Ética em IA e IA ética: prolegômenos e estudo de casos significativos

Luís C. Lamb

Freepik

resumo

A inteligência artificial é uma tecnologia de propósito geral que tem causado mudanças significativas nas atividades econômicas, com implicações em todas as atividades da vida humana. Como é característico das tecnologias disruptivas, diversas inovações decorrentes delas, bem como as suas consequências sobre a humanidade, causam impactos que vão além das atividades de base tecnológica. Em particular, a compreensão dos aspectos éticos inerentes ao desenvolvimento de tecnologias baseadas em IA apresenta desafios distintos de experiências anteriores. Neste trabalho, apresentamos uma introdução objetiva aos principais conceitos de ética que podem ser alinhados ao desenvolvimento da IA, além de delinear desafios inerentes ao desenvolvimento de sistemas que apresentem comportamento ético delineado de acordo com os valores e princípios éticos consensualizados pela sociedade.

Palavras-chave: ética em IA; inteligência artificial; aprendizado de máquina; IA ética.

abstract

Artificial intelligence is a general purpose technology that has caused significant changes in economic activities, with implications in all human life aspects. As is characteristic of disruptive technologies, various innovations resulting from them, as well as their consequences on humanity, cause impacts that go beyond technology-based activities. In particular, understanding the ethical aspects inherent to the development of basic technologies in AI presents different challenges from previous experiences. In this work, we present an objective introduction to the main ethical concepts that can be aligned with the development of AI. In addition, we highlight challenges inherent to the development of systems designed and aligned with ethical behavior outlined in accordance with the ethical values and principles agreed upon by society.

Keywords: *ethics in AI; artificial intelligence; machine learning; ethical AI.*

A inteligência artificial é hoje considerada uma tecnologia de propósito geral que tem causado implicações e disrupções mensuráveis sobre todos os aspectos da vida humana. Este amplo uso, a atenção obtida entre setores significativos da população e o impacto da IA nas últimas duas décadas sugeriram aos pesquisadores a preocupação quanto a aspectos éticos e riscos da área, que podem ir além dos resultados científicos e tecnológicos (Donnelly, 2017). Hoje, a IA transforma diretamente a economia, a educação, as relações de trabalho em diversas escalas, as relações de poder e a (geo)política internacionais (Burton, 2017; Bostrom; Yudkowsky, 2014; Brynjolfsson; Rock; Syverson, 2018). Naturalmente, devido às implicações multissetoriais da IA – como é característico de toda tecnologia de propósito geral –, pesquisadores e pensadores vinculados a esses outros domínios, além da ciência da computação, passam a investigar de forma mais sistemática as consequências das transformações causadas sobre a humanidade (Brynjolfsson; Rock; Syverson, 2018; Lipsey; Carlaw; Bekar, 2006).

Já no início do século XXI, a partir do impacto público da IA, primeiramente na academia e posteriormente causado pelas empresas de tecnologia, o físico Stephen Hawking (Universidade de Cambridge), assim como os empresários de tecnologia Bill Gates (cofundador da Microsoft) e

LUÍS C. LAMB é professor da Universidade Federal do Rio Grande do Sul (UFRGS).

Elon Musk (CEO da Tesla), expressaram e anteciparam publicamente suas preocupações sobre o impacto (inclusive ético e moral) da IA sobre a vida humana, o futuro do trabalho e da sociedade (Sainato, 2015). Nesse mesmo período, iniciaram-se, por parte de organizações científicas como a Association for the Advancement of Artificial Intelligence (AAAI), a Association for Computing Machinery (ACM), a Royal Society e o Institute of Electrical and Electronic Engineers (IEEE), entre diversas outras, a organização de workshops, conferências e grupos de trabalho sobre os impactos éticos da IA e sobre como construir IA sob o prisma da ética (Donnelly, 2017; Anderson; Anderson; Armen, 2005; Walsh, 2015; Furman et al., 2018.).

Historicamente, no entanto, desde a organização dos primeiros workshops, conferências e publicações acadêmicas sobre IA, um número reduzido de pesquisadores se dedicava à análise e investigação da temática da ética em tecnologia ou ética das máquinas[1]. *Ética em IA* e *IA ética*, por sua vez, não eram temas centrais amplamente investigados ou de preocupação principal entre os pesquisadores da área de IA. Dados acerca desta não preocupação entre os pesquisadores líderes em IA no mundo acadêmico foram sintetizados por Prates, Avelar e Lamb (2018)[2]. Igualmente, esse estudo foi utilizado como base para o influente *AI Index Report*, edições de 2019 e 2021, publicado anualmente pela Universidade de Stanford sob coordenação do HAI – Stanford Institute for Human-Centered Artificial Intelligence (Mishra; Clark; Perrault, 2020; Perrault et al., 2019; Zhang et al., 2021).

Neste breve artigo, faremos um panorama inicial sobre como as duas temáticas da ética em IA e da IA ética se conceitualizam e se relacionam atualmente

1 A primeira conferência sobre a temática da IA ocorreu na Universidade de Dartmouth, em 1956 (denominada "Dartmouth Summer Research Project on Artificial Intelligence"), organizada por pioneiros da área como John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon. Allen Newell e Herbert Simon (Nobel de Economia em 1978) também participaram do encontro. Ver, para aspectos históricos da IA: Audibert et al. (2023).

2 Nesse estudo, para avaliar o quanto a ética na IA é analisada nos artigos científicos das principais conferências e revistas de IA, os autores identificaram os termos relacionados à ética nos títulos de artigos em conferências e periódicos de referência em IA, aprendizado de máquina e robótica publicados entre 1969 e 2017. O porcentual de palavras-chave tem uma interpretação simples e direta. Para cada categoria de terminologia da palavra-chave buscada (IA clássica, tendência atual e ética) identifica-se o número de artigos para o qual o título, ou resumo do artigo, contém pelo menos uma palavra dentre as pesquisadas. No caso das conferências anuais da Association for the Advancement of AI (AAAI) ou da Neural Information Processing Systems (NeurIPS), identificam-se as palavras-chave no resumo do artigo. Pode haver classificação de um artigo em mais de uma categoria, por exemplo, um artigo que contenha uma palavra-chave clássica e uma palavra-chave classificada como ética. Os termos analisados basearam-se nas questões expostas e identificadas em trabalhos de referência na área de IA e também nos temas para discussão na First AAAI/ACM Conference on AI, Ethics and Society. As palavras-chave utilizadas foram as seguintes: *accountability, accountable, employment, ethic, ethical, ethics, fool, fooled, fooling, humane, humanity, law, machine bias, moral, morality, privacy, racism, racist, responsibility, rights, secure, security, sentience, sentient, society, sustainability, unemployment* e *workforce* (como as publicações em IA são realizadas na língua inglesa, preservei aqui o original das palavras-chave). Os conjuntos de palavras-chave clássicas e de tendência foram compilados a partir das áreas do livro mais citado sobre IA, de Stuart Russell e Peter Norvig (2020), e da curadoria de termos das palavras-chave que apareceram com mais frequência em títulos de artigos ao longo do tempo nas conferências e revistas analisadas. As análises foram realizadas sobre um conjunto de mais de 100 mil artigos.

dentro do contexto de pesquisa em IA, notadamente na academia e empresas de tecnologia, bem como ilustraremos algumas iniciativas globais que visam garantir o uso ético da inteligência artificial. O trabalho não é uma revisão sistemática ou revisão bibliográfica, apenas tem como objetivo oferecer a pesquisadores e ao público leigo uma conceitualização útil e referências que podem expandir a análise desses temas centrais em pesquisa e desenvolvimento na área de inteligência artificial. Este trabalho é direcionado ao público amplo, não apenas a especialistas. Assim, apresentaremos conceitos básicos para entendimento do texto ao longo do mesmo. Sugere-se, como literatura inicial sobre o estado da arte em IA, os textos publicados neste dossiê da **Revista USP**, o livro de Marcus e Davis (2019) e o capítulo recentemente publicado em língua portuguesa (Bazzan et al., 2023), que cobre os principais avanços da IA em termos de pesquisa na última década. Não cobriremos os aspectos da regulamentação da IA neste trabalho, embora a temática evidentemente se relacione aos aspectos, principalmente deônticos, da ética em IA.

ÉTICA NA IA E IA ÉTICA: BREVE CONCEITUALIZAÇÃO

A inteligência artificial é o domínio da ciência no qual pesquisadores analisam, desenvolvem e experimentam – sob a tutela do método científico – como construir sistemas e tecnologias que apresentem de forma mensurável habilidades ou funções cognitivas, como o raciocínio e o aprendizado de máquinas. Há inúmeras definições de IA, mas para o escopo deste trabalho esta nos é suficiente. Outros conceitos subjacentes se referem à *IA forte* (que potencialmente replicaria a cognição humana em múltiplas tarefas), *IA fraca* (que se refere a sistemas que apresentam um único foco de aplicação – como identificar imagens de uma única categoria, como animais domésticos ou um tradutor automático simples) e *IA geral* (no inglês: *artificial general intelligence* – AGI), que se refere à construção de tecnologias ou máquinas que tenham a capacidade de aplicar soluções inteligentes para problemas de qualquer natureza[3].

A *ética na IA,* um ramo da ética aplicada (assim como a bioética), recebeu atenção recente de pesquisadores devido aos potenciais *riscos* acerca de uma tecnologia de propósito geral (IA) potencialmente capaz de alterar significativamente a vida humana (Burton, 2017; Bostrom; Yudkowsky, 2014; Brynjolfsson; Rock; Syverson, 2018; Lipsey; Carlaw; Bekar, 2006; Sainato, 2015; Marcus; Davis, 2019; Bazzan, 2023; Anderson; Anderson, 2011). Aliado a estas preocupação e motivação, o desenvolvimento de tecnologias e produtos computacionais em geral e sistemas de IA em particular passou a receber especial atenção. Pesquisadores e lideranças

3 Mesmo entre especialistas, as distinções entre esses conceitos não são universalmente aceitas. Outros conceitos como *HAI (human-level AI) e superinteligência artificial* foram propostos. *Human-level AI* (inteligência no nível humano) se refere a sistemas que teriam as habilidades cognitivas humanas, capacidade de autoaprendizado e autonomia, como o ser humano. A superinteligência artificial teria habilidades cognitivas superiores à humana. No entanto, enfatizamos que esses conceitos ainda possuem intersecções e incertezas em suas definições.

globais alertam para a necessidade de construirmos sistemas que tenham como requisitos o uso ético da tecnologia. Neste caso, nos referimos à IA ética – aquela tecnologia ou produto que faz uso intensivo da IA que necessariamente passe a apresentar valores morais ou que se comporte de acordo com a moralidade (e valores) consensualizados pela sociedade.

Aqui, antes de prosseguirmos na análise, temos conceitos que precisam ser definidos para melhor leitura. Uma conceitualização geral, adotada neste trabalho, define ética como o estudo e análise dos fenômenos morais. A análise de valores morais que o ser humano ou a sociedade seguem sob o paradigma rigoroso da investigação filosófica recai sobre a ética normativa, que no caso da IA e das ciências remete à ética aplicada. O estudo da *metaética* analisa as (meta)definições dos conceitos fundamentais da área, como valores e julgamentos morais. Investiga-se, em ética normativa, aquilo que em lógica (filosófica) se associa a *modalidades* como: o *permitido* (em termos de comportamento humano, social ou das "máquinas", incluindo a IA); o *obrigatório* (que tipicamente passa a ser regulado por leis ou regulamentos nas organizações, sociedades e nações); as *proibições* (tipicamente consensualizadas pelas organizações e sociedades e também reguladas por leis, normatizadas ou regulamentadas); e também se formalizam as *omissões* (aquilo que não é obrigatório ou consensualizado dentro um sistema normativo).

Esta análise, do ponto de vista mais rigoroso, remete à deôntica no prisma da lógica filosófica, que pode ser relacionada a formulações através de um sistema denominado pelos lógicos e filósofos como lógicas modais. Em lógicas modais estudamos afirmações como "é necessário que" ou "é possível que" como qualificadores de uma afirmação ou proposição (Åqvist, 1994; Gabbay, 2013). A formalização padrão da lógica deôntica é descrita formalmente em lógicas modais no estilo de Kripke, isto é, no qual as proposições são interpretadas através dos modelos de mundos possíveis; ou interpretações aceitas, no caso da lógica deôntica[4]. Na lógica deôntica as modalidades representam, por exemplo, obrigações, permissões e proibições. No entanto, a formalização rigorosa ou lógica destes conceitos vai além do escopo deste trabalho[5]. Ainda assim, é útil mencionar a literatura básica sobre lógica deôntica ao leitor, para partirmos da mesma conceitualização ou ontologia.

OS DESAFIOS DA ÉTICA NA IA E DA CONSTRUÇÃO DE TECNOLOGIAS ÉTICAS DE IA

Como mencionamos acima, a preocupação com aspectos éticos da IA não era vista ou mensurada como prioritária até o início do século XXI. Alguns trabalhos, como o de Prates, Avelar e

4 Brevemente: sob uma interpretação ao estilo de Kripke, dizemos que uma afirmação (em lógica, chamamos afirmação de *proposição*) é obrigatória se em todos os mundos possíveis (aceitos) ou em todas as interpretações esta proposição tem de ser verdadeira. Isto é, uma obrigação tem de ser seguida ou obedecida sob todas as interpretações. Entretanto, a noção de *permissão* não implica a noção de obrigação (em todos os mundos possíveis ou interpretações aceitas). Essas noções são elegantemente formalizadas em Åqvist (1994) e Gabbay (2013).

5 Ver Åqvist (1994) e Gabbay (2013) para uma análise formal da lógica deôntica.

Lamb (2018), conforme relatado no *AI Index Report* de 2019 e 2021 (Russell; Norving, 2003; Mishra; Clark; Perrault, 2020; Perrault et al., 2019), identificaram que um porcentual abaixo de um por cento dos trabalhos publicados nas principais conferências de IA e aprendizado de máquina, entre 1969 e 2017, se referia a aspectos relacionados ao impacto ético dos resultados científicos e tecnológicos apresentados nesses trabalhos, pelo menos de forma mais explícita nos textos dos artigos. Ainda na linha de identificação e classificação de trabalhos científicos que tinham como foco as implicações éticas da IA, Avelar, Audibert e Lamb (2022) apresentaram uma metodologia que utiliza IA justamente para classificar se um determinado artigo aborda aspectos éticos. Nesse trabalho, os autores utilizaram uma base de dados rotulada por especialistas como conjunto de treinamento e avaliaram uma ampla base de artigos para realizar a classificação, mostrando, pelo menos em princípio, como a metodologia utilizada na IA pode ser utilizada para identificar estudos que lidam com aspectos éticos.

Questões fundamentais quanto à ética em (e da) inteligência artificial vão certamente além das preocupações dos pesquisadores e organizações quanto à avaliação inicial do impacto das tecnologias de IA expostas nas suas publicações (Avelar; Audibert; Lamb, 2022). Na década de 2010, trabalhos pioneiros na área de ética identificaram diversos riscos do uso indiscriminado ou pelo menos não integralmente avaliado, normatizado e mensurado da IA. Esses riscos e consequências incluíram desafios que a construção de sistemas éticos de IA tem a responder. Para que o leitor tenha uma noção concreta e realista dos desafios enfrentados na construção de IA ética, bem como da necessidade de integração de princípios éticos na IA, descrevemos quatro estudos de casos reais que ocorreram recentemente. Um desses trabalhos trata de um projeto de pesquisa visando responder a desafios inerentes ao uso da IA na seleção de pessoas para posições de trabalho. De forma inovadora, esse projeto remete à construção de uma investigação sobre questões relevantes para o mundo do trabalho por uma equipe multidisciplinar de potencial de alto impacto, liderado por pesquisadores da Universidade de Harvard. A seguir, descrevemos os casos em que se tornou explícita a necessidade de tratarmos a ética na IA e o uso da IA ética de forma sistematizada.

- Vieses de gênero, raça e nacionalidade em tecnologias de tradução automática e processamento de linguagem natural. Prates, Avelar e Lamb (2020) demonstraram que a ferramenta de tradução automática mais utilizada no mundo, Google Translate, pode exibir vieses de gênero e uma forte tendência para *defaults*/padrões masculinos, embora os autores tenham demonstrado que esses vieses resultam dos dados do mundo real utilizados no treinamento do sistema de tradução automática e há uma provável relação com a forma como a sociedade fala e escreve sobre gênero no mundo do trabalho. Através da análise sistemática da tradução de frases como “Ele/Ela é um engenheiro” (e outras similares) de línguas neutras de gênero, como o húngaro e chinês,

para o inglês, os autores coletaram estatísticas sobre a assimetria entre os pronomes de gênero feminino e masculino nos resultados da tradução. Mostraram, também, que os *defaults* masculinos, além de proeminentes, são exacerbados em áreas com estereótipos, como Stem (*science*, *technology*, *engineering and mathematics*). No artigo, os resultados também ressaltaram que o fenômeno do viés de gênero em tradução automática iria além das questões das profissões: a proporção de pronomes femininos variava significativamente nos experimentos realizados à época, de acordo com os adjetivos utilizados para descrever uma pessoa. Por exemplo, adjetivos como *shy* (envergonhada/o) e *desirable* (desejável) foram traduzidos em larga proporção e associados com pronomes femininos, enquanto adjetivos como *guilty* (culpado) e *cruel* (cruel) foram quase exclusivamente traduzidos para o masculino. Posteriormente a esse trabalho, foi percebido um notável desenvolvimento de uma linha de pesquisa em ética em IA e ética em processamento de linguagem natural (PLN), na qual pesquisadores passaram a analisar vieses em sistemas de PLN que utilizam múltiplas línguas e mecanismos de tradução, como, por exemplo, Fan (2021), Sheng (2019) e Devinney e Björklund (2022).

- Vieses na seleção de pessoas por departamentos de recursos humanos que utilizam as *hiring platforms* ou plataformas tecnológicas de contratação de pessoal (Kenthapadi; Venkataraman, 2017; Harvard, 2023).

  O trabalho de pesquisa liderado pela professora da Universidade de Harvard Cynthia Dwork (Harvard, 2023) parte das premissas de que sistemas hoje utilizados em departamentos de recursos humanos, embora sejam justos quando utilizados em isolamento, não são necessariamente justos em seu conjunto. O trabalho consiste fundamentalmente de uma equipe multidisciplinar que adotará uma abordagem holística para identificar problemas reais nesses tipos de sistemas de IA, hoje amplamente adotados nas organizações. Especificamente, sistemas cujas atividades de trabalho são descritas insuficientemente podem levar à exclusão de candidatos qualificados, bem como a própria linguagem utilizada pelo candidato pode levar à sua exclusão por conter vieses culturais.

  Nesse sentido, o projeto irá "explorar IA e outras técnicas que as plataformas de contratação (*hiring platforms*) podem utilizar para permitir uma comparação imparcial entre os candidatos". Outro desafio atual é que as plataformas que utilizam IA para contratação usam algoritmos de ranqueamento e pontuação de candidatos a empregos que acentuam as injustiças. Os pesquisadores desse projeto liderado por Dwork irão desenvolver e pesquisar algoritmos que anulem esses efeitos e melhorem o desempenho dos sistemas de IA utilizados em contratação de pessoas. Outros pontos a serem investigados nessa linha de pesquisa multidisciplinar envolvem a utilização de algoritmos preditivos para calibrar os escores preditivos do sucesso das pessoas em uma determinada posição de trabalho quando contratada. Para isso, os pesquisadores fazem uso da teoria

da pseudo-aleatoriedade para "questionar abordagens que assegurem que os modelos preditivos são tão acurados quanto possível". Também será pesquisada nesse projeto a busca de evidências sobre como modelar um ambiente de trabalho igualitário – como as plataformas de contratação que usam IA podem tornar os mecanismos de recomendações pessoais e digitais e as redes de relacionamento mais justas. Finalmente, outro ponto importante desse projeto multidisciplinar será a identificação de marcos regulatórios legais que podem vir a avançar a equidade das plataformas tecnológicas de contratação de pessoas. Em suma, esse projeto ilustra que as abordagens multidisciplinares podem ser uma resposta apropriada aos desafios que enfrentamos hoje quanto a desenvolver sistemas éticos de IA.

- Vieses raciais em algoritmos de IA que classificaram pessoas como não sendo seres humanos (Garcia, 2016; BBC News, 2015).

  Conforme relatado em 2015, Jacky Alcine, enquanto utilizava o sistema conhecido como Google Photos, percebeu que o aplicativo de reconhecimento de rostos (*faces*) rotulava ele e seu amigo como gorilas. Posteriormente, Alcine postou uma foto do sistema no Twitter que imediatamente se tornou viral nas redes sociais. Essa notícia causou grande repercussão na época, com veículos jornalísticos de todo o mundo reportando o assunto, incluindo a BBC News (2015). Além dos danos e prejuízos pessoais que podem ser múltiplos e cuja análise aprofundada vai além do escopo deste trabalho, este evento ocorrido ainda em 2015 ilustra a repercussão que os impactos éticos da IA podem ter sobre os usuários e, também, sobre organizações. No caso, uma empresa especializada em tecnologia de informação e pioneira em diversas aplicações da IA teve de se responsabilizar pelas consequências.

  Também levantamos a necessidade da reflexão, tanto por parte de profissionais e pesquisadores de IA quanto por equipes multidisciplinares sobre a relevância dos impactos éticos dessas tecnologias: nesse sentido, é relevante o exemplo do projeto relatado no item 2. Também nos parece claro que os indivíduos que fazem uso inadvertido ou desinformado dessas tecnologias podem vir a ser impactados, sendo esta ainda uma temática com múltiplas questões em aberto. Afinal, sobre quem recairá a responsabilidade sobre os impactos pessoais da utilização da IA? Sobre as empresas que produzem as tecnologias? Sobre os pesquisadores e engenheiros que desenvolveram os sistemas? Ou sobre os múltiplos agentes envolvidos na produção, certificação, padronização e posterior manutenção das tecnologias? Essas perguntas remetem ao conceito da IA responsável, que abordaremos sucintamente na última seção. Também é relevante mencionar que as iniciativas regulatórias dos países e organismos multilaterais visam compreender as melhores formas de responder a essas questões.
- Implicações do uso ético de tecnologias de IA por planos de saúde.

  No trabalho altamente citado de Obermeyer et al. (2019), os autores mos-

traram que planos de saúde que utilizam algoritmos preditivos de IA e aprendizado de máquina exibiram vieses raciais significativos. Esse é outro desafio e problema da IA ética, pois o uso de tecnologias na saúde pode afetar milhares de potenciais pacientes. Os pesquisadores demonstraram no seu artigo, publicado na revista *Science,* que o viés do sistema utilizado "surge porque o algoritmo prevê os custos dos cuidados de saúde e não a doença, mas o acesso desigual aos cuidados significa que gastamos menos dinheiro a cuidar de pacientes negros do que de pacientes brancos". Argumentam também que, "apesar dos custos dos cuidados de saúde parecerem ser um substituto eficaz para a saúde por algumas medidas de precisão preditiva, surgem grandes preconceitos raciais". Esse foi um trabalho pioneiro na área, demonstrando que aspectos éticos na IA, combinados com sistemas que devem fazer uso de valores éticos na sua construção, são fundamentais para a resolução de vieses e preconceitos que afetam uma população muito significativa.

Notadamente, como as abordagens atuais da IA são fundamentadas fortemente em bases de dados específicas para treinamento dos algoritmos de aprendizado de máquina, a utilização de bases de dados que não representam parcelas significativas e diversas da população mundial, combinada com práticas de desenvolvimento de sistemas que não consideram valores éticos, pode levar a consequências imprevistas e, principalmente, a sistemas que apresentam altos riscos para a população.

CONCLUSÃO: NA DIREÇÃO DA IA ÉTICA, CONFIÁVEL E RESPONSÁVEL

Este artigo apresentou uma breve introdução a temáticas inter-relacionadas e relevantes para o desenvolvimento da inteligência artificial ética. Abordamos aspectos relacionados à ética na IA, que se propõem a analisar como integrar valores morais consensualizados pelas sociedades nos sistemas tecnológicos de IA, bem como aspectos referentes à IA ética, que se trata do desenvolvimento de sistemas que demonstrem intenção de atuarem moral e eticamente, de acordo com os valores da sociedade. A abordagem utilizada foi a apresentação dos conceitos fundamentais que permitam ao leitor não especializado compreender os quatro estudos brevemente descritos. Esses estudos de caso ilustram os desafios atuais que pesquisadores e profissionais da área de IA enfrentam no desenvolvimento de tecnologias que incorporem valores éticos e sejam consideradas confiáveis pela sociedade.

Nesse sentido, cabe aqui apontar algumas tendências atuais de áreas de pesquisa que visam responder aos desafios pertinentes à integração entre ética e inteligência artificial. Um desses desafios diz respeito à construção de várias formas de IA que sejam consideradas *trustworthy AI*, *responsible AI* e *ethical AI*. Desse modo, um número significativo de pesquisadores se debruça sobre como construir inteligência artificial *confiável*, *responsável* e *ética*, respectivamente. Esses conceitos se inter-relacionam e apresentam desafios em comum para pesquisadores que atuam na área de IA.

Visando à construção de sistemas confiáveis, responsáveis e éticos, as pesquisas atuais apontam ser necessário responder aos desafios da *explicabilidade*, *interpretabilidade*, *semântica* e *responsabilidade*. Os avanços recentes na IA generativa[6], da qual tecnologias como ChatGPT (da empresa OpenAI) e LLaMA (da empresa Meta) são exemplos[7], acentuam ainda mais os desafios da construção de sistemas de IA confiáveis e éticos. Tais desafios são significativos pois esses sistemas não apresentam um modelo semântico formal que explique o seu funcionamento. Isto é, esses sistemas são considerados caixas-pretas com ausência de explicabilidade ou interpretabilidade. Isto significa – por mais que o leitor se surpreenda – que os pesquisadores ainda não explicaram rigorosamente se as respostas produzidas por esses sistemas estão corretas ou se são produzidas com alguma garantia de correção. Por exemplo, quando um usuário interage com grandes modelos de linguagem (Wolfram, 2023), como ChatGPT, as respostas obtidas muitas vezes são denominadas de "alucinações", ou seja, são criadas pelo sistema mas não têm nenhuma relação semântica com a pergunta do usuário. Outro fenômeno observado por pesquisadores foi a apresentação de vieses de gênero por parte dos grandes modelos de linguagem (Gordon, 2023).

Para responder a esses desafios, pesquisadores propuseram o uso de abordagens que combinem e integrem as abordagens de IA baseadas em redes neurais (que são eficientes para aprendizado sobre um grande volume de dados) com as abordagens da IA que utilizam a lógica simbólica (que permite a formulação rigorosa do processo de raciocínio). Esses sistemas, denominados *inteligência artificial neurossimbólica* (Garcez; Lamb, 2023), permitem, em princípio, interpretar e explicar de forma lógica e rigorosa o comportamento dos sistemas de IA construídos sob a abordagem das redes neurais artificiais. Isso decorre do fato de sistemas lógicos, por construção, terem semântica rigorosamente definida, além de estabelecerem os fundamentos também rigorosos da inferência e do raciocínio e por permitirem a representação formal de normas e valores, que podem ser representados nos sistemas de IA, idealmente mais éticos.

Em conclusão, é notório que os atuais sistemas de aprendizado de máquina (*machine learning*) se constituem no principal avanço da IA nas últimas duas décadas. Esses sistemas de aprendizado são a base fundamental das tecnologias da IA generativa, dos tradutores automáticos de linguagem, classificadores de imagens e fotografias, entre outras aplicações. Eles propiciaram a explosão da adoção

6 A IA generativa se refere a tecnologias de IA que produzem (geram) textos, imagens, músicas e outras expressões que representam conhecimento a partir de modelos de redes neurais artificiais (notadamente os modelos baseados em *transformers*; veja uma explicação simples no artigo de Stephen Wolfram [2023], bem como no artigo técnico de Rachel Gordon [2023]), que manipulam e aprendem a gerar aproximações sobre um grande conjunto de dados. Os sistemas se tornaram recentemente populares pela explosão de interesse dos usuários, notadamente pela tecnologia (*chatbot*) conhecida como ChatGPT. Este *chatbot* foi desenvolvido pela empresa de tecnologia OpenAI e lançado em novembro de 2022. O ChatGPT foi desenvolvido a partir das versões 3.5 e 4.0 da tecnologia chamada GPT (*generative pre-trained transformer*). Informações adicionais podem ser obtidas em: https://chat.openai.com/.

7 Grandes modelos de linguagem – do inglês *large language models* (LLMs).

das tecnologias de IA por organizações públicas, empresas e pela sociedade, com repercussão global e sobre estratégias de desenvolvimento de países e implicações em relações bilaterais, tendo em vista a relevância da IA na economia.

No entanto, os sistemas atuais de IA construídos sobre algoritmos baseados em redes neurais artificiais são, como já dito, considerados caixas-pretas pelos pesquisadores. Isso significa que não há ainda na literatura científica explicações rigorosas sobre o que essas caixas-pretas aprendem ou computam, para sermos mais precisos (Garcez; Lamb, 2023). A utilização de sistemas de IA que não sejam explicáveis ou interpretáveis pode vir a aumentar os riscos decorrentes do uso indiscriminado, inadvertido ou desinformado por parte de indivíduos, organizações e, até mesmo, por parte de governos. Para mitigar e responder a esses desafios, a construção de IA ética se apresenta como uma linha de pesquisa fundamental na academia. As pesquisas que visam identificar e analisar os princípios éticos para a construção de IA responsável, confiável e ética ainda permanecem como um grande desafio dos nossos tempos. Entretanto, tendo em vista o interesse atual por esse tema, confiamos nos resultados de pesquisas multidisciplinares na área. Nessa linha, também poderemos vislumbrar a posterior adoção desses princípios e valores éticos por profissionais e organizações de IA. Tais princípios éticos, imaginamos que talvez se tornem indispensáveis para o correto exercício profissional, bem como para a segura disseminação de tecnologias e produtos de IA em prol da humanidade.

REFERÊNCIAS

ANDERSON, M.; ANDERSON, S. L. (eds.). *Machine ethics*. Cambridge, Cambridge University Press, 2011.

ANDERSON, M.; ANDERSON, S. L.; ARMEN, C. *Machine ethics – Technical report*. FS-05-06. 2005 AAAI Fall Symposium. AAAI Press, Menlo Park, 2005.

ÅQVIST, L. "Deontic logic", in D. Gabbay; F. Guenthner, (eds.). *Handbook of philosophical logic. Volume II – Extensions of classical logic.* Dordrecht, Kluwer, 1994.

AUDIBERT, R. B. et al. "On the evolution of AI and machine learning: towards a meta-level measuring and understanding impact, influence and leadership at premier AI conferences". *Journal of Applied Logics – The IFCOLOG Journal of Logics and their Applications*, v. 10, n. 5, 2023, pp. 693-817.

AVELAR, P. H. C.; AUDIBERT, R. B.; LAMB, L. C. "Measuring ethics in AI with AI: a methodology and dataset construction". *BRACIS*, n. 1, 2022, pp. 370-84.
BAZZAN, A. L. C. et al. "'A nova eletricidade': aplicações, riscos e tendências da IA moderna", in *Escola de Computação PPGC/UFRGS 50 Anos*, 2023, pp. 167-209.
BBC NEWS. "Google apologises for photos app's racist blunder". *BBC News*, 1/jul./2015. Disponível em: https://www.bbc.com/news/technology-33347866.
BOSTROM, N.; YUDKOWSKY, E. "The ethics of artificial intelligence", in K. Frankish; W. M. Ramsey (eds.). *The Cambridge handbook of artificial intelligence*. Cambridge, Cambridge University Press, 2014, pp. 316-34.
BRYNJOLFSSON, E.; ROCK, D.; SYVERSON, C. "The productivity j-curve: How intangibles complement general purpose technologies". *NBER Working Paper*, n. 25148, oct./2018. Revised jan./2020.
BURTON, E. et al. "Ethical considerations in artificial intelligence courses". *AI Magazine*, v. 38, n. 2, 2017, pp. 22-34.
DEVINNEY, H.; BJÖRKLUND, J.; BJÖRKLUND, H. "Theories of gender", in *NLP bias research*. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.
DONNELLY, P. et al. *Machine learning: the power and promise of computers that learn by example*. The Royal Society, 2017.
FAN, A. et al. "Beyond English-centric multilingual machine translation". *Journal of Machine Learning Research,* v. 22, n. 107, 2021, pp. 1-48.
FURMAN, J. et al. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans, 2-3/feb./2018, ISBN 978-1-4503-6012-8.
GABBAY, D. et al (eds.). *Handbook of deontic logic and normative systems*, v. 1. Londres, College Publications, 2013.
GARCEZ, A. d'A.; LAMB, L. C. "Neurosymbolic AI: the 3rd wave". *Artificial Intelligence Review*, v. 56, n. 11, 2023.
GARCIA, M. "Racist in the machine: the disturbing implications of algorithmic bias". *World Policy Journal*, v. 33, n. 4, 2016, pp. 111-7.
GORDON, R. "Large language models are biased. Can logic help save them?". *MIT News,* 3/mar./2023. Disponível em: https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303.
HARVARD John A. Paulson School of Engineering and Applied Science News. "How can bias be removed from artificial intelligence-powered hiring platforms? Harvard-led institute to pursue fairness in online systems". 12/jun./2023. Disponível em: https://seas.harvard.edu/news/2023/06/how-can-bias-be-removed-artificial-intelligence-powered-hiring-platforms.
KENTHAPADI, K.; LE, B.; VENKATARAMAN, G. "Personalized job recommendation system at LinkedIn: practical challenges and lessons learned", in *Proceedings of the eleventh ACM conference on recommender systems*. 2017, pp. 346-7.
LIPSEY, R.; CARLAW, K.; BEKAR, C. "Economic transformations: general purpose technologies and long-term economic growth". *The Economic History Review*, v. 59, n. 4, 2006, pp. 881-2.
MARCUS, G.; DAVIS, E. *Rebooting AI: building artificial intelligence we can trust*. Vintage, 2019.
MISHRA, S.; CLARK, J.; PERRAULT, C. R. "Measurement in AI policy: opportunities and challenges". *CoRR*, 2020.

OBERMEYER, Z. et al. "Dissecting racial bias in an algorithm used to manage the health of populations". *Science*, v. 366, n. 6464, 2019, pp. 447-53.

PERRAULT, R. et al. *The AI index 2019 annual report*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, dec./2019.

PRATES, M. O. R.; AVELAR, P. H. C.; LAMB, L. C. "Assessing gender bias in machine translation: a case study with Google Translate". *Neural Comput. Appl., v.* 32, n. 10, 2020, pp. 6.363-81.

PRATES, M. O. R.; AVELAR, P. H. C.; LAMB, L. C. *On quantifying and understanding the role of ethics in AI research: a historical account of flagship conferences and journals*. Global Conference on AI (GCAI), 2018, pp. 188-201.

RUSSELL, S.; NORVING, P. *Artificial intelligence: a modern approach*. Pearson, 2003.

SAINATO, M. "Stephen Hawking, Elon Musk, and Bill Gates warn about artificial intelligence". *The Observer.* 19/ago./2015. Disponível em: https://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/.

SHENG, E. et al. "The woman worked as a babysitter: on biases in language generation". *EMNLP/IJCNLP*, n. 1, 2019.

VASWANI, A. et al "Attention is all you need". *NIPS*, 2017.

WALSH, T. "Artificial intelligence and ethics". *2015 AAAI Workshop*. Austin, AAAI Press, 2015.

WOLFRAM, S. "What is ChatGPT doing... and why does it work?". *Wolfram Media*, 2023.

ZHANG, D. et al. "The AI index 2021 annual report". *CoRR*, 2021.