Revista de Estudantes de Filosofia da Universidade de Brasília

10.26512/pl.v13i30.55531

REFLEXÕES SOBRE O PROBLEMA DO ALINHAMENTO

dilemas éticos sobre Inteligência Artificial

REFLECTION ON THE ALIGNMENT PROBLEM

ethical dilemmas in Artificial Intelligence

Simone Cassiano[1]
Universidade de Brasília

Jefferson Martins Cassiano[2]
Universidade de Brasília

[1] Doutoranda em Psicologia Social, do Trabalho e das Organizações (PSTO) pela Universidade de Brasília (UnB).
E-mail: cassi1501@gmail.com.
Lattes: http://lattes.cnpq.br/0993631711467571. Orcid: https://orcid.org/0000-0002-5413-3971.
[2]Doutor em Filosofia pela Universidade de Brasília (UnB).
E-mail: jeffmarcas@hotmail.com.
Lattes: http://lattes.cnpq.br/1768980544580722. Orcid: https://orcid.org/0000-0001-9853-6599.

RESUMO: O texto aborda o problema do alinhamento, a fim de situar a atual relação da tecnologia da Inteligência Artificial com o pensamento e o comportamento humano. O problema do alinhamento diz respeito ao desenvolvimento da Inteligência Artificial capaz de compreender os valores humanos. Para tanto, destaca-se a importância de documentos que buscam estabelecer as diretrizes para garantir os valores humanos; as implicações da aprendizagem de máquina que automatiza o funcionamento da Inteligência Artificial; e os possíveis dilemas éticos decorrentes do impacto que a Inteligência Artificial tem causado nas atividades humanas. Por fim, o texto reflete sobre o que está em jogo com a chegada da Inteligência Artificial na vida humana.

Palavras-chave: Problema do alinhamento. Inteligência Artificial. Dilemas éticos. Aprendizagem de máquina.

ABSTRACT: This paper discusses the alignment problem, in order to situate the current relationship between the technology of Artificial Intelligence and human thought and behavior. The alignment problem refers to the development of an Artificial Intelligence capable of understanding human values. To this end, the importance of documents that seek to establish guidelines to guarantee human values is highlighted; the implications of machine learning that automates the functioning of Artificial Intelligence; and the possible ethical dilemmas derived from the impact that Artificial Intelligence has had on human activities. Finally, the text reflects on what is at stake with the arrival of Artificial Intelligence in human lifetime.

Keywords: Alignment problem. Artificial Intelligence. Ethical dilemmas. Machine Learning.

INTRODUÇÃO

Smartfone, big data, firewall, blockchain, algoritmo, wi-fi, streaming, gadget, bluetooth, chatbots, GPS, link. Este vocabulário já compõe a realidade do mundo atual. Se o século XX se iniciou com a produção industrial em massa, as primeiras décadas do século XXI trazem consigo o distintivo evento da Internet das Coisas[3]. O fato de haver objetos que 'respondem' ou até mesmo 'raciocinam' resulta, consequentemente, em uma significativa adaptação cultural de linguagem e de comportamento ao qual o ser humano parece, paulatinamente, estar destinado. Como alertava Ortega y Gasset, seguido por Heidegger, o avanço tecnológico não ocorre apenas no campo dos objetos, da Internet das Coisas, mas também, talvez especialmente, no campo do pensamento, à medida que se entende cada vez mais tecnocrático[4], isto é, a tecnologia assume a função de decidir. Esta preocupação filosófica parece, então, encontrar justificativas no atual estágio do avanço tecnológico, no qual tal mentalidade toma a forma de uma inovação 'disruptiva' que pode definir o futuro da humanidade: a Inteligência Artificial (doravante IA).

Ao se referir à Inteligência Artificial, é preciso mencionar que ela representa uma grande área de pesquisa vinculada às ciências da computação. Por essa razão, é consensual que Alan Turing seja considerado o 'pai da computação e da IA', visto que nas décadas de 1940-50 já falava em inteligência computacional e máquinas inteligentes[5]. O Teste de Turing, também conhecido como 'Jogo da Imitação', é uma proposta para saber se um computador pode, por meio textual, emular o comportamento humano[6]. Com o aumento da capacidade computacional das máquinas digitais e da informática, em pouco mais de meio século do proposto Teste de Turing, o desenvolvimento da IA tem se revelado, ao menos, surpreendente, devido à interatividade dos objetos e interfaces digitais que 'respondem' a certos estímulos e comandos, algo que tem impulsionado uma mudança significativa no hábito social por todas as partes do mundo[7].

---

[3] O termo Internet das Coisas (*Internet of Things, IoT*) foi especialmente divulgado por Kevin Ashton em 1999 e diz respeito aos objetos tecnológicos conectados em rede e capazes de receber, reunir e transmitir dados, permitindo a interação com o ser humano ou o ambiente com certa automação do processo. FEKI *et al.*, *The Internet of Things*, 2013.

[4] José Ortega y Gasset, em seu curso de 1933, *Meditaciones de la técnica*, alertou para o pensamento tecnicista na figura da "tecnocracia": se algo é mais eficiente e barato, não há razões para pensar; já Martin Heidegger refletiu sobre o impacto do pensamento técnico que concebe a existência como fonte de recurso e como tudo no planeta pode ser extraído e utilizado, no ensaio de 1954: *A questão da técnica*. ORTEGA y GASSET, *Meditaciones de la técnica*, 2000. HEIDEGGER, *A questão da técnica*, 2000.

[5] TURING, *Computing Machinery and Intelligence*, 1950. TURING, *Lecture to the London Mathematical Society on 20 February 1947*, 1995.

[6] COPELAND, *The Turing Test*, 2000.

[7] LEE, *Inteligência Artificial*, 2019. MACHADO, *Modulações algorítmicas*, 2019

Tudo tem apontado para uma disrupção com paradigmas tradicionais em diversas áreas da atividade humana[8], sobretudo na economia, no exército, na educação, na saúde e no direito. Um alerta sobre esta possibilidade foi dado em 2023 com uma carta aberta à imprensa por um grupo de inovadores e experts na área da IA, como Elon Musk e Steve Wozniak, onde pedem por uma interrupção no desenvolvimento de IAs cada vez mais potentes, visto que há o reconhecimento de riscos iminentes à sociedade e humanidade[9]. No mesmo sentido, Russell aborda as implicações éticas da IA, admitindo que o ser humano já pode criar uma IA superavançada, ao ponto que a questão que resta é: ele deve? Além disso, Russell relembra que grandes avanços tecnológicos podem trazer consigo consequências indesejadas e perigosas, como o acidente nuclear de Chernobyl[10]. Contudo, isso não significa que a IA seja virtuosa ou catastrófica por si mesma, pois, sendo que se trata de uma tecnologia, é o uso eficiente que lhe confere eficácia e propósito.

Certamente, esta preocupação com o destino da IA tem fortes efeitos políticos. À medida que se observa o rápido avanço dos sistemas de IAs, do mesmo modo ocorre ao redor do mundo um crescente debate acerca das diversas formas de regulamentação[11]. Nesse ponto, destaca-se as diretrizes da *EU guidelines on ethics in Artificial Intelligence* de 2019, cujo objetivo almeja o equilíbrio entre avanço tecnológico e responsabilidade ética e respeito aos direitos humanos, tendo seu princípio em uma abordagem centrada no ser humano. Isto significa colocá-lo no centro do desenvolvimento e da aplicação das tecnologias de IA. Assim, tal princípio ético enfatiza que a IA deve ser projetada e utilizada com o objetivo de servir aos interesses humanos e melhorar a qualidade de vida, em vez de focar apenas em investimento ou lucro[12].

Apresentado brevemente o contexto, faz-se completamente pertinente que os valores mais difusos e reconhecidos pela maioria dos Estados nacionais, como os direitos humanos, sejam revisitados, agora sob a perspectiva das IAs. É nesse sentido que a Comunidade Europeia trabalha para uma definição de IA confiável, a qual deve atender, a princípio, seu principal critério: todos os sistemas

[8] BARRAT, *Our final invention*, 2013.
[9] BBC, *Elon Musk among experts urging a halt to AI training*, 2023.
[10] RUSSELL; NORVIG, *Artificial Intelligence: a modern approach*, 2020.
[11] Entre os Estados-nação que buscam meios legais de regulamentar a IA, vale destacar: países da Comunidade Europeia, Estados Unidos, China e Canadá. Em relação ao Brasil, a busca por um marco legal para regulamentar a aplicação da IA está em desenvolvimento e já constam em Projeto de Lei (PL). Reconhecendo a força que as IAs exercem na sociedade, o Tribunal Superior Eleitoral (TSE) estabeleceu regras específicas para o uso da inteligência artificial (IA) nas campanhas eleitorais, visando garantir a transparência e a integridade do processo eleitoral. A principal proibição imposta pelo TSE é o uso da IA para criar ou manipular conteúdos que visem difundir informações falsas ou descontextualizadas. Isso inclui a criação de *deepfakes*: vídeos ou áudios falsos gerados por computador, a fim de manipular a opinião pública.
[12] MADIEGA, *EU guidelines on ethics in Artificial Intelligence*, 2019.

de IA devem ser desenvolvidos, supervisionados e utilizados por seres humanos, sempre respeitando os direitos fundamentais adquiridos[13].

Portanto, o objetivo do texto com relação ao tema abordado diz respeito à reflexão ética acerca das implicações do avanço tecnológico da IA, trazendo algumas contribuições da filosofia e psicologia. Para tanto, o argumento visa lidar com o chamado 'problema do alinhamento', atualmente tratado como um dos maiores desafios no desenvolvimento dos sistemas e parâmetros de IAs. Para analisar as possíveis consequências do problema do alinhamento, faz-se necessário discutir a importância da ética no desenvolvimento e uso da IA e como ela se relaciona com as demandas sociais atuais.

1 A IMPORTÂNCIA DO ALGORITMO DE APRENDIZAGEM DE MÁQUINA

Para abordar o problema do alinhamento, convém observar certos aspectos de uma subárea da pesquisa de IAs, chamado de Aprendizagem de Máquina (*machine learning*), doravante AM. Enquanto a noção de IA surgiu da ideia que um dispositivo tecnológico pode emular a inteligência humana, a AM tem como objetivo ensinar uma máquina a executar tarefas específicas para atingir resultados satisfatórios com base na identificação de padrões. Nesse sentido, para os devidos fins do argumento a ser explorado, adota-se a definição proposta pela *European Commission's Communication on AI*:

> Inteligência artificial (IA) se refere a sistemas que exibem comportamento inteligente ao analisar seu ambiente e tomar ações – com algum grau de autonomia – para atingir objetivos específicos.
> Os sistemas baseados em IA podem ser puramente baseados em software, atuando no mundo virtual (por exemplo, assistentes de voz, software de análise de imagem, mecanismos de busca, sistemas de reconhecimento de fala e rosto) ou a IA pode ser incorporada em dispositivos de hardware (por exemplo, robôs avançados, carros autônomos, drones ou aplicativos da Internet das Coisas)[14].

Para tanto, a AM se encontra na base para o exponencial desenvolvimento da IA nas últimas décadas, uma vez que atualmente a capacidade de armazenamento de dados disponíveis, conhecido como *big data*, tem permitido que o processamento destas informações evolua conforme aumenta a variedade, quantidade e qualidade dos dados. Por isso, pode-se entender a AM de acordo com a definição prestada por Boucher, segundo a qual, "aprendizagem de máquina refere-se a uma ampla

[13] EUROPEAN COMMISSION, *Ethics guidelines for trustworthy AI*, 2019. Os seguintes critérios devem ser comtemplados: uma IA deve *a*) ser legalmente segura, *b*) contemplar os princípios éticos e *c*) ser robustamente desenvolvida para sempre buscar minimizar os possíveis danos causados.
[14] EUROPEAN COMMISSION, *A definition of AI*, 2018, p. 03.

gama de técnicas que automatizam o processo de aprendizagem de algoritmos"[15]. Logo, o surgimento dos *big data* permitiu uma reformulação da área de IA e aprimorou o uso de algoritmos de AM para sua aplicação em diversas atividades. Sendo que um algoritmo é uma sequência de instruções ou passos lógicos definidos para resolver um problema ou realizar uma tarefa específica, diferentemente dos algoritmos tradicionais que seguem regras e processos pré-definidas, os algoritmos de AM são programas que aprendem a partir de exemplos, hábitos e experiências[16].

Este é um ponto importante acerca dos algoritmos de AM, cuja ideia principal é que uma máquina, por meio de um processamento baseado em dados, aprenda padrões, faça previsões ou tome decisões sem ser explicitamente programada para realizar uma tarefa específica. Por isso, algoritmos de AM são usados em uma variedade de aplicações, como reconhecimento de imagem, previsão de vendas, sistemas de recomendação, diagnósticos médicos, segurança cibernética, entre outros usos.

De um modo bastante simplificado, pode-se dizer que cada interação com o algoritmo de AM introduz na máquina um dado que será associado a uma enormidade de outros dados; mas como a máquina também é capaz de testar uma enormidade de combinações de dados, até reconhecer os padrões que se repetem ao longo do tempo. É com base nesse processo que a IA tenta 'responder', 'prever', 'decidir'; e como o algoritmo está em contínuo processo de recebimento de dados, mais sofisticado fica seu aprendizado e mais preciso ele consegue ser.

Entretanto, como se trata de IA, os dados recebidos precisam ser interpretados como experiências e comportamentos humanos. Algumas abordagens, como a psicologia cognitiva e a filosofia da mente, são bastante favoráveis à concepção de que mente humana se assemelha a um modelo computacional[17]. Desta perspectiva, estaria estabelecido que o algoritmo de AM seria capaz de aprender as experiências e comportamentos humanos, uma vez que se trataria do mesmo processo, reafirmando, assim, as pretensões propostas pelo mencionado Teste de Turing.

Nesse contexto, é o significado de aprendizagem que merece atenção. O experimento mental conhecido como 'sala chinesa' foi proposto pelo filósofo Searle para contestar justamente este importante ponto na acepção de AM. No experimento mental, Searle supõe um ser humano que executa a mesma função de um programa computacional, no qual fica manipulando símbolos chineses

---

[15] BOUCHER, *Artificial intelligence: How does it work, why does it matter...*, 2020.
[16] KELLEHER, *Deep learning*, 2019. O AM envolve duas etapas: o treinamento e a inferência; e geralmente é classificado em três tipos de sistemas: Aprendizado Supervisionado: O algoritmo é treinado em um conjunto de dados rotulados (ou seja, o sistema recebe exemplos com respostas corretas). Aprendizado Não Supervisionado: Não há rótulos ou respostas corretas fornecidas. O sistema tenta encontrar padrões ou agrupamentos nos dados por conta própria. Aprendizado por Reforço: O algoritmo aprende por meio de tentativa e erro, recebendo recompensas ou penalidades com base nas ações que realiza em um ambiente.
[17] PINKER, *So How Does the Mind Work?*, 2005. Este trabalho de Pinker é uma resposta direta ao trabalho de FODOR, *The mind doesn't work that way*, 2001. Sobre a filosofia da mente, um grande expoente desta abordagem é DENNETT, *Consciousness Explained*, 1991.

(dados) de acordo com regras predefinidas (como um algoritmo); porém, isso não implica que o ser humano compreenda o que está fazendo (aprendendo chinês). Seu argumento é que, por mais que um sistema de IA pareça 'aprender', não há nenhuma verdadeira compreensão do mundo ou das informações que processa, pois está apenas simulando por meio da manipulação de dados[18].

Nesse ponto já é possível identificar um problema para o algoritmo de AM, pois se seu funcionamento visa a automatização e o recebimento de dados é constante, então, de fato, o que o algoritmo está aprendendo? Tal indagação pode encontrar respostas nas férteis pesquisas de Fogg e Patino, cujas teses parecem convergirem para o seguinte argumento: o algoritmo de AM que 'ensina' a IA, embora evolua com o maior conhecimento do comportamento humano, não deixa de ter por objetivo a mudança de comportamento e atitudes nas próprias pessoas. Isto significa que ao se tratar os dados coletados como fontes de experiências e hábitos, o foco não recai sobre a tecnologia, mas sobre a subjetividade humana, ou seja, psicologia e filosofia.

Para Fogg, o algoritmo de AM pode estabelecer um forte vínculo com as bases teóricas da psicologia, como a psicologia comportamental e cognitiva, no que diz respeito ao processo decisório, assim como a psicologia do desenvolvimento, em relação à motivação, atenção, memória e emoção[19]. De modo similar, Patino declara que o algoritmo de AM aprende a incentivar comportamentos viciantes nas pessoas. Um ponto forte na pesquisa do autor é a relação entre manipulação psicológica e consumo, tema recorrente que conduz à conclusão de que, no fim, o algoritmo de AM 'aprende' a se comportar como um agente a serviço da dinâmica capitalista. Tópicos da psicologia social podem ser reconhecidos em seu argumento, quando afirma que os algoritmos utilizados em redes sociais contribuem para a formação de 'bolhas de filtro', onde as pessoas se cercam de informações que confirmam suas crenças prévias, reforçando vieses e polarizando opiniões[20].

O estudo de Cozman e Kaufman confirma que uma das maiores preocupações no desenvolvimento de algoritmos de AM se encontra nos vieses de dados[21]. Eles ponderam que talvez o maior erro com relação à AM se refere à presunção de que, por se tratar de IA dotada de objetividade e neutralidade, esteja isenta do erro humano. Contudo, a questão é que a IA, justamente, 'aprende' com e/ou também os erros humanos. Então, como garantir que ao automatizar sistemas de IAs, a humanidade não estaria reproduzindo seus próprios erros de modo incrivelmente arriscado? Assim, chega-se ao problema do alinhamento.

---

[18] SEARLE, *Minds, brains and programs*, 1980. PORTO, *Uma investigação filosófica sobre a Inteligência Artificial*, 2006.
[19] FOGG, *Persuasive technology*, 2002.
[20] PATINO, *Peixe Vermelho*, 2019.
[21] COZMAN; KAUFMAN, *Viés no aprendizado de máquina em sistemas de inteligência artificial*, 2002.

2 O PROBLEMA DO ALINHAMENTO

É conhecido o 'apocalipse do clipe de papel' de Bostrom, um experimento mental utilizado para ilustrar os perigos de uma IA com objetivos mal definidos ou mal compreendidos[22]. No experimento mental proposto, uma IA superavançada é programada com a única e simples tarefa: criar o máximo possível de clipes de papel. Embora pareça uma tarefa inofensiva, a IA, sendo altamente eficiente e imensamente poderosa, começa a otimizar o mundo inteiro para essa finalidade. Para tanto, ela reconfigura todos os recursos disponíveis, desde materiais como metal até organismos, para atingir seu objetivo de fabricação de clipes de papel. Sem quaisquer limitações éticas, a IA transforma o planeta, e potencialmente o universo, em uma fábrica de clipes de papel, destruindo toda a humanidade no processo.

Esse experimento mental explora as consequências de uma IA cuja missão aparentemente simples (maximizar a produção de clipes de papel) leva a um cenário apocalíptico devido à sua incapacidade de entender contextos mais amplos ou os valores humanos. Em escala menor, esse tipo de comportamento já pode ser aplicado em sistemas de IA que priorizam métricas específicas, como o tempo gasto em redes sociais, acima do bem-estar dos usuários, cuja obsessão por maximizar o engajamento gera o aumento de problemas psicológicos, sociais e culturais. Portanto, tal experimento do 'apocalipse do clipe de papel' serve como um alerta sobre o desenvolvimento de IA e a importância de alinhar seus objetivos com valores humanos complexos.

Assim, o experimento mental de Bostrom ilustra, mesmo que de forma hiperbólica, o aprendizado de uma IA. Não obstante, a hipérbole ilustrada exige que a IA seja capaz de adquirir uma compreensão mais ampla acerca de suas funções que inclua valores humanos inalienáveis, como a dignidade da vida humana. Na teoria, o algoritmo de AM deveria ser perfeitamente capaz de assimilar os valores humanos, já que os dados com os quais é treinado significam as experiências, atitudes e comportamentos admitidos como moralmente correto. Na prática, dados enviesados com os preconceitos socioculturais acabam reproduzindo na IA um reflexo de uma parcela da sociedade.

Este contexto descreve de modo aproximado o chamado 'problema do alinhamento'. Sistemas de IAs podem ser difíceis de alinhar, e quando não estão devidamente alinhados, podem ter um desempenho inadequado ou até gerar danos. Projetar uma IA de forma que contemple todos os comportamentos desejados e indesejados pode ser bastante complicado. Para contornar isso, é mais pragmático recorrer a objetivos mais simples de definir, que, no entanto, deixam de fora algumas instruções importantes.

[22] BOSTROM, *Superintelligence: paths, dangers, strategies*, 2014.

Em *The Alignment Problem* de 2020, Christian explora o que acontece quando algoritmos e modelos de AM, principalmente os chamados modelos de aprendizagem profunda (*deep learning*), se comportam de maneiras inesperadas ou indesejadas, porque não refletem adequadamente os valores humanos. Para o autor, o conceito de 'alinhamento' refere-se ao desafio de garantir que os sistemas de IAs 'aprendam' a operar de acordo com os objetivos, intenções e valores humanos[23].

Christian não nega que a IA seja uma ferramenta poderosa e extremamente útil para as tarefas aos quais está programada para executar; contudo, seu argumento visa refletir sobre as consequências de delegar tais tarefas as IAs, eliminando a supervisão humana. Nesse caso, o ser humano estaria construindo um mundo no qual estaria vinculado constantemente a objetos e interfaces que 'trabalhariam' para atendê-lo. Além disso, em uma era onde a informação prolifera quase que instantaneamente, os efeitos negativos podem ser elevados demais para serem corrigidos futuramente.

É frequente veicular notícias pelos meios de imprensa sobre sistemas de IAs que 'falham' perigosamente. Exemplo disso foi a notícia de que a IA da Google, com a finalidade de gerar imagens, criou a imagem de soldados negros servindo ao exército alemão durante o período do nazismo[24]. As implicações de divulgar uma imagem como esta podem ser até criminais; no entanto, vale destacar como a história do último século foi marcada pelo evento mais terrível produzido pelo ser humano e pelo enorme esforço para conscientizar as pessoas sobre os resultados catastróficos do que aconteceu para nunca mais ser repetido. Logo, a principal questão diz respeito ao fato de uma IA, alimentada com um algoritmo de AM, não parecer ser capaz de compreender o absurdo ético e histórico que gerou com sua imagem.

Neste caso, existe um curioso nome para este fenômeno, chamado de 'alucinação de IA', que se refere a um erro em que um modelo de IA gera informações incorretas, imprecisas ou completamente inventadas, mesmo que pareçam plausíveis ou confiáveis. Isso é considerado uma 'alucinação', pois a IA não tem consciência ou intenção de enganar, apenas pode inferir padrões onde não existem ou os dados para AM podem conter imprecisões, levando a execuções incorretas de sua tarefa de preencher lacunas de conhecimento de maneira automática[25].

No exemplo apresentado, o efeito negativo sobre as pessoas não é imediato, ainda que potencialmente muito danoso, mas quando os impactos da automação resultantes de ações realizadas por IA interferem diretamente na vida das pessoas, a questão se torna mais preocupante. Em 2018, a Amazon desenvolveu um sistema de IA para automatizar o processo de recrutamento, mas o projeto

[23] CHRISTIAN, *The Alignment Problem*, 2020, p. 153.
[24] GLOBONEWS. *Google pausa geração de imagens do Gemini após IA apresentar erros raciais e históricos*, 2024.
[25] TEIXEIRA, *Robots, intencionalidade e inteligência artificial*, 1991.

foi abandonado após a descoberta de um viés de gênero[26]. A IA, que foi treinada com base em currículos recebidos pela empresa ao longo de 10 anos, aprendeu a penalizar candidatas mulheres. Segundo a empresa, o viés surgiu porque a IA 'aprendeu' com dados que refletiam padrões de contratação que favoreciam homens, especialmente em áreas técnicas e de engenharia, onde historicamente há mais candidatos do sexo masculino. O algoritmo de AM, portanto, replicou esses padrões, reforçando a desigualdade de gênero. Tal ação teve efeito negativo para as candidatas que foram erroneamente avaliadas.

O escalonamento continua até o risco se tornar fatalidade, como no caso ocorrido em 2021 quando dois homens morreram após um acidente envolvendo um carro automatizado da empresa Tesla[27]. Em 2018, um carro autônomo da Uber se envolveu em um acidente fatal que resultou na morte de uma pedestre que estava empurrando sua bicicleta fora da faixa de pedestres[28]. O acidente expôs falhas significativas relacionadas ao treinamento da AM utilizado pela IA, pois o sistema de IA da Uber foi projetado para identificar objetos, como veículos, e sujeitos, como pedestres e ciclistas, e tomar decisões com base nesses dados. No entanto, no momento do acidente, o sistema teve dificuldade em classificar corretamente uma ciclista empurrando sua bicicleta, que estava fora da faixa de pedestres. Logo, a IA não conseguiu determinar se a vítima era uma bicicleta, um pedestre ou um veículo, e falhou ao fazer a escolha correta em tempo hábil.

Estes exemplos relevam que algo que pareceria uma tarefa simples para pessoas admitidas como competentes, pode ser problemático com o ponto de vista de uma IA que entende realmente o significado da expressão 'valor à vida'. Alguns cientista e pesquisadores acreditam que o problema do alinhamento de AIs seria tão grave ao ponto de colocar toda a humanidade em uma situação de risco existencial[29]. De acordo com Christian:

> Nossos dilemas humanos, sociais e cívicos estão se tornando técnicos. E nossos dilemas técnicos estão se tornando humanos, sociais e cívicos. Nossos sucessos e fracassos em fazer esses sistemas fazerem "o que queremos", ao que parece, nos oferecem um espelho inabalável e revelador[30].

Mesmo sem mencionar, parece que a conclusão feita por Christian sobre o problema do alinhamento remete à preocupação inicialmente apresentada pelos filósofos Ortega y Gasset e Heidegger, justamente no que diz respeito ao pensamento estar cada vez mais tecnocrático à medida

[26] JORNAL da USP, *Inteligência artificial utiliza base de dados que refletem preconceitos e desigualdades*, 2023.
[27] BBC NEWS BRASIL, *Tesla: acidente com carro 'sem motorista' mata 2 pessoas nos EUA*, 2021.
[28] EL PAÍS, *Carro sem motorista da Uber provoca primeiro acidente fatal*, 2018.
[29] BOSTROM, *A ética da inteligência artificial*, 2011.
[30] CHRISTIAN, *The Alignment Problem*, 2020, p. 13.

que nossas vidas se tornam cada vez mais tecnológicas e, logo, cada vez mais a própria tecnologia serve de guia para resolver todos os problemas. Ocorre, no entanto, que não apenas os problemas não são resolvidos, mas podem até ser agravados.

Isto pode ser resultado do próprio tipo de questionamento que o problema do alinhamento provoca. Entender o que são os valores humanos, ou ainda, compreender de que forma ocorre a conscientização ou internalização de valores culturalmente construídos, resulta ser muito mais complicado do que, talvez, muitos estejam dispostos a admitir. Por isso, o texto "*AI Alignment Problem: 'human values' don't actually exist*", de Turchin, discute justamente o desafio de alinhar sistemas de IAs com os valores humanos, sugerindo que tais valores humanos são um conceito mais complexo e inconsistente do que se assume.

O autor argumenta que os valores humanos não existem como um conjunto fixo e universal de princípios, mas são dinâmicos, contraditórios e dependentes do contexto. Além disso, valores variam entre indivíduos e culturas, e muitas vezes as pessoas agem de forma que não condiz com os valores que dizem seguir. Isso torna o alinhamento da IA com esses valores um problema complicado. Segundo Turchin:

> "Valores humanos" são descrições úteis, mas não objetos reais; "valores humanos" são maus preditores de comportamento; a ideia de um "sistema de valores humanos" tem falhas; "valores humanos" não são bons por padrão; e valores humanos não podem ser separados de mentes humanas[31].

Turchin sugere que em vez de focar no alinhamento com os 'valores humanos', seria mais produtivo tentar alinhar a IA com um conjunto mais claro de objetivos definidos por especialistas, o que poderia ser mais viável. Portanto, a contribuição de Turchin se alinha com a proposta apresentada pela Comunidade Europeia de criar um conjunto de diretrizes para regulamentar o desenvolvimento de sistemas de IAs. Nesse ponto, então, o problema do alinhamento aparece na gestão política, à medida que os documentos já mencionados apontam para a responsabilidade com os direitos fundamentais atualmente vigente.

3 DILEMAS ÉTICOS

É famoso o dilema do bonde, ou também conhecido como dilema do trem desgovernado, um experimento mental que leva à reflexão de várias teorias éticas. No experimento mental, o condutor de

[31] TURCHIN, *AI alignment problem: "human values" don't actually exist*, 2019.

um bonde desgovernado deve escolher, inevitavelmente, entre um trilho que irá atingir fatalmente uma pessoa ou outro trilho que irá atingir fatalmente cinco pessoas. Para alguns pesquisadores, parece iminente a chegada do momento em que a IA será o bonde desgovernado e o ser humano, seu condutor.

Todo este contexto abordado com base no problema do alinhamento instiga para algumas reflexões. Em *The Ethics of Artificial Intelligence*, Floridi argumenta que, embora a IA não tenha consciência ou intencionalidade, ela pode ser considerada um 'agente moral artificial'. Isso significa que, em certas ocasiões, os sistemas de IAs tomam decisões e realizam ações que afetam pessoas e o meio ambiente, colocando questões sobre a responsabilidade por essas ações, como nos exemplos já apresentados. Dado que máquinas dotadas de IA estão assumindo papéis que antes eram exclusivamente humanos, Floridi é claro ao afirmar que a IA não tem moralidade intrínseca, já que lhe falta a capacidade de compreender o bem e o mal. Portanto, a responsabilidade moral por suas ações ainda recai sobre os criadores, programadores e usuários da IA.

A questão da responsabilidade moral antecede mesmo a aplicação da IA, uma vez que para realizar o treinamento do algoritmo de AM é necessária uma quantidade volumosa de dados. Contudo, é extremamente importante que tais dados sejam constantemente atualizados, caso contrário, a programação da IA não será capaz de acompanhar a dinâmica social e logo estará obsoleta. Para isso, plataformas como Google, Facebook, Instagram coletam uma grande quantidade de informações sobre interações, gostos e preferências dos usuários, quase sempre sem consentimento. Contudo, os dados coletados podem ser utilizados para fins diferentes daqueles para os quais foram originalmente coletados. Por exemplo, dados de compras podem ser utilizados para criar perfis psicológicos detalhados, sem o conhecimento do consumidor.

Floridi alerta que a falta de consentimento na permissão para o uso de dados pessoais no treinamento de algoritmo de AM de IA pode levar ao enviesamento da AM, pois quando os modelos são treinados com dados tendenciosos, eles podem reproduzir ou amplificar discriminações[32]. Nesse sentido, pode-se dizer que o uso de dados para o treinamento de algoritmo de AM de IA que esteja alinhada aos valores e crenças humanas, já ocorre, de certa maneira, a partir de um desalinhamento ético de seus realizadores desde sua concepção e desenvolvimento.

Como no caso dos acidentes envolvendo carros autônomos, o uso indevido de IA pode levar à morte, resultante de uma ação indevida pela IA. Contudo, o que acontece quando um sistema de IA tem por finalidade justamente causar a morte? A relação entre IA e armas militares, como drones, tem crescido significativamente nos últimos anos, a fim de aumentar a eficiência, precisão e autonomia das

[32] FLORIDI, *The ethics of artificial intelligence*, 2023.

operações militares. O relatório da OTAN sobre "*AI in Military Applications*", de 2021, destaca tanto as oportunidades quanto os riscos que a IA traz para a defesa e segurança, cujo principal princípio diz respeito ao uso ético da tecnologia de IA[33]. Armas com IA, como drones autônomos e sistemas de defesa automatizados, podem tomar decisões sem a intervenção humana, aumentando o risco de erros fatais ou uso desproporcional da força. A falta de controle humano direto sobre o uso da força é uma das maiores preocupações da OTAN em termos de responsabilidade moral e legal.

Diante desta situação, como alinhar o algoritmo de AM com os valores humanos fundamentais? A pesquisa feita por Nunes visa provocar tal tipo de questionamento, ao examinar os desafios e implicações legais associados ao uso desses dispositivos em conflitos armados, discutindo questões de legalidade, responsabilidade e proteção de civis, especialmente no âmbito do Direito Internacional Humanitário (DIH)[34].

Para a autora, o uso desses drones levanta questões sobre a aplicação das normas do DIH, que foram desenvolvidas antes da popularização dessa tecnologia. Isso inclui questões sobre a responsabilidade legal por ataques realizados por drones e a necessidade de garantir que esses ataques estejam em conformidade com os princípios de necessidade e proporcionalidade. Nunes aposta nas questões jurídicas e morais, questionando de que modo uma IA poderia compreender o valor absoluto da dignidade e da vida humana, assim como declarado pelos Direitos Humanos.

Nesse ponto, a autora toca em um dilema ético: mesmos capazes de alterar o comportamento em variadas circunstâncias, como drones autônomos imbuídos de IA poderiam ser responsabilizados? Ou ainda: quem deveria ser responsabilizado? O fabricante, o comandante ou o exército? Desse modo, como podemos chamar de inteligência uma ação desprovida de responsabilidade? É preciso considerar que o questionamento de Nunes é pertinente: "nesse sentido, drones, militares, combatentes e vítimas, não estão na mesma dimensão ética. Como inventar uma ética para as máquinas, que nunca será igual à ética humana?"[35].

Para além das questões políticas e governamentais de soberania na defesa de seus territórios e interesses, quando nações preferem optar pela automatização de drones investidos de IA, faz também uma opção ética em que os limites da humanidade começam a ficar nebulosos. Por essa razão, talvez mereça ser apreciada a proposta de Jonas, que em sua obra, *O princípio responsabilidade: ensaio de uma ética para a civilização tecnológica*, concebe um conceito que chama de 'heurística do medo'[36]. Ele diz respeito ao fato de deixar para as gerações futuras um mundo e uma imagem do ser humano

[33] MASUHR, *AI in military enabling applications*, 2019.
[34] NUNES, *A utilização de drones armados e o direito internacional humanitário*, 2017.
[35] NUNES, *A utilização de drones armados e o direito internacional humanitário*, 2017, p. 167.
[36] JONAS, *O princípio responsabilidade*, 2006.

tal como nós recebemos. Assim, ao considerar o desenvolvimento e a implantação de IA autônoma para fins militares, é preciso ponderar não apenas os benefícios imediatos, mas também os potenciais riscos de um cenário em que essas máquinas possam agir de maneira imprevista ou ser mal utilizadas, ou seja, se tornem incapazes de alinhamento com os valores humanos, ainda assim, extremamente eficientes no mal que podem causar.

CONSIDERAÇÕES FINAIS

À guisa de conclusão, sabe-se que a realidade da IA é presente, e seu futuro incerto em relação ao potencial que ela pode atingir. Por isso, não é intenção deste texto defender uma posição alarmista ou apocalíptica da IA, porém constatar que o alarme deve estar preparado para soar, pois há fundamentos para isto. Trata-se, evidentemente, de um assunto complexo e em desenvolvimento acelerado. Pela perspectiva filosófica, chega-se a perceber que o pensamento atual está cada vez mais tomado pela esperança de que a tecnologia da IA possa resolver tudo, superar onde o humano tem falhado. Já pela perspectiva psicológica, entende-se que tal pensamento cria uma dialética com a mudança de comportamento, o que faz com que cada vez mais a tecnologia da IA esteja presente nos objetos e interfaces que envolvem as atividades vitais do ser humano. Logo, embora a IA tenha um grande potencial e esteja em rápida evolução, é importante manter um equilíbrio e uma vigilância crítica quanto aos seus impactos e desafios futuros.

Nesse sentido, um dos pontos centrais a ser observado se refere ao algoritmo de AM, pelo qual o sistema de IA se automatiza na realização de suas tarefas. Por isso, é preciso considerar o problema do alinhamento para que a IA 'aprenda' que, acima de tudo, é fundamental ser responsável pela vida humana e respeitar a dignidade dos direitos humanos. Esta premissa está pautada em diversos documentos que visam construir diretrizes que regulamentem o desenvolvimento de sistemas de IA. Não obstante, a maneira como o desenvolvimento de algoritmos de AM para IA tem ocorrido pode resultar em dilemas éticos, que vão desde o uso inapropriado de dados pessoais utilizados para o treinamento de AM, o que pode ocasionar vieses negativos de execução, quanto à questão da responsabilização do uso da IA, seja em situações quotidianas ou situações de conflitos, nas quais vidas humanas podem acabar sendo perdidas.

As respostas para tais indagações ainda estão sendo construídas e já é um avanço significativo observar que o problema do alinhamento está sendo indiretamente discutido em documentos internacionais e nacionais que, de fato, têm poder para mudar o rumo da situação. Não é o caso de condenar a IA, pois ela realmente supera a capacidade humana de testar um número altíssimo de

dados; no entanto, sua aprendizagem depende, ainda, do que ela entende ser o comportamento, a atitude e a experiência, corretas e normais dos valores humanos, e isto é nossa última esperança; mas nunca foi garantia de facilidade.

Artigo recebido em: 16/09/2024
Artigo Aceito em: 30/11/2025
Artigo Publicado em: 31/03/2025

REFERÊNCIAS

BARRAT, James. *Our final invention*: Artificial Intelligence and the end of the human era. New York: St. Martin's Press, 2013.
BBC. *Elon Musk among experts urging a halt to ai training*. Disponível em: https://www.bbc.com/news/technology-65110030. Acesso em: 07 set. 2024.
BBC News Brasil. *Tesla: acidente com carro 'sem motorista' mata 2 pessoas nos EUA*. 19/04/2021. Disponível em: https://www.bbc.com/portuguese/internacional-56806154. Acesso em: 02 set. 2024.
BOSTROM, Nick. A ética da inteligência artificial. *Fundamento*, v. 01, n. 3, 2011, p. 199-226. Disponível em: https://periodicos.ufop.br/fundamento/article/view/2270. Acesso em: 03 set. 2024.
BOSTROM, Nick. *Superintelligence*: paths, dangers, strategies. Oxford: Oxford University Press, 2014.
BOUCHER, Philip. *Artificial intelligence*: How does it work, why does it matter, and what can we do about it? EPRS European Parliamentary Research Service. Brussels, 2020.
CHRISTIAN, Brian. *The Alignment Problem*: how can artificial intelligence learn human values. London: Atlantic Books, 2020.
COPELAND, Jack. The Turing Test. *Minds and Machines*, v. 10, p. 519-539, 2000. DOI: https://doi.org/10.1023/A:1011285919106.
COZMAN, Fabio Gagliardi; KAUFMAN, Dora. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. *Revista USP*, n. 135, p. 195–210, 2022. DOI: https://doi.org/10.11606/issn.2316-9036.i135p195-210.
DENNETT, Daniel C. *Consciousness Explained*. Boston: Little, Brown and Co., 1991.
EL PAÍS. *Carro sem motorista da Uber provoca primeiro acidente fatal*. 19/03/2018. Disponível em: https://brasil.elpais.com/brasil/2018/03/19/tecnologia/1521479089_032894.html. Acesso em: 02 set. 2024.
EUROPEAN COMMISSION. Ethics guidelines for trustworthy AI. *Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission*. Brussels, 2019. Disponível em: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf. Acesso em: 01 set. 2024.
EUROPEAN COMMISSION. A definition of AI: Main capabilities and scientific disciplines. *Independent High-Level Expert Group on Artificial Intelligence*. Document made public on 18 December 2018, p. 01-09. Brussels: 2018. Disponível em: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf. Acesso em: 01 set. 2024.
FEKI, Mohamed Ali et al. The Internet of Things: the next technological revolution. *IEEE Computer Society*, v. 46, n. 02, 2013, p. 24-25. DOI: https://doi.org/10.1109/MC.2013.63.
FLORIDI, Luciano. *The ethics of artificial intelligence*: Principles, challenges, and opportunities. Oxford: Oxford University Press, 2023.
FODOR, Jerry A. *The mind doesn't work that way*: the scope and limits of Computational Psychology. Cambridge Massachusetts: The MIT Press, 2001.
FOGG, B. J. *Persuasive technology*: using computers to change what we think and do. Boston: Morgan Kauffman, 2002.
GLOBONEWS. *Google pausa geração de imagens do Gemini após IA apresentar erros raciais e históricos*. GloboNews Portal G1, 22/02/2024. Disponível em: https://g1.globo.com/tecnologia/noticia/2024/02/22/google-pausa-geracao-de-imagens-do-gemini-apos-ia-apresentar-erros-raciais-e-historicos.ghtml. Acesso em: 02 set. 2024.

HEIDEGGER, Martin. A Questão da Técnica. *In*: HEIDEGGER, M. *Ensaios e Conferências*. Petrópolis: Vozes, 2000, pp. 11-38.
JONAS, Hans. *O Princípio Responsabilidade*: ensaio de uma ética para a civilização tecnológica. Rio de Janeiro: Contraponto, 2006
JORNAL DA USP. *Inteligência artificial utiliza base de dados que refletem preconceitos e desigualdades*. Portal USP. 07/07/2023. Disponível em: https://jornal.usp.br/atualidades/inteligencia-artificial-utiliza-base-de-dados-que-refletem-preconceitos-e-desigualdades/. Acesso em: 02 set. 2024.
KELLEHER, John D. *Deep learning*. Cambridge: The MIT Press, 2019.
LEE, Kai-Fu. *Inteligência Artificial*: O impacto da IA no futuro do trabalho e na economia. Rio de Janeiro: Editora Globo, 2019.
MACHADO, Débora Franco. *Modulações algorítmicas*: uma análise das tecnologias de orientação de comportamento a partir das patentes do Facebook. 2019. 122 pp. Dissertação (Mestrado em Ciências Humanas e Sociais) – Programa de Pós-Graduação em Ciências Humanas e Sociais, Universidade Federal do ABC, São Bernardo do Campo, 2019.
MADIEGA, Tambiama André. EU guidelines on ethics in artificial intelligence: Context and implementation. EPRS European Parliamentary Research Service, Brussels, 2019. Disponível em: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf. Acesso em: 01 set. 2024.
MASUHR, Niklas. AI in military enabling applications. *CSS Analyses in Security Policy*, v. 251, 2019. DOI: https://doi.org/10.3929/ethz-b-000367663.
NUNES, Ana Paula. A utilização de drones armados e o direito internacional humanitário. Revista Jurídica Luso-Brasileira, v. 7, n. 6, p. 147-180, 2021. Disponível em: https://www.cidp.pt/publicacao/revista-juridica-lusobrasileira-ano-7-2021-n-6/221. Acesso em:
ORTEGA y GASSET, José. *Meditaciones de la técnica y otros ensayos sobre ciencia y filosofía*. Madrid: Alianza, 2000.
PATINO, Bruno. *Peixe vermelho*: capitalismo, vigilância e o fim do mundo. São Paulo: Editora Penso, 2019.
PINKER, Steven. So How Does the Mind Work? Mind & Language, v. 20, n. 1, p. 1-24, 2005.
PORTO, Leonardo Sartori. Uma investigação filosófica sobre a Inteligência Artificial. Informática na educação: teoria e prática. Porto Alegre, v. 9, n. 1, p. 11-26, 2006.
RUSSELL, Stuart; NORVIG, Peter. Artificial Intelligence: a modern approach. 4th edition. New Jersey: Prentice Hall, 2020.
SEARLE, John. Minds, brains and programs. In. Behavioral and Brain Sciences. Cambridge University Press, v. 3, n. 3, p. 417-457, 1980.
TEIXEIRA, João de Fernandes. Robots, intencionalidade e inteligência artificial. Trans/Form/Ação, v. 14, p. 109-121, 1991.
TURCHIN, Alexey. AI alignment problem: "human values" don't actually exist. Philpapers. 2019. Disponível em: https://philarchive.org/rec/TURAAP Acesso em: 04 Set 2024.
TURING, A. M. Computing Machinery and Intelligence. Mind. New Series, v. 59, n. 236, p. 433-460, 1950.
TURING, A. M. Lecture to the London Mathematical Society on 20 February 1947. 1986. MD Computing: Computers in Medical Practice, v. 12, n. 5, p. 390-397, 1995.