e. publica
public law journal

Interpretable AI models for judicial decision-making: beyond explicability towards legal due process

Modelos de IA interpretáveis para tomada de decisão judicial: além da explicabilidade em direção ao devido processo legal

Rodrigo L. Canalli

Vol. 11 No. 1
maio 2024
e-publica.pt

ISSN 2183-184x

Com o apoio de:

Fundação para a Ciência e a Tecnologia

INTERPRETABLE AI MODELS FOR JUDICIAL DECISION-MAKING: BEYOND EXPLICABILITY TOWARDS LEGAL DUE PROCESS

MODELOS DE IA INTERPRETÁVEIS PARA TOMADA DE DECISÃO JUDICIAL: ALÉM DA EXPLICABILIDADE EM DIREÇÃO AO DEVIDO PROCESSO LEGAL

RODRIGO L. CANALLI[1]
https://orcid.org/0000-0002-4121-1395
Laboratório de Governança e Regulação de Inteligência Artificial (LIA) do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP)
rodrigo.canalli@nyu.edu

Abstract: As AI algorithms are employed to apply legal rules in determining rights and obligations, questions related to the observance of due legal process arise. The development of opaque machine learning models, whose predictions cannot be satisfactorily explained, has spurred debates around the idea of explainability of AI models for decision-making. The article argues that (i) in relation to AI models for judicial decision-making, the standard of explainability, besides proving insufficient to meet the requirements for publicity and reasoning of judicial decisions, imposes a form of nakedness not required of human judges; and (ii) a more appropriate standard would be that of interpretable models for judicial decision-making, characterized as able to offer decisions that are referred to current law (legality), internally and externally coherent (consistency), and compatible with the decision of a human judge in a similar case.

Keywords: due process; explicability; interpretability; legal reasoning; opacity

Resumo: À medida que algoritmos de IA vão sendo empregados para aplicar regras jurídicas na determinação de direitos e obrigações, surgem questões relacionadas à observância do devido processo legal. A emergência de modelos opacos de aprendizado de máquina, cujas previsões não podem ser satisfatoriamente explicadas, tem impulsionado debates em torno da ideia de explicabilidade dos modelos de IA para tomada de decisão. Argumenta-se que (i) em relação a modelos de IA para tomada de decisão judicial, o critério da explicabilidade, além de se mostrar insuficiente para satisfazer as exigências de publicidade e fundamentação das decisões judiciais, impõe uma forma de desnudamento não exigida de juízes humanos; e (ii) um critério mais adequado seria o de modelos de tomada de decisão judicial interpretáveis, entendidas, como tais, decisões referidas ao direito vigente (legalidade), interna e externamente coerentes (consistência) e compatíveis com a decisão de um juiz humano em caso análogo.

1. Master's in Law, State and Constitution; Universidade de Brasília. LL.M in Competition, Innovation and Information Law; New York University. Researcher; Ethics4AI. Researcher; Laboratório de Governança e Regulação de Inteligência Artificial (LIA) do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). ORCID ID: 0000-0002-4121-1395. Email: rodrigo.canalli@nyu.edu.

Palavras-chave: devido processo legal; explicabilidade; interpretabilidade; opacidade; racionalidade jurídica

1. Introduction: artificial intelligence in judicial decision-making

Artificial intelligence (AI) decision-making systems are being extensively employed to apply legal rules in determining rights and obligations. In the United States, for example, algorithms are used to determine the rights of individuals to social benefits, to assess the performance of employees, selecting who will be fired, and to assist judges in granting or denying bail and probation (Crawford; Schultz, 2019: 1941). In Kenya, AI-assisted legal research and predictive analysis have shown potential to reduce the processing time of cases in the Environmental and Land Court (Ogonjo et al., 2021: 59-68). In India, where the vast idiomatic variety poses a challenge for the judiciary, courts have been using AI tools for document translation into procedural acts, as well as for the management of case archives (Jauhar et al., 2021).

Probably one of the most daring experiments in judicial automation, the so-called Smart Courts system, being implemented in China, combines big data, blockchain, and determinative artificial intelligence with the alleged objective of promoting easier access to justice and enabling faster dispute resolution (Shi; Sourdin, 2023). Meanwhile, in Brazil, courts have been utilizing tools that incorporate artificial intelligence methods to manage case archives, identify similar cases, and rule on cases dealing with themes that have been decided under general repercussion effect by the Supreme Federal Court (Nascimento et al., 2022: 27).

As notable as they are controversial, these are just a few examples of a growing trend, recently intensified by the arrival of improved natural language text-generation tools powered by large language models (Markou; Deakin, 2020: 13). Most of them, however, can be fairly reduced to sophisticated systems of selection among sets of predefined options – sets of coordinated rules for the resolution of a given operation, refined by probabilistic models (Reich; Sahami; Weinstein, 2021: 82).

Despite the vigorous current debate about the potential practical roles of artificial intelligence in legal practice and assisted decision-making, its implications for general jurisprudence – and for the theory of judicial decision-making in particular – still require further development. The discussion about the theoretical possibility of a non-human entity (in this case an artificial intelligence algorithm) being capable of performing the same type of legal reasoning conducted by human judges when deciding the cases submitted to them presents the perspective of intelligent automated systems for judicial decision-making being developed as progressions from the current decision-support systems and predictive decision models.

In this context, among the multiple ethical and regulatory aspects related to the use of artificial intelligence for automated decision-making, the concept of algorithmic explainability has often been presented as an adequate or sufficient regulatory standard to ensure the transparency of decision-making models in the face of challenges posed by the opacity of AI models, which for this reason, are nicknamed *black boxes* (Pasquale, 2015: 4): how to access, audit, or understand the functioning of inscrutable and often non-intuitive algorithms (Selbst; Barocas, 2018: 1085).

Furthermore, when we talk about AI models for decision-making on rights and obligations, whether in the judicial or administrative sphere, opacity, a problem in itself, is exacerbated, from a normative standpoint, by the need to respect the due process of law, which includes the guarantee, assured to everyone in a democratic state of law, of knowing the grounds for a state decision affecting their rights.

The supposed insufficiency of the concept of explainability, as it has been conceived and developed, to effectively ensure transparency, access to relevant information, and due process of law when it comes to algorithmic models for decision-making – judicial or administrative – on rights and duties, is discussed next. The requirements of due process of law, it is argued, can only be addressed by a regulatory approach compatible with the hermeneutic nature of legal practice, which could be achieved with algorithms that are, more than explainable, interpretable.

2. Computer algorithms, cognition and legal rationality

In our current digital culture, we tend to associate the concept of an algorithm itself with one particular usage: the incorporation of algorithmic models in programming code that makes a digital computer perform tasks we assign them. But we have been using algorithms for thousands of years to perform various activities, from fishing to baking cakes. When performing standardized tasks, making a diagnosis, applying a method to solve a problem, applying a rule to a fact and checking the outcome, in all these activities we are thinking algorithmically. Generally speaking, an algorithm is nothing more than a finite set of instructions that, applied to a given set of data, produces a predictable result, "a procedure that allows us to solve a problem without having to invent a solution each time" (Abiteboul; Dowek, 2020: 6). In this sense, it is possible to understand a criminal code, with its definitions of specific crimes, aggravating factors and extenuating circumstances, requirements of culpability etc., as an algorithm that, when applied by the magistrate (computer) to a given set of data (the specific case), produces, as result, a valid verdict from a normative point of view. From a conceptual standpoint, therefore, a simple algorithm contemplating the finite set of parameters and variables present in a criminal code, operated by a human intelligence provided with the necessary information, is able in theory to accurately carry on with the judgment of a criminal case.

As is well known, a computer algorithm consists of a specific class of pre-programmed applications that, when fed a collection of data, offers a precise response. Algorithmic decision-making models, in general, are trained to find patterns in data through a process called machine learning (Reich; Sahami; Weinstein, 2021: 82). Tasks ranging from research and planning to image recognition and interpretation, through speech recognition and decision-making in situations of uncertainty, are all reducible to pattern recognition and information classification. Machine learning algorithms, in particular, are trained, validated, and tested with input from training datasets: in the case of judicial decision-making algorithms, these elements would correspond to precedents, laws, procedures, etc., from which the algorithm "learns". As more data is provided to a machine learning algorithm, the universe of errors and successes to be referenced as parameters for further refinement

expands. Thus, the mathematical model generated by algorithm running the provided data becomes capable of providing progressively more accurate responses to the class of problems it was trained to solve, improving its performance based on experience.

An especially complex class of algorithms currently undergoing a period of accelerated development includes the cognitive models: computational models that resemble human cognitive processes, simulating sophisticated behaviors. Since the architecture of these algorithms emulates the structure of a biological brain, they are also called neural networks. This structure allows mathematical representations of possibilities and probabilities to mirror human abilities of reasoning and inference (Reich; Sahami; Weinstein, 2021: 161) and, as such, perform processes that can be described as analogous to the weighing and balancing of rules and principles (Garcez; Gabbay; Lamb, 2014: 109).

On the other hand, human language and cognition have been effectively described as implementations of particularly complex algorithms for assembling hierarchical symbols (Berwick; Chomsky, 2016: 132). As a linguistic artifact, law is also a symbolic system, or mode of communication (Berwick; Chomsky, 2016: 53) and, as such, also implements algorithms (Abiteboul; Dowek, 2020: 82). Procedural law, for example, can be fairly described as an algorithm that allows lawyers and judges to judge cases in a standardized and predictable manner, as are the tests devised by courts to answer whether a given factual framework fits within the scope of a precedent. If legal rules are algorithms, they are independent of the specific language in which they are encoded (Abiteboul; Dowek, 2020: 36) and, therefore, can be translated into a machine-readable language.

To perform the same type of reasoning a judge does when applying a rule to a fact, an adjudicating entity (such as an AI) must be capable of assimilating, with a high degree of precision, the particular description of a fact to the abstract and general hypothesis conveyed in the rule, whether a rule requiring or prohibiting a conduct (imposing a duty or obligation) or a rule conferring certain rights on individuals (Hart, 2012: 27). The very structure of legal argumentation is theoretical, i.e., it is an abstraction. In this sense, applying a rule to a case is an exercise in abstraction, requiring no particular imaginative or creative skill beyond the above-mentioned cognitive process: the implementation of an algorithm to assemble hierarchical symbols (Abiteboul; Dowek, 2020: 53). Even when operating on defeasible arguments and inconsistent information, legal reasoning can still be modeled as a process of deriving and comparing arguments (Prakken, 1997: 275-280). Judges learns the algorithm when attending law schools and studying statutes and precedents. Once they learn the law's algorithm, the abstract standard, they apply it to the particular cases submitted to them.

As previously argued (Canalli, 2023: 869), in theory an AI will be capable of judicial decision-making if it produces decisions that prove to be rational, informed, and impartial by applying legal rules to identified sets of facts. In summary, a rational decision is one in which the applicable rule is identified according to consistent and comprehensible parameters in a satisfactory and explainable way (Hildebrandt, 2018: 12). An informed decision takes into the equation and properly weighs all relevant facts and affected interests. In machine learning algorithms, this ability would rest on a granular analysis of

relevant case-law (Hildebrandt, 2018: 23). An impartial decision is one in which no subjective interest or preference of the adjudicating entity, conscious or unconscious, plays any role. Theoretically, parameters provisioning for all these characteristics can be embedded into algorithmic models.[2] Such models would be effectively capable of producing valid judicial decisions, from the perspective of legal theory, even if they did not meet what has been called explainability. In such circumstances, the continued emphasis on explainable models is here questioned in favor of models that can be identified as interpretable. Although the use of the expressions explainability and interpretability interchangeably is not uncommon, we adopt, following Cynthia Rudin (2019: 206), a strict distinction between the two.

3. Opacity, explainability and interpretability

Opacity means nothing more than an absence of transparency. In part, it results from withholding access to software source code, within a strategy to protect the developer's intellectual property. Without access to the source code, it is impossible to know how decisions are made by the tool. On the other hand, opacity can also stem from a structural characteristic of complex language models, because often the mathematical models generated after training the algorithm are not comprehensible to humans, including the engineers who developed the algorithm themselves. In these cases, even if access to the source code used to implement the model is available, it remains technically impossible to identify the criterion, or set of criteria, used by an algorithm when making decisions.

The opacity of models used to make – or assist in the making of – judicial decisions raises serious questions related to the observance of due process and the rule of law. Assuming one has the right, under due process of law, to know the grounds of a decision made against her, the very legitimacy of a judicial decision is typically associated with its rationale, the expression of a certain legal rationality (Waldron, 2007: 23). In fact, the absence or deficiency of a rationale is regarded as cause for the nullity of a judicial decision (Cardoso; Fanti, 2017). In Brazil, Article 93, IX, of the Federal Constitution[3] expressly requires that all judgments be public and that judicial decisions be reasoned, a circumstance that strongly suggests the

2. Whether they can or should choose an outcome when there is no single one available from mere rule application is an issue that raises particular concerns. For a comprehensive discussion on this issue, see Williams (2018).

3. "Art. 93 (...) IX – all judgments of judicial bodies shall be public, and all decisions shall be substantiated, under penalty of nullity; in cases in which preservation of the right of intimacy of the interested parties in secrecy does not prejudice the public interest in information, the law may limit attendance at determined occasions to only the parties themselves and their attorneys, or only to the latter." ("Art. 93 (...) IX – todos os julgamentos dos órgãos do Poder Judiciário serão públicos, e fundamentadas todas as decisões, sob pena de nulidade, podendo a lei limitar a presença, em determinados atos, às próprias partes e a seus advogados, ou somente a estes, em casos nos quais a preservação do direito à intimidade do interessado no sigilo não prejudique o interesse público à informação.") (BRASIL, 2023a).

unconstitutionality of AI decision-making systems that do not meet some transparency benchmarks.

The emergence of opaque machine learning models, whose predictions cannot be explained in a way comprehensible to humans, has fueled debates around the idea of explainability of AI models for decision-making. Corporate and governmental regulation plans have adopted explainability as a normative parameter. This is the case, among others, as has been pointed out, of Article 22 of the General Data Protection Regulation in the European Union,[4] and, more incisively, of Article 20 of the General Data Protection Law (Law N.º 13.709/2018),[5] in Brazil. Both norms ensure that, in the case of a decision is taken based on automated processing of personal data, an explanation about the criteria and procedures used for the automated decision-making should be provided to the subject of the data.

Explainability is also adopted as a parameter in the final draft of the European Commission's proposal to introduce a normative and regulatory framework for artificial intelligence (European Union, 2024) as of 21st January 2024, and in Bill n.º 2338, of 2023, currently under consideration in the Brazilian Federal Senate. The latter also aims to establish general national norms for the development, implementation, and responsible use of artificial intelligence systems in Brazil, with the goal of protecting fundamental rights and ensuring the implementation of safe and reliable

4. When decisions are made exclusively on the basis of automated processing of personal data that produce legal effects concerning the data subject, Paragraph 3 of Article 22 of the General Data Protection Regulation (GDPR) requires that the data controller implement "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, including the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (European Union, 2016). Although commonly found in the literature, the claim that this provision enshrines a right to an explanation of automated decisions is controversial. In this regard, see Watcher; Mittlestadt; Floridi (2017).

5. "Article 20. The data subject has the right to request a review of decisions made solely based on automated processing of personal data that affect their interests, including decisions intended to define their personal, professional, consumer, and credit profile or aspects of their personality.

Paragraph 1. The controller must provide, whenever requested, clear and adequate information regarding the criteria and procedures used for automated decision-making, observing commercial and industrial secrets.

Paragraph 2. In the event of failure to provide information as referred to in Paragraph 1 of this article based on the observance of commercial and industrial secrets, the national authority may conduct an audit to verify discriminatory aspects in automated personal data processing."

("Art. 20. O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade.

§ 1.º O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.

§ 2.º Em caso de não oferecimento de informações de que trata o § 1.º deste artigo baseado na observância de segredo comercial e industrial, a autoridade nacional poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais." (BRASIL, 2023b.)

systems, for the benefit of human beings, the democratic regime, and scientific and technological development.

Explainable machine learning models can be described as capable of a more or less detailed *post hoc* explanation of their internal processes, which can be achieved, in the case of models generated by proprietary closed algorithms, even by a second model created to explain the first (Rudin, 2019: 206). Explainability is problematic because the information provided about the tool's functioning is often unreliable, and may even be misleading (Rudin, 2019: 207). Moreover, the concept seems to suffer from chronic imprecision, since there is no clarity on what counts as a satisfactory explanation of a decision-making model (Rudin, 2019: 214).

In this paper, it is argued that: (i) at least in relation to AI models for judicial decision-making, the criterion of explainability, while proving insufficient to meet the constitutional requirements of publicity and reasoning of judicial decisions, imposes a form of exposure that is not required of human judges; and (ii) a more suitable criterion would be that of interpretable models for judicial decision-making.

As it is a notion whose definition depends on the domain (Rudin, 2019: 206), interpretability, in the construction of an algorithmic model, is delineated according to its purpose. This means that a notion of interpretability for a decision-making algorithm about rights will not serve for health diagnostic models, or models for autonomous vehicles. For this reason, the definition of an interpretable model proposed here is specific to models of judicial decision-making (decision on subjective rights and obligations).

4. Due process of law and interpretable artificial intelligence models

We do not have access to the cognitive processes that come into play when a judge applies the law to a set of facts when deciding a case. Nevertheless, we do not have a hard time identifying this as legal reasoning. Such cognitive processes are not, so to speak, explainable. What we have are justifications, often produced after the cognitive decision-making process, by which the mental mechanisms involved are retrospectively signified – interpreted. Similarly, the internal processes of a hypothetical algorithmic entity capable of cognitive processes analogous to legal reasoning may, as seen, be equally inaccessible. In any case, the reasoning of a judicial decision – the argumentative exposition of the legal grounds that led the judge to a certain conclusion – translates an interpretation of their cognitive processes, not their explanation.

Taking the observance of due process of law as a conceptual guide for the use of artificial intelligence tools within the Judiciary, explainability proves insufficient to satisfy this normative parameter for the justification of decisions. This is because, as already seen, complex language models are not always effectively explainable.

Differently, the interpretability of the judicial decision produced by an algorithmic model means that, even if it is not possible to technically explain how a language model pointed to a certain outcome, its conclusion must, at least, be recognizable as a decision that is (i) consistent with the legal

discipline related to the matter, (ii) internally and externally coherent, and (iii) assimilable to a judicial decision made by a human judge.

Without prejudice to future developments and remembering that the notion of interpretability is specific to each domain (Rudin, 2019: 206), I suggest that interpretable models of judicial decision-making should be constructed contemplating, as the above identified characteristics, at least: legality, consistency, and compatibility. By legality, it is understood the production of decisions that reference current law, observing the hierarchy of normative species and the courts' precedents. A consistent decision presents internal coherence (absence of argumentative contradiction) and external coherence (due consideration of the facts on which it decides). Finally, the decision must be recognized as compatible with a decision that would be made by a human judge deciding an analogous case.

This third requirement operates as an *a fortiori* evaluation of the algorithmic decision's reasonability and moves us from a semantic perspective to an interpretive account of what counts as a judicial decision. Contrary to humans, who use mental models that considerably narrow the decision space when they adjudicate, machine learning algorithms do not implement true cognitive modeling (Gasser; Mayer-Schönberger, 2014: 76). While human communication relies on shared common ground enabling the interpretation of implicit meaning conveyed between individuals, a language model generates text that "is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind" (Bender; Gebru et al., 2021: 616). Machine learning algorithms do not have the communicative resources and cognitive skills to know when they are repeating something incorrect, out of context, socially inappropriate or morally disproportionate. For this reason, it is imperative that a safeguard is put in place to avoid algorithmic decision-making that, notwithstanding their logical coherence and adherence to the semantics of the law, would be perceived as effectively unjust or unreasonable. In this sense, an algorithmic judicial decision whose outcome is irreconcilable with the moral or cultural standards of the community in which it is undertaken would hardly qualify as conforming to human decision-making.

In a more prosaic manner, the compatibility requirement also works as a protection against what has been described as artificial intelligence hallucinations: outputs that are erroneous or misleading, unsupported by the source content, in spite of being coherent from a merely semantic perspective (Ji et al., 2022: 4). Such dangerous outputs are possible because the algorithmic models that generate convincing linguistic patterns do not actually understand the meaning of the language they process.

The above-mentioned three characteristics of interpretable decision-making algorithms can be decomposed into more specific procedural and material aspects. In an adversarial jurisdiction model, the external consistency of the decision could be linked, for example, to the effective consideration or evaluation by the AI model of the normative arguments conveyed by the parties (Citron, 2008: 1249).

More than the much-touted explainability, and as much as any human judge, to satisfy due process of law, an AI model for judicial decision-making must observe the conditions that ensure its interpretability.

5. Conclusion

As outlined above, explainable algorithmic models can be described as being capable of a more or less detailed *post hoc* explanation of their internal processes. The literature explored authorizes the conclusion that the concept of explainability, besides being imprecise, is limited, leading to the provision of information that is often unreliable, irrelevant, and even misleading about the process by which an algorithm arrives at a certain decision.

In this paper, I argued for the insufficiency of explainability as a criterion to ensure the observance of due process in relation to AI models for judicial decision-making. I also proposed that interpretable models would be more appropriate. Understood as a concept whose content depends on the domain in which it is applied, interpretability needs to be delineated, in the construction of an algorithmic model, according to its purpose. In the case of interpretable models of judicial decision-making, adherence to due process presupposes the production of decisions in which legality, internal and external consistency, and compatibility, as described above, are identified as legitimizing criteria.

If the use of artificial intelligence tools for judicial decision-making seems inescapable, their effects depend on how we shape the normative ecosystem in which these tools can operate. In this field, the best we can hope for are fairer and more robust decisions about rights and duties. For this to become a reality, it is essential to incorporate parameters that ensure due process into the very structure of the AI tools employed in decision-making practice.

References

Abiteboul S, Dowek G. The Age of Algorithms. Cambridge University Press, 2020.

Ash E. Judge, Jury, and EXEcute file: the brave new world of legal automation. Londres: Social Market Foundation, CAGE Research Centre, Jun. 12, 2018. Available at: https://warwick.ac.uk/fac/soc/economics/research/centres/cage/cage-final-elliot_ash.pdf. Access on Feb. 10, 2024.

Bender EM et al.. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *In*: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Canada: March 2021.

Bwewick RC, Chomsky N. Why Only Us: Language and Evolution. Cambridge: MIT Press, 2016.

BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília: Presidência da República, 2023.

BRASIL. Lei n.º 13.709/2018. Brasília: Presidência da República, 2023.

Canalli RL. Artificial intelligence and the model of rules: better than us? AI and Ethics, Vol. 1, n. 3, Aug. 2023.

Cardoso EL, Fanti F. Movimentos Sociais e Direito: o Poder Judiciário em Disputa. In Silva FG. Manual de Sociologia Jurídica. 2.ª ed. São Paulo: Saraiva, 2017.

Citron DK. Technological Due Process. Washington University Law Review, vol. 85, n. 6, 2008.

Crawford K, Schultz J. AI systems as state actors. Columbia Law Review, New York, v. 119, n. 7, 2019.

Doshi-Velez F, Kim B. Towards a Rigorous Science of Interpretable Machine Learning. arXiv.org. 2017.

EUROPEAN UNION. General Data Protection Regulation (GDPR). European Parliament, 2016.

EUROPEAN UNION. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts (final draft). European Parliament, 2024.

Garcez ASd'A, Gabbay DM, Lamb LC. A neural cognitive model of argumentation with application to legal inference and decision making. Journal of Applied Logic, vol. 12, n. 2, Jun. 2014.

Gasser U, Mayer-Schönberger V. Guardrails: guiding human decisions in the age of AI. Princeton University Press, Princeton & Oxford, 2024.

Hart HLA. The Concept of Law. Oxford University Press, Oxford, 2012.

Hildebrandt M. Law as computation in the era of artificial intelligence: speaking law to the power of statistics. University of Toronto Law Journal, vol. 68, n. 1, 2018.

Jauhar A et al.. Responsible AI for the Indian Justice System: A Strategy Paper. Vidhi Centre for Legal Policy, TCG-Crest, 2021.

Ji Z et al.. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, v. 55, n. 12, pp. 1-38, Nov. 2022. Available at: https://arxiv.org/abs/2202.03629.

Markou C, Deakin SF. Ex Machina Lex: Exploring the Limits of Legal Computability. *In*: MARKOU C, Deakin SF, (eds). Is Law Computable? Critical Perspectives on Law + Artificial Intelligence. Hart Publishing, 2020. Available at SSRN: https://ssrn.com/abstract=3407856 or http://dx.doi.org/10.2139/ssrn.3407856.

Nascimento A et al.. Aplicações da Inteligência Computacional no Judiciário. *In*: Salomão, Luis Felipe (Coord.). Inteligência Artificial: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário brasileiro. 2nd ed. Rio de Janeiro: FGV, 2022.

Ogonjo F et al.. Utilizing AI to Improve Efficiency of the Environment and Land Court in the Kenyan Judiciary: Leveraging AI Capabilities in Land Dispute Cases in the Kenyan Environmental and Land Court System. *In*: International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021), 2, 2021, São Paulo. Joint Proceedings of the Workshops on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021) & AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021), São Paulo: CEUR-WS, 2021.

Pasquale F. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, 2015.

Prakken H. Logical Tools for Modeling Legal Argument: A Study of Defeasible Reasoning in Law. Springer, Berlin, 1997.

Reich R, Sahami M, Weinstein JM. System Error: Where Big Tech Went Wrong and How We Can Reboot. Londres: Hodder and Stoughton, 2021.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, vol. 1, n. 5, May. 2019.

Selbst A, Barocas S. The Intuitive Appeal of Explainable Machines. Fordham Law Review, New York, v. 87, n. 3, Nov. 2018.

Shi C, Sourdin T, Li B. The Smart Court: A New Pathway to Justice in China? International Journal for Court Administration, [*s. l.*], v. 12, n. 1, p. 4, Mar. 2021. Available at: https://iacajournal.org/articles/10.36745/ijca.367.

Wachter S, Mittelstadt B, Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, vol. 7, n. 2, 2017.

Waldron J. The Dignity of Legislation. Cambridge University Press, 2007.

Williams R. Rethinking Deference for Algorithmic Decision-Making. Oxford Legal Studies Research Paper No. 7/2019, 2018.