http://dx.doi.org/10.23925/2179-3565.2023v15i1p133-142

RISUS - Journal on Innovation and Sustainability
volume 15, número 1 - 2024
ISSN: 2179-3565
Editor Científico: Arnoldo José de Hoyos Guevara
Editor Assistente: Vitória Catarina Dib
Avaliação: Melhores práticas editoriais da ANPAD

INTELIGÊNCIA HUMANA E INTELIGÊNCIA ARTIFICIAL E OS DESAFIOS DOS VIESES NOS ALGORITMOS DE IA

Human intelligence and artificial intelligence and the challenges of biases in AI algorithms

Erika Ribeiro Fernandes[1], Marcelo Augusto Vieira Graglia[2]
[1]Mestranda em Tecnologias da Inteligência e Design Digital pela Pontifícia, Universidade Católica de São Paulo – PUCSP, [2]Professor Doutor em Tecnologias da Inteligência e Design Digital pela Pontifícia Universidade Católica de São Paulo – PUCSP
E-mails: erikarflearn@gmail.com, mgraglia@pucsp.br

RESUMO

Este artigo reconhece as transformações profundas que a Inteligência Artificial impõe à sociedade. Estudo descritivo-exploratório, visa discutir os vieses algorítmicos e compreender seus impactos na sociedade. O artigo parte da compreensão da inteligência e aprendizado humanos sob uma perspectiva pluralista, baseada na análise de obras literárias e artigos científicos. Essa abordagem oferece um contexto no qual se possa conceber a IA e o aprendizado de máquina numa perspectiva de inovação em prol do bem-estar comum. A análise crítica ressalta a necessidade de abordagens éticas no desenvolvimento desses sistemas. Os tópicos discutidos enfatizam a importância da abordagem multidimensional na mitigação de vieses algorítmicos. Desde a seleção de dados até auditorias e responsabilização, a diversidade de perspectivas, tanto nos conjuntos de dados quanto nas equipes de desenvolvimento, é crucial. A implementação de treinamento contínuo e supervisão humana, reflete um compromisso contínuo com a transparência e equidade na inteligência artificial. Essas estratégias integradas são fundamentais para um desenvolvimento ético, transparente e equitativo da IA. Essa abordagem holística, envolvendo competências e pessoas diversificadas, treinamento contínuo e vigilância atenta, é vital para garantir o uso ético da IA em prol do bem-estar coletivo.
Palavras-chave: inteligência artificial; aprendizagem de máquina; viés algorítmico; impactos sociais; design ético.

ACEITO EM: 07/03/2024
PUBLICADO EM: 30/04/2024

RISUS - Journal on Innovation and Sustainability
volume 15, número 1 - 2024
ISSN: 2179-3565
Editor Científico: Arnoldo José de Hoyos Guevara
Editor Assistente: Vitória Catarina Dib
Avaliação: Melhores práticas editoriais da ANPAD

HUMAN INTELLIGENCE AND ARTIFICIAL INTELLIGENCE AND THE CHALLENGES OF BIASES IN AI ALGORITHMS

Inteligência humana e inteligência artificial e os desafios dos vieses nos algoritmos de IA

Erika Ribeiro Fernandes[1], Marcelo Augusto Vieira Graglia[2]
[1]Mestranda em Tecnologias da Inteligência e Design Digital pela Pontifícia, Universidade Católica de São Paulo – PUCSP, [2]Professor Doutor em Tecnologias da Inteligência e Design Digital pela Pontifícia Universidade Católica de São Paulo – PUCSP
E-mails: erikarflearn@gmail.com, mgraglia@pucsp.br

ABSTRACT

This article acknowledges the profound transformations that Artificial Intelligence imposes on society. A descriptive-exploratory study aims to discuss algorithmic biases and understand their impacts on society. The article starts from the understanding of human intelligence and learning from a pluralistic perspective, based on the analysis of literary works and scientific articles. This approach provides a context in which AI and machine learning can be conceived from an innovation perspective for the common good. The critical analysis emphasizes the need for ethical approaches in the development of these systems. The topics discussed highlight the importance of a multidimensional approach in mitigating algorithmic biases. From data selection to audits and accountability, diversity of perspectives, both in datasets and development teams, is crucial. The implementation of continuous training and human supervision reflects a continuous commitment to transparency and fairness in artificial intelligence. These integrated strategies are essential for the ethical, transparent, and equitable development of AI. This holistic approach, involving diverse skills and people, continuous training, and vigilant oversight, is vital to ensure the ethical use of AI for the collective well-being.
Keywords: artificial intelligence; machine learning; algorithmic bias; social impacts; ethical design.

INTRODUÇÃO

A Inteligência Artificial se tornou uma tecnologia presente na vida cotidiana, contribuindo para a automação de atividades lógicas, analíticas e cognitivas, gerando maior velocidade no tratamento e processamento de informações, em padrões que seriam impossíveis de atingir utilizando somente a capacidade humana. Os avanços científicos, em específico os estudos com máquinas artificiais, trouxeram ao campo da cognição humana contribuições significativas. A definição de inteligência ora considera apenas a razão, ora a combinação entre razão e emoção, ora inclui ainda aspectos biológicos e sociais no processo de inteligência e aprendizado. O artigo trata da conceituação de inteligência e aprendizado humano a partir de uma visão pluralista, utilizando diversas correntes de pensamento, com referências a Ausubel, Maturana e Varela, Morin e Piaget, e suas contribuições para entender melhor o termo inteligência e aprendizado humano. Também discute a Inteligência Artificial e o aprendizado de máquina, com uma breve descrição de *deep learning*, vieses algorítmicos e alguns dos principais impactos que causam na sociedade.

1 INTELIGÊNCIA E APRENDIZADO HUMANO

Diversos pensadores propuseram reflexões ou métodos para desenvolvimento do conhecimento e do processo de aprendizagem. Sócrates[1] propôs a Maiêutica, argumentativa, conduzindo alguém ao próprio conhecimento sem lhe acrescentar nada, onde cada resposta fornecia novas perguntas, até que se chegasse ao melhor entendimento. Ausubel (1968) trabalhou com o conceito de aprendizagem significativa, quando uma nova informação passa a adquirir significado para um aprendiz através de uma espécie de ancoragem em aspectos relevantes da estrutura cognitiva preexistente do indivíduo, por um processo de interação entre o novo conhecimento e o já existente, na qual ambos se modificam num ciclo de trocas. A teoria sociocultural de Vygotsky (1980), que estudou como as crianças aprendem à medida que se desenvolvem, considerava que, para poder aprender, as crianças necessitam da interação com outras pessoas, num ambiente colaborativo, no qual as interações sociais são necessárias tanto com o meio ambiente como com as pessoas ao seu redor. Já Piaget (2008), que também pautou seu trabalho na observação da criança, acreditava no aprendizado por descobrimento, baseado na autoconstrução do conhecimento, onde a criança pode aprender de forma autônoma, sem a necessidade de interagir com outras pessoas e, ainda que a interação possa favorecê-la, seria possível aprender assumindo um papel mais ativo e experimental. Nesse sentido, a aprendizagem é entendida como um processo que só acontece em situações de mudança; portanto, aprender seria, em parte, saber se adaptar as mudanças. Howard Gardner (1995) relata um potencial biopsicológico, que diverge do conceito tradicional de medida da inteligência, considerando que o campo da cognição humana evolui para um conjunto amplo de competências que não podem ser medidas, pois *"uma inteligência implica na capacidade de resolver problemas ou elaborar produtos que são importantes num determinado ambiente ou comunidade cultural"* (Gardner 1995, p. 21). Conforme o autor, as inteligências são um tanto independentes umas das outras, mas não funcionam sozinhas, elas precisam umas das outras e diante disso, o ser humano possui diferentes níveis de cada uma dessas inteligências e as organiza das mais diversas formas para realizar suas atividades cotidianas, dessa maneira desenvolve mais um tipo de inteligência em detrimento da outra de acordo com suas necessidades e interesses. Na abordagem de Maturana e Varela (2001), a inteligência e o aprendizado são fenômenos biológicos nas suas raízes, mentais nos seus meios e sociais nos seus fins. Para eles a aprendizagem é um sistema autopoético, ou de autoconstrução em que, ao mesmo tempo em que os sujeitos constroem conhecimentos, também são construídos por eles, sendo assim nós *"construímos o mundo em que vivemos durante nossas vidas. Por sua vez, ele também nos constrói ao longo dessa viagem comum"* (ibidem p.10). Em relação ao homem e a máquina artificial, consideram que os seres vivos são máquinas especiais que se distinguem da artificial por sua capacidade de se autor reproduzirem, sendo que a cognição humana possui função semelhante à das máquinas artificiais, ou seja, *"processar informações vindas do exterior e captadas por meio dos sentidos. Teríamos uma representação interna, localizada no cérebro como um computador"*, e a diferença existente entre a máquina artificial e o ser vivo está na capacidade auto organizativa e auto reprodutiva. (ibidem

[1] A maiêutica foi criada por Sócrates no século IV a.C.

p.10). Para os autores, esse processo de construção do conhecimento se dá de forma una, considerando que só se pode transcender quando existir uma troca entre os indivíduos:

Toda experiência cognitiva inclui aquele que conhece de um modo pessoal, enraizado em sua estrutura biológica, motivo pelo qual toda experiência de certeza é um fenômeno individual cego em relação ao ato cognitivo do outro, numa solidão que [...] só é transcendida no mundo que criamos junto com ele. (Maturana, Varela 2001 p. 22).

Os autores fazem um paralelo entre instinto e aprendizagem, onde em termos de instinto, as condutas do ser humano são determinadas pela dinâmica de estado de forma dependente da estrutura adquirida pela espécie no processo evolutivo. A aprendizagem é, portanto, fruto da história individual de acoplamento estrutural de um ser vivo, e os seres humanos são dinâmicos e produzem conhecimento a partir de seu próprio funcionamento. O antropólogo, sociólogo e filósofo Edgar Morin (1999) fala da inteligência como inteligência cega e seu oposto, a inteligência complexa, que leva ao pensamento complexo, e diz que a inteligência cega é *"incapaz de contextualizar, destruindo os conjuntos e as totalidades, isolando todos os seus objetos do seu meio ambiente"* e define a inteligência complexa como *"o tecido de acontecimentos, ações, interações, retroações, determinações, acasos, que constituem nosso mundo fenomênico"* (Morin, 1999 p. 33). O referido autor defendia um aprendizado com saberes mais sistêmicos pois, em sua concepção, a inteligência que só consegue agir quebrando o mundo complexo em pedaços, decompor o problema e transformar o multidimensional em uma única dimensão, enfraquece a possibilidade de compreensão e reflexão, *"eliminando assim as oportunidades de um julgamento corretivo ou de uma visão a longo prazo"*. Para ele, os desenvolvimentos disciplinares das ciências trouxeram as vantagens da divisão do trabalho, mas, em contrapartida, os inconvenientes do confinamento e do despedaçamento do saber, produzindo assim não somente o conhecimento e a elucidação, mas junto a ignorância e a cegueira. Sobre a diferença entre máquinas e seres vivos, Morin (1999) argumentava que ela está na complexidade do ser vivo que possui um *"princípio organizador que desenvolve suas qualidades superiores às de todas as máquinas baseando-se precisamente na desordem"*, pois o conjunto humano se reorganiza e pode funcionar a partir da degradação. Nesse sentido, o autor afirma que a diferença é perceptível, principalmente, no que tange a nossa biologia, pois nós possuímos dentro de nossas células o poder da auto-organização, enquanto a máquina artificial *"é de confiabilidade muito reduzida, ou seja, para e sofre avaria logo que um único de seus componentes se degrada. É tanto menos confiável quanto mais numerosos e interdependentes forem os seus componentes"* (MORIN, 2005, p. 298). O autor defende que as máquinas artificiais se degeneram a partir do momento em que são construídas, funcionando ou não, e a única maneira de evitar esse problema não depende da máquina, mas sim de um indivíduo do exterior que repara ou substitui as peças desgastadas. Isso significa que o poder regenerativo está fora da máquina e não depende dela, enquanto os humanos têm poder regenerativo dentro de suas células, e a máquina artificial certamente não.
A equifinalidade é a atitude dos seres vivos que lhes permite realizarem seus fins (seu "programa") por meios desviados, apesar de carências, de acidentes ou de obstáculos, enquanto a máquina, privada de um dos seus elementos ou de um dos seus alimentos, se deteriora, para ou fornece produtos errôneos. (MORIN, 2005, p. 298)

Santaella (2019) argumenta que a diferença entre inteligência humana e artificial é uma questão *"bem mais complexa, especialmente porque as definições de inteligência são muitas e nem sempre concordantes"* (Santaela, 2019, [65]). A autora esclarece que a utilização da Inteligência Artificial intensifica a era do computador e passa a funcionar, progressivamente, como uma extensão das operações mentais humanas, aumentando e acelerando nossas habilidades cognitivas. Para ela *"o conteúdo da IA é a inteligência humana, uma forma de inteligência precedente que a IA está provavelmente expandindo*". (ibidem, [83]) e aponta que a IA, por mais complexa que seja, ainda não é capaz de replicar a inteligência e aprendizado humanos.

2 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

A Inteligência Artificial (IA) é um ramo da ciência da computação que lida com a criação de máquinas inteligentes e tem se tornado mais acessível devido ao crescimento exponencial do poder de processamento, a diminuição do custo de armazenagem e gerenciamento de grandes quantidades de dados, e a capacidade de distribuir o processamento entre clusters melhorando a capacidade da análise e disponibilidade de dados, entre outros avanços tecnológicos. A primeira menção sobre uma espécie de inteligência artificial foi registrada na

década de 1840 por Lady Ada Lovelace[2] que se concentrou em símbolos e lógica, não tendo relação com as redes neurais evolutivas de hoje. A máquina que ela tinha em mente era a "máquina analítica", um dispositivo de engrenagens que, infelizmente, nunca foi totalmente construída. Os comentários de Ada Lovelace sobre um artigo do matemático italiano Luigi Menabrea são considerados a primeira sequência de instruções ou algoritmos destinados a ser executado por uma máquina, sendo assim, reputados como o primeiro programa de computador da história.

Um século depois, Alan Turing[3], em 1936, mostrou que toda computação possível pode, em princípio, ser realizada por um sistema matemático – que foi denominado como a máquina de Turing universal. Turing foi considerado o pai da ciência da computação moderna e da Inteligência Artificial. Seu trabalho sobre a teoria da computabilidade e inteligência artificial teve um grande impacto no desenvolvimento da ciência da computação, fornecendo uma formalização do que significa um algoritmo ser "inteligente". Turing colaborou com outros cientistas, como Norbert Wiener e John von Neumann. (Graglia, Lazzareschi, 2023)

Machine learning é o ramo da inteligência artificial (IA) que se concentra no uso de dados e algoritmos para mimetizar a maneira como os humanos aprendem, desenvolvendo o reconhecimento de padrões ou a capacidade de aprender de forma gradativa, fazendo ajustes sem serem especificamente programados para isso, melhorando sua precisão. Minharro (2022) explica que os algoritmos utilizam métodos computacionais para "aprender" informações diretamente dos dados, sem depender de equações predeterminadas, como modelos. Utilizando uma análise preditiva seja com o aprendizado supervisionado, não-supervisionado, semi-supervisionado ou aprendizado por reforço, os computadores conseguem identificar padrões em dados massivos e fazer previsões, sendo que seu desempenho melhora à medida que o número de amostras ou dados disponíveis para aprendizado aumenta. Portanto, ao invés de modelar e ensinar o computador em cada etapa do processo, são fornecidas instruções de como aprender a partir de exemplos e dados. Isso significa que as máquinas podem ser usadas para tarefas novas e mais sofisticadas sem que seja programado manualmente o passo a passo de solução. No aprendizado de máquina supervisionado, os algoritmos são treinados a partir de exemplos rotulados, fazendo predições alicerçadas em evidências; assim, os algoritmos trabalham com dados que já possuem respostas conhecidas. A máquina aprende de acordo com a interferência humana, onde o programador humano insere dados mostrando o que é "certo" e o que é "errado", e, então, o sistema aprende a fazer comparações e a resposta fornecida *(outputs)* será baseada nos exemplos que ele recebeu *(inputs)*. Já no aprendizado não supervisionado, o sistema age totalmente por si, descartando a necessidade de ação humana, buscando padrões ocultos nos dados que lhes foram apresentados, cabendo a ele identificar padrões e características em comum entre os dados que foram inseridos. Nesse tipo de aprendizado a clusterização é a técnica mais comum, onde diferentes dados são agrupados em categorias com características comuns. No Aprendizado semi-supervisionado, o algoritmo é treinado com base em uma combinação de dados rotulados e não-rotulados e costuma ser utilizado como fonte alternativa de informação sobre o problema a ser resolvido, garantindo maior capacidade de generalização à solução obtida, com baixo custo. No aprendizado por reforço, utiliza-se a lógica da "recompensa e punição". Dessa forma, o sistema aprende e decide quais são as melhores respostas ou ações a serem tomadas. Diante disso, Minharro (2022) explica que o *Deep Learning*, ou aprendizado profundo, é um subcampo do aprendizado de máquina que se refere ao uso de redes neurais com mais de três camadas. É um tipo de inteligência artificial que automatiza diversos processos de extração de recursos e elimina parte da intervenção humana do processo. Esse tipo de aprendizado tem sido usado em reconhecimento automatizado de fala, processamento de linguagem natural e classificação de imagens. Também é usado em outros campos, como robótica, marketing, finanças e saúde.

3 OS VIESES PRESENTES NOS ALGORITMOS

De acordo com Daniel Kahneman[4] (2012), vieses cognitivos são distorções de julgamento provocadas em diversas situações por certos atalhos de pensamento do cérebro, sendo que esses mecanismos de "pensar rápido"

[2] Ada Lovelace, disponível em www.biography.com/scholar/ada-lovelace Acesso em 17 jun. 2022 .e www.ime.unicamp.br/~apmat/ada-lovelace Acesso em 24 jun.2022
[3] Alan Turing, disponível em https://www.ufrgs.br/alanturingbrasil2012/area1.html Acesso em 14 jun. 2022.
[4] Daniel Kahneman, vencedor do Nobel de economia em 2002.

de forma instintiva, são úteis em termos de garantir a sobrevivência. Kahneman define as formas de pensar em dois sistemas cognitivos: o sistema 1, que é rápido, automático e impulsivo; e o sistema 2, que é lento, analítico e racional. O sistema 1 é paralelo, inconsciente e dirigido por emoções e associações. Ele permite realizar escolhas intuitivas, e nesse sistema estão situados os atalhos mentais associativos que proporcionam a tomada de decisão, ou seja, os vieses e as heurísticas cognitivas. Já o sistema 2 é lento, analítico, racional, sequencial, deliberativo, baseado em regras e que utiliza cálculos conscientes para chegar a decisões. Cada um dos sistemas tem a sua utilidade, porém muitas vezes se pensa estar executando um pensamento com o sistema 2 – analítico, sequencial – afinal é o lado mais consciente e racional dos dois sistemas, porém, esse pensamento está sendo executado pelo sistema 1, com base em reações emocionais e vieses cognitivos que passam desapercebidos pela maioria das pessoas. O sistema 1 é o responsável por 90% das decisões, pois gera sem esforço as impressões e sentimentos que norteiam as escolhas humanas. Vieses inconscientes são pressupostos, crenças ou atitudes aprendidas, as quais não são necessariamente cientes. Embora o viés seja um aspecto natural do funcionamento do cérebro humano, ele pode às vezes reforçar estereótipos. Um viés é uma forma tendenciosa de pensar, crenças ou generalizações estereotipadas, isto é, em peso desproporcional contra ou a favor de uma pessoa ou de um determinado grupo de pessoas[5], por exemplo. Esses vieses são padrões mentais sistemáticos que ocorrem em situações diversas que se desviam do julgamento racional, fruto de análises permeadas por pensamentos rápidos e intuitivos e que se ancoram na cultura, em experiências anteriores e nas diferentes expectativas que se tem sobre determinado assunto. Estes vieses, frequentemente, são registrados na forma de dados, armazenados em diversas bases. Estes bancos de dados, ao serem utilizados no processo de treinamento de algoritmos de inteligência artificial acabam por transmitir aos sistemas de IA certos vieses de determinados grupos humanos ou indivíduos. Assim, o sistema passa a reproduzir um padrão indesejável e, mesmo, ética e moralmente inaceitável. Esta reprodução, entretanto, pode gerar impactos negativos ainda maiores do que aqueles gerados originalmente pela ação de um indivíduo ou grupo humano, à medida que os sistemas de IA podem atingir uma maior potência pela possibilidade de atingir um número extraordinário de pessoas e de forma ultra veloz e, ainda, ocorrer de forma silenciosa, sem que tais ocorrências sejam detectáveis em seus estágios iniciais ou passíveis de identificação precoce por agentes humanos, sendo somente percebidas tempos depois, quando seus eventuais efeitos deletérios tenham se tornado irreversíveis ou causado impactos em grupos ou mesmo populações humanas (Graglia; Huelsen; Lazzareschi, 2021). Diversas ocorrências comprovam este problema, como o do sistema de recrutamento e seleção utilizado pela Amazon[6], que discriminou inscrições de mulheres em processos seletivos porque foi treinada em dados de currículos da força de trabalho vigente, em sua maioria masculina. Zhao et al (2017) realizaram uma pesquisa para estudar dados e modelos associados à classificação de objetos multicamadas e rotulagem visual de papéis semânticos. Eles descobriram que os conjuntos de dados que eles utilizaram para realizar aquelas tarefas "*continham viés de gênero significativos e que os modelos treinados nesses conjuntos de dados amplificam ainda mais o viés existente"* [7] (Zhao et al 2017). Por exemplo, quando inseridas informações acerca da atividade 'cozinhar', os resultados apresentaram 33% mais chances de que os resultados envolvessem mulheres do que homens, em um conjunto de treinamento e, um modelo treinado amplificou ainda mais a disparidade para 68% no momento do teste. A análise revelou que mais de 45% e 37% de verbos e objetos, respectivamente, exibem um viés a favor de um gênero maior que dois para um. Um outro caso, o da COMPAS[8], em que as previsões algorítmicas de reincidência de criminosos usadas pelos tribunais dos EUA foram questionadas pela Pro Publica[9], uma corporação sem fins lucrativos com sede em Nova York. Após um acompanhamento de dois anos, foi demonstrado que o algoritmo do COMPAS era tendencioso a favor de réus brancos e contra negros. O sistema é utilizado pelo Departamento de Correções de alguns estados como Flórida e Nova York, para auxiliar juízes a decidirem sobre a possibilidade de um réu responder seu processo em liberdade, atribuindo um grau de reincidência, visando analisar a probabilidade de fuga

[5] Exemplos são gerações, culturas, etnias, raças, classes sociais, orientações sexuais, idades, gêneros, religiões e outras questões comportamentais que não definem o que as pessoas são, de fato.
[6] Dastin, Jeffrey. Amazon scraps secret AI recruiting tool that showed bias against women. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[7] Primeiro o modelo recebe os dados e é treinado com eles. Se então, após o treinamento o modelo receber outros tipos de dados, ainda assim ele continuará a apresentar o mesmo comportamento que ele aprendeu anteriormente.
[8] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
[9] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

e de reincidência da pessoa. Os resultados apontaram que o sistema utilizado possuía viés algoritmo, onde dobravam as chances de atribuir uma nota negativa[10] a uma pessoa negra em comparação a uma pessoa branca. Além disso, o sistema considerava automaticamente as pessoas mais velhas como pessoas de menor risco e menor propensão de agirem de forma violenta, independentemente de quais tinham sido os crimes que o indivíduo havia cometido. Esses resultados foram amplamente divulgados e trazidos à mídia, gerando comoção e debate público quanto à utilização de análise algorítmica na determinação de liberdade condicional. Não obstante, a taxa de acerto do sistema nas análises de predição de reincidência para os casos de crimes violentos era de apenas vinte por cento. O estudo de Nikhil Garg et al. (2018) sob o título "Incorporações de palavras quantificam 100 anos de estereótipos de gênero e étnicos", traz o levantamento sobre o uso de incorporação de palavras, uma ferramenta comumente usada em processamento de linguagem natural e aprendizado de máquina, como uma estrutura para medir, quantificar e comparar crenças ao longo do tempo. Como um estudo de caso concreto, o grupo examinou a dinâmica temporal dos estereótipos de gênero e étnicos nos séculos XX e XXI nos Estados Unidos.

Integramos incorporações de palavras treinadas em 100 anos de dados de texto com o Censo dos EUA para mostrar que as mudanças na incorporação acompanham as mudanças demográficas e ocupacionais ao longo do tempo. A incorporação captura mudanças sociais – por exemplo, o movimento das mulheres na década de 1960 e a imigração asiática para os Estados Unidos – e também ilumina como adjetivos e ocupações específicos tornaram-se mais associados a certas populações ao longo do tempo (Garg et al. 2018 p.1)[11].

De acordo com os resultados, as mudanças sociais se associaram a populações específicas ao longo do tempo, como por exemplo o movimento de mulheres no período entre 1960 e 1970, que teve um efeito sistêmico e drástico nas representações das mulheres na literatura e cultura. Além disso, analisaram também adjetivos, ocupações específicas, e demonstraram que a incorporação de palavras captura estereótipos de gênero e reflete com precisão os vieses humanos. A incorporação também revela padrões interessantes, como as palavras individuais evoluem ao longo do tempo em sua associação de gênero. Por exemplo, a palavra "histérica" costumava ser, até meados dos anos 1900, um termo genérico para diagnosticar doenças mentais em mulheres, mas, desde então, se tornou uma palavra mais geral; tais mudanças são claramente refletidas nos *embedding*[12], já que histérico caiu de uma das cinco principais palavras de preconceito feminino em 1920 para não constar entre as cem mais citadas em 1990 nos *embeddings* de COHA[13]. Por outro lado, a palavra emocional tornou-se muito mais fortemente associada às mulheres ao longo do tempo nas incorporações, refletindo seu status atual como uma palavra amplamente associada às mulheres em um sentido pejorativo. Os resultados demonstram que a incorporação de palavras é uma lente poderosa através da qual se pode identificar sistematicamente estereótipos comuns e outras tendências históricas.

4 ESTRATÉGIAS PARA MITIGAR OS VIESES NOS ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL

A atenção à identificação e mitigação de vieses algorítmicos é fundamental para promover o desenvolvimento ético e responsável da IA. Vieses inadvertidos podem ser introduzidos durante o treinamento com conjuntos de dados enviesados, resultando em implicações prejudiciais e na perpetuação de desigualdades. Diante da importância de identificar e mitigar tendências históricas e estereótipos nos avanços tecnológicos, destaca-se o papel do IEEE (Institute of Electrical and Electronics Engineers) na promoção de iniciativas éticas, como a abordagem em Sistemas Autônomos e Inteligentes. Essa entidade, fundada em 1884, desempenha um papel fundamental na orientação de práticas éticas no desenvolvimento de tecnologias, visando o benefício da

[10] Quanto mais negativa a taxa mais a pessoa era considerada perigosa ou reincidente

[11] We integrate word embed-dings trained on 100 y of text data with the US Census to showthat changes in the embedding track closely with demographicand occupation shifts over time. The embedding captures societalshifts—e.g., the women's movement in the 1960s and Asian immi-gration into the United States—and also illuminates how specificadjectives and occupations became more closely associated withcertain populations over time. Our framework for temporal anal-ysis of word embedding opens up a fruitful intersection betweenmachine learning and quantitative social science

[12] Word Embeddings são métodos que fornecem boas representações vetoriais contínuas de baixa dimensão para conjuntos de textos não estruturados, quando combinados com classificadores melhoram o desempenho do modelo

[13] Google Books/COHA.Vectors trained on a combined corpus of genre-balanced by the authors. For each decade, a separate embedding is trained from the corpus data correspond-ing to that decade.

humanidade. A IEEE, em seu relatório sobre ética em pesquisa e design de Sistemas Autônomos e Inteligentes (A/IS), enfatiza de forma contundente que a tecnologia não é um ente neutro, mas sim um reflexo dos valores e preconceitos que permeiam a sociedade e os indivíduos envolvidos em seu desenvolvimento. A organização reconhece que os A/IS, longe de serem criados em um vácuo ético, incorporam as influências morais e culturais de seus criadores. Essa perspectiva desafia a noção de neutralidade na tecnologia, ressaltando que os sistemas autônomos e inteligentes refletem, inevitavelmente, as crenças, perspectivas e até mesmo os preconceitos presentes no processo de sua concepção. A compreensão da não neutralidade da tecnologia assume um papel central no relatório, especialmente ao considerar os impactos significativos dos A/IS em áreas cruciais como votação, policiamento e serviços bancários. A IEEE destaca a importância de uma abordagem ética e legal robusta no desenvolvimento desses sistemas, visando assegurar que os A/IS contribuam positivamente para a humanidade, enquanto evitam e corrigem potenciais discriminações ou consequências indesejadas. Este reconhecimento inicial da influência dos valores e preconceitos na tecnologia estabelece uma base sólida para as discussões posteriores sobre métodos éticos de pesquisa e design, destacando a responsabilidade inerente aos desenvolvedores na criação de sistemas autônomos e inteligentes socialmente conscientes. Um ponto central na mitigação de vieses algorítmicos é a cuidadosa seleção e curadoria dos conjuntos de dados utilizados no treinamento. Garantir que esses conjuntos sejam representativos e diversificados, refletindo a pluralidade da população e a inclusão equitativa de diferentes grupos é essencial para evitar distorções e para garantir a imparcialidade dos algoritmos. Além disso, a formação de equipes interdisciplinares surge como uma abordagem eficaz. Tais equipes, compostas por especialistas em ética, sociologia, psicologia e engenheiros de IA, proporcionam uma análise abrangente dos impactos sociais e culturais, contribuindo para a identificação e correção de vieses inadvertidos, já que uma diversidade de perspectivas é crucial para um desenvolvimento mais ético. (Gutierrez, 2019). Além da interdisciplinaridade, essas equipes precisam possuir também diversidade de pessoas, de modo que venham a representar uma ampla diversidade de gênero, idade, etnias, entre outros. A presença de supervisão humana se apresenta como um aspecto crucial, e deve estar presente em diversas fases do desenvolvimento de algoritmos de IA, como na implementação de comitês éticos, compostos por especialistas independentes, que oferece avaliações críticas e garante uma visão externa aos processos de tomada de decisão dos algoritmos. Essa supervisão humana contribui significativamente para a identificação proativa de vieses e correção antes da implementação completa. (Gutierrez, 2019). Conforme o referido autor, a auditoria em tipos específicos de IA pode apresentar desafios adicionais devido à diversidade e complexidade desses sistemas, e a eficácia das práticas de auditoria pode variar conforme a natureza da inteligência artificial em consideração. Para Gutierrez (2019), é factível realizar auditorias e estabelecer processos de responsabilização para sistemas de IA fundamentados em aprendizado de máquina supervisionado. Nesse cenário, os registros de treinamento e ajustes ganham destaque como objeto de auditorias, constituindo uma abordagem eficaz para a detecção de vieses. A ênfase da auditoria nos logs, que representam os parâmetros de entrada desses sistemas, apresenta-se como uma alternativa valiosa que vai além da análise do código-fonte. Um ponto crucial seria direcionar atenção à diversidade de gênero, raça, etnia, entre outros, não apenas durante o desenvolvimento, mas também na supervisão de algoritmos. O autor menciona que "a construção e revisão desses parâmetros por equipes interdisciplinares e baseadas em amplo espectro de diversidade têm sido um mecanismo alternativo por empresas para evitar *by default*[14] que esses sistemas tenham vícios de origem ou incorram em decisões ética ou legalmente condenáveis. (Gutierrez, 2019, p. 34). Outra estratégia relevante é a promoção de treinamento e aprendizagem de forma contínua, visando capacitar toda a equipe para que possam se adaptar rapidamente às mudanças sociais e culturais, garantindo a relevância contínua das práticas éticas. Esses treinamentos especializados desempenham um papel crucial ao sensibilizar os profissionais sobre os desafios éticos associados aos vieses, incluindo a compreensão das implicações sociais dos algoritmos. Essas estratégias, quando implementadas de forma integrada, contribuem para um desenvolvimento de IA mais ético, transparente e equitativo. A abordagem combinada de seleção cuidadosa de dados, formação de equipes diversas e

[14] *By default* é uma expressão em inglês que pode ser traduzida para o português como "por padrão" ou "automaticamente". Ela é usada para indicar a configuração ou o comportamento padrão de um sistema ou software, ou seja, a maneira como algo é definido ou acontece automaticamente, a menos que seja especificamente alterado ou configurado de outra forma. Em muitos contextos, "by default" refere-se à configuração ou condição que ocorre automaticamente na ausência de uma escolha explícita ou intervenção do usuário.

interdisciplinares, treinamento contínuo e supervisão humana reforça a importância de enfrentar os desafios éticos associados aos vieses algorítmicos.

CONCLUSÃO

A tomada de decisões dos seres humanos e das máquinas se ancoram em informações preexistentes. Por conta disso, a criticidade de se avaliar se decisões tomadas por algoritmos de IA estejam baseadas em dados tendenciosos. Toda a tomada de decisão envolve riscos e, no caso a IA, não seria diferente. O que se discutiu neste artigo foi o risco de as máquinas replicarem os preconceitos existentes na sociedade humana, pois o processo de aprendizagem de máquina depende dos dados com os quais ela é alimentada. O uso de conceitos de design ético (Blackman, R., 2022) de processos de desenvolvimento e aplicação de sistemas de IA passa pela compreensão da heurística, dos vieses cognitivos e da falibilidade humana que podem gerar ações discriminatórias que terminam por ser registradas em dados armazenados em bancos que podem ser utilizados para treinamento de sistemas de IA. A não neutralidade inerente à tecnologia, como destacado pelo IEEE, desafia a concepção de sistemas autônomos e inteligentes como entidades neutras. Ao contrário, esses sistemas refletem os valores, perspectivas e preconceitos presentes na sociedade e nos criadores envolvidos. Esse reconhecimento inicial estabelece uma base sólida para as discussões subsequentes sobre métodos éticos de pesquisa e design, ressaltando a responsabilidade dos desenvolvedores na criação de sistemas socialmente conscientes. No cerne da mitigação de vieses algorítmicos está a cuidadosa seleção e curadoria dos conjuntos de dados. A necessidade de garantir que esses conjuntos sejam representativos e diversificados, refletindo a pluralidade da população, é vital para evitar distorções e assegurar a imparcialidade dos algoritmos. Além disso, a formação de equipes interdisciplinares emerge como uma estratégia eficaz, envolvendo especialistas em ética, sociologia, psicologia e engenheiros de IA. Essas equipes proporcionam uma análise abrangente dos impactos sociais e culturais, contribuindo para identificar e corrigir vieses inadvertidos. A diversidade de perspectivas não apenas em termos disciplinares, mas também de gênero, idade, etnia e outros, é destacada como crucial para um desenvolvimento ético. A presença de supervisão humana, exemplificada por comitês éticos, desempenha um papel significativo ao oferecer avaliações críticas e garantir uma visão externa aos processos de tomada de decisão dos algoritmos, contribuindo proativamente para a identificação e correção de vieses. As auditorias, focalizadas nos registros de treinamento e ajustes, apresentam-se como uma abordagem valiosa para detectar vieses, especialmente em sistemas complexos. A ênfase na diversidade de gênero, raça, etnia e outras dimensões, tanto durante o desenvolvimento quanto na supervisão, é essencial para evitar vícios de origem e decisões éticas ou legalmente condenáveis. A promoção de treinamento contínuo para profissionais de IA é uma estratégia relevante, capacitando-os a manter a sensibilidade aos desafios éticos emergentes e adaptar-se às mudanças sociais e culturais. Essas estratégias, quando integradas de forma abrangente, contribuem para um desenvolvimento ético, transparente e equitativo da IA. Concluímos que a promoção da diversidade e o constante aprimoramento ético na inteligência artificial são alicerces cruciais para um desenvolvimento equitativo e transparente. Ao enfrentarmos os desafios éticos relacionados aos vieses algorítmicos, é imperativo adotar uma abordagem holística e colaborativa, valorizando todas as dimensões da diversidade. Na jornada rumo a uma IA ética, é vital reconhecer que, ao contrário das máquinas, somos dotados da capacidade cognitiva para compreender, prevenir e corrigir vieses. Conscientizar-nos de que os vieses são inerentes ao ser humano e assumir a responsabilidade por nossas ações são passos cruciais para garantir que a tecnologia seja uma força benéfica para o bem-estar humano. Nesse contexto, o conhecimento aprofundado sobre o tema não apenas nos capacita a estar atentos às nossas próprias atitudes, mas também nos permite contribuir ativamente para a construção de uma IA mais ética, livre de discriminações e alinhada com os valores humanos mais elevados, afinal, o caminho para a inovação responsável é pavimentado pela consciência e ação coletiva, sendo a ética a base para a verdadeira revolução tecnológica.

REFERÊNCIAS

AUSUBEL, D. P. Educational psychology: a cognitive view. Nova York: Holt, Rinehart and Winston, 1968.
BLACKMAN, R. Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI. Cambridge: Harvard Business Review Press, 2022.

GARDNER, H. Inteligências Múltiplas: a teoria na prática. Trad. Maria Adriana Veríssimo Veronese. Porto Alegre: Artes Médicas, 1995.
GARG, N. et al. Word Embeddings Quantify 100 Years of Genderand Ethnic Stereotypes. Proceedings of the National Academy of Sciences. Stanford. v. 115, n. 16. abr. 2018. Disponível em: https://www.pnas.org/doi/epdf/10.1073/pnas.1720347115. Acesso em: 12 jun. 2022.
GRAGLIA, M. A. V., HUELSEN, P., LAZZARESCHI, N. The growing moral challenge in the face of technologies: internet, social networks, IoT, blockchain and artificial intelligence. In: RISUS – Journal on Innovation and Sustainability, São Paulo, v. 12, n.2, p. 17-29, abr./ mai. 2021. Disponível em: https://doi.org/10.23925/2179-3565.2021v12i2p17-29. Acesso em: 14 out. 2022.
GRAGLIA, M. A. V.; LAZZARESCHI, N. Transformações no mundo do trabalho: tensões e perspectivas/ Noêmia Lazzareschi, Marcelo Augusto Vieira Graglia, orgs. - São Paulo: Educ: PIPEq, 2023.
GUTIERREZ, A. É possível confiar em um sistema de inteligência artificial? Práticas em torno da melhoria da sua confiança, segurança e evidências de accountability. In: FRAZÃO, Ana; MULHOLLAND, C (coord.). Inteligência artificial e direito: ética, regulação e responsabilidade. São Paulo: Revista dos Tribunais, 2019. p. 83-97. Disponível em: www.jusbrasil.com.br/doutrina/inteligencia-artificial-e-direito-etica-regulacao-e-responsabilidade/1196969611. Acesso em: 21 nov. 2022.
INSTITUTE FOR ELECTRICAL AND ELECTRONICS ENGINEERS. 2019. Ethically Aligned Design: a Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. First Edition. IEEE. Disponível em: https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf . Acesso em: 16 out. 2022
KAHNEMAN, D. Rápido e Devagar: duas formas de pensar. Tradução de Cássio de Arantes Leite. Rio de Janeiro: Objetiva, 2012. [E-pub].
MATURANA, H.; VARELA, F. A Árvore do Conhecimento: as bases biológicas da compreensão humana. São Paulo: Palas Athena, 2001.
MINHARRO, E. R. S. Inteligência Artificial na Justiça Brasileira. in Inteligência Artificial nas Relações de Trabalho. Leme: JH Mizuno, 2022. [E-book].
MORIN, E. O Pensar complexo: Edgar Morin e a crise da modernidade. Org. Pena-Vega, Alfredo. do Nascimento, Elimar Pinheiro. 2ª ed. Rio de Janeiro. Garamond, 1999.
_____. Ciência com consciência. Tradução de Maria 8'1 ed. D. Alexandre e Maria Alice Sampaio Dória. 8 ed. Rio de Janeiro: Bertrand Brasil, 2005.
PIAGET, J. J; INHELDER, B. The Psychology of The Child. New York: Basic Books, 2008.
SANTAELLA, L. Inteligência artificial & redes sociais, org. São Paulo: EDUC/PIPEq, 2019. E-Book.
VYGOTSKY, L. S. Mind in Society*:* The Development of Higher Psychological Processes. Reino Unido: Harvard University Press, 1980.
ZHAO, J. et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. University of Virginia. jul. 2017. Disponível em: https://arxiv.org/abs/1707.09457. Acesso em: 19 jun. 2022