the output produced, it is questioned whether humans should delegate such important decisions to AI systems, in cases where they are unable to explain how they arrived at certain conclusions. Indeed, many researchers who have delved into the subject have considered it essential to equip AI systems with functionalities capable of providing a reasonable explanation for their predictions (Villani, 2019, p. 114; Confalonieri et al., 2020). Among the options presented as capable of providing such an explanation, systems that have been called "explainable artificial intelligence" (explainable artificial intelligence – XAI) appear.

This work aims to specifically investigate XAI (explainable artificial intelligence) as a way to reduce the opacity of algorithmic models. Studies on the subject have suggested that systems enabled for XAI, by reducing opacity, assist in revealing flaws in algorithms, providing the opportunity to correct, or at least minimize, machine bias, making these systems more reliable (Ribeiro et al., 2016; Nunes; Andrade, 2022; Wells; Bednarz, 2021).

However, we would like to make it clear that we do not intend with this work to defend a purely technological solution to all the ethical and political problems brought about by the use of systems equipped with artificial intelligence. It is far from our intention to embrace a naive technophilic position, which bets simply on the possibility of technology solving by itself all the problems it raises, through just ever more sophisticated systems. We need to be aware of the limits of such an undertaking, identifying the most sensitive areas and setting clear boundaries regarding the form and, also, the very possibility of using artificial intelligence. Despite this, we understand that it is possible to improve existing systems and expand their potential for appropriate and responsible use in various domains, through, for example, more reliable and transparent mechanisms.

This work is divided into two major sections. First, the issue of fallibility and biases in algorithmic models is problematized, discussing some cases where AI systems have provided imprecise or mistaken prediction results (outputs), or that, even when correct, have undertaken

reasoning that disregarded desirable requirements. In the second stage of the work, fundamental notions about XAI are presented, explaining some techniques that act to reduce the opacity of AI systems. At the end of this stage, some examples of fallibility and biases listed in the first section are revisited, analyzing how related techniques of explainable artificial intelligence could address solutions in each of these cases. The work concludes that XAI can contribute to the identification and treatment of problems in algorithmic models, which is why it is suggested that the ability to "explain itself"—i.e., explainability—should be a requirement for the adoption of AI systems in more sensitive areas, such as assisting in judicial decision-making.

1 When ai fails: the problem of failing in the dark

1.1 AI fallibility

No matter how sophisticated, an AI system is not immune to producing imprecise, incomplete, or biased results. Several reasons can be behind an erroneous predictive outcome. Firstly, the input data may be incomplete or conflicting, creating ambiguities for the algorithm analyzing them. Additionally, the computational prediction may be poorly calibrated or insufficiently trained, failing to interpret these data and thus providing incorrect results (Ramos, 2020). Lastly, there are also cases where the algorithm "gets it right," but does so using reasoning and approximations that are not desirable.

Fallibility occurs when a system fails to correlate data in a causal manner, generating inconclusive evidence and unjustified actions (Rossetti; Angeluci, 2021, p. 8). An example of fallibility is the confusion made by image classifying algorithms trying to distinguish between wolves and Husky dogs, especially when there is snow in the picture. Since most photos of wolves in the training data contained a snowy background, the algorithm ended up using the environment—and not the animal's characteristics—to classify the image. When analyzing a Husky dog depicted against a snowy background, the system fails, misclassifying it as a "wolf" (Ribeiro et al., 2016).

Figure 1: A Husky dog (left) is mistaken for a wolf, because the snowy background (right) was wrongly associated with wolves. This failure is due to insufficient training data (many wolves photographed against the snow).

Klaus-Robert Müller and Wojciech Samek also show that an algorithm can provide an apparently "correct" result by using a flawed logical path. The authors offer the analogy of the horse Clever Hans, who became famous in the early 20th century for supposedly performing basic mathematical operations, with over 90% accuracy. Later, it was discovered that during presentations, the horse was merely responding to the body language of its trainer, who calculated the results and "whispered" the correct answer to Hans through signals. Thus, Clever Hans did indeed "get" the results right, but for reasons unrelated to mathematics.

According to Müller and Samek (2019, p. 3), the same can occur with some AI systems. A reported case involves an image classifier algorithm that won several awards in the field. Later, it was found that its prediction often did not detect the main object. Instead, it simply used correlations and indirect data to arrive at the result. While it often guessed correctly, it was discovered that the model recognized boats by the presence of water, trains by the presence of tracks, and even horses by the presence of a copyright watermark embedded in the image. These "Clever Hans predictors" may perform well in test scenarios, but will surely fail when deployed in the real world, where the recognized objects are often out of their original context. Müller and Samek (2019, p. 4 – our translation) add that, "if the AI system is a black box, it will be very difficult to unmask such predictors."

Therefore, when an algorithm produces an incorrect decision—or even a correct one but based on false premises—we are faced with fallibility, a situation in which the AI system does not operate as desired, whether due to reasons related to the design of the algorithm or the way data are encoded, collected, selected, or used to train the algorithm. Often, fallibility has innocuous effects. However, when a failure produced by AI algorithms affects groups or individuals, potentially generating biased or discriminatory results, it acquires a social

dimension, which is why it begins to be treated as algorithmic bias, as we will see in the next topic.

1.2 Algorithmic biases, prejudice, and discrimination

Algorithmic biases are tendencies occasionally produced by an AI system, generally reflecting human preference for certain values, due to social and cultural factors existing before programming and surrounding the designers (Rossetti; Angeluci, 2021). The incorporation of these tendencies into an AI system generally does not occur deliberately by programmers, but through incorrect training of the algorithm or unexpected outcomes of machine learning, compromising the system's neutrality.

Considering that AI has been used to deliberate on crucial human issues, the "contamination" of a particular algorithm by a moral tendency could reproduce prejudices and create unfair results, such as privileging one group of users over others (Najibi, 2020). In this case, biases go beyond a simple failure, as they can generate serious social repercussions, such as reinforcing social prejudices related to race, gender, sexuality, or ethnicity, leading to systematic and unfair discrimination.

As AI advances, reports of algorithmic prejudice multiply each year in the scientific literature, and it would be impossible in this brief work to exhaust all documented occurrences to date. In the following sub-items, we select and summarize three cases where algorithmic bias may have helped produce systematically discriminatory decisions. These same cases, along with the Husky example reported earlier, will be revisited at the end of the second section of this article, from the perspective of XAI, to demonstrate how explainable artificial intelligence could contribute to the detection and mitigation of these failures and biases.

1.2.1 Black individuals improperly classified with high recidivism rates

An example of an AI system capable of producing discriminatory results is Compas—correctional offender management profiling for alternative sanctions, a North American tool used to estimate the recidivism risk of prisoners in the country. A "risk rating" is developed by the system based on a questionnaire of 137 questions asked of the defendant, his criminal history, and also according to the platform's database. Generally, the data produced by Compas help establish bail amounts, inform judicial decisions regarding the defendant's freedom during

the process, and, in some states, have even greater relevance, potentially underpinning criminal sentencing (Angwin et al., 2016).

In 2016, however, an analysis conducted by the independent journalism organization ProPublica revealed that the algorithm used in Compas contained discriminatory biases. The authors of the study analyzed scores of more than 7,000 prisoners in Florida, concluding that the algorithm is more likely to mistakenly classify black defendants as "likely recidivists" and, conversely, also mistakenly frame white defendants as "individuals with low risk of recidivism" (Nunes; Marques, 2018, p. 6).

The journalists from ProPublica also found that Northpointe, the company responsible for the system, does not make public the algorithm on which the recidivism index of detainees is based, but only the questions asked of the individual and used in the calculation, so the defendant does not know why he has a high or low indicator, or even how his answers influenced the weighting of the final result (Angwin et al., 2016).

1.2.2 Prejudice in Google Photos

In 2015, a black programmer exposed a flaw in Google's photo service, which had labeled photos of him and a black friend as "gorillas." At the time, the big tech came forward to declare itself "horrified" by the failure, showing itself "committed" to ending discriminatory biases in algorithms and promising "quick correction" (Simonite, 2018).

Nearly three years later, in 2018, the American technology magazine Wired tested Google Photos using a collection of more than 40,000 images with various animal species. Although the algorithm showed remarkable performance in recognizing many creatures, such as pandas and poodles, the service curiously reported "no result" for the search terms "gorilla," "chimpanzee," and "monkey" (Simonite, 2018). It was thus discovered that Google's "solution" to the problem of Google Photos had been simplistic: despite the great repercussion of the complaint, the company merely removed gorillas and some other primates from the service's dictionary, instead of refining the recognition algorithms to correct their flaws and biases.

The exotic solution illustrates the difficulties that Google and other technology companies face in advancing image recognition technology, which, however, is already applied in extremely sensitive areas, such as migration control, protest monitoring, airport surveillance, and counter-terrorism.

1.2.3 Discrimination against women in credit granting

A recent report prepared by researchers at Oxford showed that women entrepreneurs tend to raise less funding from private shareholders, mainly due to the resistance of men (who are the majority of investors) in financing women-led companies. According to the report, many female entrepreneurs interviewed revealed that, to circumvent this problem, they prefer to ask for money from banks, since these use scoring algorithms (bank scoring), which would be automatically more "impartial" in granting credit (Sako; Parnham, 2021, p. 107).

The problem is that this supposed neutrality of algorithmic scores has been called into question by clients themselves and also by North American regulatory bodies. In 2019, the New York State Department of Financial Services opened an investigation into allegations that the Apple credit card was offering different credit limits for men and women. Several card users—including Apple co-founder Steve Wozniak—claimed on their social networks that score algorithms discriminated against women. Tech industry entrepreneur David Hansson, for example, complained that the Apple Card gave him 20 times the credit limit that his wife received, even though her formal income exceeded his. Later, Wozniak tweeted that the same thing happened to him and his wife, although they did not have separate bank accounts or assets. In its defense, Goldman Sachs, which offers the card in partnership with Apple, claimed that its credit decisions are essentially based on the credit quality of the customer, not on factors such as sex, race, age, or sexual orientation (Natarajan; Nasiripour, 2019).

Despite the uncertainty of what specifically caused the reported disparities, the problems with the AppleCard reveal a possible bias in the algorithmic models used in bank scores, which may be at the origin of discriminatory effects and differences in opportunity between men and women.

What do the three reported examples have in common? Besides the detected algorithmic bias, which ended up reproducing race and gender prejudices, all the AI systems mentioned above employed machine learning, whose predictive sequence is often incomprehensible—i.e., opaque—to humans. Algorithmic opacity amplifies the challenge of detecting and correcting biases, as we will see next.

1.3 Opacity may cover up algorithmic flaws and biases

In traditional programming, building software consisted of writing a logical model by hand, i.e., outlining a set of rules that allowed conclusions to be reached from processing individual cases. Such models are, by definition, interpretable, since their source code was previously written by a developer, making it possible to say, in each individual case, which and how the instructions were triggered to reach a result (e.g.: if an applicant's income is below "γ" per month, the financing will be denied by the system) (Villani, 2019, p. 114).

On the other hand, algorithms that incorporate machine learning, such as, for example, random forests, only return results, without, however, providing reasonable explanations for how a particular prediction was reached. In these cases, since it is not possible to clearly discern the decision-making process behind the output, the algorithm is said to be opaque—constituting a true "black box," incapable of providing reasonably understandable explanations for a human. Roberto Confalonieri and his colleagues synthesize the problem:

> Although some algorithmic models can be considered interpretable by design [...] most machine learning models behave like "black boxes." From an input, a "black box" will return the result [...] without revealing enough details about its internal logic, resulting in an opaque decision model. (Confalonieri et al., 2020, p. 7 – our translation)

However, it should be noted that not all machine learning will necessarily lead to absolute opacity. Many machine learning systems employ so-called supervised learning, through which programmers "train" the algorithm using examples and predetermined rules. In the case of supervised learning, the analysis of this set of prior instructions allows a better understanding of its reasoning stages and how data are analyzed. On the other hand, in the case of unsupervised learning, in which the number of preliminary instructions left by the developer is smaller, the algorithm operates more autonomously and, therefore, less intuitively for humans. The intelligibility of the system decreases as the algorithm begins to house not just one, but multiple neural networks that overlap, which commonly occurs in deep learning algorithms (Ghahramani, 2004).

The big problem is that, due to the increasing power and speed of AI algorithms—especially those that employ deep learning—it is almost impossible to follow their reasoning, even in cases where their code is open. For example, to recognize an image, a classifier considers millions of criteria, using millions of images from its training bank, which, in turn, contain millions of pixels (4K) (Villani, 2019, p. 114).

However, despite the intricate technical challenge of unraveling the "black boxes" of algorithms, it will be inevitable to face it, given the exponential advancement of AI and its numerous ramifications. If, in some cases, the implications are minimal, there are, on the other hand, sensitive areas in which the systematic reproduction of these mistakes cannot be admitted, especially considering that AI systems have the capacity to handle problems on a large scale. To give an idea of this scalability, the AI system called Victor, currently in operation at the Supreme Federal Court, can analyze a case in 5 seconds, while a server, to perform the same task, takes 44 minutes. Meanwhile, the Athos system, used at the Superior Court of Justice, can analyze up to 30,000 cases monthly (Andrade, 2022). In this sense, the presence of flaws or biases in algorithms of high importance and scalability can have high-risk implications, such as the biasing of hundreds or thousands of judicial analyses.

Thus, the inability to understand the relationships between input data and output data in AI systems may turn an AI system into a true "black box," to which decisions crucial to human well-being cannot (or at least should not) be entrusted, due to the simple impossibility of trusting a system whose chain of reasoning remains hidden, potentially covering up flaws or discriminatory biases. Consider the hypothetical example of a robot that assassinates, seemingly deliberately, a human being. It would be essential to unveil the internal logic that led to such an act—even for the purposes of accountability, if necessary, of the programmers, the manufacturer, or the owner of that robot. If the algorithm is a "black box," from which the internal predictive process cannot be discerned, it would be a great challenge to determine when, how, and why the algorithm erred, information without which the solution of the crime and the assignment of responsibilities become exceedingly thorny tasks.

However, it would not be the case to remove algorithms from any and all decision-making, but rather to approach artificial intelligence in a way that encourages and enables its explanation, to a greater or lesser extent, depending on its final application. This approach should aim to align AI with social values, taking into account a series of ethical questions, among which the fallibility, opacity, bias, discrimination, autonomy, responsibility, and privacy of information stand out (Rossetti; Angeluci, 2021, p. 7).

Given these reflections, we understand that the development of functionalities that enable AI to provide satisfactory explanations for its actions could resolve, in part, the problem of algorithmic opacity. It is not just about promoting transparency, but about developing an active posture, which enables AI systems to make clear their intentions, motivations, and the causal chain behind a decision, notably when it has relevant individual or social repercussions.

These "explanatory" properties have already been explored in the promising field of explainable artificial intelligence (XAI), which will be addressed next.

2 Explainable Artificial Intelligence

2.1 Explainable artificial intelligence (XAI): fundamental concepts

An intelligent system equipped with explainable artificial intelligence (XAI) is one that possesses interpretability or explainability, that is, the ability to explain its predictions through textual or visual strategies that provide qualitative understanding about its prediction process (Ribeiro et al., 2018, p. 2). The explainable AI system is enabled to provide explanations about its operation, making its behavior more intelligible to humans (Gunning et al., 2019). This means that an XAI system should be able to explain, in a way appropriate to a human being, the internal logic of its prediction: what was done, what is being done now, and what will happen next.

The literature highlights that the level of detail and characteristics of XAI should be established taking into account the target audience of the explanation. For example, software developers may understand small Bayesian networks, but they are a complete enigma to the lay user (Ribeiro et al., 2016, p. 4). Similarly, explanations that are too basic may be insufficient for experts to review or audit an algorithm. It is therefore essential that the development of XAI does not lose sight of its end consumer, providing, depending on the recipient, a) an "appropriate" amount of information (neither scant nor excessive) and b) explanations in a language comprehensible to the interlocutor.

The first attempts to generate XAI date back to the 1960s and 1970s, notably through the Mycin and Centaur systems, developed at Stanford University. However, it was encountered that the "explanations" generated by these systems were, in truth, verbalizations of the rules, not consistent interpretations of the routines or architecture of the system. A formal expression of "why we did it this way" is a justification, not an explanation (Mueller et al., 2019). Besides this limitation, it must be considered that the systems studied at the time (mostly expert systems) were substantially simpler than the machine learning and deep learning algorithms used today. Thus, despite attempts to develop explainable intelligence in the past, the challenge of providing satisfactory explanations persists, especially considering the enormous complexity of algorithmic processes brought about by advances in computational technology, particularly in the field of nanoscience.

As already seen earlier, the implementation of AI systems that use machine learning has become a central concern of scholars, as some algorithmic models constitute true "black boxes": so complex are certain predictive processes that they end up being indecipherable—even to programmers and technicians in the field. This complicates, for example: a) the experience and confidence of the end user, who may be harmed by erroneous system predictions; b) the improvement of these models by their developers, as it is not always known when and how AI fails; c) legal compliance of these tools, as their holders are subject to increased legal risk, due to opacity and also the greater fallibility of non-interpretable algorithmic models (Nunes; Andrade, 2021).

In this vein, although AI offers extraordinary possibilities in our daily lives, it is well established by scholars of the subject that its opacity, in some cases, is not desirable. Indeed, as algorithms invade our daily lives, they must reflect our laws and social standards. Faced with "algorithmic black boxes," the role of XAI will be crucial to understand, audit, and correct these systems, continually seeking their ethical and legal compliance.

Finally, it is worth noting that not every artificial intelligence system presupposes the need for XAI. Molnar (2021) offers two examples: a) when the algorithmic model and its predictions have low impact, with no social ramifications, and b) in the case where the applications of a particular system are already sufficiently studied and established—as in the case of using facial recognition to unlock cell phones.

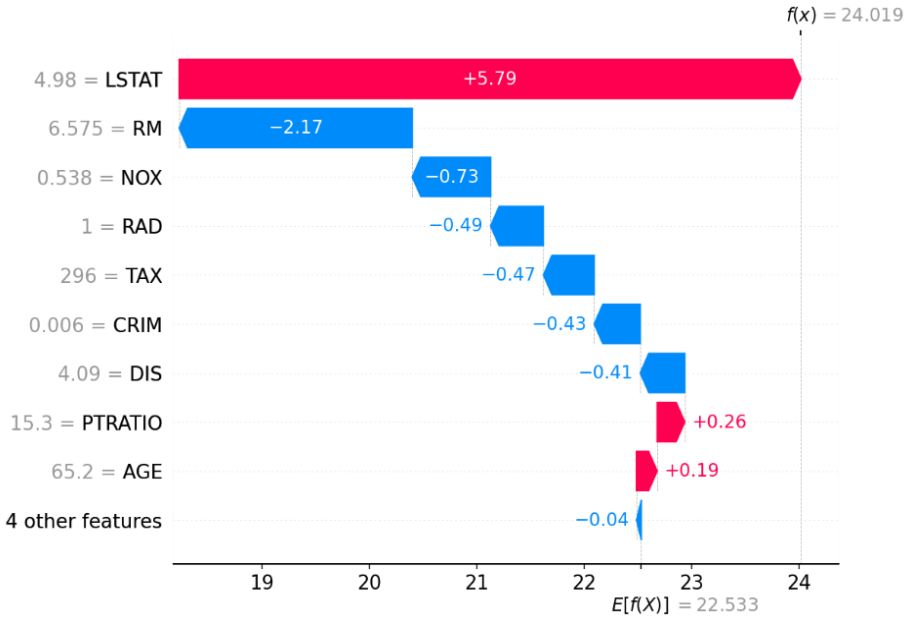
2.2 XAI in practice

The specialized literature makes a distinction between algorithmic models that are originally interpretable and those that need to be explained through specific XAI techniques. A "transparent" machine learning model is one that is explainable by itself, not requiring additional techniques for a human to understand it. In contrast to "transparent" models, there are "opaque" models, whose understanding will require an additional explanation process, called post-hoc explainability. Post-hoc explainability is directed towards models that are not readily interpretable by their design, resorting to various XAI techniques, such as text explanations, visual explanations, explanations through examples, simplification explanations, and feature relevance explanations.

Given the diversity and depth of XAI techniques, it would be impossible to discuss the specifics of each of them. But one important example allows us to clarify the functioning of XAI in general terms, which is sufficient for the purposes of this study. Consider the case of

explanation by feature relevance. This method aims to better describe an opaque algorithm, emphasizing the resources and variables crucial to the final result (output) of the algorithmic prediction. A notable contribution is the SHAPs (SHapley Additive exPlanations), which propose a kind of "score" for the influence of each predictive characteristic during the algorithmic processing. Through SHAPs, the variables used in the predictive process are ranked, presenting those that most influence the algorithm in one direction or another (Lundberg; Lee, 2017).

Figure 2: The explanation above works on the idea of relevance, detailing how each feature contributed to the "output". Features that "push" the prediction upward are shown in red, while those that "push" the prediction downward are shown in blue. Source: GitHub (https://github.com/slundberg/shap)



By revealing the "weight" of the most decisive items in the analysis, SHAPs provide crucial information about the functioning of the algorithm, allowing, for example, a developer to identify a variable that is being under or overestimated in the prediction and, consequently, make adjustments in the weights of each variable, in order to remove biases and increase the accuracy of the results provided by the system.

3 The value of explanation in addressing algorithmic biases

3.1 Detection of Unethical Correlations

In the first section of this article, it was shown through the Compas and AppleCard cases how black people and women can be harmed by prejudiced scoring systems. At the root of this discrimination are algorithms that, despite supposedly being impartial, may be imbued with their creators' subjectivity or affected by the quality of the training and data provided. In the case of very complex systems, opacity can exacerbate and conceal such biases, as it makes it difficult to understand the applied predictive logic. To resolve this impasse, XAI could be implemented as a functionality that helps detect and correct inherently biased logical chains. After all, as Nunes and Marques (2018, p. 7) lecture about Compas:

> The lack of transparency of the algorithm is especially critical in this case. How can one defend against an "index" without knowing the method of its calculation? How can the "index" be subjected to the control of due constitutional process? As much as the questions asked are disclosed, the accused do not know how their answers influence the final result (output). Thus, the defense of the accused becomes impossible due to opaque mathematical data and algorithmically biased, yet camouflaged, by the "security" of mathematics as supposedly impartial, impersonal, and fair.

XAI can even help identify the correlations (not so obvious) made by scoring systems. In this sense, although the AppleCard system did not establish a direct causality between score/race or score/gender, as the service provider claims, it is suspected that an indirect relationship may have been drawn from a broader data set. For example, over the course of a marriage, men tend to take out more loans in their own name, rather than jointly with their wives. When unadjusted, this data can lead the algorithm to infer that, generally, men take and honor loans more frequently than women, thus being more reliable (Kelion, 2019). Similarly, Compas, even without asking the race of the inmate, might be absorbing and considering this information indirectly, since the questionnaire contains questions that end up selecting poor individuals who are mostly black.

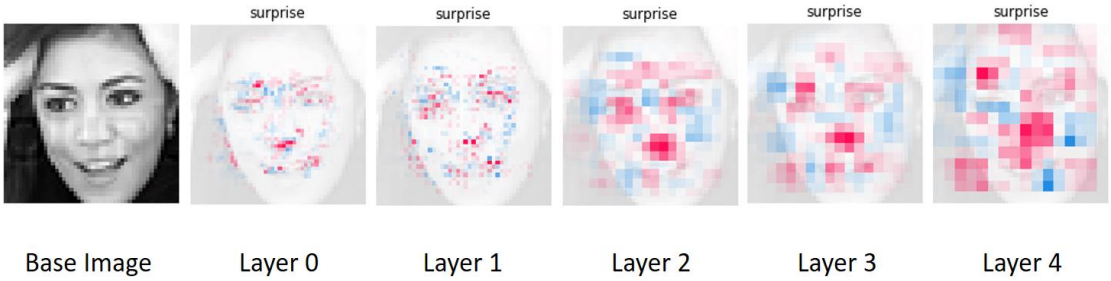
Therefore, through adequate explanations about the predictive processes of algorithms, it would be possible to identify (and eventually correct) the establishment of correlations like these which, although indirect, have a discriminatory effect, whether by favoring men, as in the case of AppleCard, or by favoring white people, in the case of Compas. Once it is established, after the issuance of the explanation, that a system ended up giving relevance – even if indirectly – to a racial or gender data in its judgment, it would be appropriate, therefore, to recalibrate the algorithmic model in question, avoiding it from expressing new prejudices in its scoring.

3.2 Enhancing Image Recognition

Serengil (2019) demonstrates that SHAPs are also capable of explaining how an algorithm distinguishes between certain emotions through facial recognition. The author used a training data set called FER-2013, which contained images of facial expressions of 28,709 people, capable of distinguishing between seven categories (0 = angry, 1 = disgust, 2 = fear, 3 = happy, 4 = sad, 5 = surprise, 6 = neutral).

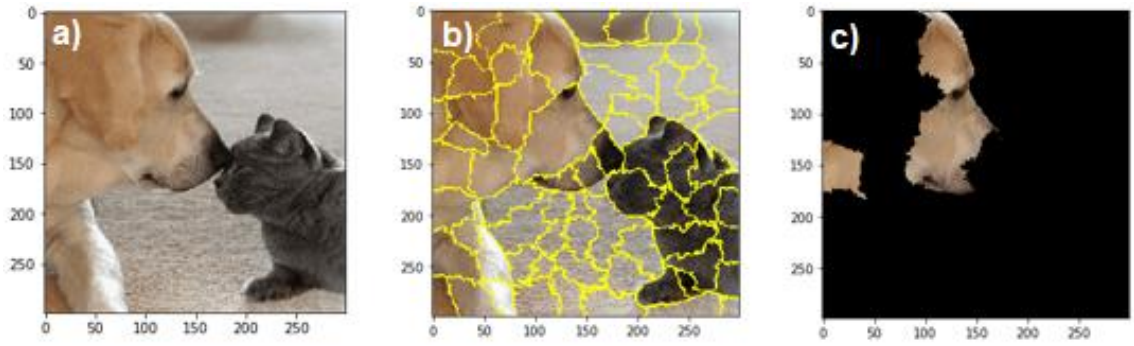
In the sequence of images below, it is possible to see how this method couples the technique of explanation by feature relevance to image classification, providing detailed and consistent information about the facial recognition process of a classifier system. Similarly, deciphering the predictive process of a classifier can be applied in other areas, to refine the algorithm's accuracy (Serengil, 2019).

Figure 3: The SHAP method (SHapley Additive exPlanations), used to explain the algorithmic detection of a facial expression: the first layers (layers 0 and 1) focus on facial features (eyes, nose, mouth, etc.), while the subsequent layers mention other areas of the face. In red, the pixels that most influence the prediction, while those with low importance are marked in blue.



Another explanation method that can be applied to image classifiers is LIME (local interpretable model-agnostic explanations), capable of explaining how input features affect the algorithmic prediction (output) (Ribeiro et al., 2016). Briefly, this system generates random perturbations in the input image, "turning off" and "on" some pixels to find "superpixels," i.e., significant image segments that resemble the cataloged database (Arteaga, 2020).

Figure 4: LIME explaining a result. Through perturbations in the original area (a), "superpixels" were created (b), and among them, the algorithm highlighted the most significant ones (c), which, compared with the database, have a stronger association with the "Labrador Retriever" breed.



As can be seen, the explanation of the predictive process, even if simplified, can help locate and adjust flaws. In the case of the Husky problem, for example, LIME is capable of revealing to the functionality user that snow acquires disproportionate significance in classification (Ribeiro et al., 2016). The same functionality could be applied to help refine Google Photos' recognition, to identify and correct the crude correlations established by the algorithm between superpixels present on the faces of black people and monkeys.

It is also noteworthy that by providing explanations about the predictive process, XAI can change users' perceptions of the reliability of a particular tool, correcting problems such as the user's "blind trust" or, also, distrust towards an algorithm. This was demonstrated in a study developed by professors from the University of Washington who asked programming students about their trust in an image classifier algorithm (Ribeiro et al., 2016, p. 9). Initially, just over a third of the individuals trusted the algorithm. After receiving the explanation elaborated by LIME – revealing that the background of the image carried considerable weight in the classification – trust in the classifier fell substantially (to approximately 10%). This evidence draws attention to the importance of coupling explanations to systems, precisely so that users do not overestimate the reliability of a possibly imprecise or biased algorithm.

This work demonstrated, in its first phase, how algorithmic failures and biases emerge, particularly by analyzing the occurrence of these problems in opaque contexts, that is, in which there is little or no understanding of the predictive process employed by a complex AI system. In a “black box” algorithm, the risk of such biases going unnoticed increases, and they can even be reproduced on a large scale. While some algorithms have minor implications, special attention should be given to those whose results can have significant individual or collective

repercussions, based on evaluations of race, gender, sexuality, and ethnicity that can lead to systematic discrimination, as in the cases of Compas and AppleCard.

Given the observation that opacity can conceal and amplify algorithmic failures and biases, the second phase of this work considered that explainable artificial intelligence (XAI) could help combat opacity by enabling the system to provide explanations about its own predictive process. This is a transformation process of the algorithmic “black box” into a genuine “glass box” – that is, transparent, easy to visualize and understand – which helps identify undesirable correlations established within the algorithm, allowing system developers to trace and correct existing flaws and biases. Moreover, the “glass box” allows for verification, auditing, and accountability when AI makes illegal decisions. Finally, XAI also promotes user and societal trust in artificial intelligence itself, as it generally shows when, how, and why an algorithm is making a particular decision.

It should be noted, however, that the solutions offered by explainable artificial intelligence are still in the experimental field, mostly concentrated in academic investigations. In this sense, there is a considerable path until XAI is satisfactorily developed and implemented to benefit its end user with useful and accessible explanations. Therefore, it will be necessary for developing companies – and ultimately, their controllers and investors – to be “motivated” to fund the development of self-explanatory functionalities for complex AI systems.

It would be naive to expect spontaneous mobilization by technology companies around ethics and transparency – which is why we believe that the true catalyst of this motivation lies precisely in the debate on the right to explanation. Hence, the discussion on guidelines related to explanation and, ultimately, the establishment and stabilization of these rights in legal frameworks around the globe (as has been occurring in the European Union) can impose sanctions on algorithmic opacity, encouraging companies to make their software more transparent and interpretable.

Although we understand that XAI constitutes a path to be explored in the coming years, it is necessary to recognize that today it is nothing more than a promise, far from being a consensus in the market. There is no guarantee of its success, much less in a broad and generalized manner. Most likely, it will be impossible to advance the implementation of explainability in certain more complex domains. In such cases, we believe that more assertive regulatory measures will be necessary, especially when more sensitive issues are at stake, such as environmental protection or fundamental rights. These measures may involve control, supervision, or real-time human oversight, even reaching the banning of the technology in certain situations.

This set of reflections leads us to conclude that the compatibility of artificial intelligence with laws and social values presupposes not only access but also the guarantee of understanding of the predictive processes employed in any AI system to which a decision with significant effects is delegated – whether at a private level, such as the granting of a personal loan, or on a broader scale, such as aiding judicial decision-making to sentence defendants.

It is not just about encouraging simple transparency but about building, at the level of public policies, a stance that recognizes XAI as a requirement for the adoption of AI systems in more sensitive areas. Considering the relevance of certain decisions, they should be delegated only to algorithms capable of clarifying their intentions and motivations, explicitly presenting their analysis in language comprehensible to humans. The more important a decision is from a social perspective, the more capable an AI system needs to be to provide detailed, precise, and understandable explanations, so there are no doubts about its neutrality and competence.

Final Considerations

This work demonstrated, in its first stage, how algorithmic biases and failures arise, particularly by analyzing the occurrence of these problems in opaque contexts, where there is little or no understanding of the predictive process employed by a complex AI system. In a "black box" algorithm, the risk that such deviations go unnoticed is elevated, potentially allowing them to be reproduced on a large scale. While some algorithms have insignificant implications, special attention should be given to those whose results can have significant individual or collective repercussions, based on race, gender, sexuality, and ethnicity assessments capable of generating systematic discriminations, as seen in the cases of Compas and AppleCard.

Given that opacity can conceal and amplify algorithmic biases and failures, the second stage of the work considered that explainable artificial intelligence (XAI) could assist in combating opacity by enabling the system to provide explanations about its own predictive process. This is a transformation process of the algorithmic "black box" into an authentic "glass box" – that is, transparent, easy to visualize *and* understand – which helps identify undesirable correlations established within the algorithm, allowing system developers to track and correct existing biases and flaws. Additionally, the "glass box" enables verifiability, auditing, and accountability when AI makes illegal decisions. Finally, XAI also promotes user and societal trust in artificial intelligence itself, as it generally shows when, how, and why an algorithm is making a particular decision.

XAI should not be reduced to a mere call for more algorithmic transparency. Thomas Berns and Tyler Reigeluth (2021, p. 156) astutely observe that the light of transparency can be blinding, either because algorithm source codes, even if made available, are unintelligible to most citizens, or due to the risk of creating a "technical" field of transparency specialists who act as regulators of the informational market. However, XAI goes a step beyond transparency, promoting an experimentation of technical realities that transcends the abstract or alienated knowledge of machine functioning (Andrade, 2024).

Also, it is important to note that the solutions offered by explainable artificial intelligence are still in the experimental field, mostly concentrated in academic investigations. In this sense, there is a considerable path until XAI is developed and satisfactorily implemented to benefit its end users with useful and accessible explanations. Therefore, it will be necessary for developing companies – and ultimately their controllers and investors – to be "motivated" to finance the development of self-explanation functionalities for complex AI systems.

It would be naïve to expect spontaneous mobilization by technology companies around ethics and transparency, which is why we imagine that the true catalyst for this motivation lies precisely in the debate about the right to an explanation. Thus, the discussion about guidelines related to explanation and, finally, the entry and stabilization of these rights in legal systems around the globe (as has occurred in the European Union) can establish sanctions for algorithmic opacity, encouraging companies to make their software more transparent and interpretable.

Although we understand that XAI constitutes a path to be explored in the coming years, it must be recognized that it remains a promise today, far from being unanimous in the market. Nothing guarantees its success, even less so in a broad and generalized manner. It will likely be impossible to advance the implementation of explainability in certain more complex domains. In these cases, we believe that more assertive regulatory measures will be necessary, especially when more sensitive issues are at stake, such as environmental protection or fundamental rights. These measures may involve control, supervision, or real-time human oversight, even reaching the point of banning the technology in certain situations.

This set of reflections leads us to conclude that making artificial intelligence compatible with laws and social values presupposes not only access but also the guarantee of understanding the predictive processes employed in any AI system to which a decision with significant effects is delegated – whether at a private level, such as granting a personal loan, or of a broader nature, such as assisting judicial decision-making to sentence defendants.

It is not only about encouraging simple transparency but building, at the level of public policies, a stance that recognizes XAI as a requirement for adopting AI systems in more sensitive areas. Considering the relevance of certain decisions, they should only be delegated to algorithms capable of clarifying their intentions and motivations, explicitly explaining their analysis in a language understandable to humans. The more important a decision is from a social perspective, the more capable an AI system needs to be to provide detailed, precise, and comprehensible explanations, so that there are no doubts about its neutrality and competence.

References

ALVES, Marco Antônio Sousa. **Cidade inteligente e governamentalidade algorítmica: liberdade e controle na era da informação**. *Philósophos*, Goiânia, v. 23, n. 2, p. 191-232, 2018. https://www.revistas.ufg.br/philosophos/article/view/52730. Access: 10 ago. 2021.

ANDRADE, Otávio Morato de. **Governamentalidade algorítmica: democracia em risco**? 1st ed. São Paulo: Dialética, 2022. 224 p.

ANDRADE, Otávio Morato de. **Open machine? Explainable Artificial Intelligence as technique realized in light of Simondon.** In press, 2024.

ANGWIN, Julia; LARSON, Jeff; SURYA, Mattu; KIRCHNER, Lauren. **Machine Bias.** *Pro Publica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Access: August 19, 2021.

ARTEAGA, Cristian. **Interpretable machine learning for image classification with LIME: increase confidence in your machine-learning model by understanding its prediction.** *Towards Data Science*, October 21, 2019. https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13. Access: 26 set. 2021.

BERNS, Thomas; REIGELUTH, Tyler. **Éthique de la communication et de l’information :** une initiation philosophique en contexte tecnologique avancé. Bruxelles : Éditions de l’Université de Bruxelles, 2021.

BRUNO, Fernanda. **Máquinas de ver, modos de ser:** vigilância, tecnologia e subjetividade. Porto Alegre: Sulina, 2013.

CONFALONIERI, Roberto; COBA, Ludovik; WAGNER, Benedikt; BESOLD, Tarek. **A historical perspective of explainable artificial intelligence.** *Wires Data Mining and Knowledge Discovery*, v. 11, e1391, 2021. https://wires.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/widm.1391. Access: 19 ago. 2021.

CORTIZ, Diogo. **Inteligência artificial: conceitos fundamentais.** In: VAINZOF, Rony; GUTIERREZ, Adriei. *Inteligência artificial*: sociedade, economia e Estado. São Paulo: Thomson Reuters, p. 45-60, 2021.

DOŠILOVIĆ, Filip; BRČIĆ, Mario; HLUPIĆ, Nikica. **Explainable artificial intelligence: a survey.** *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, p. 210-215, 2018. https://ieeexplore.ieee.org/abstract/document/8400040. Access: 19 ago. 2021.

GHAHRAMANI, Zoubin. **Unsupervised learning.** September 16, 2004. http://datajobstest.com/data-science-repo/Unsupervised-Learning-Guide-[Zoubin-Ghahramani].pdf. Access: 20 set. 2021.

GUNNING, David; STEFIK, Mark; CHOI, Jaesik; MILLER, Timothy; STUMPF, Simone; YANG, Guang-Zhong. **XAI – Explainable artificial intelligence**. *Science Robotics*, v. 4, n. 37, 2019. https://robotics.sciencemag.org/content/4/37/eaay7120/tab-article-info. Access: 15 ago. 2021.

FLORIDI, Luciano; TADDEO, Mariarosaria. How AI can be a force for good. *Science*, v. 361 n. 6404, p. 751-752, 2018. https://www.science.org/doi/10.1126/science.aat5991. Access: 26 set. 2021.

KELION, Leo. **Apple's "sexist" credit card investigated by US regulator.** *BBC News*, November 11th 2019. https://www.bbc.com/news/business-50365609. Access: 12 ago. 2021.

LUNDBERG, Scott; LEE, Su-In. **A unified approach to interpreting model predictions**. *arXiv*:1705.07874, November 25, 2017. https://arxiv.org/abs/1705.07874. Access: 26 set. 2021.

MOLNAR, Christoph. **Interpretable machine learning**: a Guide for Making Black Box Models Explainable. Leanpub, 2021. https://christophm.github.io/interpretable-ml-book/index.html. Access: 13 ago. 2021.

MÜLLER, Klaus-Robert; SAMEK, Wojciech. **Towards explainable artificial intelligence.** *arXiv*:1909.12072v1, September 26, 2019. https://arxiv.org/abs/1909.12072. Access: August 15 2021.

MUELLER, Shane; HOFFMAN, Robert; CLANCEY, William; EMREY, Abigail; KLEIN, Gary. **Explanation in Human-AI Systems:** a literature meta-review synopsis of key ideas and publications and bibliography for explainable AI. *DARPA XAI Literature Review*, february 2019. https://arxiv.org/abs/1902.01876. Access: August 14, 2021.

NAJIBI, Alex. **Racial discrimination in face recognition technology.** *Harvard Online: Science Policy and Social Justice*, October 24, 2020. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/. Access: September 26, 2021.

NATARAJAN, Sridhar; NASIRIPOUR, Shahien. **Viral Tweet About Apple Card Leads to Goldman Sachs Probe**. *Bloomberg*, November 9, 2019. https://www.bloomberg.com/news/articles/2019-11-09/viral-tweet-about-apple-card-leads-to-probe-into-goldman-sachs. Access: September 26, 2021.

NUNES, Dierle; ANDRADE, Otávio Morato de. **A explicabilidade da inteligência artificial e o devido processo tecnológico.** *Conjur*, São Paulo, July 7, 2021.

https://www.conjur.com.br/2021-jul-07/opiniao-explicabilidade-ia-devido-processo-tecnologico. Access: September 26, 2021.

NUNES, Dierle José Coelho; ANDRADE, Otávio Morato de. **O uso da inteligência artificial explicável enquanto ferramenta para compreender decisões automatizadas:** possível caminho para aumentar a legitimidade e confiabilidade dos modelos algorítmicos? Revista Eletrônica do Curso de Direito da UFSM, v. 18(1), p. e69329, Santa Maria, 2023. Disponível: https://periodicos.ufsm.br/revistadireito/article/view/69329. Access: September 5, 2021.

NUNES, Dierle; MARQUES, Ana Luiza. **Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas.** *Revista de Processo*, v. 285, p. 421-447, November 2018. https://www.academia.edu/37764508/INTELIG%C3%8ANCIA_ARTIFICIAL_E_DIREITO_PROCESSUAL_VIESES_ALGOR%C3%8DTMICOS_E_OS_RISCOS_DE_ATRIBUI%C3%87%C3%83O_DE_FUN%C3%87%C3%83O_DECIS%C3%93RIA_%C3%80S_M%C3%81QUINAS_Artificial_intelligence_and_procedural_law_algorithmic_bias_and_the_risks_of_assignment_of_decision_making_function_to_machines. Access: September 26, 2021.

RAMOS, Oscar Garcia. **"Black box":** there's no way to determine how the algorithm came to your decision. *Oscar G. Ramos Blog*, May 27, 2020. https://www.oscargarciaramos.com/blog/9gfzdns1lmwlz58k4w2yxvzjukxp22. Access: September 26, 2021.

RIBEIRO, Marco Túlio; SINGH Sameer; GUESTRIN, Carlos. **"Why should I trust you?":** explaining the predictions of any classifier. *arXiv*:1602.04938, February 16, 2016. https://arxiv.org/abs/1602.04938. Access: September 26, 2021.

ROSSETTI, Regina; ANGELUCI, Alan. **Ética algorítmica:** questões e desafios éticos do avanço tecnológico da sociedade da informação. *Galáxia*, n. 46, p. 1-18, 2021. http://dx.doi.org/10.1590/1982-2553202150301. Access: September 26, 2021.

ROUVROY, Antoinette; BERNS, Thomas. **Governamentalidade algorítmica e perspectivas de emancipação:** o díspar como condição de individuação pela relação? *Revista Eco Pós*, v. 18, n. 2, p. 35-56, 2015. https://revistaecopos.eco.ufrj.br/eco_pos/article/view/2662. Access: September 20, 2021.

SAKO, Mari; PARNHAM, Richard. **Technology and innovation in legal services:** final report for the solicitors regulation authority. University of Oxford, 2021. https://www.sra.org.uk/globalassets/documents/sra/research/full-report-technology-and-innovation-in-legal-services.pdf?version=4a1bfe. Access: September 16, 2021.

SERENGIL, Sefik. **How SHAP can keep you from black box AI.** *Blog Sefik Ilkin Serengil*, July 1st, 2019. https://sefiks.com/2019/07/01/how-shap-can-keep-you-from-black-box-ai. Access: September 10, 2021.

SIMONITE, Tom. **When it comes to Gorillas, Google Photos remains blind.** *Wired*, November 1st, 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind. Access: September 10, 2021.

SUNSTEIN, Cass. **Republic 2.0.** Princeton: Princeton University Press, 2009.

SURDEN, Harry. **Machine learning and law**. *Washington Law Review*, v. 89, n. 1, 2014. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2417415. Access: September 26, 2021.

VILLANI, Cédric. **For a meaningful artificial intelligence:** towards a French and European strategy. A parliamentary mission from 8th september 2017 to 8th march 2018. Paris, 2019. https://books.google.com.br/books?id=9cVUDwAAQBAJ&lpg=PP1&hl=pt-BR&pg=PP1#v=onepage&q&f=false. Access: September 20, 2021.

WELLS, Lindsay; BEDNARZ, Tomasz. **Explainable AI and reinforcement learning:** a systematic review of current approaches and trends. *Front Artificial Intelligence*, May 20, 2021. https://www.frontiersin.org/articles/10.3389/frai.2021.550030/full. Access: September 10, 2021.

ZUBOFF, Shoshana. **A era do capitalismo de vigilância:** a luta por um futuro humano na nova fronteira do poder. Rio de Janeiro: Intrínseca, 2020.