DOI: https://doi.org/10.31657/rcp.v5i9.453

Artigo

Inteligência Artificial e os desafios éticos: a restrita aplicabilidade dos princípios gerais para nortear o ecossistema de IA

Artificial Intelligence and the ethical challenges: the restricted applicability of general principles to guide the IA ecosystem

Inteligencia Artificial y desafíos éticos: la estricta aplicabilidad de los principios generales para guiar el ecosistema de IA

Dora Kaufman
Pontifícia Universidade Católica de São Paulo
<dkaufman@pucsp.br>

Resumo

Prolifera na sociedade o uso de tecnologias de inteligência artificial (IA). A maior parte das implementações atuais de IA é baseada na técnica de aprendizado de máquina (*machine learning*), subárea da IA, denominada de redes neurais de aprendizado profundo (*Deep Learning Neural Networks* – DLNNs) cujos algoritmos "aprendem" a partir de exemplos extraídos do big data. Nesse processo de automação de decisões, intensifica-se o debate da IA ética, concentrado em princípios gerais de aplicabilidade restrita, não traduzíveis em boas práticas para nortear o ecossistema de IA. Ademais, alguns desses princípios, como justiça e dignidade, não são universais e desconhece-se como decodificá-los em termos matemáticos. O artigo pondera sobre algumas soluções para mitigar as externalidades negativas sugeridas por Luciano Floridi, Mark Coeckelbergh e Cetric Villani.

Palavras-chave: Inteligência artificial. Ética. Externalidades negativas.

Abstract

The use of artificial intelligence (AI) technologies is widespread in society. Most current AI implementations are based on the machine learning technique, subarea of AI, called Deep Learning Neural Networks (DLNNs) whose algorithms "learn" from examples extracted from the big date. In this process of automated decisions, the debate on ethical AI is intensified focused on general principles of restricted applicability, not translatable into good practices to guide the AI ecosystem. Furthermore, some of these principles, such as justice and dignity, are not universal and it is unknown how to decode these principles in mathematical terms. The article considers some solutions to mitigate negative externalities suggested by Luciano Floridi, Mark Coeckelbergh and Cetric Villani.

Keywords: Artificial intelligence. Ethics. Negative externalities.

Resumen

El uso de tecnologías de inteligencia artificial (IA) está muy extendido en la sociedad. La mayoría de las implementaciones de IA actuales se basan en la técnica de aprendizaje automático, una subárea de la IA, llamada Deep Learning Neural Networks (DLNN) cuyos algoritmos "aprenden" de ejemplos extraídos de los grandes datos. En este proceso de decisiones automatizadas, el debate sobre la IA ética se intensifica centrado en principios generales de aplicabilidad restringida, no traducibles en buenas prácticas para orientar el ecosistema de IA. Además, algunos de estos principios, como la justicia y la dignidad, no son universales y se desconoce cómo decodificar estos principios en términos matemáticos. El artículo considera algunas soluciones para mitigar las externalidades negativas sugeridas por Luciano Floridi, Mark Coeckelbergh y Cetric Villani.

Palabras clave: Inteligencia artificial. Ética. Externalidades negativas.

1. Introdução

O Sex-Disaggregated Data Tracker,[1] maior banco de dados desagregados por sexo sobre Covid-19, apurou que, nos países onde há dados disponíveis, mais da metade de todas as mortes são entre os homens.[2] No Afeganistão, Bangladesh, Índia e Paquistão, por exemplo, os homens são responsáveis por mais de 65% das infecções e mortes por Covid-19.[3] Apesar de aparentemente não haver diferenças substanciais de gênero na probabilidade de infeção, o grau de probabilidade dos homens morrerem de Covid-19 é substancialmente maior. Segundo indicadores científicos, com evidências sugestivas e não conclusivas, as causas são específicas para este vírus.[4]

Uma das limitações dos estudos é a ausência de coleta desagregada de dados por gênero: no final de março de 2020, apenas seis dos vinte países mais afetados pela Covid-19 publicaram dados desagregados por gênero, sendo que os EUA e o Reino Unido só o fizeram plenamente em maio; em setembro de 2020, menos de 50% dos países desenvolvidos publicaram dados desagregados (PEREZ-CRIADO, 2021). Ao não contemplar as distinções de gênero na função imunológica, comprometem-se os esforços de identificar os sintomas e, consequentemente, a eficácia dos tratamentos.

Ilustrando a importância da desagregação de dados por gênero, Caroline Criado Perez (2021) cita um estudo realizado em 2016, num hospital em Long Island, NY, que correlacionou o hormônio feminino estrogênio com resultados positivos no combate ao vírus em geral; em 2020, esse mesmo hospital injetou estrogênio em seus pacientes homens com Covid-19 (os resultados não foram apurados plenamente, ou não são públicos). No livro *Invisible Women: data bias in a world designed for men*, PEREZ-CRIADO (2021) realiza um extenso levantamento histórico da "invisibilidade" feminina aditando valiosa contribuição para o debate global sobre discriminação de gênero nos dados.

Como a técnica de inteligência artificial (IA) que permeia a maior parte das aplicações atuais é baseada em dados, a sociedade está tomando decisões enviesadas por gênero em número maior do que o percebido. Na Inglaterra, por exemplo, as mulheres têm 50% mais chance de serem diagnosticadas erroneamente após um ataque cardíaco, em função da predominância de homens nos estudos sobre insuficiência cardíaca (PEREZ-CRIADO, 2021). A prática de não coletar dados desagregados por gênero, tratando os homens como neutros e/ou "padrão humano" e, a partir dessas bases de dados tendenciosas, extrair supostos padrões de comportamento humano, distorce os resultados dos modelos. O viés de gênero é apenas uma ilustração dos impactos éticos, atuais e futuros, associados a decisões automatizadas por IA.

Existem inúmeras definições de IA, reflexo das especificidades de cada campo

1 Disponível em: https://globalhealth5050.org/the-sex-gender-and-covid-19-project/the-data-tracker/. Acesso em: 3 abr. 2021.
2 Disponível em: https://globalhealth5050.org/the-sex-gender-and-covid-19-project/. Acesso em: 2 abr. 2021.
3 Disponível em: https://www.devex.com/news/opinion-we-lack-an-essential-component-to-power-covid-19-response-98054. Acesso em: 2 abr. 2021.
4 Disponível em: https://www.medrxiv.org/content/10.1101/2021.02.23.21252314v1. Acesso em: 2 abril 2021.

de conhecimento. Russell e Norvig (2009) listam oito delas, agrupadas em duas dimensões – as relativas a processos mentais e raciocínio, e as relativas a comportamento – contudo, duas definições generalistas servem ao propósito do artigo: a primeira de John McCarthy, "IA é a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes".[5] Dadas a abrangência e a imprecisão do significado de "inteligência", Davi Geiger[6] sugere considerar que a "IA é a ciência e a engenharia de criar máquinas que tenham funções exercidas pelo cérebro dos animais".

Em seus primórdios, o desafio do campo da IA era resolver tarefas executadas pelos humanos intuitivamente e com relativo grau de subjetividade, como reconhecer palavras faladas ou rostos em imagens, pela dificuldade de descrevê-las formalmente na programação computacional. Várias tentativas envolvendo linguagens formais, apoiadas em regras de inferência lógica, não foram exitosas, indicando a necessidade de os sistemas gerarem seu próprio conhecimento extraindo padrões de dados, ou seja, "aprender" com os dados sem receber instruções explícitas. Esse processo convencionou-se denominar de "aprendizado de máquina" (*machine learning*), subcampo da IA criado em 1959, três anos após a criação do próprio campo (DOMINGOS, 2015; GOODDFELOW; BENGIO; COURVILLE, 2016; ALPAYDIN, 2016).

O processo de "aprendizagem" desses sistemas é influenciado por múltiplos fatores, observáveis ou não observáveis no mundo físico, sujeitos à efeitos de fontes externas. A técnica de aprendizado de máquina que consegue lidar com a complexidade do mundo real é denominada de "aprendizado profundo" (*deep learning)*: função matemático-estatística que mapeia conjuntos de valores de entrada (*inputs*) para valores de saída (*output*) por meio de representações expressas em termos de outras representações mais simples, identificadas em distintas camadas (*layers*). A partir de 2012, os modelos empíricos com uso dessa técnica apresentaram resultados positivos – taxa de acurácia superior às alternativas disponíveis –, particularmente em visão computacional, reconhecimento de voz e imagem.

Essa relativamente nova técnica de aprendizado de máquina é também denominada de Redes Neurais de Aprendizado Profundo (*Deep Learning Neural Networks* – DLNNs) por sua inspiração no funcionamento do cérebro biológico. Para desempenhar uma tarefa específica como, por exemplo, identificar se uma imagem de um tumor numa tomografia é cancerígena ou não, é necessário primeiro treinar o sistema com base em grandes conjuntos de dados de entrada (*inputs*) – imagens de tumor cancerígeno e imagens de tumor-padrão. As DLNNs projetam cenários futuros, a probabilidade deles ocorrerem e quando, permitindo tomar medidas preventivas contra os potenciais danos.

A técnica possui limitações, tais como: (a) requer abundância de dados; (b) requer *hardware* com grande capacidade de processamento; (c) modelos opacos, não – explicabilidade (*black box*) de como os algoritmos chegaram ao objetivo/meta (*out-*

5 Disponível em: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html. Acesso em: 2 abr. 2021.
6 Disponível em: **TECCOGS – Revista Digital de Tecnologias Cognitivas**, n. 17, jan.-jun 2018, ISSN: 1984-3585 Programa de Pós-graduação em Tecnologias da Inteligência e Design Digital (TIDD) / PUC-SP.

put); (d) resultados enviesados, função das decisões do desenvolvedor do modelo e/ou da qualidade da base de dados de treinamento do modelo (KAUFMAN, 2019). Essas limitações, em parte, engendram externalidades negativas éticas e sociais.

O filósofo americano S. Matthew Liao (2020) distingue as questões éticas da IA 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 aparente "livre-arbítrio" dos indivíduos, na medida em que os algoritmos de IA são capazes de extrair dos dados conhecimento inédito sobre os usuários das plataformas e dispositivos tecnológicos e, com base nesse conhecimento, prever e elaborar estratégias para influenciar, alterar e/ou manipular o comportamento humano; a ameaça à privacidade por conta da disseminação dos sistemas de monitoramento e vigilância com o uso de técnicas de reconhecimento facial; as *deep fakes* e sua capacidade de distorcer imagem e voz, simulando falas, imagens e vídeos de pessoas reais com forte aproximação da realidade; e 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 os problemas do viés nos dados e da não explicabilidade de como os modelos chegaram ao resultado final.

O foco do artigo é o segundo conjunto de questões éticas, ponderando sobre soluções para mitigar as externalidades negativas sugeridas pelos filósofos Luciano Floridi e Mark Coeckelbergh, e pelo matemático Cetric Villani.

2. Princípios éticos gerais e externalidades negativas da IA

O debate corrente da IA ética tem se concentrado em princípios gerais, replicados em inúmeros documentos mundo afora, de aplicabilidade restrita, não traduzíveis em boas práticas para nortear o ecossistema de IA (pesquisadores, desenvolvedores, instituições, academia, empresas, governos); ademais, alguns desses princípios, como justiça e dignidade, não são universalmente aceitos, assumindo significados locais distintos e, mais desafiador, desconhece-se como traduzir esses princípios em termos matemáticos.

A origem desses princípios gerais é a *Conference on Beneficial AI*, realizada em 2017, organizada pelo Future of Life Institute (fundado em 2014 pelo cosmologista Max Tegmark do MIT). Debateu-se um conjunto de diretrizes éticas para garantir o desenvolvimento de tecnologias de IA benéficas à sociedade, no que ficou conhecido como os "*Asilomar Principles*": 23 princípios, subdivididos em três categorias – pesquisa, ética e valores, e questões de longo prazo.[7] A essência desses princípios gerais está na base fundadora de diversos institutos – além do Future of Life Institute, o Future of Humanity Institute, liderado pelo filósofo inglês Nick Bostrom; o AI

7 Disponíveis em: https://futur eoflife.org/ai-principles/. Acesso em: 3 abr. 2021.

Now Institute, Universidade de Nova York (2018); o AI for Good Institute, Universidade de Stanford (2019); e o Leverhulme Center for the Future of Intelligence, Universidade de Cambridge –, e de iniciativas de organizações multilaterais e europeia.

Em 2018, organizado pela OCDE (Organização para a Cooperação e Desenvolvimento Econômico), ocorreu em Bruxelas o AI4People,[8] primeiro fórum global da Europa sobre os impactos sociais da IA. Reunindo mais de 50 especialistas independentes, pesquisadores, tomadores de decisão e representantes da indústria e da sociedade civil, o propósito era esboçar um conjunto de diretrizes éticas destinado a facilitar o desenho de políticas favoráveis ao desenvolvimento de uma "IA benéfica", condensado no documento *"An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*.[9] Em outubro do mesmo ano, a Conferência Internacional de Comissários de Proteção de Dados e Privacidade[10] se reuniu em Bruxelas e aprovou várias resoluções que tratam do desenvolvimento de tecnologia e questões de big data. O tema da conferência foi ética digital, e uma das resoluções aprovadas foi sobre privacidade e ética em IA, visando a garantir que os sistemas baseados em aprendizado de máquina (a) respeitem os direitos e as leis de privacidade, e sejam legais e justos em suas aplicações; (b) permaneçam consistentes com seus propósitos originais; (c) prestem contas a todas as partes interessadas; (d) estabeleçam processos de governança e/ou criação de comitês de ética independentes; (e) promovam a transparência algorítmica e auditabilidade dos sistemas; (f) avaliem e documentem os impactos esperados sobre indivíduos e sociedade na partida (ética *by design*); e (g) garantam aos indivíduos o pleno exercício de direitos individuais, atenuando preconceitos ilegais ou práticas discriminatórias e investindo em pesquisas para descobrir maneiras de identificar, abordar e diminuir os vieses.

Em fevereiro de 2020, a Comissão Europeia publicou o *White Paper on Artificial Intelligence*[11] com diretrizes políticas sobre como atingir o duplo objetivo de promover o desenvolvimento e a adoção da IA e, simultaneamente, enfrentar os riscos associados a certos usos dessa tecnologia. O documento foi submetido à consulta pública, gerando contribuições para a elaboração pela Comissão Europeia, em 2021, de uma proposta de regulamentação da IA.[12] Baseada em risco, a proposta delimita os siste-

8 Comitê Científico da AI4People é composto por 12 especialistas e presidido por Luciano Floridi, tendo como propósito elaborar/propor recomendações para o desenvolvimento de uma Good AI Society com base em quatro oportunidades: quem podemos nos tornar (autorrealização autônoma); o que podemos fazer (agência humana); o que podemos alcançar (capacidades individuais e sociais); e como podemos interagir uns com os outros e com o mundo (coesão social) (Disponível em: https://link.springer.com/article/10.1007/s11023-018-9482-5 – page 1 to 25. Acesso em: 15 fev. 2021).

9 Disponível em: https://ai4people.eu/. Acesso em: 4 abr. 2021.

10 Disponível em: https://www.privacyconference2018.org/en.html. Acesso em: 15 fev. 2021.

11 European Commission, White Paper on Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final, 2020. Disponível em: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. Acesso em: 20 jun. 2021.

12 "Laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Acts) and Amending Certain Union Legislative Acts". Disponível em: https://www.europeansources.info/record/proposal-for-a-regulation-laying-down-harmonised-rules-on-artificial-intelligence-artificial-intelligence-act-and-amending-certain-union-legislative-acts/. Acesso em: 20 jun. 2021.

mas e seus usos em três categorias – *"unacceptable risk"* (risco inaceitável), *"high risk"* (risco elevado) e *"low/minimal risk"* (risco baixo/mínimo) –, prevendo procedimentos para fornecedores e usuários, com multas significativas para situações de não conformidade.

Em paralelo, as grandes empresas de tecnologia se movimentam para adequar seus produtos e serviços às leis de proteção de dados, que indiretamente afetam seus modelos de negócio, e incorporar valores éticos e sociais visando a proteger suas reputações. O *People + AI Research* (PAIR), área do Google criada em 2010, em 2019 lançou o *AI Guidebook*, conjunto de diretrizes para orientar seus desenvolvedores, *designers* e gerentes de negócio, com base em recomendações de mais de 150 especialistas internos, pesquisadores acadêmicos e consultores externos.

Além da natureza abstrata dos princípios gerais, existem outros complicadores. Não basta aplicar esses princípios, supondo ser viável, no estágio inicial de desenvolvimento e implementação dos sistemas de IA; o alinhamento aos princípios na partida, com novos conjuntos de dados, pode posteriormente desalinhar, o que requer um monitoramento contínuo (praticamente inviável nas condições atuais). Ampliando os desafios, a estruturação dos modelos reflete a subjetividade humana, o *status quo*, ou seja, os modelos codificam estruturas e padrões mentais existentes (filtrados pelos seus desenvolvedores). Além disso, como reconhece o Fórum Econômico Mundial (2021),[13] existe uma lacuna de conhecimento entre os desenvolvedores e os reguladores de IA, e a tendência é que essa assimetria informacional aumentar à medida que aumente a complexidade dos modelos.

Viés e opacidade não têm o mesmo grau de problematização nas aplicações de IA: o dano de uma recomendação equivocada em uma plataforma de *streaming*, por exemplo, é significativamente menor do que uma recomendação equivocada no contexto de saúde, de vigilância e de educação.

2.1. Viés nos resultados

A aplicação de IA à visão computacional, baseada nas DLNNs, é usada desde a identificação de imagens captadas por drones e satélites da superfície da Terra até imagens de tomografia para diagnóstico de pacientes, detecção de pragas na agricultura, personalização no trato com o gado, e nos modelos de negócio das plataformas digitais. Um dos mais controversos, contudo, tem sido o uso nos sistemas de vigilância. O relatório *Face Recognition Vendor Test (FRVT/2019)*, publicado pelo Instituto Nacional de Padrões e Tecnologia dos EUA (NIST/ *National Institute of Standards and Technology/US Departament of Commerce*),[14] descreve e quantifica os diferenciais demográficos para 200 algoritmos de reconhecimento facial de quase 100 desenvolvedores envolvendo mais de 18 milhões de imagens de mais de oito milhões de pessoas. O relatório encontrou evidências empíricas da existência de diferenças demográficas na maioria dos algoritmos avaliados, com os resultados enviesados por etnia, idade e sexo: afro-americanos e nativos americanos

13 Disponível em: https://www.weforum.org/agenda/2021/02/we-need-to-talk-about-artificial-intelligence/. Acesso em: 4 abr. 2021.
14 Disponível em: https://doi.org/10.6028/NIST.IR.8280. Acesso em: 7 abr. 2021.

ilhéus do Pacífico foram falsamente identificados, bem como crianças e idosos. Em alguns casos, os asiáticos e afro-americanos foram identificados erroneamente até 100 vezes mais do que os homens brancos; as maiores taxas de precisão foram encontradas entre homens brancos de meia idade.

Em geral, atribui-se o viés integralmente às bases de dados tendenciosas, quando o viés, como mencionado anteriormente, pode emergir antes da coleta de dados em função das decisões tomadas pelos desenvolvedores (os atributos e variáveis contemplados no modelo, inclusive, determinam a seleção dos dados). No caso de viés associado aos dados, existem duas origens: os dados coletados não representam a composição proporcional do universo objeto em questão, ou os dados refletem os preconceitos existentes na sociedade. Resultados enviesados podem decorrer, igualmente, de erros na rotulagem (*label*) da base de dados que antecede o aprendizado supervisionado e na própria geração de dados (a não desagregação por gênero, citada na introdução). A constatação do viés em um modelo, em geral, ocorre tardiamente, dificultando identificar sua origem (HAO, 2019).

2.2. Não explicabilidade (opacidade, *black-box*)

As DLNNs já têm intrinsecamente a variável incerteza dos modelos estatísticos de probabilidade; produzem conhecimento provável, mas inevitavelmente incerto, o que Floridi *et al.* (2016) denominam de "evidência inconclusiva". Os especialistas estão empenhados em reduzir o *"black box"*, ou seja, serem capazes de explicar como o sistema apurou determinado resultado e, preferencialmente, de forma acessível ao usuário. A opacidade decorre do desconhecimento de como os chamados "dados de entrada" (*inputs*) geraram o dado de saída (*output*), como o sistema correlacionou as variáveis contidas nos dados de entrada e os pesos atribuídos (denominados de "parâmetros"). As arquiteturas das DLNNs são formadas por várias camadas (*layers*), cada uma delas realiza representações mais abstratas do que na camada anterior, visando a alcançar a abstração requerida pelo *output*. Esse processo de representações cada vez mais abstratas está no cerne do problema da não explicabilidade. Os parâmetros que correlacionam os pixels no reconhecimento de imagem, por exemplo, são definidos pelo próprio sistema, portanto são as variáveis não controláveis (grau de relevância de cada pixel para atingir o objetivo final/*output*).[15] Com maiores recursos computacionais e big data, o número de fatores possíveis a serem incluídos nesses sistemas ultrapassa o nível de compreensão de um ser humano racional: incompatibilidade entre a otimização matemática na característica de alta dimensionalidade e o raciocínio e interpretação semântica em escala humana. Os especialistas em IA denominam essa opacidade de problema de "interpretabilidade". Existe uma tensão entre a necessidade de explicação e a acurácia dos resultados, um *tradeoff* entre precisão e transparência: quanto maior a precisão, menor a transparência (VILLANI, 2018).

15 Fonte: GEIGER, Davi. Computer Science do Courant Institute/NYU, comunicação pessoal direta com a autora na qualidade de seu mentor em assuntos relacionados às tecnologias de IA.

3. Propostas sugeridas *versus* limitações práticas e/ou conceituais

O debate sobre IA ética mobiliza os filósofos por ser a ética o ramo da filosofia dedicado aos assuntos morais, como o italiano Luciano Floridi, professor na Universidade de Oxford, e o belga Mark Coeckelbergh, professor da Universidade de Viena. Do campo das ciências exatas, destaca-se o matemático francês Cetric Villani, agraciado em 2010 com a Medalha Fields e desde 2017 deputado.

3.1. Auditoria dos sistemas de decisão autônomos

A convite do governo francês, Villani (2018) chefiou uma força-tarefa sobre a estratégia de IA para a França e a Europa. Um dos pontos defendidos pelo autor é a criação de mecanismos de auditoria dos sistemas de IA, com a função exercida por especialistas, preenchendo a lacuna entre os princípios gerais e a prática ética. Essa ideia é igualmente defendida por Floridi que, em diversas publicações, argumenta sobre os benefícios de um órgão dessa natureza (FLORIDI *et al.*, 2016; FLORIDI *et al.*, 2018; FLORIDI; COWLS, 2019; MOKANDER; FLORIDI, 2021). A missão da auditoria seria avaliar a consistência dos modelos relativamente aos princípios e às normas vigentes, com foco na revisão dos códigos-fonte e nos impactos das "saídas" dos algoritmos (previsões indicadas pelos modelos). Dentre suas funções, os autores incluem: fornecer suporte à tomada de decisão, visualizar e monitorar os resultados; informar os usuários por que uma decisão foi tomada e como contestá-la; aliviar o sofrimento humano antecipando e mitigando os danos; alocar responsabilidades; e equilibrar os conflitos de interesse.

Mokander e Floridi (2021) sugerem que a tarefa poderia caber a um órgão governamental, a um contratado terceirizado ou a uma função especialmente designada em organizações multilaterais. Mesmo defendendo a ideia da auditoria, os autores apontam restrições conceituais, técnicas, econômicas, sociais, organizacionais e institucionais (Quadro 1).

Quadro 1 – Restrições à auditoria como mecanismo para garantir IA confiável

Tipo	Restrições
Conceitual	Há uma falta de consenso em torno dos princípios éticos gerais. Valores normativos entram em conflito e exigem compensações. Difícil quantificar as externalidades de sistemas complexos de IA. Infalibilidade da informação perdida em meio a explicações reducionistas.
Técnica	Sistemas de IA podem ser opacos e difíceis de interpretar. Integridade e privacidade dos dados são expostas a riscos durante auditorias. Mecanismos de conformidade linear são incompatíveis com o desenvolvimento ágil de software. Testes podem não ser indicativos do comportamento dos sistemas de IA no ambiente do mundo real

Econômico e Social	Auditorias podem prejudicar ou sobrecarregar desproporcionalmente setores/grupos específicos. Garantir o alinhamento ético deve ser equilibrado com incentivos à inovação. Auditoria baseada na ética é vulnerável ao comportamento adversário. Efeitos transformadores da IA apresentam desafios sobre como acionar auditorias. Auditoria baseada na ética pode refletir e reforçar as estruturas de poder existentes.
Organizacional e institucional	Falta clareza institucional sobre quem audita quem. Auditores podem não ter o acesso às informações necessárias para avaliar os sistemas de IA. Natureza global dos sistemas de IA desafia as jurisdições nacionais.

Fonte: (MOKANDER; FLORIDI, 2021).

Complementando as restrições e/ou desafios indicados por Mokander e Floridi (2021), ressalta-se: (a) a agregação de novos dados nos sistemas baseados em DLNNs, como mencionado anteriormente, implica retreinamento dos algoritmos, gerando a necessidade de auditoria contínua; (b) a velocidade e a descentralização no desenvolvimento de novos modelos/algoritmos de IA dificultam replicar o arcabouço regulatório, por exemplo, da indústria farmacêutica (concentrada em poucos produtores, mais fácil de monitorar/fiscalizar); (c) os algoritmos de IA são, em geral, proprietários, ou seja, são protegidos por sigilo comercial; e (d) as tecnologias de IA são sofisticadas, demandando conhecimento sofisticado que, em geral, escapam aos reguladores/legisladores.

3.2. Ética *by design*

Villani (2018) defende, com ênfase, a premência de incorporar a ética à formação e ao treinamento de desenvolvedores (engenheiros, cientistas da computação, especialistas em IA), propiciando contemplar, no processo de elaboração dos modelos, os impactos éticos e socioeconômicos, no que se convencionou denominar "ética *by design*".

Sensível à pressão da sociedade, em 2019, o MIT (Massachusetts Institute of Technology) promoveu uma aproximação entre os cursos de ciências exatas e ciências humanas, incentivando os primeiros a difundir os fundamentos da computação e da IA em todo o *campus*, e incentivando que o futuro da computação e da IA seja moldado por princípios éticos agregados pelas áreas de humanas. A reitora da Escola de Humanidades, Artes e Ciências Sociais, Melissa Nobles, convidou professores de todas as cinco escolas do MIT para oferecer perspectivas sobre as dimensões sociais e éticas das tecnologias emergentes.[16] Iniciativas como essa são bem-vindas, contudo, trata-se de um processo longo, de mudança cultural e estrutural num modelo de universidade historicamente compartimen-

16 Disponível em: https://news.mit.edu/2019/ethics-computing-and-ai-perspectives-mit-0318. Acesso em: 10 mar. 2021.

talizado. A especialização foi a resposta à explosão de conhecimento (expansão e fragmentação) e à difusão do ensino superior na Europa e nos EUA a partir do século XIX; como argumenta Peter Burke (2020), "pode-se considerar a especialização como uma espécie de mecanismo de defesa, um dique contra o dilúvio de informação" (p. 204). A interdisciplinaridade surge para enfrentar as limitações da especialização, representando a migração do polímata individual (domínio de várias disciplinas) para o "polímata coletivo".

3.3. Agenciamento da IA: atribuição de responsabilidade

Coeckelbergh (2019; 2020) aborda o problema da explicabilidade e transparência dos sistemas de IA pela perspectiva de atribuição de responsabilidade: a responsabilidade dos agentes (usuários da tecnologia) decorre da expectativa dos destinatários (outro lado da relação), de que os mesmos saibam explicar as razões da decisão; por exemplo, na área de saúde o pressuposto é que o médico controle o procedimento e seja capaz de explicá-lo ao paciente (responsabilidade tratada como prestação de contas). Para o filósofo, as decisões e ações humanas precisam ser explicadas, logo a explicabilidade é importante por dois motivos: (a) para agir com responsabilidade, o agente precisa saber o que está fazendo, dar razões para sua ação e (b) para explicar as razões aos afetados pela ação ("pacientes"), que podem, e devem, exigir e merecer respostas sobre o que e como foi decidido o procedimento.

Argumentando que uma das formas de enfrentar a necessidade de explicabilidade seria por meio de medidas legais, Coeckelbergh faz uma analogia com a GDPR (lei de proteção de dados europeia) que prevê o direito de explicação ao usuário sobre o uso de seus dados. Num segundo argumento, o filósofo pondera que "se o objetivo não é fazer com que as máquinas expliquem, mas sim exigir isso de seres humanos que são capazes de explicar coisas a outros seres humanos, então há uma chance de que a IA explicável possa funcionar" (COECKELBERGH, 2019), supondo que esses humanos fossem apoiados por sistemas técnicos suficientemente transparentes.

A colocação de que a "explicabilidade" remete à demanda dos afetados é legítima, contudo, conflita com as limitações intrínsecas às tecnologias de IA: (a) a analogia com à GDPR fica comprometida pelo nível de complexidade das DLNNs; e (b) o fato de se tratar de explicações de humanos para humanos não atenua o problema da explicabilidade, a opacidade da técnica transcende a capacidade cognitiva dos humanos.

4. Conclusão

A IA representa um reservatório crescente de "agência inteligente", com potencial de multiplicar a agência humana, ou seja, tornar a ação humana melhor e mais rápida (FLORIDI *et al.*, 2018). É fundamental preservar, pelo menos parcialmente, a supervisão e as escolhas humanas para assegurar o monitoramento do desempenho dos sistemas, prevenir e/ou corrigir os danos, inclusive garantir que, em qualquer etapa do processo, a responsabilidade seja atribuída à um ser humano por meio de um procedimento predeterminado (VILLANI, 2018).

Os sistemas baseados nas DLNNs, pelas limitações abordadas no artigo, de-

vem ser encarados como "parceiros" da inteligência especializada humana, ou seja, as previsões automatizadas são subsídios para a tomada de decisão (e não soberanas). A IA pode ser uma oportunidade de ampliar o grau de transparência na sociedade e de oferecer oportunidades de decisões mais justas e mais objetivas, sem viés humano (considerando que eliminar os preconceitos humanos não parece tarefa fácil). Para que ambas as apostas se concretizem é imprescindível enfrentar as externalidades negativas com equipes multidisciplinares de desenvolvedores, favorecendo a prática da ética *by design*, e conscientizando os usuários sobre os fundamentos, a lógica e o funcionamento da IA.

Referências bibliográficas

ALPAYDIN, Ethem. **Machine Learning**. Cambridge, MA: MIT Press, 2016.

BURKE, Peter. **O Polímata**: Uma história cultural de Leonardo da Vinci a Susan Sontag. São Paulo: Editora Unesp, 2020.

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

COECKELBERGH, Mark. **AI Ethics**. Cambridge, MA: MIT Press, 2020.

DOMINGOS, Pedro. **The Master Algorithm**: How the Quest for the Ultimate Learning Machine will Remake our World. New York: Basic Books, 2015.

FLORIDI, Luciano *et al.* **The ethics of algorithms**: Mapping the debate, 2016. Disponível em: https://journals.sagepub.com/doi/full/10.1177/2053951716679679. Acesso em: 4 abr. 2021.

FLORIDI, Luciano *et al.* – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. **Minds and Machines**, v. 28, p. 689-707, 2018. Disponível em: https://link.springer.com/article/10.1007/s11023-018-9482-5. Acesso em: 4 abr. 2021.

FLORIDI, Luciano; COWLS, Josh. A United Framework of Five Principles for AI in Society. **Harvard Data Science Review**, 2019. Disponível em: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7. Acesso em: 4 abr. 2021.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

HAO, Karen. Intelligent Machines: This is how AI bias really happens - and why it's so hard to fix. **MIT Technology Review**, 2019. Disponível em: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. Acesso em: 7 abr. 2021.

KAUFMAN, Dora. **A inteligência artificial irá suplantar a inteligência humana?** São Paulo: Estação das Letras e Cores, 2019.

LIAO, S. Matthew. **Ethics of Artificial Intelligence**. New York: Oxford University Press, 2020.

MOKANDER, Jacob; FLORIDI, Luciano; Ethics – Based Auditing to Develop Trustworthy AI. **Minds and Machines**, 2021. Disponível em: https://link.springer.com/article/10.1007%2Fs11023-021-09557-8. Acesso em: 4 abr. 2021.

RUSSELL, S. J.; Nerving, P. **Artificial Intelligence**: A Modern Approach. 3. ed. New Jersey: Prentice Hall, 2009.

PEREZ-CRIADO, Caroline. **Invisible Women**: Data Bias in a World Designed for Men. US: Abrams Press, 2021.

VILLANI, Cédric. **For a Meaningful Artificial Intelligence**: Towards a French and European Strategy. ariforhumanity.fr, 2018. Disponível em: https://www.ai4eu.eu/news/meaningful-artificial-intelligencetowards-french-artificial-and-european-strategy. Acesso em: 10 abr. 2021.

Data do recebimento: 31/03/2021
Data do aceite:15/05/2021

Dados da autora:

Dora Kaufman
Lattes: http://lattes.cnpq.br/8045171889826285.
Professora – pesquisadora do Programa de Tecnologias da Inteligência e Design Digital – TIDD da Faculdade de Ciências Exatas e Tecnologia da PUC São Paulo.