1. INTRODUÇÃO

A Inteligência Artificial (IA) é aplicada em inúmeras situações do cotidiano atual, que vão desde sugestões personalizadas até tomada automatizada de decisão ou automóveis autônomos. Nesse contexto, surgem preocupações com o emprego de modelos de machine learning, que operam com pouca transparência aos olhos humanos destreinados.

A transparência de sistemas de tomada de decisão é essencial para assegurar a lisura desse processo, cuja opacidade introduz o risco de enviesamento ou imprecisões. Essas situações podem gerar violações a direitos humanos quando os vieses algorítmicos representam critérios de discriminação, perpetuação de desigualdades sociais e manutenção do poder econômico. Considerando esse contexto, o problema a ser aqui analisado é: quanto a opacidade algorítmica por barreira técnica afeta efetivamente os direitos humanos? O objetivo específico deste trabalho será explicar e desmistificar a *black-box* no campo da Inteligência Artificial, além de sugerir formas de enfrentar o desafio de modo a garantir a proteção aos direitos humanos.

Sob o ponto de vista técnico, este trabalho se volta para a opacidade algorítmica advinda do código, ou seja, da complexidade das operações internas do sistema de IA, o que costuma ser confundido com a dificuldade de explicação às pessoas leigas. Buscamos aqui desmistificar essa falácia sobre opacidade algorítmica, que surge da concepção de machine learning enquanto um conjunto de modelos não explicitamente programados.

Ocorre que, ao contrário do que é pregado com frequência, o algoritmo que produz um dado modelo em si não sofre alteração, o que ocorre tão somente com seus parâmetros. Isso em razão do resultado de um problema de otimização, que altera os parâmetros treináveis, direcionando-o a obter os melhores resultados possíveis. Dessa forma, as operações realizadas pelo modelo são definidas pela arquitetura, mas os valores operados não são conhecidos *a priori*, e sim determinados pelo protocolo de treinamento.

Todas as operações internas de um modelo de machine learning são precisas, algébricas e determinísticas, uma vez que está treinado, e podem ser investigadas dado o modelo. Isso significa dizer que a bússola utilizada para análise e regulamentação da opacidade algorítmica da Inteligência Artificial está parcialmente voltada ao Norte errado, uma vez que trata os modelos de machine learning como black boxes não significa que não possam ser testados e descobrir seus vieses, de forma a proteger e garantir direitos. Partimos, portanto, da hipótese de que é possível fazer um controle estatístico posterior de resultado de IA quando violam direitos, especialmente direitos humanos.

Metodologicamente, a investigação se voltará à análise do problema de forma à comprovação da possibilidade de testar modelos fechados de *machine learning*, visando a proteção a direitos, especialmente direitos humanos. O estudo será essencialmente bibliográfico, sendo a natureza dos dados quali-quanti. Do ponto de vista dos objetivos, é exploratório.

2. CONCEITOS BÁSICOS DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING

O termo Inteligência Artificial é conhecido pela dificuldade de ser conceituado. Enquanto ramo da ciência da computação, tradicionalmente se concentra em elaborar estruturas de sistemas computacionais que executam tarefas e tomam decisões, muitas vezes mimetizando competências humanas. Nas palavras de McCarthy (1955), pioneiro na área, é a produção de máquinas inteligentes, "especialmente programas de computador inteligentes". A IA funciona, em geral, mediante a análise de um grande volume de dados e a identificação de padrões, utilizando métodos como, por exemplo, *machine learning* e *deep learning*.

Nesse sentido frisa-se que, conforme elucidado por Edwards e Veale (2017), enquanto um algoritmo é um processo automático amplo, o *machine learning* é uma abordagem específica de aprendizado de máquina que utiliza algoritmos para criar modelos preditivos a partir de dados.

O termo *machine learning*, introduzido em 1959 por Arthur Samuel, é definido como a habilidade de um computador de realizar uma tarefa sem ter sido explicitamente programado para isso. A capacidade de realizar a tarefa é adquirida por um processo iterativo denominado de treino, que necessita da apresentação de um conjunto de dados ao modelo que está sendo treinado. O treinamento consiste em resolver um problema de otimização, que atualiza os parâmetros do modelo de modo a melhorar sua performance na tarefa de interesse.

Em resumo pode-se dizer que *machine learning* é, em essência, uma mudança de paradigma computacional. Na computação convencional um algoritmo consiste em um conjunto de regras explícitas a serem executadas com a finalidade de realizar uma tarefa, enquanto que um algoritmo de *machine learning* apresenta tanto as regras que enunciam o modelo, definindo sua arquitetura, quanto o protocolo de treinamento. O resultado da execução de tal algoritmo é a obtenção de um modelo treinado que realiza a tarefa desejada.

Esse tipo de modelo é, a rigor, uma função matemática. Um conjunto de operações algébricas definidas *à priori* (arquitetura), que contém componentes ajustáveis (parâmetros

treináveis). Isso quer dizer que, do ponto de vista matemático, absolutamente nada em um modelo de *machine learning* é opaco. A opacidade de um sistema é dependente do observador, sempre surgindo por omissão de informação e desconhecimento ou desinteresse deste sobre o referido sistema.

3. CONCEITUAÇÃO DE BLACK-BOX E EXPLICABILIDADE

A definição dos significados dos termos *black box*, opacidade algorítmica e explicabilidade é crucial para abordar adequadamente os desafios impostos pelos sistemas de IA. Para isso, antes de tudo importa destacar que o termo *black-box*, quando relacionado com o *machine learning*, tange sistemas ou modelos cujo funcionamento interno é opaco, o que indica que ele é difícil de compreender, porque não revela explicitamente como se dá o processo de tomada de decisões.

A opacidade algorítmica refere-se à falta de transparência dos processos internos de algoritmos e sistemas de IA e, segundo Burrel (2016), existem três formas de opacidade: (i) o segredo industrial; (ii) o "analfabetismo técnico", decorrente da dificuldade de compreensão dos códigos computacionais por pessoas sem conhecimento específico e, por fim, (iii) a complexidade das operações internas dos sistemas de IA, oriunda da dificuldade de compreensão acerca da análise preditiva e pela resolução de um problema de otimização feita pelo algoritmo.

O termo explicabilidade, por sua vez, se relaciona diretamente com a opacidade, na medida em que refere-se à capacidade de um sistema algorítmico de fornecer informações compreensíveis e transparentes sobre suas decisões e funcionamento interno. Nesse sentido, espera-se que a opacidade algorítmica possa ser combatida com políticas de transparência e regulamentações que exijam a divulgação de informações sobre a lógica que rege funcionamento dos algoritmos, indicando a razão do *input* ter gerado os resultados apresentados pelo sistema (Goodman & Flaxman, 2017). O avanço de sistemas lineares para modelos não paramétricos e gaussianos, assim como a utilização de redes neurais, são fatores que tendem a criar maior dificuldade de explicação e requerem avanços técnicos e metodológicos que permitam interpretar e comunicar de forma acessível os processos decisórios algorítmicos.

4. ANÁLISE DE ALGORITMOS DETERMINÍSTICOS E A POSSÍVEL INVESTIGAÇÃO EM PROL DOS DIREITOS HUMANOS

A crescente utilização de algoritmos de inteligência artificial nos processos de tomada de decisão suscita preocupações sobre os direitos humanos, particularmente em relação à falta de transparência e responsabilização associada às "*black-boxes*". Isso porque, ao ocultarem o funcionamento interno dos algoritmos, os modelos de linguagem complexos representam um risco potencial ao direito à não discriminação. De fato, a ausência de transparência e explicabilidade nos resultados dos sistemas automatizados tende a reforçar indevidamente os preconceitos existentes nos dados em que se baseiam, minando a imparcialidade e a justiça destes sistemas.

Em que pese a opacidade algorítmica, um modelo de *machine learning* pode ser analisado sem que o algoritmo seja diretamente investigado, desde que algumas informações sejam fornecidas, são elas: (a) a tarefa para a qual o modelo foi treinado e (b) quais são os parâmetros de entrada e saída do modelo. Com tais informações, somado a um conjunto de dados de teste que satisfaça a entrada do modelo, pode-se avaliar sua resposta e, assim, realizar testes para aferir eventuais vieses, por meio da estruturação de tal conjunto.

Num sistema de aprendizado de máquina, vieses são, comumente, reflexos do conjunto de dados com o qual o modelo é treinado. Além disso, outra forma recorrente de se introduzir vieses é por meio da própria arquitetura algorítmica. De qualquer maneira, a própria informação de quais são os parâmetros de entrada do modelo podem evidenciar a localização dos vieses. Sendo assim, é possível gerar conjuntos de dados que testem às cegas os modelos de *machine learning*, sem que o código fonte seja publicizado, desde que se tenha um mínimo de transparência sobre as informações sobre o sistema: *input*, tarefa e *output*.

Significa dizer que, apesar da falta de acesso direto ao algoritmo fundamental, parece ser possível rastrear os vieses utilizando uma espécie de “varredura” nos dados de entrada do modelo. Isto é feito utilizando um conjunto de dados de teste, cruzando as entradas e as saídas, a fim de localizar padrões indesejados.

Apesar da proposição solucionar parte das situações que demandam análise de enviesamento sem publicidade do código fonte, ressalva-se, todavia, que o problema da explicabilidade algorítmica não é solucionado por inteiro. Isso porque no caso de algoritmos que produzem modelos lineares, existe uma maior previsibilidade e baixa complexidade, sendo possível inferir e explicar de maneira mais clara o funcionamento do modelo a partir de um conjunto de dados - conhecendo o *output* é possível descobrir o *input*. Todavia, quando

se trata de algoritmos que geram modelos não lineares - como é a maioria dos casos de IA -, dotados de menor previsibilidade e maior complexidade, é possível tão somente se fazer um estudo estatístico de possíveis vieses, testando e verificando se em um conjunto de dados, os *outputs* apresentam algum grau discriminatório sobre os respectivos *inputs*. No segundo caso, portanto, há pequenas chances de se explicar o sistema por completo, conforme apontam Goodman e Flaxman (2017) justamente em razão da alta complexidade que, comumente, é intitulada de *black-box*.

Sendo assim, é certo que a proposta apresentada depende de análise estatística e não se presta à investigação minuciosa do código fonte visando seu desvendamento.

5. CONCLUSÃO

A opacidade algorítmica por barreira técnica de fato se mostra um empecilho à proteção e promoção de direitos humanos, na medida em que impede a compreensão e, consequentemente, a fiscalização das decisões automatizadas, de sorte a comprometer a transparência e a fiscalização que os direitos humanos exigem. Por óbvio, essa falta de clareza, além de afrontar o princípio da explicabilidade e da transparência, pode levar a discriminações e violações de direitos sem possibilidade de contestação adequada.

Dessa forma, em que pese o avanço tecnológico, especialmente no campo da inteligência artificial (IA), se mostrar extremamente valioso em diversas áreas, é de suma importância que sejam adotadas precauções rigorosas ao implementar IA em sistemas e processos de tomada de decisão que impactam diretamente a sociedade. Desta forma, até o momento, riscos significativos, como o enviesamento algorítmico e a barreira técnica que dificulta a compreensão pela sociedade do agente responsável pelas decisões, ambos com influência direta sobre as vidas dos indivíduos, continuam a existir.

A fim de mitigar os referidos riscos, a realização de testes cegos, isto é, sem acesso ao código fonte - mantendo, assim, a proteção ao segredo industrial - pode servir como um meio eficaz de validação e identificação de vieses algorítmicos. Esses testes são fundamentais para garantir a proteção dos direitos humanos e prevenir a perpetuação de discriminações. Além disso, é vital promover a transparência e a explicabilidade dos sistemas de IA, facilitando uma compreensão mais ampla e acessível dos processos decisórios envolvidos e, assim, assegurando que os sistemas de IA sejam desenvolvidos e utilizados de maneira ética e justa,

beneficiando a sociedade como um todo e respeitando os princípios fundamentais dos direitos humanos.

6. REFERÊNCIAS

BURRELL, Jenna. **How the machine 'thinks'**: Understanding opacity in machine learning algorithms. Big Data & Society, [s.l.], jan.–jun., 2016. p. 4-5.

CHOLLET, Francois. **Deep Learning with Python**. New York: Manning Publications, 2017. Disponível em: https://books.google.com.br/books?id=wzozEAAAQBAJ. Acesso em: 02 jul. 2024.

EDWARDS, Lilian; VEALE, Michael. Slave to the Algorithm?: why a 'right to an explanation' is probably not the remedy you are looking for. **Duke Law & Technology Review**, Durham, v. 16, n. 1, p. 18-84, abr. 2017. Disponível em: https://scholarship.law.duke.edu/dltr/vol16/iss1/2/. Acesso em: 03 jul. 2024.

FRAZÃO, Ana. Black box e o direito face à opacidade algorítmica. In: BARBOSA, Mafalda Miranda; NETTO, Felipe Braga; SILVA, Michael César; FALEIROS JÚNIOR, José Luiz de Moura (org.). **Direito Digital e Inteligência Artificial**: diálogos entre Brasil e Europa. Indaiatuba: Editora Foco, 2021. Cap. 2. p. 27-42.

GARCIA, Gabriel Reis. **Machine learning methods for strangeness reconstruction in ALICE**. 2023. 1 recurso online (90 p.) Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Física Gleb Wataghin, Campinas, SP. Disponível em: https://hdl.handle.net/20.500.12733/15832. Acesso em: 02 jul. 2024.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Massachusetts: The Mit Press, 2016. Disponível em: https://www.deeplearningbook.org/. Acesso em: 02 jul. 2024.

GOODMAN, Bryce; FLAXMAN, Seth. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". **Ai Magazine**, [S.L.], v. 38, n. 3, p. 50-57, set. 2017. Wiley. http://dx.doi.org/10.1609/aimag.v38i3.2741. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1609/aimag.v38i3.2741>. Acesso em: 03 jul. 2024.

MCCARTHY, John, et al. **A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence**. 1955.

MITCHELL, Tom M.. **Machine Learning**. New York: McGraw-Hill Science/Engineering/Math, 1997.

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, v. 8, n. 1, p. 1-27, 2023.. Disponível em: https://periodicos.ufsm.br/revistadireito/article/view/69329. Acesso em: 28 jun. 2024.