LIAO, Matthew. *Ethics of artificial intelligence*, por Dora Kaufman. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 157-163.

Recebido em: 22 abr. 2021
Aprovado em: 05 mai. 2021

LIAO, Matthew (org.). *Ethics of artificial intelligence*. Oxford: Oxford University Press, 2020.

dx.doi.org/10.23925/1984-3585.2021i23p157-163

Resenha do livro *Ethics of artificial intelligence*, de Matthew Liao

Dora Kaufman[1]

A empresa inglesa *DeepMind Technologies*, fundada em 2010 e adquirida pelo Google em 2014, é uma referência no campo da Inteligência Artificial (IA); em 2016, seu programa *AlphaGo* não apenas venceu por quatro a um o campeão mundial do milenar jogo chinês Go, o sul-coreano Lee Sedol, como o fez com jogadas inéditas. O feito repercutiu na comunidade de IA, impulsionando o reconhecimento do papel estratégico da tecnologia pela China. No mesmo ano, não por coincidência, o *Center for Mind, Brain, and Consciousness* da NYU, sob a coordenação dos filósofos David Chalmers e Ned Block, reuniu cerca de trinta palestrantes, dentre pesquisadores de tecnologia e de ciências humanas, na conferência "AI Ethics".

No empenho de identificar como introduzir nos sistemas inteligentes princípios éticos e valores humanos, os painéis abordaram conceitos tais como moralidade e ética das máquinas, moralidade artificial e IA amigável. Ao longo da conferência, contudo, estabeleceu-se um consenso de que a ética pertence à esfera humana, ou seja, permeia as escolhas de desenvolvedores e usuários. Coube ao filósofo sueco Nick Bostrom, autor do livro "Superintelligence" (2014), abrir o evento alertando sobre os benefícios e riscos da concretização da "máquina inteligente" no século XXI. Além de Bostrom, a conferência contou com palestras de Peter Asaro, John Basl, Meia Chita-Tegmark, Kate Devlin, Vasant Dhar, Virginia Dignum, Mara Garza, Daniel Kahneman, Adam Kolber, Yann LeCun, Gary Marcus, Steve Petersen, Francesca Rossi, Stuart Russell, Ronald Sandler, Jürgen Schmidhuber, Susan Schneider, Eric Schwitzgebel, Frans Svensson, Jaan Tallinn, Max Tegmark, Wendell Wallach, Stephen Wolfram e Eliezer Yudkowsky.

1 Doutora na Escola de Comunicações e Artes pela USP. ORCID: orcid.org/0000-0001-7060-4887. CV Lattes: lattes.cnpq.br/8045171889826285. E-mail: dkaufman@usp.br.

Essa é a origem da coletânea de artigos organizada por S. Matthew Liao (doze dos trinta colaboradores foram palestrantes na conferência). Composta de 17 ensaios inéditos, produzidos por cientistas e filósofos, agrupados em quatro seções, a coletânea aborda dilemas-chave para uma IA ética, dentre outros, os impactos da automação inteligente no mercado de trabalho; o viés contido nos dados que perpetuam os preconceitos da sociedade; a ética envolvida em aplicações como carros autônomos, sistemas de vigilância, armas autônomas; robôs sexuais; direitos e consciência da IA; e status moral. Numa perspectiva futura, os ensaios da terceira seção refletem sobre os riscos da "superinteligência".

No ensaio inicial, "A Short introduction to the ethics of Artificial Intelligence", S. Matthew Liao subdivide as abordagens éticas em dois conjuntos: (a) as associadas à eficiência da técnica em alguns domínios, implicando que os humanos podem se sentir vulneráveis ao lidar com esses sistemas, denominadas por ele de "vulnerabilidades humanas", e (b) as associadas às limitações da técnica, denominadas por ele de "vulnerabilidades no aprendizado de máquina". O primeiro conjunto de questões éticas abarca, dentre outras externalidades negativas, a ameaça ao suposto "livre arbítrio" dos indivíduos, função da capacidade dos algoritmos de IA extrair dos dados conhecimento inédito sobre os usuários das plataformas/dispositivos tecnológicos e, com base nele, elaborar estratégicas para influenciar/alterar/manipular o comportamento humano; a privacidade por conta da disseminação dos sistemas de monitoramento e vigilância com o uso de técnicas de reconhecimento facial; o aperfeiçoamento das fake news com as deepfakes e sua capacidade de distorcer imagem e voz, simulando falas, imagens e vídeos de pessoas reais com forte aproximação da realidade; o deslocamento do trabalhador humano por sistemas inteligentes mais rápidos e mais eficientes e a um custo menor. No segundo conjunto de questões éticas, destacam-se o problema do viés nos modelos de IA e o problema da não explicabilidade de como os modelos chegaram ao resultado final.

A indagação central de Liao é como criar sistemas de IA que sejam justos e não gerem resultados tendenciosos inadvertidamente; outro aspecto abordado no ensaio é se é factível atribuir status moral aos sistemas de IA. Sobre os resultados tendenciosos, é importante ter em mente que a maior parte das aplicações atuais de IA é baseada na técnica de *machine learning* denominada Redes Neurais de Aprendizado Profundo (*Deep Learning Neural Networks* – DLNNs), em que os algoritmos "aprendem" estabelecendo correlações a partir de grandes conjuntos de dados

(*big data*). Nessa técnica, o viés deriva (a) da codificação de estruturas e padrões mentais existentes, filtrados pelos desenvolvedores dos sistemas ao definir variáveis iniciais, arquiteturas, base de dados; (b) de dados tendenciosos, no caso da base de dados de treinamento dos algoritmos não representar o universo do objeto em questão; (c) da realidade ser enviesada, quando os dados refletem os preconceitos existentes na sociedade; e (d) de previsões baseadas em dados do passado, efeito minimizado em séries não temporais (*computer vision*/*image recognition* e NLP – *Natural Language Processing*). O tema do status moral dos sistemas de IA é abordado por outros autores, e retomado por Liao no último ensaio da coletânea.

Parte I: Construindo Ética em Máquinas, com cinco ensaios

No primeiro ensaio, "Ethical learning, natural and artificial", Peter Railton alerta para os impactos éticos diretos e indiretos, na medida em que a IA altera as capacidades e os potenciais benefícios e riscos de outras tecnologias. Reconhecendo o crescente protagonismo dos sistemas artificiais na tomada de decisão que afetam a vida, o autor investiga a possibilidade desses sistemas se tornarem sensíveis às questões éticas, definido essa sensibilidade como a "capacidade robusta e confiável de detectar e responder apropriadamente a características eticamente relevantes de situações, ações, agentes e resultados" (p. 45). No segundo ensaio, "The use and abuse of the Trolley Problem: self-driving cars, medical treatments, and the distribution of harm", F. M. Kamm apresenta os casos comumente utilizados para ilustrar os dilemas éticos e os julgamentos morais padrão associados a condutas permissíveis. Kamm diferencia as questões morais dos carros autônomos do padrão de *trolley problem*, atribuindo responsabilidade aos programadores por danos a pedestres, motoristas e passageiros. No terceiro ensaio, "The moral psychology of AI and the ethical opt-out problem", Jean-François Bonnefon, Azim Shariff, e Iyad Rahwan argumentam que a promessa da IA de melhorar as decisões humanas só pode se tornar realidade se incorporar os *trade-offs* morais exclusivos dos humanos, tarefa da competência dos cientistas comportamentais que terão que adaptar os métodos de psicologia moral a domínios técnicos complexos.

No quarto ensaio, "Modeling and reasoning with preferences and ethical priorities in AI Systems", Andrea Loreggia, Nicholas Mattei, Francesca Rossi e K. Brent Venable defendem a premência de construir sistemas inteligentes que se comportem moralmente (alinhados com valores humanos), pré-condição para torná-los confiáveis, particularmente no caso dos "robôs cuidadores". Os autores propõem uma modelagem

para detectar prioridades éticas, e os possíveis desvios com referência nos valores da comunidade de usuários desses sistemas. O quinto ensaio, "Computational law, symbolic discourse, and the AI constitution", Stephen Wolfram defende a viabilidade de criar uma linguagem de discurso simbólica geral e aplicá-la para construir uma estrutura para o direito computacional, incluindo no futuro as "IAs": que ética e quais princípios, e como inseri-los nos sistemas.

Os cinco ensaios atribuem um agenciamento inexistente nos sistemas atuais de IA, que são "meros" modelos estatísticos de probabilidades aos quais não pode ser atribuída a condição de agente moral, pressuposto corroborado por vários filósofos. Mark Coeckelbergh (2019, 2020) argumenta que as tecnologias de IA podem ser agentes, no sentido de atuar no mundo, mas não atendem aos critérios tradicionais de agente moral mesmo reconhecendo que as decisões automatizadas com IA podem não ser moralmente neutras. Wendell Wallach e Colin Allen (2009) cunharam o termo "moralidade funcional", para designar uma moralidade intermediária, nem plena nem neutra. Bostrom e Yudkowsky (2014) recusam conceder aos atuais sistemas de IA, ainda de competência restrita a um único domínio, o status moral, mesmo que seus algoritmos exerçam funções cognitivas anteriormente atribuídas aos seres humanos.

Parte II: O Futuro Próximo da Inteligência Artificial com quatro ensaios

No primeiro ensaio, "Planning for mass unemployment: precautionary basic income", Aaron James trata do efeito da automação sobre o emprego e potenciais iniciativas para evitar o desemprego em massa. No segundo ensaio, "Autonomous weapons and the ethics of Artificial Intelligence", Peter Asaro alerta sobre o potencial das "armas autônomas" transformarem radicalmente a guerra, o policiamento e o entendimento de direitos humanos relacionados a máquinas e algoritmos de IA. Diante da "imoralidade" dessas armas, o autor argumenta a favor da supremacia dos direitos e deveres morais sobre as razões utilitárias.

No terceiro ensaio, "Near-Term Artificial Intelligence and the ethical matrix", Cathy O'Neil e Hanna Gunn argumentam que os problemas no curso prazo dos sistemas de IA são problemas morais, presentes desde as decisões de design dos algoritmos. Por meio de estudos de caso, as autoras buscam mostrar que os interesses humanos não estão sendo contemplados no desenvolvimento e uso desses sistemas, sugerindo a criação de uma "matriz ética" aos moldes da ética de ciências de dados.

No quarto ensaio, "The ethics of the artificial lover", Kate Devlin aborda as tecnologias de sexo, ainda de consumo de nicho, mas com potencial de expansão ao promover experiências robóticas multissensoriais. Para a autora, essas tecnologias podem proporcionar vidas sexuais mais gratificantes ao romper barreiras fisiológicas, psicológicas e discriminatórias, sem negligenciar a ética (segurança de dados, privacidade e controle e consentimento do usuário).

São inúmeras as externalidades negativas, éticas e sociais, dos sistemas atuais de IA, algumas gerais e outras relacionadas ao setor e/ou tarefa de aplicabilidade (a natureza e grau de impacto ético associados a sistemas de recomendação de filmes/música são radicalmente distintos, por exemplo, de sistemas bélicos). A técnica de DLNNs, como todo modelo estatístico de probabilidade, é intrinsecamente incerta, ademais, sendo baseada em grandes conjuntos de dados, agrega os vieses contidos nos dados. Essas e outras características dos modelos de IA implicam em questões éticas a serem equacionadas, ao menos minimizadas, pela sociedade. Do âmbito social, a automação inteligente configura-se como a maior ameaça ao eliminar funções repetitivas e cognitivas em distintos setores econômicos.

Parte III: Impactos de longo prazo da superinteligência com quatro ensaios

No primeiro ensaio, "Public policy and superintelligent AI: a vector field approach", Nick Bostrom, Allan Dafoe e Carrick Flynn chamam a atenção para uma série de circunstâncias especiais que podem cercar o desenvolvimento e a implantação da "IA superinteligente"; com base em uma abordagem de "campo vetorial" para a análise normativa, os autores buscam extrair implicações de política direcional dessas circunstâncias especiais, implicações caracterizadas como um conjunto de "desiderata" (p. 313-314). "Propostas de política" referem-se a documentos oficiais do governo e planos desenvolvidos por atores privados interessados no desenvolvimentos de longo prazo da IA. O segundo ensaio, "Artificial Intelligence: a binary approach", Stuart Russell aborda o problema de controle da IA, o que envolve a construção de sistemas mais poderosos que os humanos com a garantia de que os mesmos serão benéficos. Russell distingue sistemas "melhores para tomar decisão" de sistemas que tomam as melhores decisões.

No terceiro ensaio, "Alignment for advanced Machine Learning systems", Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire e Andrew Critch identificam oito áreas de pesquisa direcionadas para projetar sistemas de IA robustos e confiáveis, ressaltando que as soluções devem contemplar os sistemas atuais e os sistemas altamente inteligentes do futuro, bem como devem funcionar na teoria e na prática. No quarto ensaio, "Moral machines: from value alignment to embodied virtue", Wendell Wallach e Shannon Vallor partem das "Três leis para robôs" de Isaac Asimov para discutir leis mais adequadas à complexidade da Inteligência Artifical Geral (*General Artificial Intelligence* - GAI), mesmo reconhecendo que a IA ainda continuará a ser projetada para contextos morais limitados nas próximas décadas, requerendo engenharia, teste, vigilância, supervisão e refinamentos iterativos. No quinto ensaio, "Machines Learning Values", Steve Petersen, considerando factível a possibilidade de os humanos criarem uma superinteligência artificial com valores próprios (capaz, inclusive, de exterminá-los), defende projetar essa superinteligência com valores fundamentais semelhantes aos humanos, conhecido como "alinhamento de valores". O autor reconhece a complexidade desses valores para serem programados explicitamente, mas não para serem "aprendidos" pelas técnicas de *machine learning,* identificando três obstáculos e suas potenciais soluções inter-relacionadas.

As questões dessa seção remetem a previsões sobre o futuro da IA que, dada as limitações atuais das técnicas de *machine learning*, carecem de evidências científicas de que serão (ou não) concretizadas.

Parte IV: Inteligência Artificial, Consciência e Status Moral com três ensaios

No primeiro ensaio, "How to catch an AI zombie: testing for consciousness in machines", Susan Schneider alerta para a premência de antecipar os futuros problemas quando a IA superará os humanos em múltiplos domínios. A autora aborda de diferentes ângulos o conceito de "consciência", inclusive o que seria uma "IA consciente" aferidos por meio de vários testes ou marcadores. No segundo ensaio, "Designing AI with rights, consciousness, self-respect, and freedom", Eric Schwitzgebel e Mara Garza ponderam que no futuro será possível criar entidades com AI que mereçam tanta consideração moral quanto os seres humanos. Os autores convocam filósofos e formuladores de políticas para antecipar a discussão dos princípios éticos associados, com o pressuposto de que a IA merecedora de consideração moral equivalente à humana deve ser projetada com valores próprios (não necessariamente humanos).

No terceiro ensaio, "The moral status and rights of Artificial Intelligence", S. Matthew Liao retoma a questão do status moral dos sistemas de IA defendendo sua aplicabilidade (a) no caso de IA "vivas, conscientes ou sencientes" (capaz de sentir dor, ter desejos); (b) no caso de ter "base física"; (c) no caso de base física por emulação do cérebro; (d) no caso de seres humanos interessados em se tornar IAs por substituição gradual (em vez de emulação do cérebro); e (e) no caso de IA com direitos autorais. Segundo o autor, alguns dos direitos serão semelhantes aos direitos dos seres humanos, como o direito à vida e à liberdade e o direito a igual proteção, além de direitos exclusivos de sua natureza, como o direito de controlar sua taxa subjetiva de tempo. Liao inclui as formas de vida artificiais em uma lista de nove entidades com potencial de ter status moral: objetos inanimados (rochas, obras de arte, edifícios, o meio ambiente); coisas vivas terrestres não humanas (plantas e animais); seres humanos com funcionamento normal; seres humanos feridos (gravemente deficientes mentais); seres humanos no início da vida (fetos, bebês, crianças pequenas); possíveis seres humanos futuros (gerações futuras); seres humanos não vivos (seres humanos mortos); espécies extraterrestres não humanas de seres vivos (alienígenas, seres do espaço sideral); e formas de vida artificiais (androides, robôs, computadores, algoritmos).

São múltiplas as externalidades negativas dos sistemas atuais de IA, éticas e sociais. O desafio é como mitigá-las preservando as externalidades positivas intrínsecas aos modelos de negócio baseados em dados (*data-driven models*). Esse cenário recomenda manter o foco na busca por soluções de curto-médio prazo, deixando o longo prazo para a esfera da ficção científica.

Referências

BOSTROM, Nick; YUDKOWSKY, Eliezer. The Ethics of Artificial Intelligence. In: FRANKISH, Keith; RAMSEY, William (eds.). *The Cambridge Handbook of Artificial Intelligence*. New York, NY: Cambridge University Press, 2014. Disponível em: cambridge.org/core/books/cambridge-handbook-of-artificial-intelligence/ethics-of-artificial-intelligence B46D2A9DF7CF3A9D92601D9A8ADA58A8. Acesso em: 12 maio 2021.

COECKELBERGH, Mark. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 2019. Disponível em: link.springer.com/article/10.1007/s11948-019-00146-8. Acesso em: 5 abril 2021.

_____. AI Ethics. Cambridge, MA: MIT Press, 2020.

WALLACH, Mendell; ALLEN, Colin. Moral Machines: Teaching Robots Right from Wrong. New York, NY: Oxford University Press, 2009.