ARTIGO

2965-6931 10.35699/2965-6931.2023.47673

REVISTA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ética e inteligência artificial: Desafios e melhores práticas

Carolina de Melo Nunes Lopes
Universidade Federal de Ouro Preto (UFOP)
https://orcid.org/0000-0003-2634-5536
carolina.nunes@aluno.ufop.edu.br

Júlia Castro Mendes
Universidade Federal de Juiz de Fora (UFJF)
https://orcid.org/0000-0002-6323-5355
juliacastro.mendes@ufjf.br

RESUMO

O rápido avanço da Inteligência Artificial (IA) trouxe consigo a necessidade de compreensão de seu impacto social e implicações éticas. Nesse sentido, este artigo levantou e discutiu as principais questões éticas relacionadas à IA, os principais obstáculos atuais no desenvolvimento de algoritmos de aprendizado de máquina, e as melhores práticas para desenvolver algoritmos éticos e justos. Os vieses podem se perpetuar facilmente através do uso de dados desbalanceados e de correlações infundadas. Diante disso, a colaboração entre desenvolvedores de algoritmos e outros especialistas torna-se fundamental para compreender diversas perspectivas e identificar as sutis formas de propagação dos preconceitos. A ética dos algoritmos não é uma questão que será solucionada apenas através de uma abordagem tecnológica - essa temática envolve também assuntos de ordem social, cultural, jurídica e política. Portanto, o desenvolvimento tecnológico e a responsabilidade social devem caminhar lado a lado, a fim de evitar o agravamento das diferenças sociais.

Palavras-chave: inteligência artificial; ética; aprendizado de máquina; aprendizado profundo; viés algoritmo.

Ethics and artificial intelligence: Challenges and best practices

ABSTRACT

The rapid advancement of artificial intelligence (AI) brought with it the need to understand its social impact and ethical implications. In this sense, this article raised and discussed the main ethical issues related to AI, the current primary obstacles in the development of machine learning algorithms, and the best practices to develop ethical and fair algorithms. Biases can easily perpetuate themselves through the use of unbalanced datasets and unsubstantiated correlations. Therefore, collaboration between algorithm developers and other specialists becomes essential to understand different perspectives and identify the subtle forms of the propagation of prejudices. The ethics of algorithms is not an issue that will be solved only through a technological approach - this theme also involves social, cultural, legal, and political aspects. Therefore, technological development and social responsibility must go hand in hand to avoid the aggravation of social differences.

Keywords: artificial intelligence; ethics; machine learning; deep learning; algorithm bias.

Submissão em: 14/08/2023 | **Aprovação em:** 25/11/2023

1. INTRODUÇÃO

A quarta revolução industrial impulsionou a aplicação de técnicas de Inteligência Artificial (IA) em diversos setores da sociedade. Muitas vezes não nos damos conta que interagimos com sistemas inteligentes o tempo todo, seja na sugestão de palavras do corretor ortográfico, ou nos anúncios das redes sociais (Garcia, 2020). Cada vez mais, os algoritmos mediam processos, transações bancárias, decisões governamentais e interferem na forma como nós nos percebemos e interagimos socialmente (Mittelstadd *al*., 2016).

É inegável que a transformação digital e as aplicações proporcionadas pela IA trouxeram inúmeros benefícios para diversos setores da sociedade (Winfield & Jirotka, 2018; Doneda *et al.*, 2018). No entanto, o desenvolvimento dessas tecnologias precisa vir acompanhado de reflexões éticas e morais (Rosetti; Angeluci, 2021). Como exemplos, o uso indevido de dados e informações pessoais, os vieses propagados pelos algoritmos, e a própria ética das máquinas, como nos carros autônomos, são questões que surgiram com o avanço da Inteligência Artificial nos últimos anos (Winfield; Jirotka, 2018).

Um dos principais dilemas éticos reside no viés algoritmo. Os algoritmos de aprendizado de máquina são treinados a partir de um conjunto de dados (Garcia, 2020). Se esses dados estiverem impregnados de vieses de gênero, raça, ou qualquer outro, o algoritmo irá perpetuar esses vieses (Garcia, 2020). Em 2018, por exemplo, a empresa *Amazon* contou com um sistema inteligente para fazer uma pré-seleção de currículos em um processo seletivo (Garcia, 2020). Para treinar esse sistema, a empresa utilizou a base de dados dos próprios funcionários, a fim de que o algoritmo selecionasse pessoas com características semelhantes às que já trabalhavam na empresa. O que não foi levado em conta pelos desenvolvedores é que a base de dados era majoritariamente masculina. Dessa forma, nenhuma mulher foi selecionada pelo algoritmo (Garcia, 2020).

Além disso, a transparência e "explicabilidade" (ou interpretabilidade) dos sistemas de IA são desafiadoras, uma vez que muitos modelos de aprendizado de máquinas são como "caixas pretas", ou seja, não permitem o acompanhamento das etapas intermediárias, ou dos cálculos que levam a um dado resultado ou decisão (Mittelstadt *et al.*, 2016). As lacunas entre o design, a operação de algoritmos e a nossa compreensão de suas implicações éticas podem ter consequências graves que afetam indivíduos, bem como toda a sociedade (Mittelstadt *et al.*, 2016). Para Garcia (2020), mais grave ainda é que as respostas geradas por esses algoritmos estão revestidas de mérito por serem resultado de testes estatísticos e matemáticos, o que confere a eles uma falsa imparcialidade.

Diante deste contexto, este estudo tem como objetivo levantar e discutir as questões éticas predominantes associadas à aplicação de tecnologias de IA, bem como relacionar as melhores práticas a serem adotadas neste contexto. Para tanto, foi realizada uma pesquisa bibliográfica de natureza exploratória, não-sistemática, utilizando as palavras-chave: Ética; Inteligência Artificial e Aprendizado de máquina, na ferramenta de pesquisa convencional do *Google*, no *Google* Acadêmico, *Scopus* e Scielo, no período de março a maio de 2023. Assim, a seguir, serão discutidas as principais preocupações em relação à ética dos algoritmos, alguns obstáculos do aprendizado de máquinas, do aprendizado profundo e as melhores práticas a serem adotadas para mitigar essas questões em projetos futuros, de qualquer área, envolvendo IA.

1.1 A ÉTICA DOS ALGORITMOS

Determinar o impacto ético de um algoritmo é um trabalho árduo (Mittelstadt *et al.*, 2016). Segundo Rosseti e Angeluci (2021), identificar a influência da subjetividade humana no projeto e configuração de algoritmos geralmente requer uma investigação multidisciplinar e de longo prazo, uma vez que a incerteza e a opacidade sobre como funcionam alguns algoritmos dificultam o seu entendimento (Winfield; Jirotka, 2018). Para Mittelstadt *et al.* (2016), as principais preocupações relacionadas à ética dos

algoritmos são: evidências inconclusivas; evidências incompreensíveis; evidências equivocadas; resultados injustos; privacidade e rastreabilidade.

1.2 EVIDÊNCIA INCONCLUSIVAS E AÇÕES INJUSTIFICADAS

Os algoritmos de decisão e a mineração de dados buscam estabelecer padrões e correlações em grandes conjuntos de dados (Mittelstadt *et al.*, 2016). No entanto, nem toda correlação encontrada pelos algoritmos tem uma relação de causa e efeito, o que pode gerar ações injustificadas (Doneda *et al.*, 2018; Rossetti; Angeluci, 2021). Assim, a decisão determinada por um algoritmo pode se basear em uma correlação estatística que não representa necessariamente uma relação causal, sem qualquer relação com o assunto objeto da decisão (Doneda *et al.*, 2018).

Nesse sentido, destaca-se a polêmica envolvendo a empresa alemã, Schufa, que presta serviços de proteção ao crédito (Doneda *et al.*, 2018). A empresa, ao avaliar o risco de inadimplência, classificava como critério negativo o pedido de consumidores para acessar seus próprios dados. Ou seja, a empresa penalizava consumidores que queriam contratar crédito simplesmente pelo exercício de um direito do consumidor. Existia aí uma relação de correlação (é mais provável que pessoas que tenham dificuldades financeiras queiram consultar seu próprio *score* de crédito), mas não de causa (essa consulta não é uma razão das dificuldades financeiras ou da inadimplência). Dessa forma, a Alemanha vedou essa prática na reforma da Lei Federal de Proteção de Dados (LGPD), de 2009 (Doneda *et al.*, 2018).

1.3 EVIDÊNCIAS INCOMPREENSÍVEIS E OPACIDADE

Quando os dados são processados para produzir evidências para uma conclusão, é esperado que exista uma relação compreensível e acessível entre esses dados e as conclusões obtidas (Mittelstadt *et al.*, 2016; Winfield; Jirotka, 2018; Rossetti; Angeluci, 2021). No entanto, muitos algoritmos de aprendizado de máquina, principalmente de aprendizagem profunda (*deep-learning*), atuam como uma caixa-preta - ou seja, é possível

visualizar os dados de entrada e a saída, mas o funcionamento interno é obscuro e não facilmente interpretável pelos humanos (Winfield; Jirotka, 2018).

A opacidade dos algoritmos pode ser problemática em várias áreas, especialmente em setores críticos, como saúde, justiça, transporte e finanças. Quando um algoritmo de aprendizado de máquina toma uma decisão importante, como aprovar um empréstimo, sugerir tratamentos médicos, selecionar dosagens otimizadas de fármacos, dimensionar estruturas civis ou até mesmo guiar um carro autônomo, é essencial que haja uma explicação compreensível sobre como a decisão foi tomada. Isso é especialmente relevante quando há potenciais consequências negativas, como a integridade física, a saúde mental e o bem-estar de pessoas.

Alguns programadores alegam que a opacidade se deve à necessidade de proteger seus trabalhos da concorrência e à privacidade dos dados - o que cria um conflito ético entre os princípios da transparência e da privacidade (Rossetti; Angeluci, 2021). No entanto, para que os algoritmos de decisão sejam bem aceitos pela sociedade, entender seu funcionamento é essencial (Winfield; Jirotka, 2018). Por essa razão, a interpretabilidade e a transparência de algoritmos têm sido focos crescentes de pesquisa em IA. No caso dos carros autônomos, por exemplo, que já geraram acidades fatais (Stilgoe, 2018), há, claramente, uma necessidade urgente de transparência para descobrir como e por que esses acidentes ocorreram (Winfield; Jirotka, 2018).

1.4 EVIDÊNCIAS EQUIVOCADAS E OS VIESES

A automação das tomadas de decisões pelos algoritmos muitas vezes é justificada por uma suposta ausência de viés nesses processos, diferente das decisões humanas (Mittelstadt *et al.*, 2016). Essa crença é insustentável, conforme relatam estudos que demonstram que os algoritmos, inevitavelmente, tomam decisões tendenciosas (Doneda *et al.*, 2018; Garcia, 2020; Karale, 2021; Rossetti e Angeluci, 2021). A qualidade das decisões automatizadas baseadas em algoritmos está intimamente relacionada com a qualidade dos dados processados (Garcia, 2020; Mittelstadt *et al.*, 2016; Rossetti e Angeluci, 2021). Dessa forma, se o algoritmo é desenvolvido a partir de dados

preconceituosos, ele reproduz, de forma automatizada, os mesmos padrões preconceituosos (Doneda *et al.*, 2018).

A Inteligência Artificial tem o potencial de revolucionar a área da saúde, incluindo a detecção de câncer de pele (Adamson; Smith, 2018). No entanto, se não for desenvolvida de forma inclusiva, a IA pode agravar as diferenças de gênero, raça, entre outras. Segundo Adamson e Smith (2018), na área da dermatologia, a maioria dos programas de detecção de doenças de pele é treinado com dados de pele clara, o que pode resultar em resultados enviesados e menor desempenho em imagens de pele mais escura. Essa limitação pode ter consequências preocupantes no diagnóstico de melanoma, que pode se apresentar de forma diferente em peles mais escuras (Adamson; Smith, 2018). De acordo com os autores, para solucionar essa questão, os algoritmos devem ser treinados para reconhecer o melanoma em todos os tipos de pele, e os repositórios devem incluir mais fotos com condições de pele diversificadas. Caso contrário, as tecnologias de aprendizado de máquina e aprendizado profundo correm o risco de agravar ainda mais as disparidades na área da saúde.

Outro exemplo do viés algoritmo é o caso do *Google* Fotos, que em 2015 etiquetava pessoas negras como gorilas (Rossetti; Angeluci, 2021). A diversidade e a representatividade dos dados são fundamentais para evitar a reprodução de preconceitos e desigualdades no desenvolvimento de tecnologias de IA. Isso é essencial para diversas áreas que usam imagens de seres humanos além do diagnóstico de doenças, como as de monitoramento de segurança em canteiros de obras, identificação de criminosos, classificação de massa corporal, leitura de linguagem de sinais, reconhecimento de gestos, entre outros. Assim, a busca por modelos mais justos e equitativos é um dos maiores desafios no desenvolvimento de algoritmos de aprendizado de máquina éticos e confiáveis.

1.5 RESULTADOS INJUSTOS E DISCRIMINAÇÃO

Decisões automatizadas que contenham vieses podem levar à discriminação (Rossetti; Angeluci, 2021). O viés está relacionado à própria formulação da decisão,

enquanto a discriminação refere-se aos resultados negativos desproporcionais resultantes das decisões algorítmicas (Mittelstadt; Allo, *et al.*, 2016). Em outras palavras, o viés está presente na forma como o algoritmo toma decisões, e a discriminação é o efeito prejudicial que essas decisões podem ter em termos de impacto desigual.

Obermeyeret *et al.* (2019) publicaram um estudo mostrando o viés racial em sistemas da área da saúde nos Estados Unidos. Segundo o estudo, uma empresa de seguro saúde norte-americana, com objetivo de reduzir seus custos, decidiu oferecer tratamentos preventivos para pacientes com doenças crônicas em estado grave. Para selecionar os casos mais graves, a empresa considerou o número de vezes que o paciente utilizava o sistema de saúde. Quanto mais o paciente usava o sistema, mais crítico era considerado o seu caso. Para evitar vieses, dados como raça e gênero foram excluídos da base de dados do algoritmo. No entanto, a grande maioria dos selecionados foram pessoas brancas, o que fez os pesquisadores se perguntarem: se a base de dados não continha informações sobre raça, por que o algoritmo ainda privilegiava os brancos? O que se conclui é que pessoas negras utilizavam menos o sistema de saúde por não poderem arcar com os custos de coparticipação. Dessa forma, o conjunto de dados não continha muitas informações de negros que utilizavam o sistema de saúde, gerando um algoritmo racista.

A mera exclusão direta de características como raça e gênero da base de dados não é suficiente para garantir a equidade das decisões algorítmicas (Obermeyer *et al.*, 2019). Portanto, a conscientização sobre a interação crítica entre viés e discriminação é essencial para que os desenvolvedores e tomadores de decisão compreendam não apenas as nuances técnicas, mas também as implicações sociais mais amplas desses sistemas automatizados. Assim, ressalta-se a importância da elaboração e verificação da representatividade dos bancos de dados e das premissas de decisão dos algoritmos, que devem ser tão rigorosas quanto os possíveis efeitos da discriminação causados pelos modelos.

1.6 AUTONOMIA E PRIVACIDADE

Os algoritmos de personalização coletam dados pessoais dos usuários para criar experiências altamente personalizadas. Esses algoritmos filtram quais informações são apresentadas ao usuário com base na compreensão de preferências, comportamentos e talvez até vulnerabilidades a serem influenciadas – uma linha tênue entre auxiliar e controlar decisões (Mittelstadt *et al.*, 2016). A personalização é nítida no funcionamento das redes sociais. Mas se o que vemos é manipulável, será que temos autonomia para tomar nossas decisões?

Os algoritmos também estão transformando a nossa noção de privacidade (Mittelstadt et al., 2016). Com a crescente utilização de *smartphones*, *smartwatches*, sensores, dispositivos de monitoramento médico, entre outros, manter nossa privacidade tornou-se uma tarefa desafiadora (Karale, 2021). A popularização de dispositivos como *Amazon Echo* e *Google Home* levantam questões se esses dispositivos ouvem nossas conversas e quem tem acesso a esses dados (Karale, 2021).

Para Rossetti e Angeluci (2021), a autonomia do indivíduo é desrespeitada quando sua escolha pode estar vinculada a interesses de terceiros. Os usuários podem não estar cientes do grau de detalhamento em que suas informações são coletadas e como são utilizadas pelas empresas. Isso levanta preocupações sobre o consentimento informado e a transparência no uso de dados pessoais. Além disso, a personalização pode criar "bolhas", já que os usuários são expostos apenas a informações e opiniões que reforçam suas visões existentes. Isso pode levar à polarização e à disseminação de desinformação, uma vez que os algoritmos tentam manter os usuários engajados mostrando-lhes conteúdos que eles já concordam, em vez de oferecer perspectivas diversificadas (Knobloch-Westerwick; Westerwick, 2023).

A onipresença dos algoritmos de personalização desencadeia uma reflexão profunda sobre os limites da nossa capacidade de tomar decisões autônomas. Assim, é imperativo que avancemos em direção a abordagens mais éticas, transparentes e responsáveis na criação e implementação de algoritmos de personalização, assegurando

que o progresso tecnológico respeite os princípios fundamentais de autonomia, privacidade e diversidade.

1.7 RASTREABILIDADE E RESPONSABILIDADE

Tradicionalmente, os desenvolvedores têm controle sobre o funcionamento de algoritmos, na medida em que podem explicar seu design e função a terceiros (Mittelstadt *et al.*, 2016). Esse entendimento de responsabilidade no *design* de *softwares* pressupõe que o programador é capaz de prever possíveis erros e de corrigir resultados indesejáveis (Mittelstadt *et al.*, 2016). No entanto, com o advento dos algoritmos de aprendizado de máquina "caixas pretas", principalmente dos de aprendizagem não supervisionada, o modelo tradicional de responsabilidade deixa de existir. Não há controle suficiente sobre as tomadas de decisões dos algoritmos, e, portanto, não há quem possa assumir a responsabilidade por elas (Mittelstadt *et al.*, 2016).

Os carros autônomos ilustram bem esta preocupação (Winfield e Jirotka, 2018; Doneda *et al.*, 2018; Stilgoe, 2018). No caso de acidentes envolvendo carros autônomos, quem é o responsável? O(a) programador(a)? A empresa que desenvolveu o carro? Segundo Mittelstadt *et al.* (2016), ainda não há um consenso sobre a responsabilidade dos algoritmos. Porém, há uma crescente opinião de que os algoritmos do futuro precisarão, no mínimo, serem projetados para refletir as normas éticas e culturais de seus usuários e da sociedade (Winfield; Jirotka, 2018).

2. OS PRINCIPAIS OBSTÁCULOS EM APRENDIZADO DE MÁQUINA

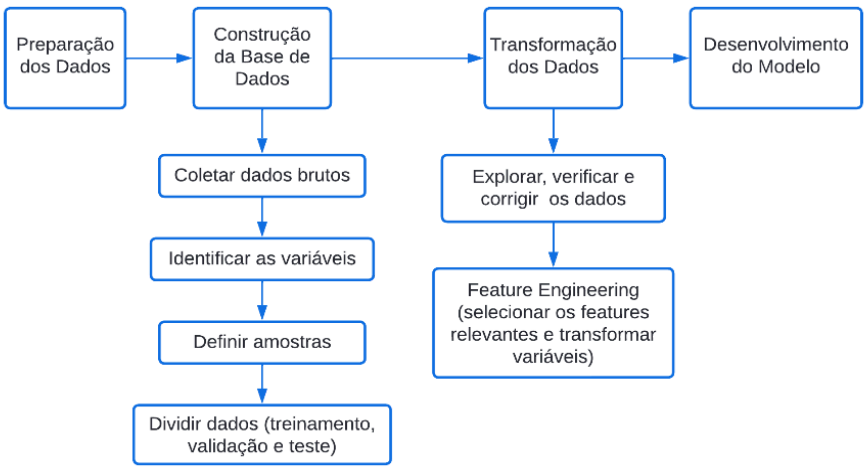
No desenvolvimento de algoritmos de aprendizado de máquina, superficial ou profundo, os(as) programadores(as) precisam evitar algumas armadilhas que podem ter efeitos graves no desempenho dos modelos. Portanto, esta seção abordará cinco obstáculos comuns no desenvolvimento de algoritmos de aprendizado de máquina: dados de baixa qualidade, sobreajuste (*overfitting*), subajuste (*underfitting*), má seleção da métrica de avaliação e o vazamento de dados (*data leakage*).

2.1 DADOS DE BAIXA QUALIDADE

O aprendizado de máquina nos ajuda a encontrar padrões em um conjunto de dados, que são utilizados para fazer previsões sobre novos dados (*Google* AI, 2023). Como vimos, dados de baixa qualidade são um desafio significativo no desenvolvimento de algoritmos de aprendizado de máquina, uma vez que a qualidade dos dados usados para treinar, validar e testar modelos de aprendizado de máquina tem um impacto direto na precisão, eficácia e confiabilidade dos algoritmos.

Dados de baixa qualidade podem significar: dados insuficientes; dados desbalanceados (existência de categorias sub-representadas); dados incompletos; e *outliers* (pontos fora da curva) (Nex Software Development Company, 2023). Esses dados de baixa qualidade podem levar a resultados imprecisos, viés algorítmico e consequências indesejáveis (modelo pouco acurado ou discriminação). Portanto, é essencial realizar uma análise cuidadosa dos dados antes do desenvolvimento de modelos. De acordo com um curso de preparação de dados para aprendizado de máquina, disponibilizado pelo Google (Google AI, 2023), a construção de um bom conjunto de dados deve se dar conforme o fluxograma da Figura 1.

Figura 1 - Fluxograma da preparação de dados



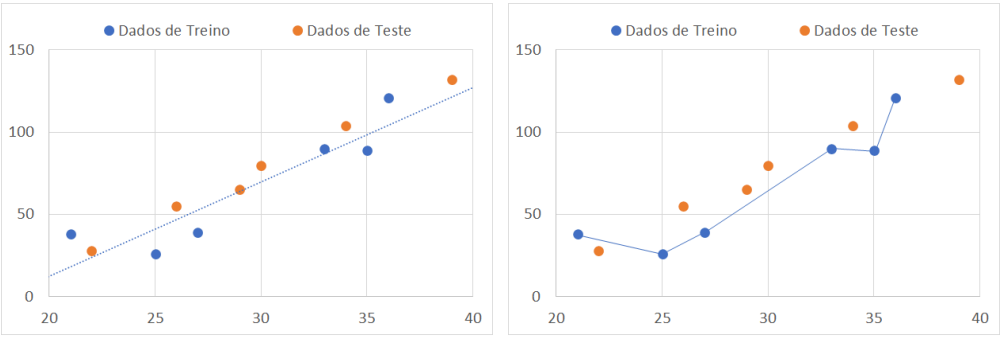
Fonte: Adaptado pelo Autor (Google AI, 2023).

2.2 SOBREAJUSTE - OVERFITTING

O sobreajuste - *overfitting* - ocorre quando o modelo se ajusta tão bem aos dados de treinamento que perde a capacidade de generalização (Santos *et al.*, 2018). Dessa forma, quando validado junto ao conjunto de teste, o desempenho é significativamente inferior quando comparado ao desempenho com o conjunto de treinamento (Santos *et al.*, 2018). Pode-se dizer, neste caso, que o algoritmo "decorou" os dados de treino, e, portanto, quando os dados de teste são introduzidos, o modelo tenta aplicar as mesmas regras decoradas, obtendo resultados piores (já que os dados são diferentes) (Santos *et al.*, 2018). Segundo Correa (2021), o *overfitting* é comparável a um estudante que estuda para uma prova decorando informações ou exercícios, e quando na prova surgem questões diferentes das decoradas, o estudante não sabe resolvê-las.

A Figura 2 explica a ocorrência do *overfitting*. Na Figura 2, no gráfico à esquerda, os dados estão divididos em dados de treino (azul) e de teste (laranja). A linha pontilhada azul representa a linha de tendência do modelo, considerando os dados de treino. Aparentemente, esta linha também se ajusta aos dados de teste (laranja). Já no gráfico à direita, a linha de tendência do modelo para os dados de treino (azul), não se ajusta bem aos dados de teste (laranja). Ou seja, o modelo é muito adequado para os dados de treino (azul), mas perde a capacidade de generalização para os dados de teste (laranja), demonstrando o *ovefitting*.

Figura 2 - Sobreajuste *(overfitting)*



Fonte: Elaborado pelos Autores (2023).

Uma das maneiras mais adotadas de se combater o *overfitting* é por meio da validação cruzada (*k-fold*) (Wagner e Rondinelli, 2016). Neste processo, o(a) cientista de dados divide os dados em partições iguais, usando uma fração dos dados para testar o desempenho do modelo e o restante para treinamento, em um processo repetitivo (Santos *et al.*, 2018). Os dados são particionados em k subconjuntos, chamados *folds* (Santos *et al.*, 2018). As partições são então iteradas para que cada partição tenha sido usada como conjunto de teste uma vez e os erros sejam calculados (Santos *et al.*, 2018), conforme ilustra a Figura 3. Isso tem o efeito de simular a precisão com que o modelo lidará com novas observações (Wagner e Rondinelli, 2016). Assim, o uso da validação cruzada pode reduzir os vieses (agrupamentos desbalanceados) que podem ocorrer ao se separar simplesmente o banco de dados em conjunto de treino e conjunto de teste.

Figura 3 - Esquema do processo de validação cruzada (k-fold) para k = 10

Fonte: Elaborado pelos Autores

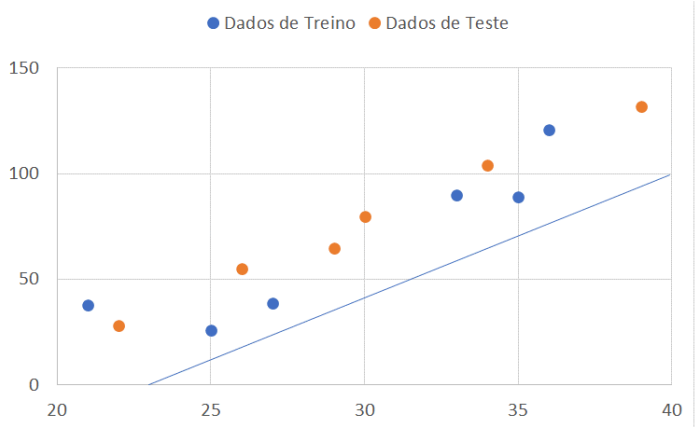
2.3 SUBAJUSTE - UNDERFITTING

Já o *underfitting* (subajuste) ocorre quando o modelo não é capaz de aprender os padrões presentes nos dados de treinamento de forma adequada. Isso significa que o modelo não consegue se ajustar bem aos dados, resultando em baixo desempenho tanto nos dados de treinamento quanto nos dados de teste. Nesse problema, os resultados tanto dos dados de treinamento quanto dos dados de teste são ruins (Nex Softsys Software Development Company, 2023). Dessa forma, é mais fácil descobrir se o modelo

está sofrendo de *underfitting* do que *overfitting*, já que no *underfitting* as métricas de validação serão ruins, ao passo que métricas boas podem tanto significar um bom desempenho geral do modelo quanto *overfitting* ou a ocorrência de vazamento de dados (que será visto nas próximas seções) (Nex Softsys Software Development Company, 2023).

Uma das formas de combater o *underfitting* é aumentar o tamanho do conjunto de dados de treino, ou adicionar mais atributos *(features)* ao modelo, tornando o modelo um pouco mais complexo (Nex Softsys Software Development Company, 2023). A Figura 4 representa uma situação de *underfitting*. A linha de tendência (em azul) não representa bem os dados de treino e de teste. Ou seja, o modelo não está prevendo adequadamente o comportamento desse fenômeno, o que pode derivar de diversas razões, como a baixa qualidade do banco de dados, uso de algoritmos com mecanismos incompatíveis com o problema em questão, má calibração dos parâmetros de ajuste dos algoritmos (*tunning*), entre outros.

Figura 4 - Linha de tendência subajustada (*underfitting*)



Fonte: Elaborado pelos Autores

2.4 MÁ SELEÇÃO DA MÉTRICA DE AVALIAÇÃO

As métricas de avaliação ajudam a avaliar o desempenho dos modelos de aprendizado de máquina. A seleção adequada da métrica de avaliação é uma etapa crucial no desenvolvimento de modelos, pois influencia diretamente as decisões tomadas

durante o processo de treinamento e ajuste do modelo. Uma má escolha da métrica de avaliação pode levar a conclusões enganosas sobre o desempenho do modelo e resultar em decisões prejudiciais.

Existem muitos tipos de métricas de avaliação disponíveis (Altexsoft, 2023). No entanto, não existe uma regra a respeito de quais métricas devem ser utilizadas, já que isso depende do tipo de modelo que está sendo desenvolvido (Nex Softsys Software Development Company, 2023). Os modelos de regressão, por exemplo, podem ser avaliados pelas métricas Mse (Erro Quadrado Médio), Rmse (Raiz do Erro Quadrado Médio), Mae (Erro Médio Absoluto), Mape (Erro médio percentual absoluto), coeficiente de determinação (R^2), entre outras (Altexsoft, 2023). Já os modelos de classificação, que tem como objetivo classificar dados (exemplo: classificar um e-mail como spam ou não-spam), geralmente utilizam métricas como acurácia, precisão, *recall e F1-score* (Altexsoft, 2023).

A acurácia, por exemplo, uma das métricas de avaliação mais simples, determina o número de previsões corretas do modelo (Altexsoft, 2023). No entanto, se o conjunto de dados utilizado estiver desbalanceado, alguma categoria estiver sub-representada, essa métrica é inadequada para avaliar o modelo (Nex Softsys Software Development Company, 2023; Altexsoft, 2023). Por exemplo, em um conjunto de treinamento em que 10% dos dados referem-se a pessoas com alguma doença e 90% a pessoas saudáveis, se o modelo não for bem elaborado e classificar todas as pessoas como "saudáveis", errando a classificação de todos os doentes, a acurácia é de 90% (Altexsoft, 2023). Isso causa uma falsa impressão de bom desempenho do modelo. Portanto, a acurácia não é uma métrica adequada nesse caso (Nex Softsys Software Development Company, 2023).

2.5 VAZAMENTO DE DADOS - DATA LEAKAGE

O vazamento de dados - *data leakage* - em aprendizado de máquina não está relacionado ao contexto de segurança, mas ao vazamento de informações do conjunto de treinamento para o conjunto de teste (Neoway, 2023). Nesse caso, o desempenho nos conjuntos de teste será semelhante ao obtido nos conjuntos de treinamento, não porque o

modelo seja capaz de prever corretamente, mas porque existem padrões semelhantes nas amostras de treinamento e teste (Santos, Soares, *et al.*, 2018). Se o modelo estiver alcançando resultados muito bons em todas as avaliações, pode ser que esteja ocorrendo vazamento de dados (Nex Softsys Software Development Company, 2023). Nesse caso, o modelo é chamado de excessivamente otimista (*overly optimistic*) (Santos, Soares, *et al.*, 2018).

O vazamento de dados pode ser ilustrado com a situação de um aluno que está estudando para uma prova e tem acesso ao gabarito (Neoway, 2023). O conhecimento do gabarito certamente influenciará os seus estudos. Provavelmente o aluno terá ótimos resultados na prova, mas isso não significa que ele estará preparado para aplicar o conhecimento em situações diferentes (Neoway, 2023). No aprendizado de máquina é semelhante: se ocorre o vazamento de dados, há um bom desempenho de treinamento e teste, mas não há um bom desempenho em situações reais (Nex Softsys Software Development Company, 2023). Portanto, estratégias rigorosas devem ser impreterivelmente adotadas na divisão dos dados em dados de treinamento e teste.

3. AS MELHORES PRÁTICAS EM IA

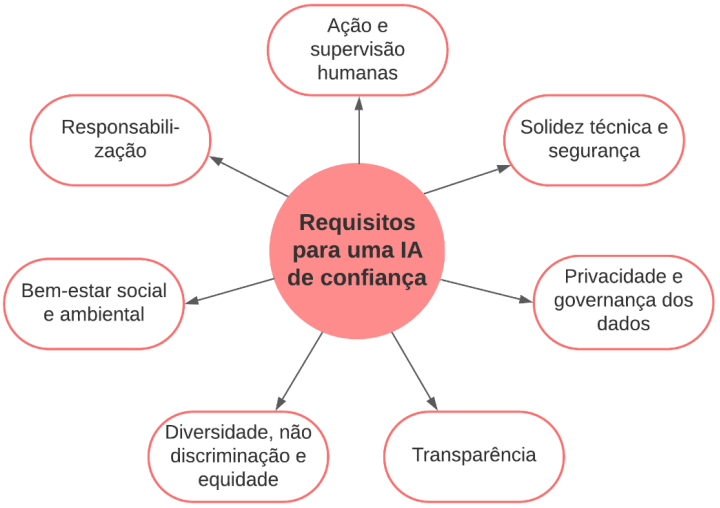
O desenvolvimento acelerado da Inteligência Artificial nos últimos anos e o seu potencial impacto em diversos setores da sociedade estão provocando um amplo debate sobre os princípios e valores que devem guiar o seu desenvolvimento (Jobin, Ienca e Vayena, 2019). O conhecimento das implicações legais, éticas e de segurança do uso da IA são fundamentais para grupos de pesquisa e entidades públicas e privadas na atualidade (Koshiyama *et al.*, 2021; *Google AI*, 2022). Para Koshiyama *et al.* (2021), assim como na auditoria financeira, eventualmente governos, empresas e sociedade exigirão a auditoria de algoritmos, ou seja, a garantia formal de que os algoritmos são legais, éticos e seguros.

No que diz respeito à pesquisa científica, não se tem conhecimento, até o momento, de uma diretriz universal específica para pesquisa científica em IA publicada

por comitês de ética que se aplique a todos os contextos e jurisdições. No entanto, as considerações éticas em torno da pesquisa em IA estão se tornando cada vez mais importantes. Nesse sentido, várias instituições e órgãos reguladores publicaram relatórios e documentos a respeito das melhores práticas a serem adotadas no desenvolvimento das IAs nos últimos anos (Jobin, Ienca e Vayena, 2019). Organizações profissionais, como a *Association for Computing Machinery* (ACM) e o *Institute of Electrical and Electronics Engineers* (IEEE), desenvolveram diretrizes éticas e códigos de conduta para pesquisadores de IA (Mittelstadt, 2019). Essas diretrizes geralmente enfatizam princípios como transparência, responsabilidade, justiça e prevenção de danos (Mittelstadt, 2019). No entanto, segundo Mittelstadt (2019), esses documentos permanecem comparativamente curtos, teóricos e carentes de aconselhamento e normas comportamentais específicas.

Em 2019, a Comissão Europeia nomeou um Grupo de Especialistas de Alto Nível em Inteligência Artificial para elaboração do documento "Orientações Éticas para uma IA de Confiança" (High-Level Expert Group On AI, 2019). De acordo com o documento, uma IA de confiança deve ser: legal - garantindo o respeito à legislação existente, ética - observando princípios e valores éticos; e sólida do ponto de vista técnico e social. O documento também prescreve quatro princípios éticos imprescindíveis no desenvolvimento das IAs: respeito pela autonomia humana, prevenção de danos, justiça e explicabilidade. A partir desses princípios, o documento elabora sete requisitos que devem ser aplicados neste contexto, conforme ilustra a Figura 5.

Figura 5 - Requisitos para uma IA de confiança



Fonte: Adaptado de High-Level Expert Group On AI, 2019.

O Google (2023) também disponibiliza em seu site "Google AI" uma lista de recomendações a serem adotadas no desenvolvimento de algoritmos de IA. Nesta lista há recomendações gerais, para o desenvolvimento de algoritmos justos, para garantir a interpretabilidade, para garantir a privacidade e para garantir a segurança. Essas recomendações serão abordadas a seguir de forma resumida.

3.1 RECOMENDAÇÕES GERAIS

- **usar uma abordagem de design centrado no ser humano:** a experiência de usuários(as) reais com o sistema é essencial para avaliar o verdadeiro impacto de suas previsões, recomendações e decisões. Recomenda-se a utilização de recursos de design que garantam uma boa experiência do(a) usuário(a) e a incorporação de *feedbacks* antes e durante o desenvolvimento do projeto;
- **utilizar várias métricas para avaliar o modelo**: o uso de várias métricas ajudará a entender diferentes tipos de erros e experiências;
- **examinar diretamente os dados brutos:** os modelos refletirão os dados em que são treinados, portanto, sempre que possível, deve-se analisar os dados brutos

para verificar a existência de dados faltosos, dados desbalanceados, dados redundantes, possíveis vieses, entre outros **-** sempre respeitando-se a privacidade;

- **entender as limitações do conjunto de dados e do modelo:** um modelo treinado para detectar correlações estatísticas não deve ser utilizado para fazer inferências causais. Embora duas variáveis possam estar correlacionadas entre si, isso não significa que uma variável seja a causa da outra. Portanto, é importante entender o escopo e as limitações dos modelos, bem como comunicar as limitações aos usuários;
- **testar o modelo várias vezes:** testar os modelos de aprendizado de máquina superficial e profunda é uma prática crucial para garantir a eficácia, robustez e segurança dos algoritmos em aplicações reais. Portanto, recomenda-se a realização de testes rigorosos para testar cada componente do sistema isoladamente e testes de integração para entender como os componentes do sistema interagem com outras partes;
- **monitorar a atualizar os modelos:** atualizar e monitorar o desempenho dos modelos de aprendizado de máquina são práticas fundamentais para garantir que os algoritmos de IA continuem eficazes e relevantes ao longo do tempo.

3.2 RECOMENDAÇÕES PARA UM ALGORITMO JUSTO

- **projetar o modelo para que seja imparcial:** o envolvimento de uma equipe multidisciplinar no desenvolvimento de algoritmos, como advogados(as), cientistas sociais, engenheiros(as), cientistas da computação, entre outros, é essencial para que se leve em conta diferentes perspectivas;
- **usar conjunto de dados representativos:** recomenda-se o uso de dados representativos para treinar e testar os modelos ou deixar claro ao(a) usuário(a) e demais desenvolvedores quando existirem grupos sub-representados;
- **verificar a existência de vieses:** recomenda-se selecionar conjuntos de dados de teste diversificados que possam testar o sistema de forma adversa. Além disso, recomenda-se atualização contínua dos sistemas e o encorajamento de *feedback* de usuários;

- **analisar o desempenho:** recomenda-se avaliar o desempenho continuamente e em um amplo espectro de usuários.

3.3 RECOMENDAÇÕES PARA GARANTIR A INTERPRETABILIDADE

- **planejar as ações para evitar a opacidade:** trabalhar em colaboração com diferentes especialistas contribui para identificação de quais recursos de interpretabilidade são necessários e por quê;
- **tratar a interpretabilidade como uma parte central da experiência do usuário:** recomenda-se a interação com os usuários no ciclo de desenvolvimento para testar e refinar suposições sobre as necessidades e metas do usuário;
- **projetar o modelo para ser interpretável:** recomenda-se a utilização do menor número de entradas possíveis para que se entenda os fatores que influenciam o modelo e o uso preferencial de algoritmos que permitam a interpretabilidade dos cálculos, decisões e etapas intermediárias de desenvolvimento;
- **comunicação com o usuário:** é importante que os modelos ofereçam explicações compreensíveis e adequadas para os diferentes usuários – sempre que possível.

3.4 RECOMENDAÇÕES PARA GARANTIR A PRIVACIDADE

- **coletar e manipular os dados com responsabilidade:** recomenda-se que os modelos sejam treinados sem o uso de dados confidenciais. Se não for possível, os dados confidenciais devem ser manuseados com cuidado e deve-se respeitar a legislação existente;
- **proteger a privacidade dos dados:** recomenda-se verificar se o modelo está memorizando ou expondo involuntariamente dados confidenciais ou permitindo inferências e associações.

3.5 RECOMENDAÇÕES PARA GARANTIR A SEGURANÇA

- **identificar possíveis ameaças ao sistema:** recomenda-se o desenvolvimento de um modelo de ameaças rigoroso para entender todos os vetores de ataque possíveis;
- **desenvolver uma abordagem para combater ameaças:** testar o desempenho dos sistemas em cenários adversos e ter um plano em caso de problemas;
- **continuar aprendendo:** manter-se atualizado sobre os últimos avanços na área é imprescindível para melhorias no desempenho de defesas.

Há algumas críticas a respeito desses documentos que propõe diretrizes para o desenvolvimento de IAs éticas. Segundo Larsson (2020), o documento "Orientações Éticas para uma IA de Confiança" da Comissão Europeia enfatiza questões de responsabilidade, transparência e proteção de dados no desenvolvimento de tecnologias de IA confiáveis, mas não aborda a interação entre a legislação e essas diretrizes. De acordo com Jobin, Ienca e Vayene (2019), traduzir os princípios éticos para a prática e aliar a ética à legislação são próximos passos importantes para a comunidade global.

Jobin, Ienca e Vayene (2019) analisaram 84 documentos que tratavam sobre princípios e diretrizes para o desenvolvimento de tecnologias de IA éticas e observaram que há uma convergência global em torno de cinco princípios éticos: transparência, justiça e equidade, não-maleficência, responsabilidade e privacidade. No entanto, existem divergências relacionadas à interpretação, importância e implementação desses princípios. Essas diferenças podem prejudicar a criação de uma agenda global para IA ética, já que há incerteza sobre quais princípios priorizar e como resolver conflitos entre eles (Jobin, Ienca e Vayena, 2019). Por exemplo, a necessidade de conjuntos de dados cada vez maiores e diversificados para evitar vieses pode entrar em conflito com a exigência de dar aos indivíduos maior controle sobre seus dados, a fim de respeitar sua privacidade e autonomia. Além disso, os autores observaram que os países do norte global, principalmente os EUA, têm uma representação proeminente no debate a respeito

da IA ética. Essa sub-representação do sul-global levanta preocupações a respeito da equidade e justiça global na discussão da ética em IA (Jobin, Ienca e Vayena, 2019).

O debate da ética em IA deve envolver uma força de trabalho diversificada, capaz de incorporar conhecimentos críticos variados e de verificar possíveis vieses (GOOGLE AI, 2023). Segundo o Grupo de Especialistas de Alto Nível em Inteligência Artificial, que elaborou o documento "Orientações Éticas para uma IA de Confiança", mais do que a definição de um conjunto de regras, assegurar a confiança na IA requer a construção e manutenção de uma cultura e uma mentalidade ética por meio do envolvimento em debates públicos, da educação e da aprendizagem prática (High-Level Expert Group On AI, 2019).

4. CONSIDERAÇÕES FINAIS

Com o rápido e crescente desenvolvimento da IA nas mais variadas áreas de estudo, entender seu impacto social e implicações éticas se tornou uma necessidade da sociedade atual. Nesse sentido, este artigo abordou as principais questões éticas relacionadas à IA, os principais obstáculos no desenvolvimento de algoritmos de aprendizado de máquina superficial e profundo, e as melhores práticas para desenvolver algoritmos éticos, interpretáveis, seguros e justos.

Vieses e preconceitos podem se perpetuar facilmente através do uso de dados desbalanceados e de correlações infundadas. Diante disso, a colaboração entre desenvolvedores de algoritmos e outros especialistas (especialmente cientistas sociais) torna-se fundamental para compreender diversas perspectivas e identificar as sutis formas de propagação dos preconceitos. Além disso, é imprescindível o uso de conjuntos de dados representativos e balanceados para o treinamento e teste dos modelos de aprendizado de máquina.

A opacidade dos algoritmos é um dos principais obstáculos para a aplicação e aceitação das técnicas de IA. A falta de transparência e auditabilidade em algoritmos que tomam decisões governamentais, mediam transações bancárias ou fornecem diagnósticos

médicos parece injusta e preocupante. A compreensão lógica dessas decisões é essencial para detectar falhas e atribuir responsabilidades. Portanto, é urgente a implementação de governança ética, transparente, inclusiva e ágil pelas organizações que desenvolvem e operam as IAs, já que a ampla aceitação futura desses algoritmos dependerá dessas questões. Assim, é possível que, em breve, exija-se que os algoritmos sejam auditáveis, a fim de garantir a ética, legalidade e segurança dos modelos. Pesquisadores(as) e desenvolvedores(as) devem se preparar para esse cenário o quanto antes.

A ética dos algoritmos não é uma questão que será solucionada apenas através de uma abordagem tecnológica. Essa temática envolve também assuntos de ordem social, cultural, jurídica e política. A implantação de algoritmos éticos requer uma abordagem holística, que leve em conta não apenas a eficácia e a precisão dos modelos, mas também suas implicações sociais e impactos na sociedade como um todo. O desenvolvimento tecnológico e a responsabilidade social devem caminhar lado a lado, a fim de evitar que as diferenças sociais cresçam ainda mais. Portanto, mais do que discutir os princípios éticos da IA, é preciso que esses princípios sejam colocados em prática por meio das legislações.

Por fim, com o rápido desenvolvimento da Inteligência Artificial e dos algoritmos de aprendizado de máquina, algumas das técnicas apresentadas nesse artigo podem se tornar obsoletas. Por isso, recomenda-se a contínua atualização a respeito das melhores práticas e avanços científicos relacionados ao tema.

5. AGRADECIMENTOS

Agradecemos à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Código 001 – bolsa de doutorado para Carolina Lopes) e à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, projeto número APQ-01838-21 para Julia Mendes) pelo apoio financeiro. Agradecemos também a colaboração do Grupo de Pesquisa em Ciência de Dados aplicada à Engenharia (CIDENG – CNPq).

REFERÊNCIAS

ADAMSON, Adewole S.; SMITH, Avery. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, v. 154, n. 11, p. 1247-1248, 2018.

ALTEXSOFT. *Machine Learning Metrics:* How to Measure the Performance of a Machine Learning Model, 16 jun. 2023. Disponível em: https://www.altexsoft.com/blog/machine-learning-metrics/. Acesso em: 26 abr. 2023.

CORREA, Danielle M. *Feature Engineering*: preparando dados para aprendizado de máquina. Ateliware, 12 maio 2021. Disponível em: https://ateliware.com/blog/feature-engineering. Acesso em: 29 jun. 2023.

GARCIA, Ana C. *Ética e Inteligência Artificial*. Computação Brasil, v. 43, p. 14-22, 2020.

GOOGLE AI. *Data Preparation and Feature Engineering in ML,* 2023. Disponível em: https://developers.google.com/machine-learning/data-prep. Acesso em: 11 maio 2023.

GOOGLE AI. *Responsible AI practices,* 2023. Disponível em: https://ai.google/responsibilities/responsible-ai-practices/. Acesso em: 17 maio 2023.

HIGH-LEVEL EXPERT GROUP ON AI. *Ethics guidelines for trustworthy AI.*

JOBIN, Anna; IENCA, Marcello; VAYENA, Effy. The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* v. 1, p. 389-399, 2019.

KARALE, Ashwin. The challenges of IoT addressing security, ethics, privacy, and laws. *Internet of Things*, v. 15, p. 100420, 2021.

KNOBLOCH-WESTERWICK, Silvia; WESTERWICK, Axel. Algorithmic personalization of source cues in the filter bubble: Self-esteem and self-construal impact information exposure. *New Media & Society,* v. 25, n. 8, 2023.

KOSHIYAMA, Adriano *et al. Towards algorithm auditing:* a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN, 2021.

LARSSON, Stefan. On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society,* v. 7, n. 3, p. 437-451, 2020.

MITTELSTADT, Brent. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence,* v. 1, p. 501-507, 2019.

MITTELSTADT, Brent D. *et al.* The ethics of algorithms: Mapping the debate. **B***ig Data & Society*, v. 3, n. 2, p. 1-21, 2016.

NEOWAY. *O que é data leakage e como essa falha pode contaminar as decisões*, 12 ago. 2023. Disponível em: https://blog.neoway.com.br/data-leakage/. Acesso em: 04 jul. 2023.

NEX SOFTSYS SOFTWARE DEVELOPMENT COMPANY. *Common Pitfalls Which May Mislead Results In Machine Learning,* 2023. Disponível em: https://www.nexsoftsys.com/articles/common-pitfalls-in-machine-learning.html. Acesso em: 13 abr. 2023.

OBERMEYER, Ziad. e*t al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, p. 447-453, 2019.

ROSSETTI, Regina; ANGELUCI, Alan. Ética Algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. *Galáxia,* v. 46, p. 1-18, 2021.

SANTOS, Miriam S. *et al.* Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. **I***EEE Computational Intelligence Magazine*, v. 13, n. 4, p. 59-76, 2018.

STILGOE, Jack. Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, v. 48, n. 1, p. 25-26, 2018.

WAGNER, Nicholas; RONDINELLI, James M. Theory-guided machine learning in materials science. *Frontiers in Materials,* v. 3, p. 28, 2016.

WINFIELD, Alan F.; JIROTKA, Marina. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences,* v. 376, n. 2133, p. 20180085, 2018.