DEEP LEARNING AND THE RIGHT TO EXPLANATION: TECHNOLOGICAL CHALLENGES TO LEGALITY AND DUE PROCESS OF LAW IN TECHNOLOGY

APRENDIZADO PROFUNDO E O DIREITO À EXPLICAÇÃO: DESAFIOS TECNOLÓGICOS À LEGALIDADE E AO DEVIDO PROCESSO LEGAL

Mateus de Oliveira Fornasier

Professor do Programa de Pós-Graduação Stricto Sensu (Mestrado e Doutorado) em Direito da Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI).
Doutor em Direito pela Universidade do Vale do Rio dos Sinos (UNISINOS), com Pós-Doutorado pela University of Westminster (Reino Unido).
E-mail: mateus.fornasier@gmail.com

Recebido em: 30/06/2021
Aprovado em: 18/03/2022

RESUMO: This article studies the right to explainability, which is extremely important in times of fast technological evolution and use of deep learning for the most varied decision-making procedures based on personal data. Its main hypothesis is that the right to explanation is totally linked to the due process of Law and legality, being a safeguard for those who need to contest automatic decisions taken by algorithms, whether in judicial contexts, in general Public Administration contexts, or even in private entrepreneurial contexts.. Through hypothetical-deductive procedure method, qualitative and transdisciplinary approach, and bibliographic review technique, it was concluded that opacity, characteristic of the most complex systems of deep learning, can impair access to justice, due process legal and contradictory. In addition, it is important to develop strategies to overcome opacity through the work of experts, mainly (but not only). Finally, Brazilian LGPD provides for the right to explanation, but the lack of clarity in its text demands that the Judiciary and researchers also make efforts to better build its regulation.

Palavras-chave: Explainability. Opacity. Artificial Intelligence. LGPD. GDPR.

ABSTRACT: Este artigo estuda o direito à explicabilidade, importantíssimo em tempos de rápida evolução tecnológica e uso de aprendizado profundo para os mais variados procedimentos decisórios a partir de dados pessoais. Sua hipótese principal é de que o direito à explicação está totalmente ligado ao devido processo legal e à legalidade, sendo uma salvaguarda para quem necessita contestar decisões automáticas tomadas por algoritmos, seja em contextos judiciais, na Administração Publica em geral, ou mesmo em contextos empresariais privados. Valendo-se do método de procedimento hipotético-dedutivo, da abordagem qualitativa e transdisciplinar e da técnica de revisão bibliográfica, concluiu-se que a opacidade algorítmica, característica dos mais complexos sistemas de aprendizado profundo, pode prejudicar o acesso à justiça, o devido processo legal e o contraditório. Ademais, é importante desenvolver estratégias de superação da opacidade mediante o trabalho de experts, principalmente (mas não unicamente). Por fim, a LGPD brasileira

prevê o direito à explicação, porém, a falta de clareza da sua redação demanda que também o Judiciário e pesquisadores empreendam esforços para melhor construir sua regulação.

Keywords: Explicabilidade. Opacidade. Inteligência artificial. LGPD. GDPR.

SUMÁRIO: Introduction. 1 Opacity (or the "black box" problem).2 From explainability to Responsible AI. 3 The right to explanation — in GDPR and in LGPD. Conclusion. References.

INTRODUCTION

Explainability in artificial intelligence (AI) is fundamental for solving legal issues arising with the increased frequency of the use of AI systems, especially with regard to the analysis of compliance with legislation — and an urgent need in situations of attribution of responsibility for failure of the system.[1] It is also essential for detecting flaws in the algorithmic model and discriminatory biases in the data — enabling learning from the system, as well as its verification and improvement.

But there are opposite positions, depending on the application context. Robbins,[2] for example, considers that there should be no requirement for explanation in tasks in which the capacity of deep learning algorithms surpass that of humans (for example, in the diagnosis of tumor malignancy from images) — especially because when performed by humans, such a task would not need to be explained (doctors do not need to explain their diagnostic techniques exhaustively, for example). The biggest problem would not be the explanation of the decision itself, but in the discriminatory bias that the algorithm might assume due to its initial programming or to the data with which it was trained (having been programmed or trained with data derived from judgments originated with racism, sexism, etc.). It is in situations where the prejudice of the decision maker can affect the rights of the subjects over who decides something that should be required to be explained, therefore.

The High Level Expert Group on AI created by the European Commission[3] considers that explainability is fundamental to maintaining users' confidence in AI. AI processes, therefore, must be transparent and explainable to those who are directly and indirectly affected by them, otherwise their decisions are indisputable. Explainability therefore requires that AI decisions are humanly understandable and traceable, which may make it necessary to choose between improving the explicability of a system (reducing its accuracy) or increasing its accuracy (at the expense of explainability). Whenever an AI system has a significant impact on people's lives, it must be possible to demand an adequate explanation of the process, which must be timely and adapted to the specialized knowledge of the interested parties (lay people, public policy makers, lawyers, researchers, etc.). In short, it is important for explainability:[4] (I) it may be imposed on different opportunities (in the process in general and in the final decision); (II) in the current evolutionary moment of technology, it is common to reason that the precision of the AI system and its explainability are inversely proportional; (III) explainability obligations must be based on risk and

[1] SAMEK, Wojciech; WIEGAND, Thomas; MÜLLER, Klaus-Robert. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. **arXiv preprint arXiv:1708.08296**, 2017. Available at: https://arxiv.org/abs/1708.08296. Access in: 15 jul. 2020.

[2] ROBBINS, Scott. A Misdirected Principle with a Catch: Explicability for AI. **Minds and Machines**, v. 29, n. 4, p. 495-514, 2019. DOI: https://doi.org/10.1007/s11023-019-09509-3, p. 511-512.

[3] EUROPEAN COMMISION. High-Level Expert Group on Artificial Intelligence. **Ethics Guidelines for Trustworthy AI.** Brussels: European Commission, 2019. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation. Access in: 15 jul. 2020, p. 13-18.

[4] DE STREEL, Alexandre et al. Explaining the Black Box: when Law controls AI. Brussels: Centre on Regulation in Europe (CERRE), 2020. Available at: http://www.crid.be/pdf/public/8578.pdf. Access in: 15 jul. 2020.

depend on the impacts that the algorithmic decision has on users' lives; (IV) explainability must be adapted to the level of technical understanding of the users.

That is an extremely relevant research topic to understand fundamental rights concerning to contradictory and wide-ranging defense, access to justice and due process in the context of new technologies that are increasingly being used for the most diverse types of decision-making processes based on personal data. Procedures related to medical diagnoses, risk analysis for granting credit (in civil scope) or evaluation of the possibility of committing illegal acts or recidivism (in criminal scope), use of algorithms by the Public Administration and the Judiciary, are just some significant examples of situations in which the need for explanation could be glimpsed, so that the *modus decidendi* of judges, administrators (public and private) and companies can be challenged and examined in the light of the Law, when citizens may be harmed in their dignity in digital scope (databases, internet of things, clouds, etc.).

The problem that guided this research can be expressed in the following question: how could the right to explanation be expressed? Its main hypothesis is that the right to explanation is totally linked to the due process of Law and legality, being a safeguard for those who need to contest automatic decisions taken by algorithms, whether in judicial contexts, in general Public Administration contexts, or even in private entrepreneurial contexts.

Done through hypothetical-deductive procedure method, qualitative and transdisciplinary approach and literature review, the general objective of this article is to study the right to explanation in decision-making procedures that use AI, especially with regard to deep learning. In order to ensure that this objective is operationalized, the development of the article was divided into three sections, each one corresponding to a specific objective of the research. In this sense, the first section aims to understand the algorithmic opacity of decisions and the need for their transparency, the main issues related to explainability. The second section, on the other hand, attempts to relate explainability to the development of responsible AI, a more comprehensive and significant notion for the standardization of explainable AI (XAI). Finally, the third section attempts to investigate the meaning (and existence) of the right to explanation of algorithms in the Brazilian legal system, from a compared perspective to European Union's GDPR.

1 OPACITY (OR THE "BLACK BOX" PROBLEM)

An atavism in a certain degree, when one is in relation to technology, is relatively normal, since technology (and the companies connected to it) is generally presented as "black boxes", closed and opaque, that interfere in human lives. There are even scholars who ignore the digital transformation; others (usually the most innovative ones) try to fit the black boxes in the world as known, "squeezing" them into existing ways of thinking and producing. Along this path, they argue about granting rights to robots, try to justify the regulation of technology, etc. But the worldviews in which technological innovations are been fit into are ending and, in parallel, digital transformation becomes ubiquitous, creating positive expectations — in the form of "opportunities", usually — and negative — to those who feel threatened by technological changes.[5]

The most correct attitude in this regard is the adoption of a transdisciplinary view (as opposed to the unilateral, disciplinary one), in order to understand the internal workings of complex automated systems and their applications, thus orienting the technology in favor of the advantages for mankind. Concentrating on the design of the technology and continuously reflecting on its functioning is essential for that. But such understanding is only possible if education is radically redesigned, focusing on the basic understanding of technology, moving away from knowledge based exclusively on the past, introducing new skills in educational programs in general.

[5] FENWICK, Mark; VERMEULEN, Erik PM. It Is Time for Regulators to Open the 'Black Box' of Technology. **Lex Research Topics in Corporate Law & Economics Working Paper**, n. 2019-2, p. 1-17, 2019. Available at: https://ssrn.com/abstract=3379205. Access in: 15 jul. 2020, p. 12-13.

Including, excluding and classifying are attitudes that constitute a new power, which guarantees what public impressions will be permanent or not. Because of that, services related to search engines (provided by companies like Google) have become essential for advertisers and users. They make profound inroads into cultural, economic and political spheres of influence previously dominated by the traditional press. But their mastery is so complete, and their technology so complex, that pressurizes transparency and trust that held traditional media accountable to the public — and people have long since ceased to know about those services, to which they give their lives' data.[6]

The "box" is "black" because of the human inability to understand decision-making processes of AI, or to minimally predict its decisions. Technologies that use AI can be considered, opaque, unintelligible, "black boxes", because they are based on machine learning algorithms that internalize information in an inaudible or incomprehensible way by humans. Such an opacity can result from two main causes: the structural complexity of the algorithm — which occurs in deep neural networks, consisting of thousands of artificial neurons working together diffusely to solve problems; or the dimensionality of AI, that is, because it may be using a learning algorithm based on geometric relationships that humans cannot visualize, as with support vector machines.[7]

(I) *Structural complexity:* deep neural networks are based on mathematical models called artificial neurons — which, despite the name, do not simulate, on a computer, biological neurons, but aim to achieve the same ability to learn from the experience that such cells work with. Training methods for those networks have been developed since the 1980s, and the ability to connect layers of neural networks produces surprising results. This connection capability, in which several layers of interconnected neurons are used to find patterns throughout data, or to make logical connections between data, has become known as Deep Neural Networks — a technology that is used even for cancer detection more accurately than experienced doctors.

In networks like that, thousands of neurons work together to decide in a complex way, since each neuron does not work alone for a well-defined part of the task, and often what is encoded by such networks is unintelligible to humans. Thus, a kind of "machine intuition" is developed, learned from trial and error — which, however much they may be described and detailed in stages, such descriptions do not allow full explanations of the process from the resulting information. They are similar to learning to ride a bicycle: as logically explainable as the process is, the real learning of balance is intuitive, resulting of trial and error, being that such explanation would never be enough for a neophyte.

(II) *Dimensionality*: machine learning algorithms often decide by observing many variables at once and finding geometric patterns among the variables that humans could not visualize. This is what happens with support vector machines ("SVMs"). An example of the principle underlying SVM can be given below.

Suppose an SVM has to determine the fittest individuals to form a basketball team from classifying them by the height and weight of many individuals and, from there, determine whether an individual is would be a good athlete or not. If each person's height and weight are written in a two-dimensional graph, it is possible to draw a dividing line from the data usable to predict something. If a height / weight combination is found on one side of the line, the information is expected to belong to a fit player; otherwise, he/she is not. There are several ways to establish this dividing line, but some of them will clearly be the best one to predict (due to various statistical, pragmatic factors, etc.).

[6] PASQUALE, Frank. **The Black Box Society:** the secret algorithms that control money and information. Cambridge; London: Harvard University Press, 2015, p. 61.

[7] BATHAEE, Yavar. The Artificial Intelligence Black Box and the Failure of Intent and Causation. **Harvard Journal of Law and Technology**, v. 31, n. 2, p. 889-938, 2018. Available at: https://jolt.law.harvard.edu/volumes/volume-31. Access in: 15 jul. 2020, p. 901-905.

The dividing parameter is linear when there are only two variables supplying the model. When there are three, instead of a line (e.g., the number of points the individual marked in the last season), there will be a plan. Thus, successively, if the mathematical model is supplied with seventeen, a thousand, millions of variables, it will be humanly impossible to visualize the dividing parameter, because the human mind cannot process high dimensionalities. In addition, not all SVMs use non-linear divisions. An AI that uses an SVM to process dozens or thousands of variables would therefore be a black box due to the dimensionality of the mathematical model.[8]

In addition, black boxes can also be divided into two other types, due to their ability to understand from reverse engineering processes applied after obtaining results: strong and weak black boxes. Thus, strong black boxes do not allow understanding after reverse engineering, and are related to AIs that decide in ways that are totally opaque to humans because they make it impossible to determine (a) how AI arrived at a result; (b) what information is decisive for such result; or because (c) they classify the processed variables by the AI in order of their importance. Those are black boxes that cannot be analyzed by reverse engineering AI results. In contrast, weak black boxes, although they are also opaque to humans, can be reverse engineered, allowing the importance of the variables considered to be classified (and thus, subsequently, to predict in a limited way how the model will decide).

Adadi and Berrada[9] provide a more succinct explanation as to the levels of opacity - preferring to use the terms "black box", "gray box" and "white box" as to the various levels of closure of the internal essence of a content. Thus, while a black box does not reveal anything about its content and functioning, the inside of a white box is completely exposed to the user. And in between, several levels of gray boxes might exist, depending on how much detail is available. Commercially, the concept of “black box” has been explored by technology companies, generally in relation to their efforts to protect intellectual property and maintain competitiveness. As for AI, however, the difficulty of the system in providing an adequate explanation for the way through which it came to an answer is called “black box problem”. And Rai[10] puts yet another type of differentiation between machine learning algorithms, as to their explainability:

(I) *Inherently interpretable algorithms*: machine learning algorithms (decision trees, Bayesian classifiers, additive models and spare linear models) are inherently interpretable when the components of the mathematical model can be directly inspected in order to understand the model's predictions. Those algorithms use a relatively limited number of internal components (paths, rules, resources, etc.), but they are traceable and transparent as to their decision-making processes.

(II) *Deep learning algorithms*: algorithms that, for the sake of precision, sacrifice transparency and explainability. Such algorithms are currently employed in consumer behavior prediction applications based on high-dimensional inputs (in pixels, generally), speech and image recognition (faces, mostly), and natural language processing. In those cases, the mathematical model learns the important resources, being the programmer not required to design it with the relevant resources. As it involves pixel-level inputs and complex connections between layers of the network of artificial neurons that produce highly nonlinear associations, the model is inherently incomprehensible to humans.

Currently, however, significant advances in post-hoc interpretability techniques have brought black box models closer to simpler interpretable models, which can be inspected to explain black box models. When such techniques transform black box models into transparent models, they

[8] BATHAEE, Yavar. The Artificial Intelligence Black Box and the Failure of Intent and Causation. **Harvard Journal of Law and Technology**, v. 31, n. 2, p. 889-938, 2018. Available at: https://jolt.law.harvard.edu/volumes/volume-31. Access in: 15 jul. 2020, p. 901-905.

[9] ADADI, Amina; BERRADA, Mohammed. Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, v. 6, p. 52138-52160, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2870052, p. 52141.

[10] RAI, Arun. Explainable AI: from black box to glass box. **Journal of the Academy of Marketing Science**, v. 48, n. 1, p. 137–141, 2020. DOI: https://doi.org/10.1007/s11747-019-00710-5, p. 138.

are called explainable AI (XAI), and offer a way to achieve predictive accuracy and the possibility of interpretation with AI applications.

Current debate about AI transparency have emphasized the epistemic constraints that AI-based decision systems pose to outside observers — being that lay citizens in general, public servants and even many programmers are unable to understand even the simplest machine learning algorithm.[11] But those difficulties do not obliterate the fact that the logic behind AI is generally well-understood by experts, much better than many other complex phenomena (climate change, nanotechnology or financial markets, for example). In addition, software companies have recently developed tools to test complex systems — including AI-based systems for that purpose. A team of auditors equipped with such devices could then check the programming code, reconstruct algorithms, analyze training processes, evaluate training and results databases, feed them fictitious data, etc. Thus, although the current understanding of AI is quite advanced in theory, the literature on its transparency points to three serious challenges faced by AI audits in practice:

(I) The greater the volume, variety and speed of data processing, the more difficult it becomes to understand and predict the behavior of a system or to reconstruct its computed correlations. But this problem is not specific to AI-based systems. Today, even fully determinable programs process an almost unlimited number of objects in a very short time. And technology often also provides a solution to the problem, with at least parts of the audit being automated. The accessibility of data-based systems, therefore, does not depend on the "greatness" of the data, but on the way it is processed by the algorithms.

(II) in some AI-based algorithms, it is practically impossible to retroactively connect a specific input to a specific output and vice versa. All AI systems using machine learning algorithms are often considered to elude such causal explanations. But such learning comprises a wide variety of techniques — from linear regression models on SVMs to decision tree algorithms for different types of neural networks. The difficulty of establishing an ex post causal link between a specific input and output of data is very different, depending on the technique being used. Although some algorithms allow causal explanations, others do not. Even so, the difficulty of identifying causal relationships in neural networks does not render AI audits on them or on other forms of AI useless, as opacity does not affect the possibility of collecting information about the system and its operations. The search for explanations for opaque phenomena from gathering data and testing hypotheses is the standard way of producing knowledge and understanding data in science and society.

(III) Many AI systems constantly readjust the importance of the variables they use, depending on the effects of their algorithms on users. Thus, all their training operations update the system and, in such a dynamic architecture, the possible explanations are only valid for a brief moment in time. Then, even with the transparency in the programming code of an algorithm, in its complete set of training data and in its test data, only instant and specific reports would be possible. But even this dynamism does not preclude audits, as a single instant can still contain valuable information.

The solution to the lack of access to information alone involves two proposals, at least: the fiduciary model and the infomedial model, which can operate complementarily and jointly.[12] Both suggest that individuals must relate in a principal-agent way to another entity to achieve a division of labor that can produce results. The individual (principal) would delegate, thus, to the trustee or the infomediary (agent), the power to choose the way to arrive at a scenario where more

[11] WISCHMEYER, Thomas. Artificial Intelligence and Transparency: Opening the Black Box In: WISCHMEYER, Thomas; RADEMACHER, Timo (eds.). **Regulating Artificial Intelligence**. Cham: Springer, 2020, p. 75-102. DOI: https://doi.org/10.1007/978-3-030-32361-5, p. 80-82.

[12] OBAR, Jonathan A. Sunlight alone is not a disinfectant/ Consent and the futility of opening Big Data black boxes (without assistance). **Big Data & Society**, v. 7, n. 1, p. 1–5, 2020. DOI: https://doi.org/10.1177/2053951720935615, p. 4-5.

simplified, perceptible options and in less technocratic and more individualized ways, which allow users to make decisions based on their own experience and interest.

(I) *Fiduciary model:* the agent who receives the delegation (social networks, banks, e-mail provider, cloud storage, etc.) is also the administrator who will operate in the individual's best interest. The trustee not only protects personal information in general, therefore, because where complications arise (such as in the consent processes, for example), it builds a scenario that predetermines how the best interest of the user is served, and ensures that communication with the principal (individual) help to achieve this result. Thus, in the context of using applications containing algorithms, individuals must be able to not only consent, but also to evaluate opportunities for consent.

(II) *Infomediary model*: infomediaries, agents specialized in protecting privacy and reputation, would operate between the individual (principal) and the entities involved in the practice of data (fiduciary and others). The delegation to an infomediary would involve both for-profit and non-profit entities to ensure principal-agent relationships that could provide meaningful consent scenarios. Four potential strategies can be used, according to which infomediaries may i) help to overcome information asymmetries (between suppliers and consumers, for example); ii) help to overcome time constraints; iii) providing the principal with opportunities of less extensive choices; and iv) ensure that such choices, located in a smaller range, lead to clear results.

Intelligent systems that operate under normative systems generally have standards monitoring structures, which do not restrict the behavior of the monitored system, but record compliance or violation and, as a result, take actions resulting from the system's behavior — thus, monitoring is an observation mechanism, not an active participant.[13] When a decision is challenged, the decision itself and the procedure leading to it are checked against specific rules for the system, analyzing precisely the existence of violations.

Those mechanisms may also be programmed to enforce compliance to norms. But their adoption would greatly slow down the evolution of computing technologies, and would be very economically demanding, because with each decision and action, the system would have to verify whether a standard applies and, then, how to act accordingly. Obviously, some entities may be willing to pay such costs, allowing the implementation of such an approach. But it may be more useful to just monitor and bear the possible penalties later.

2 FROM EXPLAINABILITY TO RESPONSIBLE AI

Although XAI research focuses on developing methods that make machine learning indicators more interpretable and explainable, there are no unambiguous definitions of the interpretation and explanation notions — which are mistakenly treated as synonyms.[14] Not all black boxes are equally interpretable — which means that some types are more susceptible to human understanding than others. In fact, there are positions according to which the best performing black boxes are the most difficult to interpret ones.

Some examples might help to understand the difference between both notions. When an agent (human or software) assigns a subjective meaning to an object, the result of that assignment is understood to be an "interpretation". In that sense, an object may be considered interpretable to an agent if it is easy for him/her to interpret the object — where "easy" means requiring low computational or cognitive effort to understand (for example, road signs are intuitive and quickly

[13] TUBELLA, Andrea Aler et al. Contestable Black-Boxes. **Arxiv**, 2020. Available at: https://arxiv.org/abs/2006.05133. Access in: 15 jul. 2020.

[14] CIATTO, Giovanni et al. **Agent-Based Explanations in AI: Towards an Abstract Framework**, 2020. Available at: https://www.researchgate.net/profile/Davide_Calvaresi/publication/341509975_Agent-Based_Explanations_in_AI_Towards_an_Abstract_Framework/links/5ec5020b299bf1c09acc036d/Agent-Based-Explanations-in-AI-Towards-an-Abstract-Framework. Access in: 15 jul. 2020.

interpretable because they contain symbols, not programming codes). Thus, when a human agent compares a simple decision tree to a deep neural network, being both algorithms trained with the same examples to solve the same types of problems (and with similar predicative results), that person will consider the decision tree more interpretable, because it brings semantically more understandable information than that of complex neural networks.

While"explanation" is the activity of producing a more interpretable object for an agent, from a less interpretable one. However, it must be considered that the meanings of the expressions "more interpretable" and "less interpretable" still depend very much on the agent (human or artificial), subjectively.

But the use of machine learning for decisions within the scope of Public Administration might increase efficiency without disregarding the Law, if the data learned by the algorithm are sufficient and respect the correct legal foundations. Learning from the history of cases, and reproducing those reasons in new situations, with connected legal results and descriptions of facts, is not very different from what humans would do.[15] The difference between humans and algorithms here is that and algorithm tends to be more rigorous than humans, who, in turn, respond more organically to past cases because they have broader horizons of understanding, contextualizing their tasks more richly and, therefore, adjusting their decisions to a broader spectrum of facts — including those hidden in the literal interpretation of the norm. It is precisely this phenomenon that explains why new practices can develop under the same law. But algorithms operate without wide context, relating only to explicit texts. Thus, having human beings in the legal-administrative circuit relativizes strict compliance with rules, increasing possibilities for adapting the interpretation of the norm to more diverse contexts (new or unforeseen situations by the legislator when editing the norm, for example).

Such limited contextualization of algorithmic reasoning is problematic if all new decisions are made based on an algorithm that reproduces only the past, and those past decision models are subjected to little (if any) change by a human collaborator. This is due to the fact that, when the algorithm's initial stage of learning is finished, and it starts to be used to produce decision models, new decisions will be based on those models. With that, one of two different situations may then occur (neither of which is ideal for maintaining an updated algorithmic support system):
(I) *new decisions are fed back into the machine learning stage*: thus, a feedback loop, in which the algorithm receives its own decisions, is created;
(II) *the machine learning stage is blocked after the initial training phase*: and with that, every new decision is based on what the algorithm captured from the original training set.

The performance of an AI-based decision system may be considered good (measured by its ability to produce decision models of good legal quality) — if it is constantly maintained by new information, which can be done in several ways, depending on how the algorithmic system is implemented in the administrative body and its procedures for issuing decisions. Thus, collaborative models between humans and AI must be developed, enabling technology companies to develop systems that improve the effectiveness and quality of Public Administration. Collaboration provides working conditions in which AI, in the long term, guarantees the detection of hidden biases and other bureaucratic deficiencies. Furthermore, collaboration can dispel black box fear, as it allows for the review of both pure human decisions and those of machines — maintaining, thus, the human standard of decisions as well.

Due to the ubiquity of technological systems in human daily life, their responsibility has been demanded by the public — and, with that, explanations about how algorithmic systems decide come in such a wake, in the same way that information is requested in the most varied areas.

[15] OLSEN, Henrik Palmer et al. What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration. **Legal Studies Research Paper Series**, n. 2019-84, p. 1-27, 2019. Available at: http://jura.ku.dk/icourts/working-papers/. Access in: 15 jul. 2020.

Nicholas,[16] however, understands that explainability in decision-making systems is not just a technical issue: algorithm developers will only optimize explainability when they have reason to do so. In this sense, currently existing laws encourage explanation to some extent in selected areas - but other legal areas, such as Intellectual Property Law, encourage the opposite. The explanation is a starting point for responsibility, not a final destination. Explanations describe how an algorithm decides, but it does not provide internal means of correction when explanations contradict larger social values. This is not a technical problem and, therefore, lawyers and politicians must also discuss algorithmic explainability to decide how and when these explanations can address socio-legal damage.

It is also important to consider that a decision process depends on much more than technical elements and socio-legal rationality: intuition is fundamental in normative reasoning — which can even be helped by non-intuitive reasoning, but not replaced.[17] It is a basic component of human reasoning and, therefore, it would be foreign to the human mind to use mathematical models in algorithms that do not admit any intuitive explanation. And the desire to regulate machine learning seems natural, so that the results of their decisions are compatible with intuition — including that the AI regulations approach their explainability.

Inexplainability has been legally and technically approached for the usefulness of explanations — mainly because of its inherent value (because respect for the other, being transparent to him, would be part of the very understanding of what the dignity of others is), enabling action (here that explainability allows understanding the decision-making process in full) and the evaluation of the fundamentals of decisions. But Law has more substantial concerns about explainability, and believing that it provides a path to normative evaluation by itself, and which is insufficient if intuition is not recognized as something important for the decision. Humans and mathematical models, in their basic day-to-day operations in Public Administration (and in the Judiciary), work in a similar way: applying similar precedents to the present case. But there is a fundamental difference between both: the richer contextualization capacity and the broader horizon of understanding of the human, unlike the machine, which applies models rigorously. Therefore, explainability alone will not lead to satisfactory respect for the use of AI in the public decision-making context in relation to due legal process, the fundamental right to information, in short, to what is understood by the Democratic Rule of Law and human dignity: it is necessary the collaborative combination between the human and the artificial elements in a certain margin, so that the learning of the latter is also adapted to the context of the former.

The possibility that AI will present innovations that can be used for improvements in all fields of industry, commerce, services (including the publicly provided ones), and knowledge, demands the solution of the explainability problem. The paradigms underlying this problem fall within the field of developing explainable AI. In this sense, Arrieta et al.[18] developed a definition of XAI focusing on the public for which explainability is sought. Based on this definition, they proposed the concept of Responsible Artificial Intelligence. XAI would be responsible when based on a series of principles in their view: justice, privacy, responsibility, ethics, transparency, and security — which could be translated into meeting a complex of technical, normative and epistemic requirements, which could be summarized in the following list:

(I) *Critics to the reasoning about Explanability and Performance*: it is necessary to overcome the idea that explainability impairs the performance of deep learning. In fact, more complex models

[16] NICHOLAS, Gabriel. Explaining Algorithmic Decisions: A Technical Primer. **Georgetown Law Technology Review**, Forthcoming, 2020. Available at: https://ssrn.com/abstract=3523456. Access in: 15 jul. 2020.

[17] SELBST, Andrew D.; BAROCAS, Solon. The Intuitive Appeal of Explainable Machines. **Fordham Law Review**, v. 87, n. 3, p. 1085-1139, 2018. Available at: http://fordhamlawreview.org/issues/the-intuitive-appeal-of-explainable-machines/. Access in: 15 jul. 2020, p. 1138-1139.

[18] ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. **Information Fusion**, v. 58, p. 82-115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

provide more flexibility to mathematical models, allowing them to develop more complex functions — but more complex systems are not inherently more accurate (as accuracy is totally context-depending). And even when performance improvement depends on complexity, the emergence of more sophisticated interpretation methods might mitigate the losses of accuracy eventually resulting from interpretability.

(II) *Definition of XAI concepts and metrics*: a unified concept of explainability must be implemented, thus providing common ground for the development of new techniques and methods. Arrieta et al.[19] define explainability as the ability of a model to make its operation understandable to the public. And a common metric, established in a transdisciplinary way, would allow meaningful measurements on how explainable a model can be considered, in relation to different contexts.

(III) *Achievement of explainability in deep learning*: the intelligibility of the functioning of systems does not need to be at the same level of understanding for AI development specialists, public policy makers or totally lay users. Such a subjectivity can be reduced with inspiration from experiments in Psychology, Sociology or Cognitive Sciences to create objectively convincing explanations.

The relevant findings to be considered when creating a XAI model can be presented as follows: i) explanations are better when they not only indicate why a model decided in a certain way, but also why it decided that way and not other; ii) satisfactory explanations must translate the probabilistic and quantitative reasoning of the black boxes into causal and qualitative relationships; iii) explanations must be selective, being sufficient to focus only on the main causes of a decision-making process; iv) the use of counterfactual explanations can help the user to understand the decision of a model; v) a good explanation needs to influence the user's mental model, that is, the representation of reality external to the algorithm (using common sense reasoning, natural language, etc.); vi) an explainable model does not delegate the explanation to users, as different explanations can be deduced depending on the prior knowledge of each user.

(IV) *Obtaining security in XAI*: confidentiality in AI systems must also be ensured, including for reasons of Intellectual Property Law. The development of a mathematical model for deep learning may have demanded a lot of time and economic resources — and industrial secrecy could be compromised because of explainability, even at a minimum level.

(V) *Development of rational explanations and Critical Data Studies*: in order to transform data into a valuable asset, individuals must collaboratively engage, sharing the context in which they produce their discoveries, in which context it refers to sets of narratives about the modes through which data were processed and analyzed. XAI techniques should also be adopted because of their ability to understandably describe black boxes for scientists of Social, Human and Legal Sciences. This transdisciplinary confluence in projects related to Data Science and the search for ethical evaluation methodologies have been called Critical Data Studies, a field in which XAI can increase the exchange of information between heterogeneous audiences.

(VI) *Theoretically oriented Data Science*: a theoretically guided synergy between XAI and Data Science must be developed, by combining the principles underlying the application/context in which the data is produced. Thus, the model to be adopted must be chosen according to the type of relationship sought. Its structure must follow knowledge already mastered previously. The training approach should not allow the algorithm to enter non-plausible fields of information. And the results on the model's output should completely inform what was learned by the model, allowing to reason and combine new knowledge with what was previously known.

(VII) *XAI Implementation and Guidelines*: making AI-based models interpretable requires a great deal of transdisciplinary effort — which means examining and considering the interests, demands and requirements of all stakeholders who interact with the system to be explained (from algorithm

[19] ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. **Information Fusion**, v. 58, p. 82-115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

developers to consumers, also comprehending suppliers, lawyers and politicians). For that, it is important to consider a series of guidelines, mainly: i) contextual factors, potential impacts and specific needs of each field of application; ii) interpretable techniques should be preferred whenever possible; iii) if a black box model has been chosen, the ethical, legal and security impacts must be weighed (in particular, the responsibility for the design and implementation of the AI system); iv) interpretability must be rethought in terms of the cognitive abilities, capacities and limitations of the human individual.

(VIII) *Confidence in the results from XAI*: in several scenarios in which deep learning is used (vehicle perception, driving autonomous vehicles, automated surgery, data-based medical diagnosis, insurance risk assessment, etc.), erroneous results might be catastrophic, which requires comprehensive regulatory efforts that ensure that no decision is made solely on the basis of data processing. Ways are also needed to minimize the risk and uncertainty of damage from automated decisions. For that reason, the use of XAI to expose in which region of the input data the model is focused when producing a certain output may discriminate possible sources of uncertainty.

(IX) *XAI and Data Fusion*: the growing amount of information currently available in almost all domains requires approaches to data fusion in order to explore them simultaneously for learning. The combination of heterogeneous information is proven to improve the performance of algorithmic machine learning models in many applications (industrial forecasts, cyber-physical social systems, internet of things, etc.). Data fusion techniques should also be explored to enrich the explainability of machine learning models, therefore.

3 THE RIGHT TO EXPLANATION — IN GDPR AND IN LGPD

It has been debated, since the adoption of European Union's General Data Protection Regulation (GDPR)[20] in 2016 (but enforced only from 2018), that the “right to explanation” of AI-based decisions is legally mandatory. But there are scholars who stand by denying the existence (and even the viability) of such a right. The GDPR (in its articles 13 to 15) presents a requirement that data subjects receive significant, but adequately limited, information about the logic involved, the meaning and the consequent objective of automated decision-making systems — being that Wachter, Mittelstadt and Floridi[21] call it “the right to be informed”, and not the right to explanation itself.

Article 22 of the GDPR establishes an ambiguous and scope-limited “right to not to be subject to automated decision-making”, from which, according to some, is possible to deduct the supposed “right to explanation”, however. Thus, the authors understand that there is a lack of precision in the language of the GDPR, which obliterates the clarity and definition of the rights instituted therein, and consequently impairs the identification of safeguards against automated decision-making and, therefore, citizens under such rule are at the mercy of a flawed regulatory system. For that reason, they propose a series of legislative steps that, if implemented, could improve transparency and accountability of automated decision-making within the scope of GDPR.

The first one concerns to the meaning of the expressions "existence of significance", "expected consequences" and "logic involved", contained in article 15 (1) h, which should be clarified in some way. The others concern to article 22 and its subdivisions. The right to an explanation should be added explicitly and legally to art. 22 (3). Furthermore, article 22 should be

[20] EUROPEAN UNION. **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.** On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=PT#d1e1797-1-1. Access in: 15 jul. 2020.

[21] WACHTER; Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. **International Data Privacy Law**, v. 7, n. 2, p. 76-99, 2017. DOI: 10.1093/idpl/ipx005.

elucidated by clearly indicating prohibitions pertaining to such standard. Such a clarification should be extended to art. 22 (1), so that it is possible to indicate when decisions are based exclusively on automated processing, as well as what may be considered "legal or significant effect of the automated decision, including the creation of profiles". And also the expression "necessary for the execution or execution of a contract" in article 22 (2) a should be clarified.

As a balance to the protection given to trade secrets, the mandatory introduction of an external audit mechanism for automated decision-making, or the definition of internal audit requirements for data controllers, should be introduced in GDPR. Finally, there should be support in the legislation for further research on the feasibility of explanations of alternative accountability mechanisms.

Selbst and Powles[22] have the same opinion, in the sense that they do not interpret the GDPR as having a single and clear legal provision that establishes the "right to explanation". However, they stand that GDPR is interpretable as containing such a right. The authors read articles 13 to 15, which provide rights to "meaningful information about the logic involved" in automated decisions, as being ballasts for the right to explanation (although the expression does not appear literally in the regulation). Thus, they believe that the right to explanation should be interpreted in a functional and flexible way, and should, at the very least, allow the data subject to exercise it under the aegis of both the GDPR and human rights.

Malgieri and Comandé[23] also understand the articles 13 to 15 and 22, GDPR, as providing a right to explanation, when a systemic interpretation of such devices and their subdivisions is carried out, particularly considering that: (I) some form of minimal human intervention is necessary in automated decision making (article 22 (1)), which can also include nominal human intervention; (II) the "meaningful" expected for individuals (art. 22 (1)) may also include marketing manipulation, price discrimination, etc; (III) "meaningful information", which must be provided to data subjects on the logic, meaning and consequences of decision making (art. 15 (1) (h)) should be read as "legibility" of "architecture" and "implementation" of algorithmic processing; (IV) although the protection of trade secrets limits the right of access for data subjects, there is a general legal favor for data protection rights that should reduce the impact of the protection of trade secrets.

Furthermore, Casey, Farhangi and Vogl[24] understand that GDPR not only provides for a right to explanation, but has also been revolutionary in conferring broad enforcement powers given to European data protection competent authorities (DPA), in chapters 6 (articles 51 to 59) and 8 (articles 77 to 84) of the Regulation. Those powers concern to investigation, consultation, correction and punishment, which make DPAs the de facto authorities for interpreting the controversial "right to explanation" of the Regulation.

GDPR provides, thus, a "right to explanation" with wide-ranging legal implications for designing, prototyping, field testing and implementing automated data processing systems. The legally enshrined protections may not require transparency in the form of a complete individualized explanation, but a holistic understanding of the interpretation of DPAs reveals that the true legal power conferred by the right to explanation in the GDPR derives from its synergistic effects when combined with algorithmic audits and "data protection by design" methodologies. Consequently, the combination of audits and regulation by design will become a standard for companies that implement machine learning systems inside and outside the European Union.

---

[22] SELBST, Andrew D.; POWLES, Julia. Meaningful information and the right to explanation. **International Data Privacy Law**, v. 7, n. 4, p. 233-242, 2017. DOI: 10.1093/idpl/ipx022.

[23] MALGIERI, Gianclaudio; COMANDÉ, Giovanni. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. **International Data Privacy Law**, v. 7, n. 4, p. 243-165, 2017. DOI: 10.1093/idpl/ipx019.

[24] CASEY, Bryan; FARHANGI, Ashkon; VOGL, Roland. Rethinking explainable machines: the GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise. **Berkeley Technology Law Journal**, v. 34, n. 1, p. 143-188, 2019. DOI: https://doi.org/10.15779/Z38M32N986.

According to Kaminski,[25] GDPR establishes a system of "qualified transparency" on algorithmic decision-making, which gives laypeople one type of information, and experts and regulators, another. This means that there is, in the GDPR, an individual right to explanation — however, deeper than counterfactuals or a superficial and broad systemic overview, being associated with other transparency measures that provide regulatory and third-party supervision in decision-making algorithmic. Such transparency provisions are just one of the ways in which the GDPR's algorithmic accountability system is potentially broader, deeper and stronger than the previous EU regime.

But the system of algorithmic accountability that GDPR and its official interpretations have significant implementation obstacles — mainly limited individual access to justice; limited technical capacity of individuals and regulators; and high costs for companies and regulators. Because of its strong reliance on collaborative governance, therefore, in the absence of significant oversight by the public or third parties, GDPR may fail, leading to the capture or under-representation of individual rights.

Judges are also starting to use machine learning algorithms in criminal, administrative and civil cases — and as a result of that, Deeks[26] argues that judges should demand explanations for those algorithmic results. One of the proposals that the author presents to provide more explanability to the Judiciary is to design systems that explain how the algorithms arrive at results. Thus, Courts might contribute by developing XAI meanings and definitions for different legal contexts. Judicial reasoning that develops in a downward fashion, using case-by-case considerations of the facts to produce approved decisions, is a pragmatic way of developing rules for XAI. In addition, Courts may encourage the production of different forms of XAI that respond to different legal and public contexts. It is a way of involving more public actors in the elaboration of XAI, which has, until today, been left almost exclusively in private hands.

Good examples of that contextualization of XAI by the Courts could occur in matters such as product liability involving autonomous cars or the Internet of Things; use of algorithms by the Public Administration to evaluate teachers; cases of medical negligence against doctors who depend on medical algorithms for diagnosis; government decisions to freeze assets based on algorithmic recommendations; police approaches based on the use of "automated suspicion" algorithms; and questioning of algorithm-driven forensic tests. Those cases may have questions based on due process, on the necessary review, or on the testimonies and expertise of experts on how a particular algorithm works.

It is clear that the Legislative must also participate in the regulation of XAI, requiring and formatting its use in specific sectors, or within the government. However, any statute that regulates the use of XAI must be drafted with a high level of generality — capturing basic values, but allowing the adequacy of legal interpretation in the face of rapid technological and social changes related to AI. Furthermore, the likelihood that Parliaments and Congresses around the world will be able to act is limited, if their recent actions on complicated technological issues serve as a guide.

With regard to the Brazilian data protection legislation — Lei nº 13.709/ 2018, the General Law for the Protection of Personal Data (LGPD[27] in Portuguese abbreviation) — its article 20 assures the data subject to request a review of the exclusively automated decision that may affect his/her interests (digital profiling, for example).[28] The inclusion of this norm reveals the concern

[25] KAMINSKI, Margot E. The Right to Explanation, Explained. **Berkeley Technology Law Journal**, v. 34, n. 1, p. 189-218, 2019. DOI: https://doi.org/10.15779/Z38TD9N83H.

[26] DEEKS, Ashley. The Judicial Demand for Explainable Artificial Intelligence. **Columbia Law Review**, v. 119, n. 7, p. 1829-1850, 2019. Available at: https://columbialawreview.org/content/the-judicial-demand-for-explainable-artificial-intelligence/. Access in: 15 jul. 2020.

[27] BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados (LGPD). Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Access in: 15 jul. 2020.

[28] TEIXEIRA, Tarcisio; ARMELIN, Ruth Maria Guerreiro da Fonseca. **Lei Geral de Proteção de Dados Pessoais:** comentada artigo por artigo. Salvador: Editora JusPodivm, 2019, p. 85-87.

with the limit of influence of machine learning decisions to people's lives, considering the frequency with which data analysis is carried out in an automated way, which may lead to erroneous assumptions and, consequently, to abuse through such a decision-making process. It is, therefore, the legal provision in favor of positivizing, in the Brazilian LGPD, the right to explanation, in a similar way to that of the GDPR.[29]

However, Almeida[30] sees the rule in question as imprecise, as there is no clarity as to which decisions may be considered to be made "solely on the basis of automated treatment", which may affect "the interests of the holders", nor the degree of explanation and transparency required in such situations. In addition, the audit to be carried out by the national authority may not be effective, since the system may decide based on humanly incomprehensible circumstances. In this sense, Bioni and Luciano's[31] approach to explainability is interesting. The authors understand that explainability is not to be confused with pure and simple transparency, since the right to explanation is related to the requirement for information about the rationality of a specific decision. In other words, it is necessary to clarify what explainability is in order to make it efficient, making sense to those who request it.

It should also be noted that three bills are being processed in the Brazilian Federal Legislature: PL 21/2020,[32] in the Chamber of Deputies; and PL 5051/2019[33] — which seeks to establish principles for the use of AI in Brazil — and PL 5691/2019[34] — which seeks to institute the National AI Policy, both in the Senate. Regarding the proposal pending before the Chamber, its article 6, IV attempts to define transparency and explainability as principles for the responsible use of AI in Brazil. As for the proposals pending in the Senate, one of the fundamentals of the use of AI brings "transparency, reliability and the possibility of auditing systems" (art. 2, IV, of PL 5051/2019). There is also a duty, for AI solutions, to be "intelligible, justifiable and accessible", and also to "provide traceable decisions and without discriminatory or prejudiced bias (article 4, IV and VIII, PL 5691/2019).

It is observed the concern of the legislator to bring specific regulation for the use of AI in Brazil, and that such regulation brings with it the duties of transparency and explainability, therefore. However, it is noted that such projects, in addition to suffering from high imprecision in language, are too short and redundant, because what they bring in their text only repeats, in other words, what article 20 of the LGPD already provides. Thus, there is a high probability of archiving those proposals.

CONCLUSION

"Black box" is how opacity of a machine learning algorithm became known — that is, to the human inability to understand the logic through which an algorithm arrived at a given result.

[29] MULHOLLAND, Caitlin; FRAJHOF, Isabella Z. Inteligência Artificial e a Lei Geral de Proteção de Dados Pessoais: breves anotações sobre o direito à explicação perante a tomada de decisões por meio de *machine learning* In: FRAZÃO, Ana; MULHOLLAND, Caitlin (coord.). **Inteligência Artificial e Direito**. São Paulo: Thomson Reuters Brasil, 2019, p. 265-291; p. 271.

[30] ALMEIDA, Daniel Evangelista Vasconcelos. Direito à Explicação em Decisões Automatizadas. In: ALVES, Isabella Fonseca (org.). **Inteligência Artificial e Processo.** Belo Horizonte; São Paulo: D 'Plácido, 2020, p. 95-114; p. 100.

[31] BIONI, Bruno; LUCIANO, Maria. O princípio da precaução na regulação da inteligência artificial: seriam as leis de proteção de dados o seu portal de entrada? In: FRAZÃO, Ana; MULHOLLAND, Caitlin (coord.). **Inteligência Artificial e Direito**. São Paulo: Thomson Reuters Brasil, 2019, p. 207-231; p. 220-221.

[32] BRASIL. Câmara dos deputados. **Projeto de Lei 21/2020**. Available at: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340. Access in: 15 jul. 2020.

[33] BRASIL. Senado Federal. **Projeto de Lei 5091/2019**. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/138790. Access in: 15 jul. 2020.

[34] BRASIL. Senado Federal. **Projeto de Lei 5691/2019**. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/139586. Access in: 15 jul. 2020.

Such opacity can occur due to the complexity of the relationship between artificial neurons (being that they learn due to a kind of "machine intuition", explainable and detailed by human logic sometimes, but never completely), or to the dimensionality of parameters used by the algorithms (which establish interpretive mathematical logics according to logics that human beings also cannot fully understand, due to the limitations of the human mind itself).

Opacity of automated decision-making procedures can damage the democratic and dignitary degree of decisions, since, with respect to the principle of due legal process, public decisions must have their logical *iter*, their context and their foundations — that is, their reasons, grounding and circumstances — explained. But the possibilities for reducing AI's algorithmic opacity should not be elevated to absurd generalizations — that is, requiring that any user could fully understand the learning algorithm. It is a set of very hermetic techniques, and even computer graduates are often unable to decipher more advanced algorithms.

The regulation of algorithms for the sake of transparency must therefore focus, firstly, on the possibility of audits by experts — who, despite the inherent difficulties of the complexity of neural networks, the dimensionality of SVMs and similar systems, and the dynamics of deep learning algorithms, they can take readings (at least of important parts) of the systems to analyze their decision procedures. In addition to the importance of expert auditors, who would analyze the code available *ex post*, in order to obtain useful information regarding the use of personal data, it is important to develop other models of overcoming opacity — and in this sense, the models of fiduciary and infomediators would preventively assist lay users of deep learning algorithms. These models, which delegate technical expertise, would make important cognitive mediation for users, so that the transparency of the algorithm — which would allow one to previously analyze the consequences of its use in a useful way — translates into access to the information itself. Algorithm monitoring technologies are also useful for the concomitant assessment of machine learning operations — in an observant manner, rather than active participation, that will force the algorithms to decide according to certain standards.

But it is necessary to expand the normative concepts beyond just explainable AI as well: it must be made an element in favor of the development of responsible AI. This can be achieved if the explainable AI (XAI) meets normative principles of justice, privacy, responsibility, ethics, transparency, and security. Technically, those principles can be translated into meeting a complex of technical and normative requirements, summarized as developing a critique of the reasoning about the clash between interpretability and performance; definition of XAI concepts and metrics; development of explanatory techniques for deep learning and security at XAI (translatable here as respect for confidentiality, when necessary); development of rational explanations (usefully understandable by the most varied types of interested parties); establishment of Critical Data Studies (from a transdisciplinary approach), theoretically oriented Data Science, and normative guidelines on XAI; and development of techniques that will give greater confidence to the results from the use of XAI (including, through the development of the notions pertaining to knowledge about Data Fusion).

European GDPR has been criticized due to the lack of clarity of its text, which engenders clarification of its language, and the insertion of an explicit"right to explanation" (which cannot be directly read in GDPR. But it must be considered that its articles 13 to 15 and 22 provide a right to "meaningful information about the logic involved" — which, when read in a broader sense, allow an interpretation in the sense of a right to explanation. And a synergy between the right to explanation, the powers given to data protection authorities (arts. 51 to 59 w/ art. 77 to 84, GDPR) and the protection methodologies for the design of the algorithms, creates a true standard to be followed not only within the EU, but also by other countries as a good normative exemple.

Brazilian legislation already provides for transparency and explainability, mainly in its LGPD (article 20 and paragraphs, mainly). But the regulation of XAI must have the primary participation not only of the Legislative, but also of the Judiciary branch — which has great power

to contextualize the XAI to the cases in which its demand may occur. This is because, although the Legislative can contribute to the elaboration of statutes containing the values that support the use of XAI, such devices must be of high generality, with the power of the legislation to define very specifically highly dynamic issues (such as implications of technological evolution for Law and society).

REFERENCES

ADADI, Amina; BERRADA, Mohammed. Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access, v. 6, p. 52138-52160, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2870052.

ALMEIDA, Daniel Evangelista Vasconcelos. Direito à Explicação em Decisões Automatizadas. In: ALVES, Isabella Fonseca (org.). Inteligência Artificial e Processo. Belo Horizonte; São Paulo: D' Plácido, 2020, p. 95-114.

ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, v. 58, p. 82-115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

BATHAEE, Yavar. The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law and Technology, v. 31, n. 2, p. 889-938, 2018. Available at: https://jolt.law.harvard.edu/volumes/volume-31. Access in: 15 jul. 2020.

BIONI, Bruno; LUCIANO, Maria. O princípio da precaução na regulação da inteligência artificial: seriam as leis de proteção de dados o seu portal de entrada? In: FRAZÃO, Ana; MULHOLLAND, Caitlin (coord.). Inteligência Artificial e Direito. São Paulo: Thomson Reuters Brasil, 2019, p. 207-231.

BRASIL. Câmara dos deputados. Projeto de Lei 21/2020. Available at: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340. Access in: 15 jul. 2020.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados (LGPD). Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Access in: 15 jul. 2020.

BRASIL. Senado Federal. Projeto de Lei 5091/2019. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/138790. Access in: 15 jul. 2020.

BRASIL. Senado Federal. Projeto de Lei 5691/2019. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/139586. Access in: 15 jul. 2020.

CASEY, Bryan; FARHANGI, Ashkon; VOGL, Roland. Rethinking explainable machines: the GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise. Berkeley Technology Law Journal, v. 34, n. 1, p. 143-188, 2019. DOI: https://doi.org/10.15779/Z38M32N986.

CIATTO, Giovanni et al. Agent-Based Explanations in AI: Towards an Abstract Framework, 2020. Available

at:https://www.researchgate.net/profile/Davide_Calvaresi/publication/341509975_Agent-Based_Explanations_in_AI_Towards_an_Abstract_Framework/links/5ec5020b299bf1c09acc036d/Agent-Based-Explanations-in-AI-Towards-an-Abstract-Framework. Access in: 15 jul. 2020.

DEEKS, Ashley. The Judicial Demand for Explainable Artificial Intelligence. Columbia Law Review, v. 119, n. 7, p. 1829-1850, 2019. Available at:https://columbialawreview.org/content/the-judicial-demand-for-explainable-artificial-intelligence/. Access in: 15 jul. 2020.

DE STREEL, Alexandre et al. Explaining the Black Box: when Law controls AI. Brussels: Centre on Regulation in Europe (CERRE), 2020. Available at:http://www.crid.be/pdf/public/8578.pdf. Access in: 15 jul. 2020.

EUROPEAN COMMISION. High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission, 2019. Available at:https://ec.europa.eu/futurium/en/ai-alliance-consultation. Access in: 15 jul. 2020.

EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. Available at:https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=PT#d1e1797-1-1. Access in: 15 jul. 2020.

FENWICK, Mark; VERMEULEN, Erik PM. It Is Time for Regulators to Open the 'Black Box' of Technology. Lex Research Topics in Corporate Law & Economics Working Paper, n. 2019-2, p. 1-17, 2019. Available at:https://ssrn.com/abstract=3379205. Access in: 15 jul. 2020.

KAMINSKI, Margot E. The Right to Explanation, Explained. Berkeley Technology Law Journal, v. 34, n. 1, p. 189-218, 2019. DOI: https://doi.org/10.15779/Z38TD9N83H.

MALGIERI, Gianclaudio; COMANDÉ, Giovanni. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. International Data Privacy Law, v. 7, n. 4, p. 243-165, 2017. DOI: 10.1093/idpl/ipx019.

MULHOLLAND, Caitlin; FRAJHOF, Isabella Z. Inteligência Artificial e a Lei Geral de Proteção de Dados Pessoais: breves anotações sobre o direito à explicação perante a tomada de decisões por meio de *machine learning* In: FRAZÃO, Ana; MULHOLLAND, Caitlin (coord.). Inteligência Artificial e Direito. São Paulo: Thomson Reuters Brasil, 2019, p. 265-291.

NICHOLAS, Gabriel. Explaining Algorithmic Decisions: A Technical Primer. Georgetown Law Technology Review, Forthcoming, 2020. Available at: https://ssrn.com/abstract=3523456. Access in: 15 jul. 2020.

OBAR, Jonathan A. Sunlight alone is not a disinfectant/ Consent and the futility of opening Big Data black boxes (without assistance). Big Data & Society, v. 7, n. 1, p. 1–5, 2020. DOI: https://doi.org/10.1177/2053951720935615.

OLSEN, Henrik Palmer et al. What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration. Legal Studies Research

Paper Series, n. 2019-84, p. 1-27, 2019. Available at:http://jura.ku.dk/icourts/working-papers/. Access in: 15 jul. 2020.

PASQUALE, Frank. The Black Box Society: the secret algorithms that control money and information. Cambridge; London: Harvard University Press, 2015.

RAI, Arun. Explainable AI: from black box to glass box. Journal of the Academy of Marketing Science, v. 48, n. 1, p. 137–141, 2020. DOI: https://doi.org/10.1007/s11747-019-00710-5.

ROBBINS, Scott. A Misdirected Principle with a Catch: Explicability for AI. Minds and Machines, v. 29, n. 4, p. 495-514, 2019. DOI: https://doi.org/10.1007/s11023-019-09509-3.

SAMEK, Wojciech; WIEGAND, Thomas; MÜLLER, Klaus-Robert. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017. Available at:https://arxiv.org/abs/1708.08296. Access in: 15 jul. 2020.

SELBST, Andrew D.; BAROCAS, Solon. The Intuitive Appeal of Explainable Machines. Fordham Law Review, v. 87, n. 3, p. 1085-1139, 2018. Available at:http://fordhamlawreview.org/issues/the-intuitive-appeal-of-explainable-machines/. Access in: 15 jul. 2020.

SELBST, Andrew D.; POWLES, Julia. Meaningful information and the right to explanation. International Data Privacy Law, v. 7, n. 4, p. 233-242, 2017. DOI: 10.1093/idpl/ipx022.

STRANDBURG, Katherine J. Rulemaking and Inscrutable Automated Decision Tools. Columbia aw Review, v. 119, n. 7, p. 1851-1886, 2019. Available at:https://columbialawreview.org/content/rulemaking-and-inscrutable-automated-decision-tools/. Access in: 15 jul. 2020.

TEIXEIRA, Tarcisio; ARMELIN, Ruth Maria Guerreiro da Fonseca. Lei Geral de Proteção de Dados Pessoais: comentada artigo por artigo. Salvador: Editora JusPodivm, 2019.

TUBELLA, Andrea Aler et al. Contestable Black-Boxes. Arxiv, 2020. Available at:https://arxiv.org/abs/2006.05133. Access in: 15 jul. 2020.

WACHTER; Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, v. 7, n. 2, p. 76-99, 2017. DOI: 10.1093/idpl/ipx005.

WISCHMEYER, Thomas. Artificial Intelligence and Transparency: Opening the Black Box In: WISCHMEYER, Thomas; RADEMACHER, Timo (eds.). Regulating Artificial Intelligence. Cham: Springer, 2020, p. 75-102. DOI: https://doi.org/10.1007/978-3-030-32361-5.