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

From "Black Box" to "Glass Box": using Explainable Artificial Intelligence (XAI) to Reduce Opacity and Address Bias in Algorithmic Models

MARCO ANTÔNIO SOUSA ALVES[1]
Universidade Federal de Minas Gerais (UFMG).

OTÁVIO MORATO DE ANDRADE[2]
Universidade Federal de Minas Gerais (UFMG).

RESUMO: A inteligência artificial (IA) tem sido utilizada em larga escala em variados domínios, com cada vez mais implicações sociais, éticas e de privacidade. À medida que suas potencialidades e aplicações são expandidas, surgem dúvidas sobre a confiabilidade dos sistemas equipados com IA, particularmente aqueles que empregam técnicas de *deep learning* que podem torná-los verdadeiras "caixas-pretas". A XAI (*explainable artificial intelligence*), ou inteligência artificial explicável, objetiva oferecer informações que ajudam a explicar o processo preditivo de determinado modelo algorítmico. Este artigo se volta especificamente para o estudo da XAI, investigando seu potencial para explicar decisões de modelos algorítmicos e combater o enviesamento dos sistemas de IA. Na primeira etapa do trabalho, é discutida a questão da falibilidade e enviesamento da IA, e como a opacidade agrava esses problemas. Na segunda parte, apresenta-se a inteligência artificial explicável e suas potenciais contribuições para tornar os sistemas mais transparentes, auxiliando no combate aos erros e vieses algorítmicos. Conclui-se que a XAI pode colaborar para a identificação de vieses em modelos algorítmicos, razão pela qual se sugere que a capacidade de "se explicar" – ou seja, a *explicabilidade* – seja um requisito para a adoção de sistemas de IA em searas mais sensíveis, como, por exemplo, o auxílio à tomada de decisão judicial.

PALAVRAS-CHAVE: XAI; inteligência artificial explicável; opacidade algorítmica; transparência.

1 Orcid: http://orcid.org/0000-0002-4885-8773.
2 Orcid: http://orcid.org/0000-0002-0541-7353.

ABSTRACT: Artificial intelligence (AI) has been used on a large scale in different domains, with increasing social, ethical and privacy implications. As their potential and applications expand, concerns arise about the reliability of AI systems, particularly those that use deep learning techniques that can make them true "black boxes". XAI (explainable artificial intelligence) aims to offer information that helps explain the predictive process of a given algorithmic model. This article focuses specifically on the study of XAI, investigating its potential to explain algorithmic model decisions and combat bias in AI systems. In the first stage of the work, the issue of AI fallibility and bias is discussed – and how opacity intensifies these problems. The second part presents explainable artificial intelligence and its potential contributions to make systems more transparent, helping to combat algorithmic errors and biases. It is concluded that XAI can contribute to the identification of biases in algorithmic models, then it is suggested that the ability to "explain" should be a requirement for the adoption of AI systems in sensitive areas, such as help in court decisions.

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

INTRODUÇÃO

Nos últimos anos, sistemas dotados de inteligência artificial (IA) têm invadido nossas vidas e otimizado processos nos mais variados campos, como, por exemplo, nas recomendações de buscas, produtos e preferências do usuário, no atendimento automático feito por *chatbots*, em diagnósticos médicos e em dispositivos "inteligentes" (*smart devices*), como os carros autômatos.

O avanço e a popularização da IA suscitam extensa gama de preocupações, desde aspectos relacionados à privacidade – vulnerável frente ao poder preditivo das *big techs* – passando pelo comprometimento de traços característicos da subjetividade humana (Rouvroy; Berns, 2015), até implicações éticas sobre vieses discriminatórios. Também se trava efervescente debate sobre como novos fenômenos relacionados à IA poderiam afetar a democracia e o pluralismo político, por meio da disseminação de *fake news*, radicalização ideológica ou mesmo da sofisticação de técnicas de censura e vigilância em massa (Sunstein, 2009; Bruno, 2013).

Tais preocupações são intensificadas pelo fato de que o funcionamento interno de determinados algoritmos – em particular aqueles dotados de aprendizado de máquina profundo (*deep learning*[3]) – pode ser um mistério completo para o usuário médio de tecnologia e, não raro, até para aqueles com competências avançadas na área. Enquanto no aprendizado de máquina (*machine learning*[4]) existe uma estrutura de aprendizado estatístico mais enxuta entre a entrada de dados (*input*) e a saída (*output*), no caso do *deep learning*, existem múltiplas camadas de redes neurais que se sobrepõem umas às outras, tornando mais complexa a compreensão do seu raciocínio. Desta forma, quando se trata de sistemas envolvendo múltiplas redes neurais artificiais, até o momento são escassas as ferramentas e técnicas capazes de facilitar o entendimento das decisões tomadas por um algoritmo (Cortiz, 2021).

É preciso considerar que a IA não é uma tecnologia em si, mas sim uma área do conhecimento, formada por diferentes vertentes, incluindo a *machine learning*, que não deve ser vista como sinônimo de "caixa preta". Afinal, nem toda técnica baseada em aprendizado de máquina padece de opacidade algorítmica. As técnicas de *machine learning* supervisionadas, por exemplo, são facilmente explicáveis. Neste artigo, nosso foco será direcionado para as técnicas não supervisionadas, especialmente o *deep learning* ou a aprendizagem profunda, que pode ser compreendida como uma subárea do aprendizado estatístico ou *machine learning*. Nesse domínio, a questão da explicação e a da regulação assumem um contorno bem mais delicado e complexo.

Diante dessa "opacidade algorítmica", ou seja, a incapacidade de se enxergar além do *output* produzido, questiona-se se o ser humano deveria delegar decisões tão importantes a sistemas de IA, nos casos em que estes são incapazes de explicar como chegaram a algumas conclusões. Com efeito, muitos pesquisadores que se debruçam sobre assunto têm considerado que é fundamental equipar os sistemas de IA com funcionalidades capazes de fornecer uma explicação razoável sobre suas predições

3 *Deep learning* é um método de aprendizado de máquina que usa redes neurais artificiais com várias camadas intermediárias entre a camada de entrada (*input*) e a camada de saída (*output*) e, portanto, uma extensa estrutura interna. Esse modelo tem a capacidade de ampliar suas camadas de redes neurais para solucionar o problema enfrentado (Cortiz, 2021).

4 *Machine learning* é um processo no qual um sistema artificial utiliza métodos estatísticos para aprender a partir de exemplos. Por apresentarem uma estrutura mais simples, os algoritmos de *machine learning* tendem a ser mais passíveis de entendimento do que os algoritmos que usam aprendizagem profunda (Cortiz, 2021).

(Villani, 2019, p. 114; Confalonieri *et al.*, 2020). Dentre as opções apresentadas como aptas a fornecer tal explicação, figuram os sistemas que têm sido chamados de "inteligência artificial explicável" (*explainable artificial intelligence – XAI*).

O presente trabalho tem por objetivo investigar, especificamente, a XAI (*explainable artificial intelligence*) como forma de reduzir a opacidade dos modelos algorítmicos. Estudos sobre o tema têm sugerido que sistemas habilitados para XAI, ao reduzir a opacidade, auxiliam na revelação de falhas nos algoritmos, proporcionando a oportunidade de se corrigir, ou ao menos minimizar, o enviesamento de máquina, tornando estes sistemas mais confiáveis (Ribeiro *et al.*, 2016; Nunes; Andrade, 2021; Wells; Bednarz, 2021).

Gostaríamos de deixar claro, contudo, que não pretendemos com este trabalho defender uma solução puramente tecnológica para todos os problemas éticos e políticos trazidos pelo uso de sistemas dotados de inteligência artificial. Está longe de nossa intenção abraçar uma posição tecnófila ingênua, que aposta simplesmente na possibilidade de a tecnologia resolver por si mesma todos os problemas que ela mesmo suscita, por meio apenas de sistemas cada vez mais sofisticados. Precisamos estar atentos aos limites de tal empreitada, identificando as áreas mais sensíveis e fixando balizas claras quanto à forma e, também, quanto à possibilidade mesma do uso da inteligência artificial. Apesar disso, entendemos que é possível aprimorar os sistemas existentes e ampliar sua possibilidade de uso adequado e responsável em diversos domínios, por meio, por exemplo, de mecanismos mais confiáveis e transparentes.

Este trabalho divide-se em duas grandes seções. Primeiramente, é problematizada a questão da falibilidade e dos vieses nos modelos algorítmicos, discorrendo-se sobre alguns casos em que os sistemas de IA forneceram resultados de predição (*outputs*) imprecisos ou equivocados, ou que, mesmo quando corretos, empreenderam um raciocínio que desconsiderou requisitos desejáveis. Na segunda etapa do trabalho, são apresentadas noções fundamentais sobre a XAI, explicitando-se algumas técnicas que atuam para reduzir a opacidade dos sistemas de IA. Ao final desta etapa, são retomados alguns exemplos de falibilidade e vieses elencados na primeira seção, analisando-se como técnicas correlatas de inteligência artificial explicável poderiam endereçar soluções em cada um desses casos. O trabalho conclui que a XAI pode contribuir para a identificação e o tratamento de problemas em modelos algorítmicos, razão pela qual se sugere que a capa-

cidade de "se explicar" – ou seja, a *explicabilidade* – seja um requisito para a adoção de sistemas de IA em searas mais sensíveis, como, por exemplo, o auxílio à tomada de decisão judicial.

1 QUANDO A IA FRACASSA: O PROBLEMA DE FRACASSAR NO ESCURO

1.1 Falibilidade da IA

Por mais sofisticado que seja, um sistema de IA não está isento de produzir resultados imprecisos, incompletos ou tendenciosos. Várias razões podem estar por trás de um resultado preditivo errôneo. Antes de tudo, os dados de entrada (*input*) podem ser inacabados ou conflitantes entre si, gerando ambiguidades para o algoritmo[5] que os analisa. Além disso, a predição computacional pode estar mal calibrada ou insuficientemente treinada, falhando na interpretação desses dados e terminando por fornecer resultados incorretos (Ramos, 2020). Por fim, há ainda os casos em que o algoritmo "acerta" a resposta, lançando mão, no entanto, de raciocínios e aproximações não desejáveis.

A *falibilidade* ocorre quando um sistema fracassa ao correlacionar os dados de modo causal, gerando evidências inconclusivas e ações injustificadas (Rossetti; Angeluci, 2021, p. 8). Um exemplo de falibilidade é a confusão feita por algoritmos classificadores de imagens ao tentarem distinguir entre lobos e cães da raça Husky, principalmente quando há presença de neve na imagem. Como a maioria das fotos de lobos nos dados de treinamento continha um fundo de neve, o algoritmo acabava se pautando pelo ambiente – e não pelas características do animal – para classificar a imagem. Ao analisar um cão Husky retratado contra um fundo de neve, o sistema falhava, classificando-o incorretamente como "lobo" (Ribeiro *et al.*, 2016).

5 Por algoritmo entende-se "um recurso crescente de agência de autoaprendizagem, interativa, autônoma, que permite que artefatos computacionais executem tarefas que, de outra forma, exigiriam que a inteligência humana fosse executada com sucesso" (Floridi; Taddeo, 2018, p. 751 – tradução nossa).

Figura 1: Um cão Husky (à esquerda) é confundido com um lobo, porque o fundo nevado (à direita) era relacionado erroneamente aos lobos. Essa falha se deve a dados de treinamento insuficientes (muitos lobos fotografados contra a neve), o que acabou induzindo o sistema ao erro.

Klaus-Robert Müller e Wojciech Samek também mostram que um algoritmo pode fornecer um resultado aparentemente "correto" ao se valer de um percurso lógico equivocado. Os autores oferecem como analogia o caso do cavalo Clever Hans, que ficou célebre no início do século XX por supostamente realizar operações matemáticas básicas, com mais de 90% de acertos. Posteriormente, descobriu-se que, nas apresentações, o cavalo apenas respondia à linguagem corporal do seu treinador, que fazia os cálculos e "soprava" o resultado correto a Hans por meio de sinais. Ou seja, Clever Hans de fato "acertava" os resultados, mas por motivos estranhos à matemática.

De acordo com Müller e Samek (2019, p. 3), o mesmo pode ocorrer com alguns sistemas de IA. Relata-se o caso de um algoritmo classificador de imagens que venceu vários prêmios na área. Mais tarde, foi averiguado que sua predição, muitas vezes, não detectava o objeto principal. Em vez disso, apenas utilizava correlações e dados indiretos para chegar ao resultado. Embora acertasse frequentemente, verificou-se que o modelo reconhecia barcos pela presença de água, trens pela presença de trilhos e até cavalos pela presença de uma marca d'água de direitos autorais embutida na imagem. Esses "preditores Clever Hans" podem ter bom desempenho em cenários de teste, mas certamente falharão quando implementados no mundo real, onde os objetos reconhecidos muitas vezes estão fora do seu contexto original. Müller e Samek (2019, p. 4 – tradução nossa) acrescentam

que, "se o sistema de IA for uma caixa-preta, será muito difícil desmascarar tais preditores".

Portanto, quando um algoritmo produz uma decisão incorreta – ou mesmo correta, mas apoiada em premissas falsas –, estaremos diante da falibilidade, situação em que o sistema de IA não opera da maneira desejável, seja por razões ligadas ao *design* do algoritmo, seja pela forma como os dados são codificados, coletados, selecionados ou usados para treinar o algoritmo. Muitas vezes, a falibilidade tem efeitos inócuos. No entanto, quando uma falha produzida por algoritmos de IA afeta grupos ou indivíduos, potencialmente gerando resultados tendenciosos ou discriminatórios, ela adquire uma dimensão social, razão pela qual passa a ser tratada como *viés algorítmico*, como veremos no próximo tópico.

1.2 Vieses algorítmicos, preconceito e discriminação

Os *vieses algorítmicos* são tendências eventualmente produzidas por um sistema de IA, refletindo, em geral, a predileção humana por determinados valores, devido a fatores sociais e culturais preexistentes à programação e que circundam os projetistas (Rossetti; Angeluci, 2021). A incorporação dessas tendências em um sistema de IA geralmente não se dá deliberadamente pelos programadores, mas pelo treinamento incorreto do algoritmo ou por desdobramentos inesperados do aprendizado de máquina, resultando no comprometimento da neutralidade do sistema.

Considerando que a IA tem sido usada para deliberar a respeito de questões humanas cruciais, a "contaminação" de determinado algoritmo por uma tendência moral poderia reproduzir preconceitos e criar resultados injustos, como privilegiar um grupo de usuários em detrimento de outros (Najibi, 2020). Neste caso, os vieses vão além de uma simples falha, pois podem gerar repercussões sociais graves, como, por exemplo, o reforço de preconceitos sociais relacionados à raça, ao gênero, à sexualidade ou à etnia, gerando discriminações sistemáticas e injustas.

À medida que a IA avança, relatos de preconceito algorítmico se multiplicam a cada ano na literatura científica e seria impossível, neste breve trabalho, esgotar todas as espécies de ocorrências documentadas até o momento. Nos subitens a seguir, selecionamos e resumimos três casos nos quais o *viés algorítmico* pode ter ajudado a produzir decisões sistematicamente discriminatórias. Esses mesmos casos, assim como o exemplo do Husky relatado anteriormente, serão retomados ao final da segunda seção

deste artigo, a partir da perspectiva da XAI, de modo a demonstrar como a inteligência artificial explicável poderia contribuir para a detecção e mitigação dessas falhas e vieses.

1.2.1 Negros indevidamente classificados com altos índices de reincidência

Um exemplo de sistema de IA capaz de produzir resultados discriminatórios é o Compas – *correctional offender management profiling for alternative sanctions* (em português: perfil de gerenciamento corretivo de infratores para sanções alternativas), ferramenta norte-americana utilizada para estimar o risco de reincidência de prisioneiros no país. Uma "classificação de risco" é elaborada pelo sistema com base em um questionário de 137 perguntas feitas ao réu, em seu histórico criminal, e também segundo a base de dados da plataforma. De modo geral, os dados produzidos pelo Compas ajudam a estipular valores de fiança, informam decisões judiciais sobre a liberdade do réu durante o processo, e, em alguns Estados, possuem ainda maior relevância, podendo embasar a sentença criminal (Angwin *et al.*, 2016).

Em 2016, no entanto, uma análise feita pela organização de jornalismo independente *ProPublica* revelou que o algoritmo utilizado no Compas continha vieses discriminatórios. Os autores do estudo analisaram pontuações de mais de 7.000 prisioneiros na Flórida, concluindo que o algoritmo está mais propenso a classificar, equivocadamente, acusados negros como "prováveis reincidentes" e, por outro lado, enquadrar, também de forma equivocada, acusados brancos como "indivíduos com baixo risco de reincidência" (Nunes; Marques, 2018, p. 6).

Os jornalistas da *ProPublica* apuraram ainda que a Northpointe, empresa responsável pelo sistema, não disponibiliza ao público o algoritmo no qual se baseia o índice de reincidência dos detentos, mas tão somente as perguntas feitas ao indivíduo e utilizadas no cálculo, de modo que o réu não sabe por qual motivo possui um alto ou baixo indicador, ou sequer como suas respostas influenciaram na ponderação do resultado final (Angwin *et al.*, 2016).

1.2.2 Preconceito no Google Fotos

Em 2015, um programador negro expôs uma falha no serviço de fotos da Google, que havia rotulado fotos dele e de um amigo negro como "gorilas". À época, a *big tech* veio a público para se declarar "horrorizada" com a

falha, mostrando-se "comprometida" com o fim dos vieses discriminatórios em algoritmos e prometendo "rápida correção" (Simonite, 2018).

Quase três anos depois, em 2018, a revista estadunidense de tecnologia *Wired* testou o Google Fotos usando uma coleção de mais de 40.000 imagens com diversas espécies de animais. Muito embora o algoritmo tenha apresentado desempenho notável ao reconhecer muitas criaturas, como pandas e poodles, o serviço curiosamente relatou "nenhum resultado" para os termos de pesquisa "gorila", "chimpanzé" e "macaco" (Simonite, 2018). Descobriu-se, portanto, que a "solução" dada pela Google ao problema do Google Fotos havia sido simplória: apesar da grande repercussão da denúncia, a empresa apenas removeu gorilas e alguns outros primatas do dicionário do serviço, em vez de refinar os algoritmos de reconhecimento para corrigir suas falhas e vieses.

A exótica solução ilustra as dificuldades que o Google e outras empresas de tecnologia enfrentam no avanço da tecnologia de reconhecimento de imagem, que, todavia, já é aplicado em áreas extremamente sensíveis, como, por exemplo, no controle de migração, no monitoramento de protestos, na vigilância de aeroportos e no combate ao terrorismo.

1.2.3 Discriminação contra mulheres na concessão de crédito

Um recente relatório elaborado por pesquisadores de Oxford mostrou que mulheres empreendedoras tendem a captar menos financiamento junto a acionistas privados, sobretudo devido à resistência dos homens (que são a maioria dos investidores) em financiar empresas lideradas por mulheres. Ainda de acordo com o relatório, muitas empreendedoras entrevistadas revelaram que, a fim de contornar este problema, preferem solicitar dinheiro aos bancos, já que, como estes utilizam algoritmos de pontuação (*score* bancário), seriam automaticamente mais "imparciais" na concessão do crédito (Sako; Parnham, 2021, p. 107).

O problema é que essa suposta neutralidade dos *scores* algorítmicos tem sido colocada em xeque pelos próprios clientes e também por órgãos reguladores norte-americanos. Em 2019, o Departamento de Serviços Financeiros do Estado de Nova York abriu uma investigação sobre denúncias de que o cartão de crédito da Apple estava oferecendo limites de crédito distintos para homens e mulheres. Vários usuários do cartão – incluindo do cofundador da Apple, Steve Wozniak – alegaram, em suas redes sociais, que algoritmos de *score* discriminavam mulheres. O empresário do setor de

tecnologia David Hansson, por exemplo, reclamou que o *Apple Card* deu a ele 20 vezes o limite de crédito que sua esposa obteve, muito embora a renda formal da esposa ultrapassasse a dele. Mais tarde, Wozniak tuitou que a mesma coisa aconteceu a ele e sua esposa, embora não tivessem contas bancárias ou ativos separados. Em sua defesa, o banco Goldman Sachs, que oferece o cartão em parceria com a Apple, alegou que suas decisões de crédito são pautadas essencialmente na qualidade de crédito do cliente, e não em fatores como sexo, raça, idade ou orientação sexual (Natarajan; Nasiripour, 2019).

Apesar de não se ter a certeza do que especificamente ensejou as disparidades relatadas, os problemas com o *AppleCard* revelam um possível enviesamento dos modelos algorítmicos usados em *scores* bancários, que pode estar na origem de efeitos discriminatórios e diferenças de oportunidade entre homens e mulheres.

O que os três exemplos relatados têm em comum? Além do viés algorítmico detectado, que acabou por reproduzir preconceitos de raça e de gênero, todos os sistemas de IA mencionados acima empregavam aprendizado de máquina, cuja sequência preditiva é frequentemente incompreensível – ou seja, opaca – para seres humanos. A *opacidade algorítmica*[6] amplia o desafio de se detectar e corrigir vieses, como veremos a seguir.

1.3 A opacidade pode encobrir falhas e vieses algorítmicos

Na programação tradicional, construir um *software* consistia em redigir um modelo lógico à mão, ou seja, traçar um conjunto de regras que permitiam atingir conclusões a partir do processamento de casos individuais. Tais modelos são, por definição, *interpretáveis*, uma vez que seu código-fonte foi previamente escrito por um desenvolvedor, sendo possível dizer, em cada caso individual, *quais* e *como* as instruções foram acionadas para se chegar a um resultado (*e.g.*: se a renda de um solicitante for inferior a "*y*" por mês, o financiamento será recusado pelo sistema) (Villani, 2019, p. 114).

6 Opacidade algorítmica, de acordo com Harry Surden (2014, p. 158 – tradução nossa), é "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".

Por outro lado, algoritmos que incorporam o aprendizado de máquina, como, por exemplo, as *random forests*[7], apenas retornam resultados, sem, contudo, oferecer explicações razoáveis sobre como se chegou a determinada predição. Nesses casos, já que não é possível vislumbrar com clareza o processo decisório por trás do *output*, diz-se que o algoritmo é *opaco* – por constituir uma verdadeira "caixa-preta", incapaz de fornecer explicações razoavelmente compreensíveis para um ser humano. Roberto Confalonieri e seus colegas sintetizam o problema:

> Embora alguns modelos algorítmicos possam ser considerados interpretáveis por design [...] a maioria dos modelos de *machine learning* comporta-se como "caixas-pretas". A partir de uma entrada (*input*), uma "caixa-preta" retornará o resultado [...] sem revelar detalhes suficientes sobre sua lógica interna, resultando em um modelo de decisão *opaco*. (Confalonieri *et al.*, 2020, p. 7 – tradução nossa)

Deve-se pontuar, entretanto, que nem todo aprendizado de máquina conduzirá, necessariamente, à opacidade absoluta. Muitos sistemas de *machine learning* empregam a chamada aprendizagem *supervisionada*, através da qual os programadores "treinam" o algoritmo por meio de exemplos e regras predeterminadas. No caso da aprendizagem supervisionada, a análise deste conjunto de instruções prévias permite conhecer melhor suas etapas de raciocínio e a forma como os dados são analisados. Já no caso da aprendizagem *não supervisionada*, na qual a quantidade de orientações preliminares deixadas pelo desenvolvedor é menor, o algoritmo tem uma operação mais autônoma e, portanto, menos intuitiva para seres humanos. A inteligibilidade do sistema diminui à medida que o algoritmo passa a abrigar não apenas uma, mas múltiplas redes neurais que se sobrepõem, o que comumente ocorre nos algoritmos de *deep learning* (Ghahramani, 2004).

O grande problema é que, em razão da crescente potência e velocidade dos algoritmos de IA – sobretudo os que empregam *deeplearning* –, é quase impossível acompanhar o seu raciocínio, mesmo nos casos em que seu código é aberto. Por exemplo, para reconhecer uma imagem, um classificador pondera milhões de critérios, utilizando milhões de imagens do seu banco de treinamento, as quais, por sua vez, contêm milhões de pixels (4K) (Villani, 2019, p. 114).

---

7 *Random forests* (em português: "florestas aleatórias") são versões mais avançadas das árvores de decisão. Consistem em agrupar grande número de árvores relativamente não correlacionadas, cuja previsão "coletiva" será mais precisa que qualquer uma de suas previsões individuais.

No entanto, em que pese o intricado desafio técnico de se desvendar as "caixas-pretas" algorítmicas, será inevitável enfrentá-lo, tendo em vista o exponencial avanço da IA e seus inúmeros desdobramentos. Se, em alguns casos, as implicações são mínimas, existem, em contrapartida, áreas sensíveis nas quais não se pode admitir a reprodução sistemática desses equívocos, sobretudo se considerarmos que os sistemas de IA possuem a capacidade de tratar problemas em larga escala. Para se ter uma noção dessa escalabilidade, o sistema de IA chamado Victor, em operação no Supremo Tribunal Federal atualmente, é capaz de analisar um processo em 5 segundos, enquanto um servidor, para realizar a mesma tarefa, gasta 44 minutos. Por sua vez, o sistema Athos, utilizado no Superior Tribunal de Justiça, pode analisar até 30 mil processos mensalmente (Andrade, 2021). Nesse sentido, a presença de falhas ou vieses em algoritmos de elevada importância e escalabilidade pode ter implicações de alto risco, como, por exemplo, o enviesamento de centenas ou milhares de análises judiciais.

Desta forma, a incapacidade de se compreender as relações entre dados de entrada (*input*) e dados de saída (*output*) em sistemas de IA pode tornar um sistema de IA uma verdadeira "caixa-preta", à qual não se pode (ou pelo menos não se deveria) entregar decisões cruciais, pela simples impossibilidade de se confiar em um sistema cuja cadeia de raciocínio permanece oculta, podendo encobrir falhas ou vieses discriminatórios. Tome-se o exemplo imaginário de um robô que assassina, de forma aparentemente deliberada, um ser humano. Seria fundamental desvelar a lógica interna que o levou a tal ato – até mesmo para efeitos de responsabilização, se for o caso, dos programadores, do fabricante ou do proprietário daquele robô. Se o algoritmo é uma "caixa-preta", da qual não se pode depreender o processo preditivo interno, seria um grande desafio determinar quando, como e por que o algoritmo errou, informações sem as quais a solução do crime e a atribuição de responsabilidades tornam-se tarefas eminentemente espinhosas.

Não seria o caso, contudo, de afastar os algoritmos de toda e qualquer tomada de decisão, mas sim de abordar a inteligência artificial de forma a incentivar e viabilizar sua explicação, em maior ou menor grau, a depender da sua aplicação final. Essa abordagem deve ter por objetivo a compatibilização da IA com valores sociais, levando-se em conta uma série de questões éticas, entre as quais se destacam a falibilidade, a opacidade, o

viés, a discriminação, a autonomia, a responsabilidade e a privacidade de informações (Rossetti; Angeluci, 2021, p. 7).

Diante de tais reflexões, entendemos que o desenvolvimento de funcionalidades que habilitem a IA a fornecer explicações satisfatórias para seus atos poderia resolver, em parte, o problema da opacidade algorítmica. Não se trata apenas de fomentar a transparência, mas de desenvolver uma postura ativa, que possibilite aos sistemas de IA deixar claras as suas intenções, motivações e o encadeamento causal por trás de uma decisão, notadamente quando esta tem repercussões individuais ou sociais relevantes. Essas propriedades "explicativas" já vêm sendo exploradas no promissor campo da *explainable artificial intelligence (XAI)*, que será tratado a seguir.

2 A INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

2.1 Explainable artificial intelligence (XAI): conceitos fundamentais

Um sistema inteligente dotado de inteligência artificial explicável (XAI) é aquele que possui *interpretabilidade ou explicabilidade*[8], quer dizer, capacidade para explicar suas predições, por meio de estratégias textuais ou visuais que forneçam compreensão qualitativa sobre seu processo de predição (Ribeiro *et al.*, 2018, p. 2). O sistema de IA explicável está habilitado a fornecer explicações sobre sua operação, tornando seu comportamento mais inteligível para os humanos (Gunning *et al.*, 2019). Significa dizer que um sistema XAI deve estar apto a explicar, de maneira apropriada a um ser humano, a lógica interna de sua predição: o que foi feito, o que está fazendo agora e o que acontecerá a seguir.

A literatura destaca que o nível de detalhes e as características da XAI devem ser estabelecidos levando-se em conta o público-alvo da explicação. Por exemplo, desenvolvedores de *software* podem compreender pequenas redes bayesianas, mas elas são um completo enigma para o usuário leigo (Ribeiro *et al.*, 2016, p. 4). Da mesma maneira, explicações muito básicas devem ser insuficientes para que *experts* revisem ou auditem um algoritmo.

8 Alguns especialistas entendem que *interpretabilidade* e *explicabililidade* são sinônimos. Para outros, contudo, a *interpretabilidade* limita-se à simples transparência do sistema (que sua lógica decisória seja, por si só, ser inteligível para o ser humano), enquanto a *explicabilidade* está ligada a uma postura mais ativa do sistema, à sua capacidade de elaborar e apresentar esclarecimentos humanamente compreensíveis quanto à sua deliberação (Došilović *et al*., 2018, p. 2). Embora essa distinção revele um cuidado conceitual bastante útil para explicações mais técnicas, para o presente trabalho, ela não tem grande relevância, razão pela qual optamos por usar os termos de modo intercambiável.

É primordial, portanto, que o desenvolvimento da XAI não perca de vista o seu consumidor final, devendo fornecer, a depender do destinatário, a) quantidade "apropriada" de informações (nem informações escassas, tampouco em demasia) e b) explicações em linguagem compreensível para o interlocutor.

As primeiras tentativas de gerar XAI remontam às décadas de 1960 e 1970, notadamente por meio dos sistemas Mycin e Centaur, desenvolvidos no âmbito da Universidade Stanford. Contudo, deparou-se com o problema de que as "explicações" engendradas por estes sistemas eram, em verdade, verbalizações das regras, não interpretações consistentes das rotinas ou arquitetura do sistema. Uma expressão formal de "por que fizemos assim" é uma justificativa, não uma explicação (Mueller *et al.*, 2019). Além dessa limitação, há de se considerar o fato de que os sistemas estudados à época (em sua maioria, *sistemas especialistas*[9]) eram substancialmente mais simples do que algoritmos de *machine learning e deeplearning* empregados atualmente. Sendo assim, apesar das tentativas de desenvolvimento da inteligência explicável no passado, ainda persiste o desafio de se fornecer explicações satisfatórias, sobretudo considerando-se a enorme complexificação dos processos algorítmicos trazida pelos avanços da tecnologia computacional, em especial no campo da nanociência.

Como já visto anteriormente, a implementação de sistemas de IA que utilizam aprendizado de máquina tornou-se ponto fulcral de preocupação dos estudiosos, na medida em que alguns modelos algorítmicos constituem verdadeiras "caixas-pretas": de tão complexos, determinados processos preditivos acabam sendo indecifráveis – até mesmo para programadores e técnicos da área. Isso dificulta, por exemplo: a) a experiência e a confiança do usuário final, que pode ser prejudicada por predições equivocadas do sistema; b) o aperfeiçoamento desses modelos por seus desenvolvedores, visto que nem sempre se sabe quando e como a IA falha; c) conformidade legal dessas ferramentas, pois seus detentores estão sujeitos a um risco legal aumentado, decorrente da opacidade e, também, da maior falibilidade de modelos algorítmicos não interpretáveis (Nunes; Andrade, 2021).

9 Os *sistemas especialistas* originais eram *softwares* que objetivavam simular o raciocínio de um humano *expert* em determinada área. Dotados de raciocínio lógico estrito, estes *softwares* logo se mostraram impraticáveis no mundo real, por ignorarem a incerteza. Uma segunda geração de sistemas especialistas passou a empregar técnicas probabilísticas, que, contudo, não salvariam tais *softwares* do fracasso, dado o elevado número de probabilidades existentes e as limitações computacionais da época.

Nessa linha, embora a IA ofereça extraordinárias possibilidades em nosso cotidiano, está bem estabelecido pelos estudiosos do tema que sua opacidade, em alguns casos, não é desejável. De fato, com os algoritmos invadindo nossas vidas diárias, é preciso que reflitam nossas leis e padrões sociais. Diante de "caixas-pretas algorítmicas", o papel da XAI será fundamental para entender, auditar e corrigir esses sistemas, buscando-se permanentemente a sua conformidade ética e legal.

Por fim, vale ressaltar que nem todo sistema de inteligência artificial pressupõe a necessidade de XAI. Molnar (2021) oferece dois exemplos: a) quando o modelo algorítmico e suas predições têm baixo impacto, não havendo desdobramentos sociais, e b) na hipótese em que as aplicações de um determinado sistema já estejam suficientemente estudadas e estabelecidas – como no caso do uso do reconhecimento facial para desbloqueio de celulares.

2.2 A *XAI* na prática

A literatura especializada faz uma distinção entre modelos algorítmicos que são *originalmente interpretáveis* e aqueles que *necessitam ser explicados por meio de técnicas específicas de XAI*. Um modelo "transparente" de aprendizado de máquina é aquele que é explicável por si só, não requerendo técnicas adicionais para que o humano possa compreendê-lo. Em oposição aos modelos "transparentes", existem os "modelos opacos", cuja compreensão irá demandar um processo de explicação adicional, chamado de *explicabilidade post-hoc*. A *explicabilidade post-hoc* se volta para modelos que não são prontamente interpretáveis pelo seu *design*, recorrendo a diversas técnicas de XAI, como explicações de texto, explicações visuais, explicações por meio de exemplos, explicações por simplificação e explicações de relevância de recurso.

Dada a diversidade e profundidade de técnicas XAI, seria impossível discorrer sobre as especificidades de cada uma delas. Mas um exemplo importante nos permite esclarecer o funcionamento da XAI em suas linhas gerais, o que é suficiente para os fins deste estudo. Veja-se o caso da *explicação por relevância de recurso*. Esse método tem por objetivo descrever melhor um algoritmo opaco, enfatizando-se os recursos e variáveis decisivos para o resultado final (*output*) da predição algorítmica. Uma contribuição de destaque são as SHAPs (*SHapley Additive exPlanations*), que propõem uma espécie de "pontuação" para a influência de cada característica preditiva

durante o processamento algorítmico. Por meio das SHAPs, as variáveis usadas no processo preditivo são ordenadas, apresentando-se aquelas que mais influenciam o algoritmo para uma ou outra direção (Lundberg; Lee, 2017).

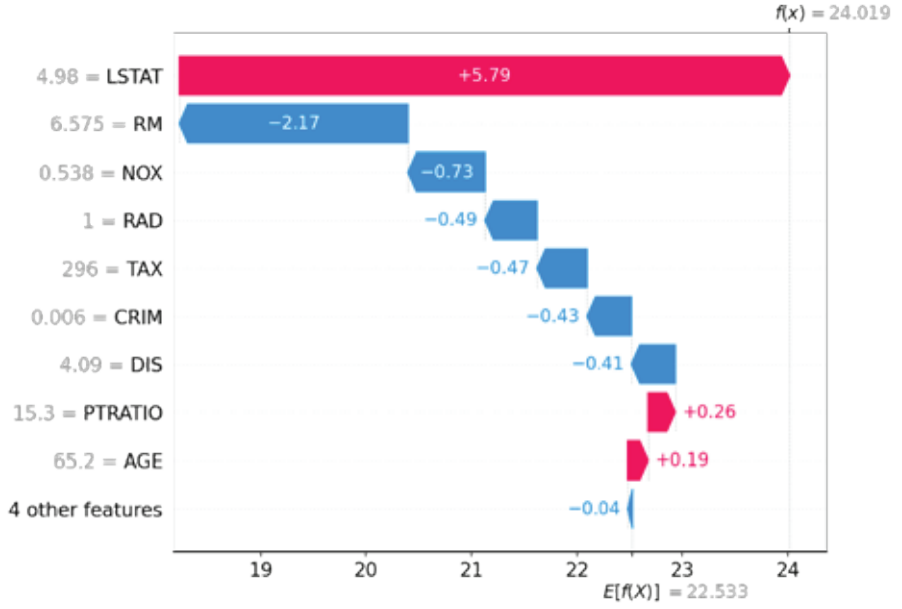

Figura 2: A explicação acima trabalha a ideia de relevância, detalhando como cada recurso contribuiu para o "*output*". Os recursos que "empurram" a previsão para cima são mostrados em vermelho, enquanto os que "empurram" a previsão para baixo são mostrados em azul. Fonte: GitHub (https://github.com/slundberg/shap).

Ao revelar o "peso" dos itens mais decisivos na análise, as SHAPs fornecem informações cruciais sobre o funcionamento do algoritmo, permitindo, por exemplo, que um desenvolvedor identifique uma variável que está sendo sub ou superestimada na predição e, consequentemente, faça regulagens nos pesos de cada variável, de forma a remover vieses e aumentar a precisão dos resultados fornecidos pelo sistema.

3 O VALOR DA EXPLICAÇÃO NO ENFRENTAMENTO DOS VIESES ALGORÍTMICOS

3.1 Detecção de correlações antiéticas

Na primeira seção deste artigo, mostrou-se, por meio dos casos *Compas* e *AppleCard*, como negros e mulheres podem ser prejudicados por

sistemas de pontuação preconceituosos. Na base dessa discriminação, encontram-se algoritmos que, mesmo pretensamente imparciais, podem estar imbuídos da subjetividade de seus criadores ou afetados pela qualidade do treinamento e dos dados fornecidos. No caso de sistemas muito complexos, a opacidade pode agravar e encobrir tais preconceitos, na medida em que dificulta a compreensão da lógica preditiva aplicada. Para resolver tal impasse, a XAI poderia ser implementada, enquanto funcionalidade que ajuda a detectar e corrigir encadeamentos lógicos inerentemente tendenciosos. Afinal, como prelecionam Nunes e Marques (2018, p. 7), a respeito do Compas:

> A ausência de transparência do algoritmo é especialmente crítica nesse caso. Como defender-se de um "índice" sem saber o método de seu cálculo? Como submeter o "índice" ao controle do devido processo constitucional? Por mais que sejam divulgadas as perguntas realizadas, os acusados não sabem como suas respostas influenciam no resultado final (*output*). Dessa forma, a defesa do acusado torna-se impossibilitada por dados matemáticos opacos e algoritmicamente enviesados, mas camuflados, pela "segurança" da matemática, como supostamente imparciais, impessoais e justos.

A XAI pode, inclusive, ajudar a identificar as correlações (não tão óbvias) feitas pelos sistemas de pontuação. Nesse sentido, ainda que o sistema do *AppleCard* não tenha estabelecido uma causalidade direta entre pontuação/raça ou pontuação/gênero, como alega o provedor do serviço, suspeita-se que uma relação indireta pode ter sido traçada a partir de um conjunto de dados mais amplo. Por exemplo, ao longo do casamento, os homens tendem a contrair mais empréstimos em seu nome, em vez de os tomarem em conjunto com suas esposas. Quando não ajustado, esse dado pode levar o algoritmo a inferir que, de modo geral, os homens tomam e honram empréstimos com mais frequência que as mulheres, sendo, portanto, mais confiáveis (Kelion, 2019). De igual forma, o Compas, mesmo sem perguntar a raça do detento, poderia estar absorvendo e considerando essa informação indiretamente, já que, no questionário, existem perguntas que acabam por selecionar indivíduos pobres que são, em sua maioria, negros.

Portanto, por meio de explicações adequadas sobre os processos preditivos dos algoritmos, seria possível identificar (e eventualmente corrigir) o estabelecimento de correlações como essas que, embora indiretas, possuem efeito discriminatório, seja por favorecer homens, como no caso do *AppleCard*, seja por favorecer pessoas de raça branca, no caso do Compas. Constatado, após a emissão da explicação, que um sistema acabou por dar

relevância – mesmo que indiretamente – a um dado racial ou de gênero em seu julgamento, seria o caso, portanto, de recalibrar o modelo algorítmico em questão, evitando-se que ele expresse novos preconceitos em sua pontuação.

3.2 Aperfeiçoando o reconhecimento de imagens

Serengil (2019) demonstra que as SHAPs também são capazes de explicar como um algoritmo distingue entre determinadas emoções por meio do reconhecimento facial. O autor utilizou um conjunto de treinamento de dados chamado FER-2013, que continha imagens de expressões faciais de 28.709 pessoas, sendo capaz de distinguir entre sete categorias (0 = zangado, 1 = nojo, 2 = medo, 3 = feliz, 4 = triste, 5 = surpresa, 6 = neutro).

Na sequência de imagens abaixo, é possível notar como esse método acopla a técnica de *explicação por relevância de recurso* à classificação de imagens, fornecendo informações detalhadas e consistentes sobre o processo de reconhecimento facial de um sistema classificador. De igual forma, a decifração do processo preditivo de um classificador pode ser aplicada em outras áreas, de modo a refinar a precisão do algoritmo (Serengil, 2019).

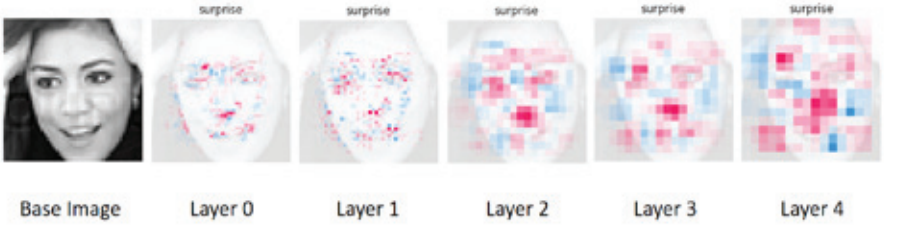

Figura 3: O método SHAP (SHapley Additive exPlanations), utilizado para explicar a detecção algorítmica de uma expressão facial: as primeiras camadas (*layers* 0 e 1) focam nas características do rosto (olhos, nariz, boca etc.), enquanto as camadas seguintes mencionam demais áreas do rosto. Em vermelho, os *pixels* que mais influenciam a previsão, enquanto aqueles com baixa importância estão marcados de azul.

Outro método de explicação que pode ser aplicado a classificadores de imagem é o LIME (*local interpretable model-agnostic explanations*), capaz de explicar como recursos de entrada (*input*) afetam a previsão algorítmica (*output*) (Ribeiro *et al.*, 2016). Resumidamente, esse sistema gera perturbações aleatórias na imagem de entrada, "desligando" e "ligando" alguns *pixels* para encontrar "*superpixels*", ou seja, segmentos de imagem significativos que se assemelham à base de dados catalogada (Arteaga, 2020).

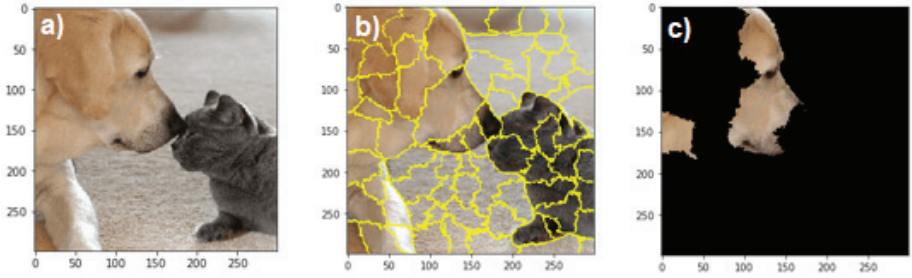

Figura 4: LIME explicando um resultado. Por meio de perturbações na área original (a), "*superpixels*" foram criados (b), e, entre eles, o algoritmo destacou os mais significativos (c), que, comparados com a base de dados, possuem associação mais forte com a raça "Labrador Retriever".

Como se percebe, a explanação acerca do processo preditivo, mesmo que simplificada, pode contribuir para localizar e ajustar falhas. No caso do problema do Husky, por exemplo, o LIME é capaz de revelar, ao usuário da funcionalidade, que a neve adquire significância desproporcional na classificação (Ribeiro *et al.*, 2016). A mesma funcionalidade poderia ser aplicada para ajudar a refinar o reconhecimento do Google Fotos, a fim de identificar e corrigir as grosseiras correlações estabelecidas pelo algoritmo entre *superpixels* presentes na face de negros e de macacos.

Destaca-se ainda que, ao fornecer explicações sobre o processo preditivo, a XAI pode mudar a percepção dos usuários sobre a *confiabilidade* de determinada ferramenta, corrigindo problemas como a "confiança cega" do usuário ou, também, a desconfiança em relação a um algoritmo. Isso ficou demonstrado em estudo desenvolvido por professores da University of Washington que perguntaram a estudantes de programação qual era a sua confiança em um algoritmo classificador de imagem (Ribeiro *et al.*, 2016, p. 9). A princípio, pouco mais de um terço dos indivíduos confiavam no algoritmo. Após receberem a explicação elaborada pelo LIME – revelando que o fundo da imagem continha um peso considerável na classificação –, a confiança no classificador caiu substancialmente (para aproximadamente 10%). Essa evidência chama atenção para a importância de se acoplar explicações aos sistemas, justamente para que usuários não superestimem a confiabilidade de um algoritmo possivelmente impreciso ou enviesado.

CONSIDERAÇÕES FINAIS

Este trabalho mostrou, em sua primeira etapa, como surgem as falhas e os vieses algorítmicos, em especial, analisando-se a ocorrência desses problemas em contextos opacos, ou seja, nos quais se tem pouca ou nenhuma compreensão sobre o processo preditivo empregado por um sistema de IA complexo. Em um algoritmo "caixa-preta", eleva-se o risco de que tais desvios passem despercebidos, podendo ser, inclusive, reproduzidos em larga escala. Se, por um lado, alguns algoritmos têm implicações pouco significativas, por outro, merecem especial atenção aqueles cujos resultados podem ter repercussão individual ou coletiva relevante, a partir de avaliações com base em raça, gênero, sexualidade e etnia capazes de gerar discriminações sistemáticas, como nos casos do Compas e do *AppleCard*.

Diante da constatação de que a opacidade pode encobrir e potencializar falhas e vieses algorítmicos, ponderou-se, na segunda etapa do trabalho, que a inteligência artificial explicável (XAI) poderia auxiliar no combate à opacidade, na medida em que habilita o sistema a fornecer explicações sobre seu próprio processo preditivo. Esse é um processo de transformação da "caixa-preta" algorítmica em uma autêntica "caixa de vidro" – ou seja, transparente, fácil de visualizar e entender – que contribui para a identificação de correlações indesejáveis, estabelecidas no interior do algoritmo, permitindo que desenvolvedores de um sistema rastreiem e corrijam falhas e vieses ali presentes. Ainda, a "caixa de vidro" permite a verificabilidade, auditoria e apuração de responsabilidade quando a IA toma decisões ilegais. Por fim, a XAI também promove a confiança dos usuários e da sociedade na própria inteligência artificial, pois mostra, de maneira geral, quando, como e por que um algoritmo está tomando determinada decisão.

Não se pode esquecer, contudo, que as soluções oferecidas pela inteligência artificial explicável ainda estão situadas no campo experimental, concentrando-se, em sua maioria, em investigações acadêmicas. Nesse sentido, há um percurso considerável até que a XAI seja desenvolvida e implementada satisfatoriamente, de modo a poder beneficiar seu usuário final com explicações úteis e acessíveis. Portanto, será necessário que empresas desenvolvedoras – e, em última análise, seus controladores e investidores – estejam "motivados" a financiar o desenvolvimento de funcionalidades de autoexplicação para sistemas de IA complexos.

Seria ingênuo, no entanto, esperar uma mobilização espontânea por iniciativa das empresas de tecnologia em torno da ética e da transparência

– razão pela qual imaginamos que o verdadeiro catalisador desta motivação repousa, justamente, no debate sobre o direito à explicação. Por isso, a discussão sobre diretrizes relacionadas à explicação e, finalmente, a entrada e estabilização desses direitos nos ordenamentos jurídicos ao redor do globo (a exemplo do que tem ocorrido na União Europeia) podem estabelecer sanções à opacidade algorítmica, incentivando as empresas a tornarem seus *softwares* mais transparentes e interpretáveis.

Apesar de entendermos que a XAI constitui uma via a ser explorada nos próximos anos, é preciso reconhecer que ela não passa, hoje, de uma promessa, que está longe de ser uma unanimidade no mercado. Nada garante seu sucesso, menos ainda de forma ampla e generalizada. Muito provavelmente, será impossível avançar na implementação da explicabilidade em certos domínios mais complexos. Nesses casos, entendemos que expedientes regulatórios mais assertivos serão necessários, especialmente quando estiverem em jogo questões mais sensíveis, como proteção ao meio ambiente ou direitos fundamentais. Esses expedientes podem envolver medidas de controle, de supervisão ou de tutela humana em tempo real, chegando até mesmo ao banimento da tecnologia em determinadas situações.

Esse conjunto de reflexões nos leva a concluir que a compatibilização da inteligência artificial com as leis e valores sociais pressupõe não apenas o acesso, mas também a garantia de compreensão dos processos preditivos empregados em todo sistema de IA ao qual é delegada uma decisão com efeitos significativos – sejam eles a nível privado, como a concessão de um empréstimo pessoal, ou de cunho mais abrangente, como o auxílio à tomada de decisão judicial para sentenciar acusados.

Não se trata somente de incentivar a simples transparência, mas de construir, a nível de políticas públicas, uma postura que reconheça a XAI enquanto requisito para adoção de sistemas de IA em searas mais sensíveis. 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.

REFERÊNCIAS

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

ANDRADE, Otávio Morato de. Utilizando inteligência artificial para combater a morosidade processual e democratizar o acesso ao judiciário. *Duc in Altum: Cadernos de Direito*. No prelo.

ANGWIN, Julia; LARSON, Jeff; SURYA, Mattu; KIRCHNER, Lauren. Machine Bias. *Pro Publica*, 23 de maio de 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 19 ago. 2021.

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

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

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

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

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

GHAHRAMANI, Zoubin. *Unsupervised learning*. 16 de setembro de 2004. Disponível em: http://datajobstest.com/data-science-repo/Unsupervised-Learning-Guide-[Zoubin-Ghahramani].pdf. Acesso em: 20 set. 2021.

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

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

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

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

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

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

MÜLLER, Klaus-Robert; SAMEK, Wojciech. Towards explainable artificial intelligence. *arXiv*:1909.12072v1, 26 de setembro de 2019. Disponível em: https://arxiv.org/abs/1909.12072. Acesso em: 15 ago. 2021.

NAJIBI, Alex. Racial discrimination in face recognition technology. *Harvard Online: Science Policy and Social Justice*, 24 de outubro de 2020. Disponível em: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/. Acesso em: 26 set. 2021.

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

NUNES, Dierle; ANDRADE, Otávio. A explicabilidade da inteligência artificial e o devido processo tecnológico. *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: 26 set. 2021.

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

algorithmic_bias_and_the_risks_of_assignment_of_decision_making_function_to_machines. Acesso em: 26 set. 2021.

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

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

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

ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação? *Revista Eco Pós*, v. 18, n. 2, p. 35-56, 2015. Disponível em: https://revistaecopos.eco.ufrj.br/eco_pos/article/view/2662. Acesso em: 26 set. 2021.

SAKO, Mari; PARNHAM, Richard. *Technology and innovation in legal services*: final report for the solicitors regulation authority. University of Oxford, 2021. Disponível em: https://www.sra.org.uk/globalassets/documents/sra/research/full-report-technology-and-innovation-in-legal-services.pdf?version=4a1bfe. Acesso em: 26 set. 2021.

SERENGIL, Sefik. How SHAP can keep you from black box AI. *Blog Sefik Ilkin Serengil*, 1º de julho de 2019. Disponível em: https://sefiks.com/2019/07/01/how-shap-can-keep-you-from-black-box-ai. Acesso em: 26 set. 2021.

SIMONITE, Tom. When it comes to Gorillas, Google Photos remains blind. *Wired*, 1º de novembro de 2018. Disponível em: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind. Acesso em: 26 set. 2021.

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

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

VILLANI, Cédric. *For a meaningful artificial intelligence*: towards a French and European strategy. A parliamentary mission from 8th september 2017 to 8th march 2018. Paris, 2019. Disponível em: https://books.google.com.br/books?id=9cVUDwAAQBAJ&lpg=PP1&hl=pt-BR&pg=PP1#v=onepage&q&f=false. Acesso em: 26 set. 2021.

WELLS, Lindsay; BEDNARZ, Tomasz. Explainable AI and reinforcement learning: a systematic review of current approaches and trends. *Front Artificial Intelligence*, 20 de maio de 2021. Disponível em: https://www.frontiersin.org/articles/10.3389/frai.2021.550030/full. Acesso em: 26 set. 2021.

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

Sobre os autores:

Marco Antônio Sousa Alves | *E-mail:* marcofilosofia@gmail.com

Professor Adjunto de Teoria e Filosofia do Direito e do Estado da Faculdade de Direito da Universidade Federal de Minas Gerais. Doutor em Filosofia pela UFMG, com estágio de pesquisa na École des Hautes Études en Sciences Sociales (EHESS/Paris). Membro Permanente do Programa de Pós-Graduação em Direito (PPGD/UFMG). Coordenador do Grupo SIGA/UFMG (Sociedade da Informação e Governo Algorítmico).

Otávio Morato de Andrade | *E-mail:* otaviomorato@gmail.com

Mestrando em Direito pela UFMG, com bolsa do CNPq. Especialista em Direito Civil pela PUC-MG. Graduado em Direito (UFMG) e Administração (PUC-MG). Membro-Monitor do Grupo SIGA/UFMG (Sociedade da Informação e Governo Algorítmico).

Data de submissão: 27 de setembro de 2021.
Data de aceite: 10 de janeiro de 2022.