ISSN 1981-3694
(DOI): 10.5902/1981369469329


# O USO DA *INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL* ENQUANTO FERRAMENTA PARA COMPREENDER DECISÕES AUTOMATIZADAS: POSSÍVEL CAMINHO PARA AUMENTAR A LEGITIMIDADE E CONFIABILIDADE DOS MODELOS ALGORÍTMICOS?

## THE USE OF EXPLAINABLE ARTIFICIAL INTELLIGENCE AS A TOOL TO UNDERSTAND AUTOMATED DECISIONS: A POSSIBLE WAY TO INCREASE THE LEGITIMITY AND RELIABILITY OF ALGORITHMIC MODELS?

## EL USO DE LA INTELIGENCIA ARTIFICIAL EXPLICABLE COMO HERRAMIENTA PARA ENTENDER LAS DECISIONES AUTOMATIZADAS: ¿UNA POSIBLE FORMA DE AUMENTAR LA LEGITIMIDAD Y CONFIABILIDAD DE LOS MODELOS ALGORITMICOS?

**DIERLE JOSÉ COELHO NUNES**
https://orcid.org/0000-0003-4724-5956 / http://lattes.cnpq.br/6141886803125625 / dierlenunes@gmail.com

**OTÁVIO MORATO DE ANDRADE**
https://orcid.org/0000-0002-0541-7353 / http://lattes.cnpq.br/5811976298311056 / otaviomorato@gmail.com

**RESUMO**
Tendo em vista que a falta de transparência em modelos de inteligência artificial (IA) representa um risco para a sua aplicação em áreas sensíveis, este trabalho objetiva investigar a *explicabilidade*, que se dedica a fornecer explicações satisfatórias sobre decisões de modelos algorítmicos. A partir de uma revisão de literatura atual sobre o tema, é empreendida uma análise indutiva. Conclui-se que a inteligência artificial explicável deve ser elemento constitutivo da transparência dos sistemas de IA, uma vez que atua como importante contrapeso à opacidade, transformando "caixas-pretas" algorítmicas em "caixas de vidro". Neste sentido, a criação de sistemas mais transparentes e interpretáveis deve ser considerada e estimulada na formulação de políticas públicas, de modo a elevar a legitimidade das decisões produzidas por sistemas inteligentes.

**Palavras-chave**: *deep learning*; inteligência artificial; inteligência artificial explicável; *machine learning*.

**ABSTRACT**
Considering that the lack of transparency in artificial intelligence (AI) models represents a risk for its application in sensitive areas, this work aims to investigate explainable artificial intelligence (XAI), which is dedicated to providing satisfactory explanations about algorithmic model decisions. From a review of current literature on the subject, an inductive analysis is undertaken. It is concluded that XAI must be a constitutive element of the transparency of AI systems, since it acts as an important counterweight to opacity, transforming algorithmic "black boxes" into "glass boxes". In this sense, the creation of more transparent and interpretable systems should be considered and encouraged in the formulation of public policies, in order to increase the legitimacy of decisions produced by intelligent systems.

**Keywords**: *deep learning*; artificial intelligence; explainable artificial intelligence; *machine learning*.

**RESUMEN**
Considerando que la falta de transparencia en los modelos de inteligencia artificial (IA) representa un riesgo para su aplicación en áreas sensibles, este trabajo tiene como objetivo investigar la inteligencia artificial explicable (XAI),

que se dedica a proporcionar explicaciones satisfactorias sobre las decisiones del modelo algorítmico. A partir de una revisión de la literatura actual sobre el tema, se emprende un análisis inductivo. Se concluye que la XAI debe ser un elemento constitutivo de la transparencia de los sistemas de IA, ya que actúa como un importante contrapeso a la opacidad, transformando las "cajas negras" algorítmicas en "cajas de cristal". En este sentido, la creación de sistemas más transparentes e interpretables debe ser considerada y fomentada en la formulación de políticas públicas, a fin de incrementar la legitimidad de las decisiones producidas por los sistemas inteligentes.

**Palabras clave**: aprendizaje profundo; inteligencia artificial; inteligencia artificial explicable; aprendizaje automático.

## SUMÁRIO



# INTRODUÇÃO

A expansão da Inteligência Artificial (IA) suscita uma multiplicidade de questões: desde aspectos sobre a privacidade – vulnerável frente às grandes plataformas digitais – passando por implicações éticas decorrentes do uso de algoritmos, como os vieses discriminatórios e a possibilidade de as máquinas reproduzirem preconceitos, até debates sobre como esses novos fenômenos podem afetar as estruturas das relações e o nosso sistema político[1]. Dentre tantas implicações possíveis no uso da IA, este trabalho propõe um recorte metodológico, *objetivando discutir a explicabilidade – ou seja, a melhor compreensão humana sobre o processo decisório das máquinas inteligentes – enquanto requisito para o desenvolvimento dessas novas tecnologias*.

A hipótese colocada é de que, se os usuários quiserem gerenciar e confiar nos sistemas artificialmente inteligentes, será fundamental oferecer mais transparência em relação aos processos internos que levaram os sistemas de IA a tomarem as suas decisões. Em última instância, o desenvolvimento de uma inteligência artificial explicável aumentaria a compreensão

[1] NUNES, Dierle José Coelho; MARQUES, Ana Luiza. Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. **Revista de Processo**, São Paulo, v. 43, p. 421 – 447, nov. 2018. Disponível em: https://bd.tjdft.jus.br/jspui/handle/tjdft/43025. Acesso em: 18 jun. 2023; ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação? **Revista Eco Pós**, v. 18, n. 2, p. 35-56, 2015. Disponível em: https://revistaecopos.eco.ufrj.br/eco_pos/article/view/2662. Acesso em: 18 jun. 2023; MOROZOV, Evgeny. **Big tech**: a ascensão dos dados e a morte da política. São Paulo: Ubu, 2018.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


e a legitimação das ações tomadas pelos sistemas autônomos[2]. Neste sentido, o objetivo geral do trabalho é investigar a *explicabilidade* enquanto requisito para a justificação e legitimação das decisões tomadas pelos modelos de IA. Para articular esta investigação, parte-se de *objetivos específicos*, sendo eles: a) conceituação dos institutos relativos à IA; b) apresentação das noções essenciais à compreensão da opacidade dos sistemas autômatos e c) análise de atributos e técnicas inerentes à inteligência artificial explicável.

Em termos metodológicos, o presente construto apresenta abordagem qualitativa, já que o problema em questão não pode ser quantificável. Em relação à lógica utilizada, a pesquisa é definida como indutiva, pois através da revisão de literatura, busca-se chegar a uma conclusão genérica a partir da análise um conjunto teórico específico acerca do tema, gerando conhecimento sobre um tema novo e relativamente pouco desenvolvido. Quanto ao objeto, a pesquisa é descritiva e exploratória.

O artigo é apresentado em quatro etapas. Primeiramente, será formulada uma síntese dos conceitos basilares sobre a IA, tais como redes neurais, *machine learning* e *deep learning*, apresentando-se também um breve histórico da evolução da IA. Em segundo lugar, analisa-se a questão da opacidade, a partir da qual é discutida a necessidade de melhor compreensão dos processos internos de modelos algorítmicos. Na terceira etapa do trabalho, são estudados a trajetória evolutiva e as características da *explicabilidade*, bem como alguns métodos utilizados para interpretar os sistemas de IA. Na quarta etapa, desenvolve-se uma reflexão crítica sobre a importância e os benefícios de se identificar, o mais detalhadamente possível, os passos ou mecanismos que levam determinados algoritmos a construírem suas decisões.

Conclui-se que a promoção da inteligência artificial explicável pode aumentar a confiabilidade e a legitimidade dos modelos algorítmicos, seja através do direito à explicação, seja por meio de normas e leis que incentivem o desenvolvimento desta funcionalidade em determinados sistemas. Neste sentido, é desejável que a explicabilidade seja considerada um elemento constitutivo da transparência dos sistemas de IA, atuando como contrapeso à opacidade característica de certos modelos algorítmicos. Desta forma, a criação de sistemas mais transparentes, interpretáveis e autoexplicativos deve ser considerada e estimulada na formulação de leis e políticas públicas.

[2] VILLANI, Cédric. **Donner uns sens à li'intelligence artificielle**: pour une stratégie nationale et européenne. Paris, 2018. Disponível em: https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/184000159.pdf. Acesso em: 18 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


## 1 INTELIGÊNCIA ARTIFICIAL: DO APRENDIZADO SIMBÓLICO AO *DEEP LEARNING*

As primeiras pesquisas sobre IA (inteligência artificial[3]), feitas a partir da década de 1940, visavam solucionar problemas a partir de *métodos simbólicos*, que consistiam em mecanismos matemáticos menos sofisticados, como: o aprendizado por analogia/instâncias (exemplo: sistemas baseados em casos); o aprendizado por indução (e.g.: árvores de decisão) e o aprendizado por evolução/seleção (e.g.: algoritmos genéticos). Nestas abordagens, as máquinas eram orientadas para manipular informações simbólicas (qualitativas), gerando limitações para manipular valores numéricos e tratar os problemas com completude[4].

Em contraponto aos métodos simbólicos, desenvolveu-se o estudo das redes neurais artificiais (RNAs) a partir do *método conexionista*, baseado na modelagem dos neurônios humanos. Em 1983, a agência norte-americana DARPA fundou um segmento focado em neurocomputação, impulsionando pesquisas sobre RNAs, que acabaram prevalecendo sobre os métodos simbólicos. Como resultado, começou a emergir, ao longo dos anos 80, o aprendizado de máquina (*machine learning*[5]). Essa área da computação desenvolve algoritmos que se

[3] A inteligência artificial geralmente é definida como a capacidade da máquina de interpretar dados de forma racional ou humana, tomando decisões autênticas com base em informações preexistentes. Morais explica que o termo inteligência artificial geralmente é empregado em um sentido amplo, abarcando quaisquer programas computacionais aptos a reproduzir alguma habilidade humana. No entanto, estudos comportamentais têm sublinhado a presença de vieses cognitivos nas decisões humanas, fazendo que com que muitos pesquisadores rejeitem a aptidão de "pensar humanamente" como traço definidor da IA. Como resultado, parte expressiva da doutrina tem definido a IA como a capacidade da máquina de "agir racionalmente". *In*: MORAIS, Fausto Santo de. O uso da inteligência artificial na repercussão geral: desafios teóricos e éticos. **Revista de Direito Público**, Brasília, v. 18, n. 100, p. 306-326, 2021. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/6001/pdf. Acesso em: 18 jun. 2023; ANDRADE, Otávio Morato de. **Governamentalidade algorítmica**: democracia em risco? 1. ed. São Paulo: Dialética, 2022; RUSSEL, Stuart; NORVIG, Peter. **Artificial intelligence**: a modern approach. New Jersey: Prentice-Hall, 1995. Ademais, importante notar que a inteligência artificial vai muito além da mera automação. Nesta última, a máquina é meramente pré-configurada por ser humanos, que tendem a operá-la, monitorá-la e reconfigurá-la continuamente. Na inteligência artificial, por sua vez, a máquina mantém um grau de independência maior, interpretando o ambiente, produzindo raciocínios completos e, não raro, remodelando por conta própria suas avaliações e decisões ao longo do tempo.

[4] OSÓRIO, Fernando. Redes Neurais - Aprendizado Artificial. **Forum de I.A**. [S.l.]. Disponível em: http://osorio.wait4.org/oldsite/IForumIA/fia99.pdf. Acesso em: 18 jun. 2023. p. 6.

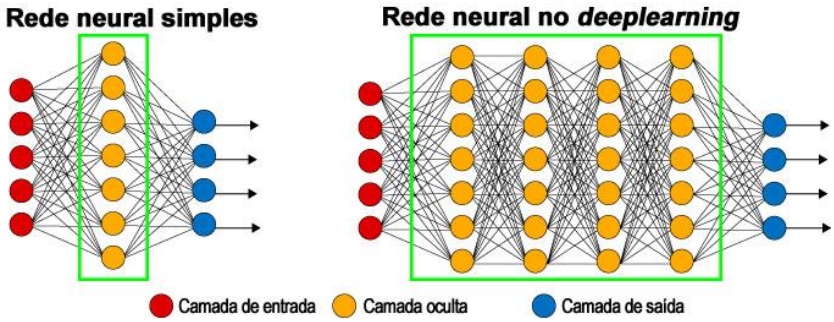
[5] Para Harry Surden, *Machine learning* relaciona-se ao aprendizado e aperfeiçoamento computacional a partir da experiência (...) se tiver um bom desempenho, os algoritmos de *machine learning* podem produzir resultados automatizados que se aproximam daqueles que teriam sido encontrados por uma pessoa em situação semelhante. *In*: SURDEN, Harry. Machine learning and law. **Washington Law Review**, [s.l.], v. 89, N. 1, mar. 2014. Disponível em: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5/. Acesso em: 18 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


aprimoram automaticamente por meio da experiência e do uso de dados, construindo modelos baseados em dados de amostra ou dados de treinamento (do inglês: *training data*). No *machine learning*, o computador é desenvolvido para "se autoprogramar" com base em sua própria experiência, reunindo dados, interpretando informações e tomando decisões de modo similar aos humanos[6].

Por sua vez, o aprendizado profundo (*deep learning*) é uma classe mais recente e avançada de algoritmos de aprendizado de máquina, valendo-se de múltiplas camadas para extrair progressivamente recursos de nível superior da entrada bruta. Neste sentido, enquanto o aprendizado de máquina usa algoritmos para analisar dados, aprender com esses dados e tomar decisões informadas com base no que aprendeu, o aprendizado profundo estrutura algoritmos em camadas para criar uma "rede neural artificial" que pode aprender e tomar decisões inteligentes por conta própria.

**Figura 1** - Diferença entre a rede neural simples e a rede neural no *deep learning*. Destaca-se, em verde, a complexidade e a multiplicidade dos processos de aprendizagem profunda.




**Fonte**: Adaptado pelos autores a partir de DeepAI[7].

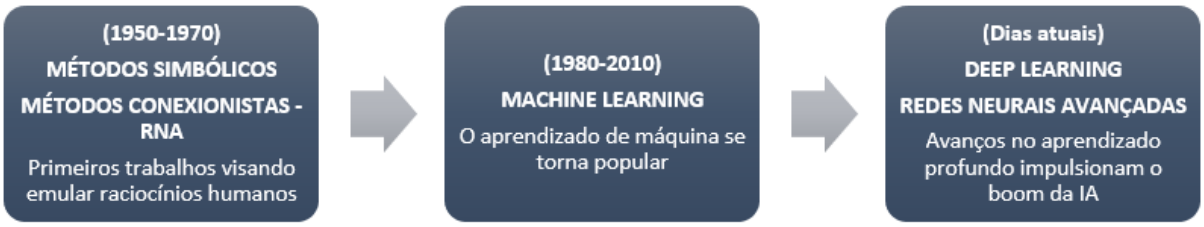
Um exemplo de *deep learning* é um trabalho recente de cientistas liderados por Sebastian Thrun, que programaram uma única rede convolucional neural (do inglês: *convolutional neural network - CNN*) para detectar o câncer de pele, depois de treinar o algoritmo com um banco de dados de 129.450 imagens de manchas dérmicas. A performance de uma única CNN na detecção do câncer de pele foi confrontada com a de 21 médicos

[6] ARENS, Bob. **Cognitive computing**: under the hood. Thomson Reuters, [s.l.], jan. 2017.
[7] HIDDEN LAYER. DeepAI, [s.l.]. Disponível em: https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning. Acesso em: 18 jun. 2023.

dermatologistas e, em todas as tarefas de identificação propostas, o desempenho do algoritmo foi equivalente ao dos especialistas humanos[8].

Em 2016, a *DeepMind Technologies* (empresa pertencente à *Google*) também demonstrou o poder do *deep learning* por meio do algoritmo *AlphaGo*, que aprendeu a jogar um *Go*, um jogo de tabuleiro conhecido por exigir intelecto e intuição aguçados. O sistema foi além do *machine learning*, sem precisar ser informado de quando deveria fazer movimentos específicos e, exibindo extraordinário refinamento, derrotou vários mestres mundialmente renomados do *Go* – depois de estudar, aprender e reformular suas técnicas mais complexas enquanto jogava[9].

**Figura 2** - Evolução histórica dos modelos de inteligência artificial.




**Fonte**: Elaboração própria.

Considerando tais avanços, o desenvolvimento da IA e sua aplicação no auxílio à tomada de decisões em áreas sensíveis (como justiça criminal, saúde, educação, finanças, etc.) atrai uma gama de dúvidas e desafios éticos e jurídicos[10]. Uma das preocupações é a dificuldade em se entender uma decisão algorítmica, já que normalmente recebe-se apenas o *output* – ou seja, o resultado de sua predição – sem que, contudo, se conheça a cadeia de operações que conduziu a ele. Quando isso ocorre, surge o problema da *opacidade*, ou da "caixa preta" algorítmica[11].

[8] ESTEVA, Andre *et al*. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, [s.l.], n. 542, p. 115–118, 2017. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28117445/. Acesso em: 18 jun. 2023.
[9] SILVER, David et al. Mastering the game of go without human knowledge. **Nature**, [s.l.], out. 2017. Disponível em: https://www.nature.com/articles/nature24270. Acesso em: 18 jun. 2023.
[10] NUNES, Dierle José Coelho; ANDRADE, Otávio. A explicabilidade da inteligência artificial e o devido processo tecnológico. **Revista Conjur**, São Paulo, 7 de julho de 2021. Disponível em: https://www.conjur.com.br/2021-jul-07/opiniao-explicabilidade-ia-devido-processo-tecnologico. Acesso em: 18 jun. 2023.
[11] ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato. Da "caixa-preta" à "caixa de vidro": o uso da Explainable Artificial Intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Revista de Direito Público**, Brasília, v. 18, n. 100, 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 17 jul. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


## 2 O PROBLEMA DA OPACIDADE

Quando se pensa na disseminação de algoritmos de inteligência artificial pelos diversos ramos da vida humana, irrompem preocupações relativas à eventual opacidade de determinados sistemas, sobretudo aqueles dotados de aprendizado de máquina profundo *(deep learning)*.

O fato é que o funcionamento interno de alguns algoritmos pode ser um mistério completo para o usuário médio de tecnologia e, não raro, até para aqueles com competências avançadas na área[12]. Disso decorre que o processo decisório de algoritmos complexos muitas vezes é de difícil compreensão para o ser humano, o que pode comprometer aa legitimidade dessas decisões. Para entender melhor este problema, serão explorados os tipos de opacidade e as preocupações levantadas diante da sua existência.

### 2.1 Opacidade: conceito e formas

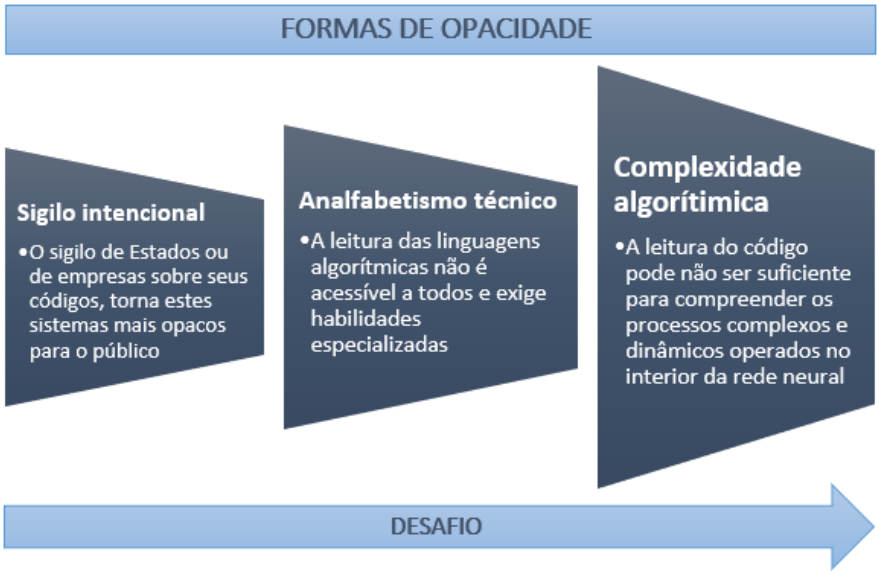
Coloquialmente, diz-se que um material é "opaco" quando não ele permite a passagem da luz de forma adequada, tornando difícil a visualização do que há em seu interior. Transpondo esta ideia para a presente análise, um modelo de IA opaco seria aquele em que não se consegue visualizar, com clareza, *como* e *porque* ele toma determinada decisão. Henry Surden[13] utiliza o termo *opacidade tecnológica* para definir "qualquer momento que um sistema tecnológico se engaja em comportamentos que, embora apropriados, podem ser difíceis de entender ou prever, do ponto de vista humano".

Burrell[14] descreveu três formas de opacidade. A primeira é o "sigilo corporativo ou de estado intencional", imposto por uma empresa ou Estado, cujo objetivo é manter uma vantagem competitiva em relação a seus pares. Um segundo nível de opacidade diz respeito ao "analfabetismo técnico", já que escrever e ler um código de algoritmos é uma habilidade especializada. Como a sintaxe da língua humana é diferente das linguagens de programação, há uma dificuldade natural, para o cidadão comum, em compreender códigos computacionais, tornando estes obscuros para a maior parte da população. Por fim, há uma terceira forma de

[12] ANDRADE, Otávio Morato de. **Governamentalidade algorítmica**: democracia em risco? 1. ed. São Paulo: Dialética, 2022.
[13] SURDEN, Harry. Machine learning and law. **Washington Law Review**, [s.l.], v. 89, N. 1, mar. 2014. Disponível em: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5/. Acesso em: 18 jun. 2023. p. 158.
[14] BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, [s.l.], jan.-jun., 2016. p. 4-5.

opacidade, mais desafiadora que as duas anteriores: a complexidade das operações internas dos sistemas de IA. O desafio aqui não consiste na dificuldade de *acesso* ou *leitura do código*, mas *na incapacidade de entender a análise preditiva feita pelo algoritmo*. Isso porque, além de processar um volume incomensurável de dados, o modelo algorítmico frequentemente (re)adapta sua lógica de decisão interna, à medida que "aprende" com os dados de treinamento.

**Figura 3** - Formas de opacidade.




**Fonte**: Elaboração própria a partir de Jenna Burrell[15].

Para as duas primeiras formas de opacidade há soluções palpáveis, como: regulamentação, acesso parcial ao código e a auditoria algorítmica (no caso do sigilo intencional), além dos esforços educacionais e de informação, que visam facilitar a leitura ou traduzir os códigos para o público em geral (no caso do analfabetismo técnico). No caso da complexidade algorítmica, contudo, a simples abertura do código e o aumento da capacidade do público de ler a linguagem computacional estão longe de assegurar a transparência do sistema. Isso porque as operações dos algoritmos podem ser muito complexas, volumosas e heterogêneas, não sendo suficiente o mero acesso ou a capacidade de leitura para a compreensão das decisões ali produzidas *(outputs)*.

[15] BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, [s.l.], jan.-jun. 2016.

## 2.2 Preocupações sobre a opacidade, ética e transparência

As diversas formas de opacidade presentes nestes sistemas geram preocupações de ordem ética e legal[16]. Imagine-se o caso de um veículo autômato dirigido por IA que atropela um transeunte. Ou de um robô que elabora julgamentos enviesados e reproduz comentários racistas. Tal qual ocorre nos erros humanos, será preciso adentrar a esfera das decisões para perquirir a culpabilidade das ações. Todavia, se o processo de decisão de um sistema não é acessível, isso pode dificultar a investigação e até mesmo a responsabilização nos casos em que a IA comete ou colabora para um erro, crime ou injustiça[17].

De acordo com especialistas, a solução não seria afastar os algoritmos da tomada de decisão, mas sim expandir os princípios éticos que norteiam o aprendizado de máquina, de forma "a exigir que eles incorporem - de forma quantitativa, mensurável e verificável – muitos dos valores éticos com os quais nos preocupamos como indivíduos e como sociedade"[18]. Por isso, uma abordagem ética da IA deve estar alinhada com valores sociais como segurança, moralidade e responsabilidade e, em especial, a transparência.

Neste caminho, é impossível tratar a questão da opacidade e dos vieses algorítmicos sem a transparência e, em especial, a necessidade no âmbito jurídico de agregar princípios e normas orientadas para modular a atuação algorítmica nas diversas searas em que são aplicados. Afinal, se os padrões de raciocínio que envolvem os processos de inteligência artificial são "opacos" ao decidir até mesmo para seus programadores, eles constituem uma verdadeira "caixa preta" para a sociedade em geral[19].

[16] Pontue-se, contudo, que a opacidade é irrelevante quando as operações algorítmicas não acarretam riscos ou consequências maiores, como por exemplo nos modelos de desbloqueio de smartphones, eis que nos basta saber se ele libera o dispositivo ao se deparar com nossa biometria e não o faz diante de outro indivíduo.

[17] ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato. Da "caixa-preta" à "caixa de vidro": o uso da Explainable Artificial Intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Revista de Direito Público**, Brasília, v. 18, n. 100, 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 17 jul. 2023.

[18] KERNS, Michael; ROTH, Aaron. **The etichal algorithm**: the science of socially aware algorithm design. Oxford University Press, 2020. p. 15.

[19] Neste sentido, o matemático francês Cédric Villani destaca que: "[e]ste é o famoso problema da caixa preta: sistemas algorítmicos a partir dos quais é possível observar os dados de entrada, os dados de saída, mas cujo funcionamento interno é mal compreendido. Esta falta de conhecimento se deve principalmente hoje à mudança de paradigma introduzida pelo advento da aprendizagem, especialmente a aprendizagem profunda". *In*: VILLANI, Cédric. **Donner uns sens à li'intelligence artificielle**: pour une stratégie nationale et européenne. Paris, 2018. Disponível em: https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/184000159.pdf. Acesso em: 18 jun. 2023. p. 70.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


Com o desenvolvimento do *machine learning*, a necessidade de se interpretar as ações dos sistemas de inteligência artificial tem ficado cada vez mais evidente nos dias atuais. Isso porque o notável desempenho e os "feitos" das mais diversas máquinas que utilizam IA suscitam inseguranças sobre transparência e a confiabilidade destes sistemas[20]. Incertezas essas, intensificadas pela presença de vieses e preconceitos em alguns algoritmos, no âmbito de uma sociedade cada vez mais atenta às questões raciais, inclusão, diversidade e igualdade de gênero[21]. Tais constatações conduzem ao desafio de como explicar as decisões da IA, o qual será enfrentado no capítulo a seguir.

## 3 *EXPLICABILIDADE*

### 3.1 O desenvolvimento da inteligência artificial explicável

No contexto da IA, *explicabilidade* significa compreender melhor os motivos e detalhes por trás de uma decisão algorítmica. Os sistemas de inteligência artificial explicável seriam, portanto, aqueles "capazes de explicar seus fundamentos, caracterizar seus pontos fortes e fracos e transmitir uma compreensão acerca das suas condutas futuras"[22]. Como sucede com as explicações em outros campos da ciência, o entendimento dos processos inerentes à IA precisa usar representações comunicáveis, como, por exemplo, locuções linguísticas ou lógicas, sentenças matemáticas e diagramas visuais[23]. De tal forma, enquanto a opacidade cria uma "caixa preta", que limita o entendimento humano acerca das decisões de um sistema de IA, a

[20] MARCUS, Gary; DAVIS, Ernest. How to build artificial intelligence we can trust. **The New York Times**, [s.l.], 6 de setembro de 2019. Disponível em: https://www.nytimes.com/2019/09/06/opinion/ai-explainability.html. Acesso em: 18 jun. 2023.
[21] NUNES, Dierle José Coelho; MARQUES, Ana Luiza. Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. **Revista de Processo**, São Paulo, v. 43, p. 421 – 447, nov. 2018. Disponível em: https://bd.tjdft.jus.br/jspui/handle/tjdft/43025. Acesso em: 18 jun. 2023.
[22] GUNNING, David. Explainable Artificial Intelligence (XAI) DARPA/I2O. **DARPA** - Defense Advanced Research Projects Agency, [s.l.], 2016. Disponível em: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf. Acesso em: 18 jun. 2023.
[23] MOLNAR, Christoph. **Interpretable machine learning**: a guide for making black box models explainable. [S.l.], 2023. Disponível em: https://christophm.github.io/interpretable-ml-book/index.html. Acesso em: 18 jun. 2023.

*explicabilidade* produziria o oposto, ou seja, uma "caixa de vidro" que permite a compreensão dos processos internos por trás de um *output* algorítmico[24].

Gerar explicações sobre IA tem sido uma questão-chave desde as décadas de 1960-1970. Na década de 1970, o *software MYCIN*, criado na Universidade de Stanford para identificar bactérias e recomendar antibióticos, permitia ao usuário perguntar "Por quê?" e "Como?" o programa chegou em determinadas conclusões[25]. Mais adiante, em 1980, pesquisadores de Stanford desenvolveram o *CENTAUR*, sistema arquitetado para oferecer "representação explícita" dos problemas resolvidos, visando fornecendo detalhes sobre as hipóteses, as inconsistências e os erros encontrados durante a operação algorítmica[26]. De acordo com Alun Preece[27], isso fez com que o CENTAUR fosse "o primeiro sistema de inteligência artificial projetado para ser explicado".

Visando ampliar as explicações dadas pela IA, a DARPA desenvolveu, na década de 1990, um projeto dirigido à geração de explicação nos sistemas de IA. O *Explainable Expert Systems* (EES) era enriquecido com "conhecimento estratégico explícito" sobre diferentes áreas do conhecimento humano, e além de fornecer a cadeia reversa da lógica utilizada para chegar em uma solução, este conhecimento também era utilizado para fornecer explicações sobre o raciocínio do sistema. Desta forma, era capaz de produzir uma representação abstrata do conhecimento estratégico ("como uma ação X do sistema se relaciona com o objetivo geral?") e na representação da lógica do design ("por quê a ação Y é razoável em vista do objetivo?")[28].

### 3.2 A explicação depende do tipo de modelo algorítmico

[24] NUNES, Dierle José Coelho; ANDRADE, Otávio. A explicabilidade da inteligência artificial e o devido processo tecnológico. **Revista Conjur**, São Paulo, 7 de julho de 2021. Disponível em: https://www.conjur.com.br/2021-jul-07/opiniao-explicabilidade-ia-devido-processo-tecnologico. Acesso em: 18 jun. 2023.

[25] PREECE, Alun. Asking 'Why' in ai: explainability of intelligent systems – perspectives and challenges. **Intelligent Systems**, [s.l.], 2018. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.1422. Acesso em: 18 jun. 2023.

[26] AIKINS, Janice. **Prototypes and production rules**: a knowledge representation for computer consultations. 1980. 112 f. Tese (Doutorado em Ciência da Computação) – Departament of Computer Science, Stanford University, California. 1980. Disponível em: https://apps.dtic.mil/sti/pdfs/ADA091177.pdf. Acesso em: 17 jun. 2023.

[27] PREECE, Alun. Asking 'Why' in ai: explainability of intelligent systems – perspectives and challenges. **Intelligent Systems**, [s.l.], 2018. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.1422. Acesso em: 18 jun. 2023.

[28] BARZILAY, Regina *et al*. A new approach to expert system explanations. Associoation for Computational Linguistics, Canadá, natual language generation, 1998. Disponível em: https://aclanthology.org/W98-1409/. Acesso em: 18 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


Existem modelos algorítmicos que são "auto"-interpretáveis e existem aqueles que necessitam ser explicados por meio de técnicas externas de *explicabilidade*. Um modelo interpretável (ou transparente) de *machine learning* é aquele que não requer técnicas adicionais para que o humano possa compreendê-lo. Entre estes sistemas encontram-se: regressão linear/logística, árvores de decisão, k-vizinhos mais próximos, RBML (*Rule-based machine learning*), GAM (*generalized additive model*) e os modelos bayesianos[29].

O fato de um modelo algorítmico ser interpretável, não significa, todavia, que ele dispense a *explicabilidade*. Pelo contrário, sua transparência tornará a *explicabilidade* ainda mais viável. Tome-se como exemplo a árvore de decisão. Os modelos baseados em árvore dividem os dados várias vezes de acordo com determinados valores de corte nos recursos. Por meio da divisão, diferentes subconjuntos do conjunto de dados são criados, com cada instância pertencendo a um subconjunto. As previsões individuais de uma árvore de decisão podem ser explicadas decompondo-se o caminho de decisão em um componente por recurso. Pode-se rastrear uma decisão por meio da árvore e explicar uma previsão pelas contribuições adicionadas em cada nó de decisão.

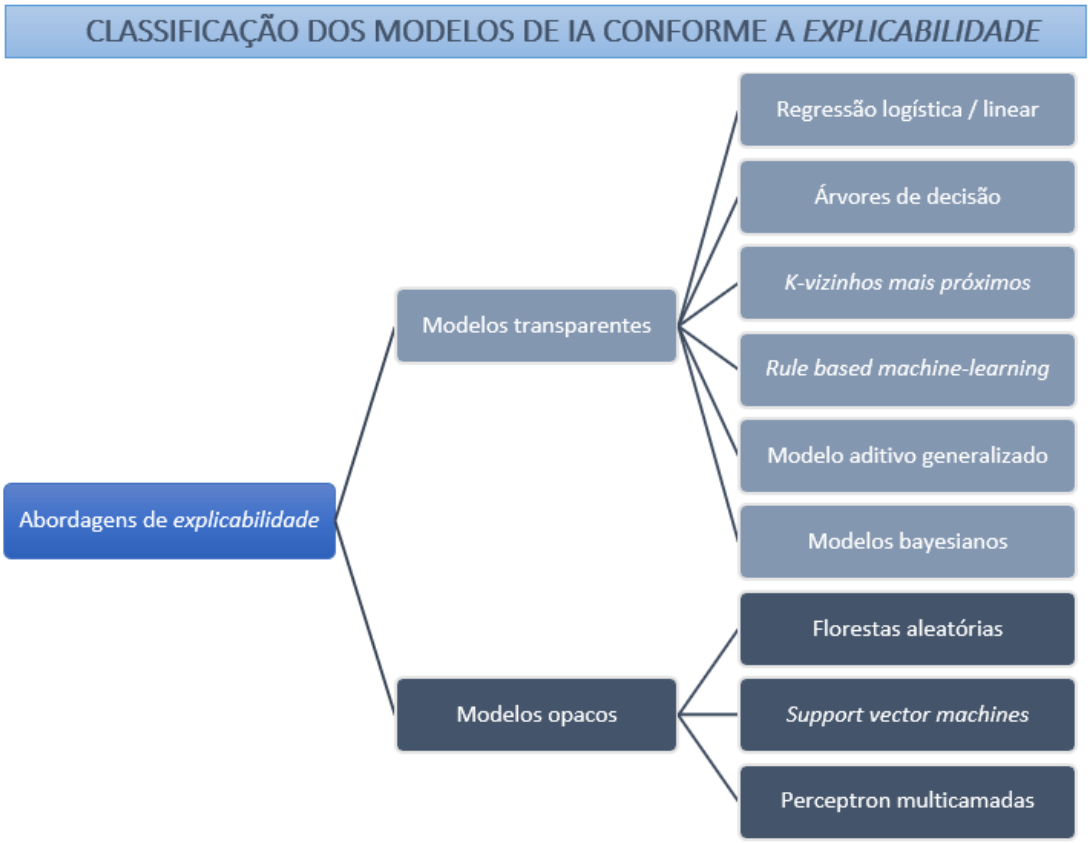
**Figura 4** - A árvore de decisão simples é um exemplo de 'modelo transparente'. A cada pergunta, ela classifica os elementos analisados.


**Fonte**: Elaboração própria.

[29] MOLNAR, Christoph. **Interpretable machine learning**: a guide for making black box models explainable. [S.l.], 2023. Disponível em: https://christophm.github.io/interpretable-ml-book/index.html. Acesso em: 18 jun. 2023.

Por outro lado, em oposição aos "modelos transparentes", existem os "modelos opacos", cuja compreensão irá requerer um processo de explicação adicional, chamado de "*explicabilidade post-hoc*". Entre esses modelos opacos, é possível apontar as *Support Vector Machines* (SVM), o Perceptron multicamadas e as florestas de decisão aleatórias (*random forests*). Esta divisão está esquematizada no diagrama a seguir:

**Figura 5** - Classificação dos modelos de IA conforme as abordagens de *explicabilidade* utilizadas.




**Fonte**: Elaborado a partir de Ionnis Papantonis e Vaishak Belle[30].

[30] PAPANTONIS, Ioannis; BELLE, Vaishak. Principles and practice of explainable machine learning. **ArXiv**, [s.l.], v. 1, set. 2020. Disponível em: https://arxiv.org/pdf/2009.11698.pdf. Acesso em: 18 jun. 2023. p. 6.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


### 3.3 Métodos de *explicabilidade post-hoc*

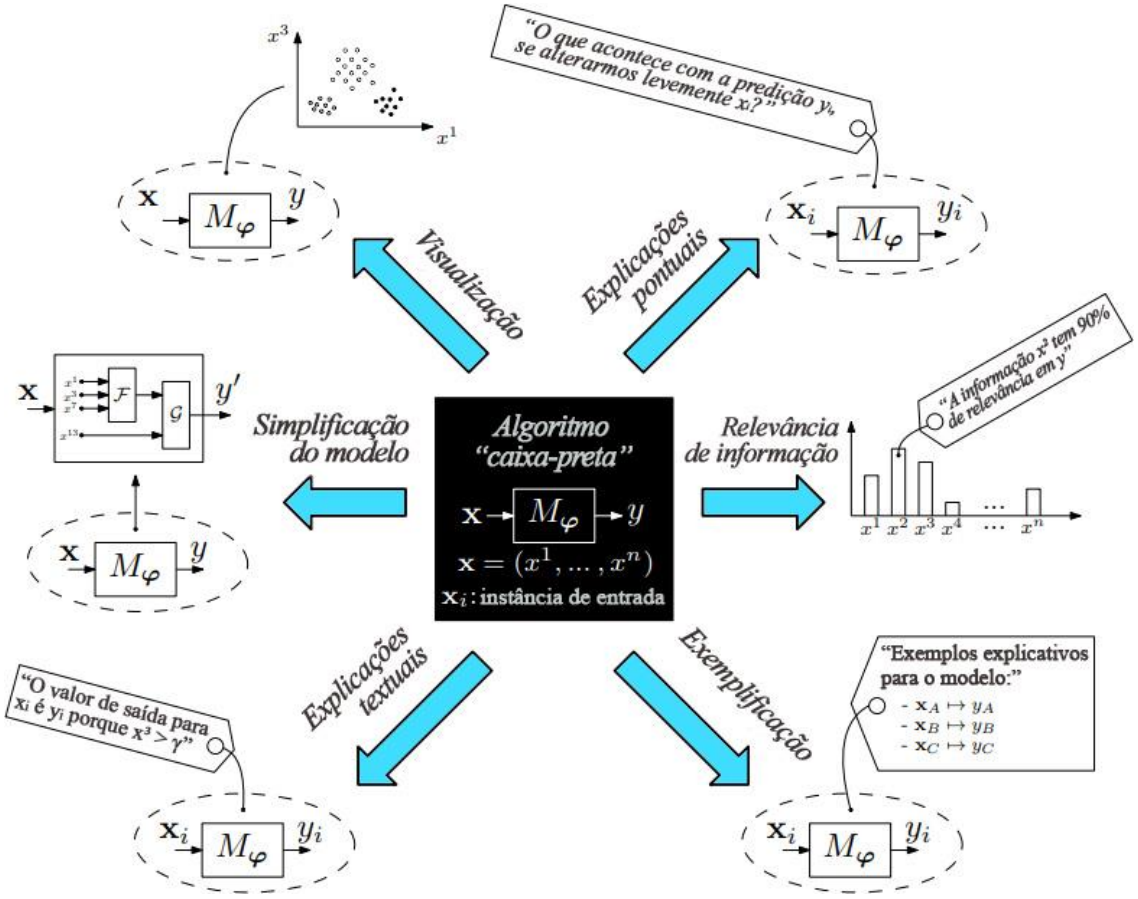
Alguns modelos de *machine learning* são bem mais difíceis de interpretar. Nestes casos, será necessária a aplicação de um ou mais métodos para que o modelo seja capaz de explicar suas próprias decisões. Este é o propósito das técnicas de *explicabilidade post-hoc* (também conhecida como "*explicabilidade* pós-modelagem"), que visam comunicar informações compreensíveis sobre como um modelo já desenvolvido produz suas previsões para qualquer entrada. A *explicabilidade post-hoc* se volta para modelos que não são prontamente interpretáveis pelo seu design, recorrendo a diversos meios e ferramentas efetuar a *explicabilidade*, tais como: explicações de texto, explicações visuais, explicações através de exemplos, explicações por simplificação e explicações de relevância de recurso técnicas[31].

As técnicas de *explicabilidade post-hoc* ainda se dividem em duas categorias: as técnicas *gerais* ou *agnósticas*, ou seja, aquelas que se aplicam genericamente a qualquer modelo de *machine learning*, e as técnicas *específicas*, desenvolvidas para interpretar um único modelo, razão pela qual não é possível extrapolá-las para interpretar outros modelos. O presente trabalho enfoca somente as técnicas *agnósticas*, pois a partir delas já é possível se obter uma noção geral sobre o funcionamento da inteligência artificial explicável, que é bastante satisfatório para subsidiar a discussão proposta por este artigo. Ficam fora desta análise, portanto, as técnicas específicas de *explicabilidade*, as quais são desenhadas considerando as particularidades de cada modelo sua respectiva linguagem de programação.

[31] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.

**Figura 6** - Diagrama conceitual mostrando as diferentes abordagens de *explicabilidade* post-hoc disponíveis para um modelo *machine learning*.

**TÉCNICAS GERAIS USADAS NA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL**




**Fonte**: Elaborado a partir de Alejandro Arrieta *et all*[32].

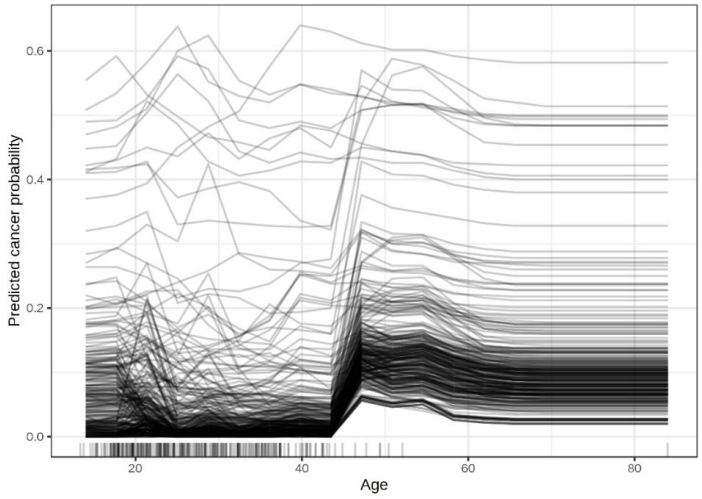
Dentre as técnicas agnósticas de *explicabilidade* – ou seja, aquelas de podem ser utilizadas de maneira mais genérica, aplicando-se a diferentes modelos – as *explicações textuais* visam ilustrar o processo de funcionamento do algoritmo através da linguagem escrita ou de símbolos. Não se trata de exibir a íntegra do código, mas de apresentar, em semântica mais compreensível para o humano (por exemplo, através de frases ou pequenos comentários), os resultados e o funcionamento do modelo a ser interpretado[33].

[32] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023. p. 13.
[33] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


Em segundo lugar, as *explicações por simplificação* consistem num dos métodos mais populares de *explicabilidade*, tendo por estratégia a extração de regras para resumir o processo preditivo. Se um algoritmo usou correlações complexas e várias camadas de raciocínio, uma explicação por simplificação eficaz tende a selecionar as mais importantes, exibindo ao usuário uma versão mais descomplicada que permite entender melhor o processo preditivo[34]. A terceira técnica a ser apresentada, *explicação visual*, é aquela na qual o algoritmo produz um gráfico, tabela ou outro esquema visual que torne mais compreensível as suas previsões. A explicação visual é menos comum do que as demais, sendo geralmente associada a outras técnicas de explicação. Apesar disso, há exemplos de sua utilização, como os gráficos ICE (*Individual Conditional Expectation*), que mostram visualmente a relação entre um recurso e uma previsão.

**Figura 7** - Gráfico ICE da probabilidade de câncer cervical por idade. Cada linha representa uma mulher. Para a maioria das mulheres, há aumento na probabilidade prevista de câncer com o aumento da idade. Para algumas mulheres com probabilidade prevista de câncer acima de 0,4, a previsão não muda muito com o aumento da idade.




**Fonte**: Retirado de Christoph Molnar[35].

[34] Entre suas principais modalidades estão o G-REX e as LIME *(Local Interpretable Model-Agnostic Explanations)* e suas variações, que constroem modelos lineares localmente em torno das previsões de um modelo opaco para explicá-lo. O LIME, por exemplo, utiliza recursos matemáticos como a regressão linear para simplificar o modelo a ser explicado, criando um "modelo substituto" que auxilia no entendimento do processo preditivo original. *In*: RIBEIRO, Marco Túlio; SINGH Sameer; GUESTRIN, Carlos. "Why should i trust you?" explaining the predictions of any classifier**. arXiv**, [s.l.], v. 1, fev. 2016. Disponível em: https://cardiacmr.hms.harvard.edu/files/cardiacmr/files/ribeiro_et_al._arxiv_2016.pdf. Acesso em: 18 jun. 2023.

[35] MOLNAR, Christoph. **Interpretable machine learning**: a guide for making black box models explainable. [S.l.], 2023. Disponível em: https://christophm.github.io/interpretable-ml-book/index.html. Acesso em: 18 jun. 2023.

Uma outra técnica de *explicabilidade* que pode ser aplicada a vários modelos é a de *relevância da informação*, que consiste em revelar o peso de cada variável em uma decisão algorítmica. Por exemplo: imagine-se o caso em que o sistema de um banco nega um empréstimo a determinado cliente. Usando a técnica de relevância da informação, o algoritmo deveria revelar quais fatores foram decisivos para esta decisão. Por exemplo: "a informação de que 'o cliente possuía um financiamento não quitado' foi 70% relevante para a não concessão do crédito". Essa técnica pode ser extremamente útil para revelar vieses em modelos algoritmos preconceituosos[36].

A quinta técnica passível de ser usada em inteligência artificial explicável é o método das *explicações pontuais*. As explicações pontuais buscam fornecer explicações a segmentos especificamente importantes (chamados "pontuais" ou "locais") do modelo, destacando uma parte do funcionamento do sistema que é relevante para os resultados. Isso porque, em alguns casos, a explicação de um setor "crucial" do sistema pode ser suficiente para a compreensão do funcionamento do sistema[37]. A sexta e última técnica de *explicabilidade* é a exemplificação. Neste método, o algoritmo extrai amostras de dados análogos ou similares aos resultados gerados pelo modelo, permitindo que o usuário compreenda, por analogia, as relações e correlações encontradas pelo modelo algorítmico[38].

## 4 UMA INTELIGÊNCIA ARTIFICIAL MAIS TRANSPARENTE E CONFIÁVEL

A explicação das decisões mediadas por IA traz vantagens significativas, tanto para a sociedade como para as empresas que utilizam sistemas inteligentes. No âmbito empresarial, a *explicabilidade* assegura, por exemplo, a melhoria da conformidade legal dos sistemas utilizados, reduzindo os riscos jurídicos associados ao descumprimento de normas regulatórias sobre inteligência artificial. A *explicabilidade* também aumenta a confiança dos funcionários e

[36] ANDRADE, Otávio Morato de. **Governamentalidade algorítmica**: democracia em risco? 1. ed. São Paulo: Dialética, 2022.
[37] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.
[38] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.

clientes na IA, na medida em que possibilita um melhor entendimento dos processos autômatos, demonstrando uma postura respeitosa da empresa para com seus *stakeholders*[39].

Da mesma maneira, a *explicabilidade* é essencial para os usuários e para a sociedade. Primeiro, porque o maior conhecimento público acerca dos processos algorítmicos propicia um debate mais informado sobre o implemento e o avanço das novas tecnologias. Em segundo lugar, porque ela possibilita a melhoria das decisões de IA, ajudando a eliminar *outputs* discriminatórios e a mitigar vieses de modelos algorítmicos. Em artigo recente sobre o tema, Marco Antônio Sousa Alves e Otávio Morato de Andrade[40] demonstraram aplicações práticas da inteligência artificial explicável no combate ao preconceito algorítmico, concluindo que:

> Considerada a relevância de determinadas decisões, elas devem ser delegadas apenas a algoritmos aptos a esclarecer suas intenções e motivações, explicitando sua análise em linguagem compreensível para o ser humano. Quanto mais importante for uma decisão do ponto de vista social, mais capacitado precisa estar um sistema de IA para fornecer explicações detalhadas, precisas e compreensíveis, para que não pairem dúvidas sobre a sua neutralidade e competência[41].

Entendimento similar é o de Fausto Santos de Morais, que assevera que "quanto maior for o grau de potência lesiva aos direitos fundamentais da tecnológica utilizada, maior deve ser a extensão da clareza sobre o seu uso e a explicação sobre o seu funcionamento"[42]. De fato, com o avanço das novas tecnologias e a expansão dos sistemas de apoio de decisão de IA na sociedade, há quem defenda que a *explicabilidade* precisaria ser condição *sine qua non* para a legitimação de tais decisões. Nesta linha, pondera Cédric Villani:

[39] EXPLAINING decisions made with AI. ICO - INFORMATION COMMISSIONER'S OFFICE. Londres, 2020. Disponível em: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/. Acesso em: 18 jun. 2023. p. 16-17.

[40] ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato. Da "caixa-preta" à "caixa de vidro": o uso da Explainable Artificial Intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Revista de Direito Público**, Brasília, v. 18, n. 100, 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 17 jul. 2023.

[41] ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato. Da "caixa-preta" à "caixa de vidro": o uso da Explainable Artificial Intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Revista de Direito Público**, Brasília, v. 18, n. 100, 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 17 jul. 2023. p. 369.

[42] MORAIS, Fausto Santo de. O uso da inteligência artificial na repercussão geral: desafios teóricos e éticos. **Revista de Direito Público**, Brasília, v. 18, n. 100, p. 306-326, 2021. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/6001/pdf. Acesso em: 18 jun. 2023. p. 323.

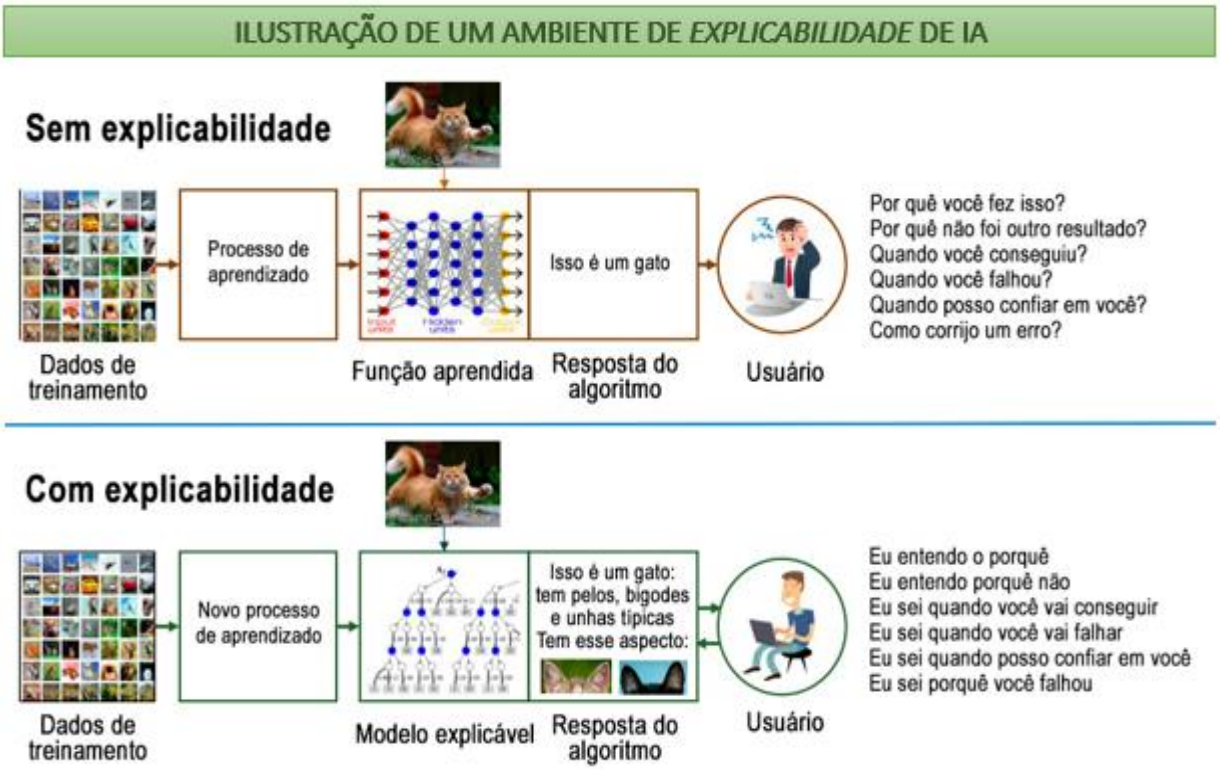
> A longo prazo, a explicabilidade dessas tecnologias é um dos requisitos de sua aceitação social. No que concerne a certos temas, é mesmo uma questão de princípio: como sociedade, não se pode aceitar que certas decisões importantes sejam tomadas sem explicação. Com efeito, sem a oportunidade de explicar as decisões tomadas por sistemas autônomos, parece difícil justificá-las. Mas como aceitar o injustificável em áreas tão decisivas para a vida de um indivíduo quanto o acesso ao crédito, emprego, moradia, justiça ou saúde? Parece inconcebível[43].

A *explicabilidade*, desse modo, oferece uma gama de benefícios ao usuário final, às empresas, aos desenvolvedores e à sociedade de modo geral, principalmente na atual conjuntura de "digitalização da vida", em que os algoritmos têm um imenso volume de dados à sua disposição e grande poder para influenciar condutas individuais e estruturar o campo das ações possíveis[44]. O esquema a seguir oferece uma noção do que ocorre em um ambiente com e sem inteligência artificial explicável. Em ambos os casos, o algoritmo analisa a mesma fotografia, concluindo tratar-se de um gato. No entanto, no segundo cenário a inteligência artificial explicável oferece informações adicionais sobre como a IA chegou em tal conclusão. O usuário é claramente beneficiado pela explicação fornecida pelo algoritmo, que torna a resposta da máquina mais segura e confiável.

**Figura 8** - A imagem ilustra a diferença de ambientes com/sem *explicabilidade* para o usuário final. Uma interface explicável oferece mais segurança sobre os resultados do modelo algorítmico, permitindo, ao mesmo tempo, uma compreensão acerca das suas vulnerabilidades ou imprecisões.

[43] VILLANI, Cédric. **Donner uns sens à li'intelligence artificielle**: pour une stratégie nationale et européenne. Paris, 2018. Disponível em: https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/184000159.pdf. Acesso em: 18 jun. 2023. p. 78.
[44] ALVES, Marco Antônio Sousa. Cidade inteligente e governamentalidade algorítmica: liberdade e controle na era da informação. **Philósophos**, Goiânia, v. 23, n. 2, 2019. Disponível em: https://revistas.ufg.br/philosophos/article/view/52730. Acesso em: 17 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


**Fonte**: Elaborado a partir de David Gunning[45].

Em revisão bibliográfica, Arrieta *et al*[46] identificaram os principais objetivos a serem perseguidos para se alcançar a *explicabilidade*. São eles: a) causalidade, que explicita a correlação causal entre as variáveis envolvidas no processo de decisão; b) transferibilidade, relacionada à capacidade de aplicação de um mesmo modelo de IA para diferentes sistemas e quais as limitações disso; c) informatividade, que consiste em oferecer informações detalhadas sobre o problema enfrentado pela máquina, uma vez que o problema resolvido pelo modelo nem sempre coincide com o problema enfrentado pelo usuário; d) confiança, na qual o modelo deve fornecer informações para que o usuário avalie a robustez e a estabilidade do regime de trabalho e das decisões do sistema; e) equidade, que permite uma análise "justa" ou ética das decisões tomadas pelo modelo; f) acessibilidade, relacionada à capacidade dos usuários finais de

[45] GUNNING, David. Explainable Artificial Intelligence (XAI) DARPA/I2O. **DARPA** - Defense Advanced Research Projects Agency, [s.l.], 2016. Disponível em: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf. Acesso em: 18 jun. 2023.

[46] ARRIETA, Alejandro *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.

se envolverem no processo e melhoria e desenvolvimento do sistema; g) interatividade, no qual um modelo explicável pode interagir com o usuário final, aperfeiçoando o processo de compreensão; e h) conscientização da privacidade, que possibilita ao usuário ter um entendimento sobre possíveis violações de seus dados pessoais pelo algoritmo.

Paralelamente, entende-se que a *explicabilidade* é capaz de fornecer seis diferentes tipos de abordagem, a depender da perspectiva a ser privilegiada na explicação: i) de justificativa, relacionada aos motivos que levaram a uma decisão; ii) de responsabilidade, que se refere aos agentes envolvidos no desenho, gestão e implementação do sistema; iii) de dados, que explica quais foram os dados processados; iv) de imparcialidade, relativa aos cuidados tomados no projeto para garantir que as decisões fossem neutras e justas; v) de segurança e desempenho, concernentes aos processos que visam otimizar a precisão e confiabilidade das decisões e comportamentos; vi) de impacto, atinente às precauções de monitoramento dos impactos que o uso de um sistema de inteligência artificial e suas decisões têm ou podem ter sobre um indivíduo e na sociedade em geral[47].

Muito se discute acerca do direito de obtenção de informações mais completas sobre as decisões de um modelo algorítmico, garantia que tem sido chamada de "direito à explicação". Neste sentido, o Regulamento Geral sobre a Proteção de Dados da União Europeia, subscrito em 2016 e aplicado desde 2018, trouxe diversas garantias sobre o direito do cidadão à explicação, dentre as quais os arts. 13 e 14 – que asseguram ao titular de dados acesso às informações significativas sobre a lógica envolvida quando seus dados são tratados de forma automatizada – e o art. 22, que protege o titular de dados de decisões exclusivamente automatizadas. Ainda, é relevante destacar que nos *Recitals* (espécie de "exposição de motivos"), o Item 71 reforça que o titular de dados tem o direito de não ser submetido a decisões automatizadas, a não ser em casos expressamente permitidos pelo Estado, como nos casos de combate a fraudes e evasão fiscal[48].

[47] EXPLAINING decisions made with AI. ICO - INFORMATION COMMISSIONER'S OFFICE. Londres, 2020. Disponível em: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/. Acesso em: 18 jun. 2023. p. 21-32.

[48] O dispositivo ainda acrescenta que: Em qualquer caso, esse tratamento deve estar sujeito a salvaguardas adequadas, que devem incluir informações específicas sobre o titular dos dados e o direito de obter intervenção humana, de expressar o seu ponto de vista, de **obter uma explicação da decisão tomada após tal avaliação e contestar a decisão** *In*: UNIÃO EUROPEIA. **Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016**. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0001.01.POR&toc=OJ%3AL%3A2016%3A119%3AFULL. Acesso em: 18 jun. 2023. (Grifo nosso).

ISSN 1981-3694
(DOI): 10.5902/1981369469329


Na legislação brasileira, destaca-se a recente entrada da matéria no ambiente regulatório nacional, por meio da Resolução n. 332/2020, do CNJ – Conselho Nacional de Justiça, que passou a exigir, em seu art. 5º, inciso VI, o "fornecimento de explicação satisfatória e passível de auditoria por autoridade humana", estabelecendo a *explicabilidade* como atributo intrínseco à transparência dos sistemas de IA[49]. Além disso, o Projeto de Lei n. 21/2020, aprovado pela Câmara dos Deputados e atualmente em tramitação no Senado, visa estabelecer a transparência a *explicabilidade* como princípios para o uso responsável de inteligência artificial[50]. O desenvolvimento do "direito à explicação" e o aperfeiçoamento das técnicas e premissas da *explicabilidade* podem beneficiar os indivíduos e à sociedade como um todo, haja vista estimularem a criação de sistemas mais transparentes e interpretáveis, elevando a confiabilidade e a legitimidade das decisões produzidas por máquinas. Além disso, também pode beneficiar desenvolvedores e empresas, pois auxilia na otimização dos próprios sistemas, possibilitando a correção de erros e a otimização de suas funcionalidades.

## CONCLUSÃO

Muito embora o *machine e o deep learning* possibilitem a automatização de uma série de tarefas, eles nem sempre geram decisões perfeitas, equânimes e imparciais. Exemplo disso é a susceptibilidade dos sistemas de IA aos chamados vieses de modelo algorítmico, ou seja, "preconceitos cognitivos" que comprometem a imparcialidade de suas decisões. Atualmente, a maioria dos sistemas de IA são opacos, o que impede que o processo de formação decisória de seus algoritmos seja facilmente compreensível por seres humanos.

[49] CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Resolução n. 332, de 21 de agosto de 2020**. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências. Brasília: Conselho Nacional de Justiça, [2020]. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429. Acesso em: 18 jun. 2023.

[50] De acordo com o art. 6º, do Projeto de Lei n. 21/2020: "são princípios para o uso responsável de inteligência artificial no Brasil: [...] IV - transparência e *explicabilidade*: garantia de transparência sobre o uso e funcionamento dos sistemas de inteligência artificial e de divulgação responsável do conhecimento de inteligência artificial, observados os segredos comercial e industrial, e de conscientização das partes interessadas sobre suas interações com os sistemas, inclusive no local de trabalho". *In*: BRASIL. **Projeto de Lei n. 21 de 2020**. Estabelece pincípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá proviências. Brasília: Câmara dos Deputados, [2020]. Disponível em: https://www.camara.leg.br/propostas-legislativas/2236340. Acesso em: 18 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


Tendo em vista as preocupações discutidas neste trabalho, sugere-se a transformação dessas "caixas-pretas" (ou seja, dos algoritmos opacos) em "caixas de vidro" nas quais prevalece a transparência, de modo que os modelos computacionais possam, de forma ativa, explicar e justificar seus processos de raciocínio. Neste cenário, acredita-se que a *explicabilidade* pode ajudar a reconhecer vieses, apurar responsabilidades e tornar os sistemas mais confiáveis e seguros. Os primeiros estudos e experimentos sobre *explicabilidade* surgiram nas décadas de 1960-1970, e o avanço nesta área possibilitou o desenvolvimento de princípios e abordagens para favorecer a interpretação dos modelos de IA. Contudo, o desafio de compreender as decisões da IA amenta na medida que os algoritmos se complexificam por meio de novas possibilidades computacionais.

Considerando a complexidade de técnicas de *machine learning*, diferenciam-se modelos transparentes (aqueles que são originalmente interpretáveis) da *explicabilidade post-hoc* (que compreende técnicas que visam explicar satisfatoriamente modelos mais complexos). No interior da *explicabilidade post-hoc*, concentramo-nos nas "técnicas agnósticas", que podem ser aplicadas, genericamente, à vários modelos de aprendizado de máquina. Não se deve olvidar, contudo, que além dos métodos agnósticos existem também as "técnicas específicas" de *explicabilidade*, que criam explicações partindo das particularidades de cada modelo algorítmico.

Após a conceituação dos institutos relativos à IA, a apresentação das noções essenciais à compreensão da opacidade dos sistemas autômatos e de uma análise de atributos e técnicas inerentes à inteligência artificial explicável, conclui-se que o desenvolvimento de uma melhor compreensão acerca do funcionamento dos modelos algorítmicos pode subsidiar o debate público acerca da conciliação das novas tecnologias com a ordem democrática, contribuindo na regulação dos sistemas existentes e na fixação de parâmetros legais e éticos para o design daqueles que ainda serão concebidos.

Neste sentido, a criação de sistemas inteligentes que incorporem a *explicabilidade* deve ser considerada e estimulada na formulação de políticas públicas, de modo a elevar a confiabilidade e a legitimidade das decisões produzidas por sistemas inteligentes. Portanto, é indispensável pautar a discussão sobre *explicabilidade* no meio acadêmico e no debate público, tendo em vista o crescente impacto da IA na sociedade e as inúmeras consequências éticas e jurídicas que sua implementação acarreta.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


# REFERÊNCIAS

AIKINS, Janice. **Prototypes and production rules**: a knowledge representation for computer consultations. 1980. 112 f. Tese (Doutorado em Ciência da Computação) – Departament of Computer Science, Stanford University, California. 1980. Disponível em: https://apps.dtic.mil/sti/pdfs/ADA091177.pdf. Acesso em: 17 jun. 2023.

ALVES, Marco Antônio Sousa. Cidade inteligente e governamentalidade algorítmica: liberdade e controle na era da informação. **Philósophos**, Goiânia, v. 23, n. 2, 2019. Disponível em: https://revistas.ufg.br/philosophos/article/view/52730. Acesso em: 17 jun. 2023.

ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato. Da "caixa-preta" à "caixa de vidro": o uso da Explainable Artificial Intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Revista de Direito Público**, Brasília, v. 18, n. 100, 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 17 jul. 2023.

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

ARENS, Bob. **Cognitive computing**: under the hood. Thomson Reuters, [s.l.], Jan. 2017.

ARRIETA, Alejandro et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [s.l.], v. 58, 2019. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 17 jun. 2023.

BARZILAY, Regina *et al*. A new approach to expert system explanations. **Associoation for Computational Linguistics**, Canadá, v. natual language generation, 1998. Disponível em: https://aclanthology.org/W98-1409/. Acesso em: 18 jun. 2023.

BRASIL. **Projeto de Lei n. 21 de 2020**. Estabelece pincípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá proviências. Brasília: Câmara dos Deputados, [2020]. Disponível em: https://www.camara.leg.br/propostas-legislativas/2236340. Acesso em: 18 jun. 2023.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, [s.l.], jan.–jun., 2016.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Resolução n. 332, de 21 de agosto de 2020**. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências. Brasília: Conselho Nacional de Justiça, [2020]. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429. Acesso em: 18 jun. 2023.

ESTEVA, Andre *et al*. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, [s.l.], n. 542, p. 115–118, 2017. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28117445/. Acesso em: 18 jun. 2023.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


EXPLAINING decisions made with AI. ICO - INFORMATION COMMISSIONER'S OFFICE. Londres, 2020. Disponível em: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/. Acesso em: 18 jun. 2023.

GUNNING, David. Explainable Artificial Intelligence (XAI) DARPA/I2O. **DARPA** - Defense Advanced Research Projects Agency, [s.l.], 2016. Disponível em: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf. Acesso em: 18 jun. 2023.

HIDDEN LAYER. DeepAI, [s.l.]. Disponível em: https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning. Acesso em: 18 jun. 2023.

KERNS, Michael; ROTH, Aaron. **The etichal algorithm**: the science of socially aware algorithm design. Oxford University Press, 2020.

MARCUS, Gary; DAVIS, Ernest. How to build artificial intelligence we can trust. **The New York Times**, [s.l.], 6 de setembro de 2019. Disponível em: https://www.nytimes.com/2019/09/06/opinion/ai-explainability.html. Acesso em: 18 jun. 2023.

MOLNAR, Christoph. **Interpretable machine learning**: a guide for making black box models explainable. [S.l.], 2023. Disponível em: https://christophm.github.io/interpretable-ml-book/index.html. Acesso em: 18 jun. 2023.

MORAIS, Fausto Santo de. O uso da inteligência artificial na repercussão geral: desafios teóricos e éticos. **Revista de Direito Público**, Brasília, v. 18, n. 100, p. 306-326, 2021. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/6001/pdf. Acesso em: 18 jun. 2023.

MOROZOV, Evgeny. **Big tech**: a ascensão dos dados e a morte da política. São Paulo: Ubu, 2018.

NUNES, Dierle José Coelho; ANDRADE, Otávio. A explicabilidade da inteligência artificial e o devido processo tecnológico. **Revista Conjur**, São Paulo, 7 de julho de 2021. Disponível em: https://www.conjur.com.br/2021-jul-07/opiniao-explicabilidade-ia-devido-processo-tecnologico. Acesso em: 18 jun. 2023.

NUNES, Dierle José Coelho; MARQUES, Ana Luiza. Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. **Revista de Processo**, São Paulo, v. 43, p. 421–447, nov. 2018. Disponível em: https://bd.tjdft.jus.br/jspui/handle/tjdft/43025. Acesso em: 18 jun. 2023.

OSÓRIO, Fernando. Redes Neurais - Aprendizado Artificial. **Forum de I.A**. [S.l.]. Disponível em: http://osorio.wait4.org/oldsite/IForumIA/fia99.pdf. Acesso em: 18 jun. 2023.

PAPANTONIS, Ioannis; BELLE, Vaishak. Principles and practice of explainable machine learning. **ArXiv**, [s.l.], v. 1, set. 2020. Disponível em: https://arxiv.org/pdf/2009.11698.pdf. Acesso em: 18 jun. 2023.

PREECE, Alun. Asking 'Why' in ai: explainability of intelligent systems – perspectives and challenges. **Intelligent Systems**, [s.l.], 2018. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.1422. Acesso em: 18 jun. 2023.

RIBEIRO, Marco Túlio; SINGH Sameer; GUESTRIN, Carlos. "Why should i trust you?" explaining the predictions of any classifier. **arXiv**, [s.l.], v. 1, fev. 2016. Disponível em: https://cardiacmr.hms.harvard.edu/files/cardiacmr/files/ribeiro_et_al._arxiv_2016.pdf. Acesso em: 18 jun. 2023.

ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação? **Revista Eco Pós**, v. 18, n. 2, p. 35-56, 2015. Disponível em: https://revistaecopos.eco.ufrj.br/eco_pos/article/view/2662. Acesso em: 18 jun. 2023.

RUSSEL, Stuart; NORVIG, Peter. **Artificial intelligence**: a modern approach. New Jersey: Prentice-Hall, 1995.

SILVER, David *et al*. Mastering the game of go without human knowledge. **Nature**, [s.l.], out. 2017. Disponível em: https://www.nature.com/articles/nature24270. Acesso em: 18 jun. 2023.

SURDEN, Harry. Machine learning and law. **Washington Law Review**, [s.l.], v. 89, N. 1, mar. 2014. Disponível em: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5/. Acesso em: 18 jun. 2023.

SUSSKIND, Richard. **The end of lawyers**: rethinking the nature of legal services. [S.l.]: Oxford Uniersity Press, 2010.

UNIÃO EUROPEIA. **Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016**. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0001.01.POR&toc=OJ%3AL%3A2016%3A119%3AFULL. Acesso em: 18 jun. 2023.

VILLANI, Cédric. **Donner uns sens à li'intelligence artificielle**: pour une stratégie nationale et européenne. Paris, 2018. Disponível em: https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/184000159.pdf. Acesso em: 18 jun. 2023.

**Recebido em: 18.02.2022 / Aprovado em: 03.06.2023 / Publicado em: 20.06.2023**

**COMO FAZER REFERÊNCIA AO ARTIGO (ABNT):**

NUNES, Dierle José Coelho; ANDRADE, Otávio Morato de. O uso da inteligência artificial explicável enquanto ferramenta para compreender decisões automatizadas: possível caminho para aumentar a legitimidade e confiabilidade dos modelos algorítmicos? **Revista Eletrônica do Curso de Direito da UFSM**, Santa Maria, RS, v. 18, n. 1, e69329, 2023. ISSN 1981-3694. DOI: http://dx.doi.org/10.5902/1981369469329. Disponível em: https://periodicos.ufsm.br/revistadireito/article/view/69329 Acesso em: dia mês. ano.

ISSN 1981-3694
(DOI): 10.5902/1981369469329


## SOBRE OS AUTORES

**DIERLE JOSÉ COELHO NUNES**
Doutor em direito processual (PUC/MG - Università degli Studi di Roma "La Sapienza"). Mestre em direito processual (PUC/MG). Professor permanente do PPGD da PUC/MG e colaborador do PPGD da UFMG. Professor adjunto na PUC/MG e na UFMG. Secretário Adjunto do Instituto Brasileiro de Direito Processual. Membro fundador do ABDPC. Membro da International Association of Procedural Law, Instituto Iberoamericano de Derecho Procesal, Instituto Panamericano de Derecho Procesal. Diretor executivo do Instituto de Direito Processual–IDPro. Membro da Comissão de Juristas que assessorou no Novo Código de Processo Civil na Câmara dos Deputados. Diretor do Instituto de Direito e Inteligência Artificial (IDEIA). Advogado sócio de CRON Advocacia.

**OTÁVIO MORATO DE ANDRADE**
Doutorando em Direito na UFMG. Mestre em Direito pela UFMG. Pós-graduado em Direito Civil pela PUC/MG. Bacharel em Direito pela UFMG. Bacharel em Ciências Contábeis pela PUC/MG e Bacharel em Administração pela PUC/MG. Advogado inscrito na OAB/MG. Autor do livro 'Governamentalidade algorítmica: democracia em risco?', além de diversos artigos nas áreas de Direito Civil, bioética, tecnologia, inovação, inteligência artificial e economia comportamental. Membro da Comissão de Inteligência Artificial da OAB/MG, sob presidência do Dr. Dierle Nunes. Membro do grupo de estudos SIGA/UFMG, sob coordenação do Prof. Dr. Marco Antônio de Sousa Alves. Ministrou oficinas e palestras no campo do Direito Civil. É editor-chefe da Revista do CAAP, no âmbito da Faculdade de Direito da UFMG.