A TRANSPARÊNCIA NO CENTRO DA CONSTRUÇÃO DE UMA IA ÉTICA

http://dx.doi.org/10.25091/s01013300202000020008

Glauco Arbix*

RESUMO

A Inteligência Artificial (IA) se configura como uma constelação de tecnologias capaz de gerar outras tecnologias, novas metodologias e aplicações e, por isso, suas características são de natureza distinta da de outras inovações. Seu desenvolvimento propõe questões relacionadas à ética de seus algoritmos, o que envolve processos de decisão nem sempre pautados pela transparência, contrastando por vezes com direitos de indivíduos e valores das sociedades.

PALAVRAS-CHAVE: *Inteligência Artificial; ética; transparência;* Black Box*; IA responsável*

Transparency at the Heart of Building an Ethical AI

ABSTRACT

AI is a constellation of technologies able to generate other technologies, new methodologies and applications. Its characteristics are different in nature from other innovations that emerge in society. This essay approaches AI as the brain and engine of the current technological wave. And it addresses issues related to the ethics of algorithms, whose decision processes are not always transparent, in contrast to the rights of individuals and the values of societies.

KEYWORDS: *Artificial Intelligence; ethics; transparency; Black Box; responsible AI*

[*] Universidade de São Paulo, São Paulo, SP, Brasil. E-mail: garbix@usp.br

When we program morality into robots, are we doomed to disappoint them with our very human ethical inconsistency?
Ian McEwan

Desde os primeiros anos do século XXI, o mundo acompanha o crescimento acelerado de um conjunto de tecnologias chamadas de Inteligência Artificial (IA). Seu peso e influência é maior a cada dia e seus impactos já podem ser sentidos em praticamente todas as esferas da vida econômica e social.

Hoje em dia, a IA anima cotidianamente bilhões de *smartphones* e determina o ritmo das redes sociais. Como em uma procissão de milagres, a IA parece dar vida a assistentes virtuais que satisfazem nossa vontade, amenizam a babel linguística do mundo com traduções instantâneas e oferecem sugestões de leitura, filmes, vídeos e músicas. Pela eficiência, ganharam a naturalidade dos rituais diários que marcam a vida cotidiana. Criada pela engenhosidade humana, a IA modifica hábitos pessoais e coletivos que agitam o universo da diversão, das artes e do trabalho.

Graças ao poder dos processadores e da análise de imensos volumes de dados, a IA consegue identificar padrões e avançar na previsão de eventos diversos — tanto nas cidades como no campo — nos serviços, na indústria, no comércio e na agricultura. Assim, neste final da segunda década do século XXI, a IA exibe potencial de definir parâmetros inovadores para a remodelagem das cidades, para a mobilidade e a vida urbana, para a busca de fontes limpas de energia e para o crescimento sustentável de economias de baixo carbono. De um modo mais simples, o potencial oferecido pelo ciclo tecnológico atual sugere o desenho de um futuro com mais qualidade de vida para populações ao redor do mundo.

A novidade é que esses avanços têm implicações profundas para a economia, para a elevação da produtividade, para o emprego e o desenvolvimento dos países. Mais ainda, a IA tem potencial para comandar os processos de inovação e remodelar toda a Pesquisa e Desenvolvimento (P&D) nas empresas. Ou seja, distinta do passado, a IA que se desenvolve hoje exibe características semelhantes às que marcaram a computação digital, a eletricidade e a máquina a vapor, que reviraram o modo de produzir, consumir, comercializar e viver (Agrawal; Gans; Goldfarb, 2018, p. 5).

Por isso mesmo, os estudos mais recentes tendem a tratar a IA como uma constelação de tecnologias de propósito geral (TPG), por causa de quatro características básicas: (i) encontram aplicação em praticamente todos os setores da economia; (ii) geram mais inovações e melhoram o desempenho nas áreas em que são aplicadas; (iii) potencializam a P&D; e (iv) credenciam-se cada vez mais como ferramentas básicas para novas descobertas, invenções e inovações (Cockburn; Henderson; Stern, 2018).

Em outras palavras, a IA de hoje se configura como um agrupamento de tecnologias capaz de gerar outras tecnologias, novas metodologias e aplicações e, por isso mesmo, suas características são de natureza distinta da de outras inovações que chegam ao mercado. Seu impacto no crescimento da economia e na melhoria da vida social é potencialmente maior do que o impacto de outras tecnologias — e é isso que justifica a atenção especial à IA e que fundamentou este artigo.

Tecnologias, no entanto, não operam no vácuo, nem determinam a história de povos e países apenas com o desdobrar de sua própria lógica. São, por natureza, artefatos sociais, criados por agentes morais submetidos às tensões da política, das contradições econômicas, das imposições de poder, das desigualdades, dos vieses cognitivos, dos hábitos arraigados e de todos os constrangimentos que alicerçam as sociedades modernas. Esse reconhecimento não diminui o peso específico da tecnologia, pelo contrário, identificar seus traços humanos ajuda a visualizar seus limites, ainda que nem sempre eles estejam ao alcance de nossa compreensão.

Este artigo apresenta um breve relato das características da IA hoje, cérebro e motor da onda tecnológica atual, e discute questões relacionadas à ética dos algoritmos, cujos processos de decisão nem sempre se pautam pela transparência, em contraste com direitos de indivíduos e valores das sociedades. No fim deste texto são sugeridas também algumas referências para a construção de um arcabouço de proteção da sociedade sem, no entanto, inibir a criatividade científica.

Como pequeno alerta, é preciso dizer que a IA não será tratada aqui como um novo Frankenstein que, sem amarras, passou a assombrar seus criadores. Tampouco será abordada como mais uma tecnologia, semelhante a tantas outras que ajudaram a tecer a história da humanidade. Diferentemente, a IA é apresentada como uma poderosa força transformadora; seu protagonismo fez aflorar problemas inéditos no campo da ética, cujo equacionamento está longe de ser fácil: as soluções não obedecem a receitas prontas ou construídas apenas como extensões do passado.

As preocupações éticas procuram balizar a IA e garantir que seu curso esteja sempre voltado para melhorar a sociedade, e não para exacerbar seus desequilíbrios, preconceitos, desigualdades ou até mesmo corroer sua democracia (Eubanks, 2018, p. 14).

IA E SEUS DESEQUILÍBRIOS

São grandes as dificuldades para avaliar o impacto ético dos algoritmos, ainda mais quando eles são capazes de aprender, exatamente os mais promissores. A identificação dos traços de subjetividade nos parâmetros de aprendizagem e no tratamento dos dados nesses ordenamentos matemáticos não é simples, o que significa que nem sempre é possível conhecer *how and why decisions are made due to their capacity to tweak operational parameters and decision-making rules 'in the wild'* (Burrell, 2016). Há, portanto, uma distância grande a separar, de um lado, o *design* e a operação de um algoritmo e, de outro, a visualização plena de suas consequências éticas, o que pode acarretar sérios danos a indivíduos, grupos ou setores da sociedade (Mittelstadt et al., 2016).

As implicações éticas da IA percorrem o mundo da pesquisa, de empresas, de governos e de todas as instituições globais e nacionais preocupadas com o bem-estar das pessoas. Episódios de discriminação e quebra de privacidade promovidos por algoritmos enviesados surgiram com força em épocas recentes e inundaram o ano de 2019 com as mais diferentes preocupações éticas (Luccioni; Bengio, 2019).

Nada mais justificável quando os indivíduos e as sociedades se encontram sem instrumentos e mecanismos claros de proteção diante de sistemas automatizados, que operam de modo opaco e mostram-se arredios até mesmo quando solicitados a fornecer informações básicas sobre seus critérios de decisão. Expostos a falhas, empresas, governos e pesquisadores passaram a ser instados a tratar dos impactos sociais da IA e a explicar que muitos dados que alimentam os algoritmos têm *bias*, que os algoritmos falham e que em processos de alta complexidade nem mesmo seus criadores conseguem compreender totalmente como as conclusões são construídas.

Questões éticas afloram em situações desse tipo, como as que resultam em discriminação contra mulheres, negros e pobres — o que expõe imprecisões e desequilíbrios. As circunstâncias são agravadas quando se sabe que ferramentas de IA passaram a frequentar áreas públicas hipersensíveis e sem a adequada supervisão humana, como segurança, defesa e saúde.

De início apresentados como mecanismos objetivos e matematicamente distantes das imprevisíveis emoções dos indivíduos, os algoritmos não reduziram o comportamento tendencioso ou distorcido que caracteriza a ação humana, mas em alguns casos até o amplificaram. Pesquisas apontam que muitos algoritmos oferecem resultados equivocados, seja por causa dos valores escolhidos pelo *designer*, por distorções dos bancos de dados, por falhas na arquitetura ou mesmo por ambiguidades dos sistemas reguladores. Imprecisões e lacunas em normas e sistemas de controle, de auditagem e de interpretação legal prolongam a permanência de sistemas inadequados, o que não raramente provoca danos à sociedade.

Quando o alvo são os algoritmos de alta complexidade, que constituem o foco de interesse deste texto, é importante estabelecer que são "construções matemáticas com uma estrutura de controle finita, eficaz e efetiva, que cumpre seus objetivos a partir de determinadas orientações definidas pelo programador" (Hill, 2016). Algoritmos assim concebidos processam dados para resolver problemas, oferecer resultados, interpretar a realidade, prever e agir. Graças à sua capacidade de aprendizagem, atuam com eficiência e certo grau de

autonomia, o que provoca o surgimento de problemas éticos complexos e muitos imprevistos que exigem ainda muita pesquisa para serem solucionados. Se é que o serão.

UMA CONSTELAÇÃO DE TECNOLOGIAS

É preciso ampliar e aprofundar um pouco mais os termos da discussão. O conjunto de tecnologias que responde pelo nome de IA não conta com uma definição consensual. Trata-se de um conceito de várias faces, que se transformou ao longo do tempo, e por isso mesmo sua definição é polêmica e diversificada.

A dificuldade de cravar uma definição está mais ligada ao conceito de *inteligência* do que ao termo *artificial*. E os obstáculos principais à construção de uma IA ética não "se devem à natureza mais ou menos inteligente da tecnologia, mas à natureza própria da ética, que a ação humana cria e recria no decorrer do tempo". Por isso mesmo, "o lugar dos sistemas de IA nas sociedades é matéria para uma ética normativa, não descritiva" (Bryson, 2018, p. 15). É o que permite avançar recomendações para o tratamento ético da IA.

Nomear o que é IA e realçar suas características é fundamental para formatar as recomendações que pretendem proteger indivíduos e sociedades. Nesse sentido, a afirmação de que só os humanos são inteligentes eleva o nível de tensão que o debate naturalmente acumula.

John McCarthy, que cunhou o termo IA nos anos 1950, afirmou que uma definição sólida de inteligência só poderia estar relacionada à inteligência humana, porque seria difícil "caracterizar que tipo de procedimento computacional queremos chamar de inteligente" (McCarthy, 2007, p. 5). A psicologia e a neurociência mostraram que os conceitos de inteligência também são fartos e variados. Muitos surgem ligados a consciência, autoconsciência, uso da linguagem, aprendizagem e raciocínio, para citar algumas características humanas nem sempre fáceis de conceituar e plenas de ambiguidades. É essa sobreposição e esse cruzamento de dificuldades que pesam tanto quando se procura definir a IA e a ética dos algoritmos.

Apenas como exemplo de partida, Stuart Russell e Peter Norvig (2010, pp. 27-8), autores de um dos livros mais citados em cursos de ciência da computação, registram oito definições de IA, agrupadas em categorias como o pensar e agir de modo humano e o pensar a agir racionalmente. Evidentemente, não oferecem uma conclusão, mas sim referências para a construção de um conceito.

O termo IA sugere que as máquinas podem pensar. Podem mesmo? O que de fato fazem quando resolvem problemas que, em princípio, caberiam somente aos humanos?

Perguntas desse teor, que intrigam gerações de cientistas, foram minimizadas por um dos fundadores da ciência da computação, Alan Turing (1950). Mesmo tendo vivido antes da formulação da expressão "Inteligência Artificial", Turing não via sentido nessas perguntas porque diferenciava o procedimento das máquinas do pensamento humano. Dito por ele, o debate parece simples. Não é.

Esse debate se tornou ainda mais complicado com a evolução dos computadores, que passaram a desenvolver algumas atividades tidas como tipicamente humanas. Uma das visões de IA que marca a produção atual procura aproximar as máquinas dos humanos ao realçar que são capazes de resolver problemas e de perseguir objetivos — duas características do agir racional. Porém, quando as máquinas — mesmo as mais autônomas e capazes de tomar decisões — são confrontadas com a realidade, a noção de intencionalidade, que está no coração da racionalidade humana, mas não no das máquinas, confunde a atividade reflexiva dos cientistas.

Diante dessas dificuldades, é razoável afirmar que o entendimento sobre o que é IA apresenta-se diversificado e, em geral, dependente das circunstâncias que envelopam essas tecnologias.

Grande parte das definições atuais realça as características computacionais da IA, que permitem detectar padrões e indicar soluções a partir de dados. Essa IA tem na sua base processos chamados de aprendizagem de máquina, intensivos em procedimentos sustentados pelas ciências dos dados, e seus algoritmos mais avançados buscaram inspiração, ainda que distante, no funcionamento das redes de neurônios humanos. Essas tecnologias se desenvolveram rapidamente em anos recentes e tornaram-se dominantes na pesquisa acadêmica e nos negócios, ainda que suas origens remontem aos anos 1950.[1]

Para este artigo, a opção foi o uso de uma forma mais flexível e despretensiosa, dada a preocupação com a ética e as limitações do autor. Nesse sentido, a IA será abordada como um sistema interativo, capaz de operar com alguma autonomia e apto à autoaprendizagem. Essa IA é a que se constrói no âmbito das ciências da computação e que se dedica a fazer máquinas e sistemas complexos atuarem de modo a parecerem ser dotados de inteligência humana (Taddeo; Floridi, 2018b, p. 16).

O avanço recente da computação e dos sistemas foi tão grande que se tornaram capazes de resolver problemas de alta complexidade, de cumprir tarefas com precisão, de prever, decidir e agir como se fossem humanos, habilidades que estão na base de sua aceitação e seu uso por grande parte da humanidade. Sua propagação denota força. Sua opacidade sugere uma fraqueza que ganha sentido quando a não humanidade de suas conclusões, ou seja, quando sua intencionalidade e suas decisões forem identificadas como valores selecionados e implantados pelo *designer* durante sua concepção.

[1] Em 1951, Marvin Minsky, cientista cognitivo norte-americano e um dos fundadores da IA moderna, concebeu uma das primeiras redes neurais para a arquitetura de algoritmos com capacidade de aprendizagem.

Muitos pesquisadores e historiadores optaram por distinguir a IA entre *narrow* ou *weak AI* (estreita ou fraca) e a *general* ou *strong AI* (geral ou forte).[2]

A *narrow AI* foi a que avançou rapidamente no mundo de hoje e responde pelos resultados positivos em várias áreas da sociedade. Algumas técnicas que se encaixam na *narrow AI* são capazes de realizar operações de alta complexidade, mas com um escopo limitado, como a identificação de padrões, a redação de textos, a composição de músicas, a análise de documentos e a elaboração de diagnósticos de algumas doenças com enorme precisão. Seus algoritmos alimentam-se de dados (estruturados ou não) e, dessa forma, "aprendem" ou são "treinados"[3] para realizar tarefas específicas, em faixas predeterminadas e predefinidas. Quanto mais dados, maior seu aprendizado. Por isso essas técnicas foram chamadas de *machine learning* (aprendizagem de máquina, ML) (Corea, 2019; Domingos, 2015).

Apesar da versatilidade e da aplicabilidade em praticamente todos os domínios da economia e da sociedade, a *narrow AI* ainda engatinha quando é confrontada com emoções, pensamento ou autoconsciência, que continuam sendo atribuições tipicamente humanas.[4] Os assistentes de tradução, de voz, de classificação, de seleção e mesmo de decisão que existem nos *smartphones* ou em computadores são expressões de sistemas de *narrow AI*. Mostram-se fortes para algumas tarefas, mas muito fracos se comparados à inteligência humana.

Por sua vez, as pesquisas sobre *strong AI* (forte) têm como foco as máquinas que buscam desenvolver inteligência similar à humana. Seriam aptas a executar tarefas intelectuais, como as exibidas em filmes como *Her*, dirigido por Spike Jonze,[5] e em outras peças de ficção nas quais os humanos interagem com máquinas que exibem consciência, emoção e motivação. Pesquisadores dessa área, a exemplo de Nick Bostrom,[6] trabalham com a hipótese de que sistemas avançados poderiam projetar, desenvolver e implementar seus próprios códigos de tal forma que teriam condições até mesmo de se desdobrarem em uma *super AI*, como nos estudos de Raymond Kurzweil,[7] cuja inteligência seria superior à dos humanos, tanto em conhecimento, raciocínio, julgamento, como em discernimento, livre-arbítrio e sabedoria. Máquinas desse calibre, segundo esses pensadores, seriam tão poderosas que ameaçariam a própria existência humana.

Pesquisas com esse teor são mais do que polêmicas e sua argumentação básica, assim como a de seus críticos, merece aprofundamento em outros trabalhos. Para este artigo, é suficiente indicar que a humanidade está muito distante desse tipo de IA, ainda que uma perspectiva semelhante a essa tenha estado presente nos primórdios da IA, nos

[2] Há várias distinções e classificações de IA. Neste texto, o objetivo é somente oferecer uma base mínima que permita a compreensão de alguns problemas éticos ligados a esse grupo de tecnologias.

[3] As aspas foram usadas desta vez apenas para diferenciar a atuação das máquinas da atuação humana, que aprende e é treinada. São termos de uso corriqueiro entre pesquisadores.

[4] Apesar das enormes dificuldades, uma das subáreas que mais crescem nas pesquisas em IA recebeu o nome de *sentiment analysis*. O objetivo é identificar e extrair sentimentos em textos e imagens de fontes como mídias sociais, blogs, comentários sobre produtos, fotos e filmes. Como é de esperar, é enorme o desafio de capturar sinais de ironia, sarcasmo, alegria, tristeza e uma série de outros sentimentos presentes no dia a dia da atividade humana. Como breve referência, ver: Feldman (2013) e Liu (2012).

[5] Filme de 2014 em que um escritor solitário se apaixona pelo assistente virtual de seu computador, que atende pelo nome de Samantha.

[6] Filósofo da Universidade de Oxford (Reino Unido), dirigente do Future of Humanity Institute.

[7] Formado pelo Massachusetts Institute of Technology (MIT), é diretor de engenharia do Google, cofundador da Singularity University. De modo distinto de Bostrom, Kurzweil pensa que as máquinas evoluirão por simbiose com humanos.

debates no Dartmouth Summer Research Project on Artificial Intelligence (EUA),[8] em 1956, quando a expressão foi cunhada. De fato, a busca por uma IA começou a ganhar corpo durante a Segunda Guerra Mundial, com os trabalhos de Alan Turing, e elevaram seu estatuto na década de 1950. Sua trajetória, porém, esteve longe de uma ascensão linear. Viveu oscilações fortes, com surtos positivos e retração de investimento, tanto financeiro como humano, sempre ligados aos resultados prometidos e nem sempre realizados (Nilsson, 2010, pp. 408-10).

Jordan (2019) refere-se a essa fase como a busca de uma *human-imitative AI*. Nessa época, como iniciativa acadêmica, a IA pretendia capacitar-se para desenvolver o raciocínio, o pensamento e a cognição típica dos humanos. Sua sintonia e convergência com disciplinas correlatas terminariam por impulsionar muitos dos avanços que permitiram o salto da IA (Candès; Duchi; Sabatti, 2019).

A partir de 2010, os algoritmos de *machine learning* (ML, subárea da ciência da computação) e os de *deep learning* (DL, subárea da ML) receberam impulso de novos avanços, em *hardware* e *software*. Sinteticamente, as mudanças no ambiente de IA estiveram vinculadas: (i) ao aumento rápido e contínuo dos bancos de dados de fala e imagem nos últimos dez anos, basicamente em decorrência da proliferação de *smartphones* e de uma gama de navegadores (como Chrome, Edge, Explorer, Firefox e outros) e de aplicativos como WeChat, Skype, WhatsApp; (ii) à ampliação do poder de processamento dos computadores e à consolidação do *cloud computing*, que possibilitou o armazenamento de dados e o treinamento dos novos algoritmos; (iii) a uma verdadeira revolução na ciência de dados, que ampliou o campo da estatística e viabilizou os tradutores da Google e os mecanismos de *touch* ID e de reconhecimento de voz, por exemplo (Donoho, 2017).

As técnicas de DL que se apoiam em redes neurais estiveram na base de um enorme avanço da IA. Por similaridade, tentam aproximar-se do que se imagina ser o funcionamento dos neurônios humanos, o que, como se sabe, permanece um campo ainda nebuloso para a ciência. Os processos de redes neurais foram (e continuam sendo) alvo de muitas críticas, basicamente porque não conseguem explicitar os motivos que levaram às suas previsões. Por isso, muitos especialistas caracterizaram os algoritmos de DL que operam com redes neurais como *black boxes*, por serem tão opacos quanto o funcionamento do cérebro (Castelvecchi, 2016).

A força das redes neurais decorre de sua capacidade de aprendizagem. A partir de um conjunto de dados disponíveis para seu treinamento, as redes podem melhorar progressivamente seu desempenho, aperfeiçoando a força de cada conexão até que seus resultados também sejam corrigidos. Esse processo tenta simular

[8] Em 1956, um pequeno grupo de cientistas se reuniu no Dartmouth College (Hanover, Nova Hampshire, EUA) para discutir o que muitos chamavam de *automata theory* ou *thinking machines*. Liderados por John McCarthy, Marvin Minsky e Claude Shanon, a expressão "Inteligência Artificial" foi criada para o seminário que durou pouco mais de um mês e envolveu cerca de 12 cientistas.

como o cérebro humano aprende, fortalecendo ou enfraquecendo suas sinapses, e seu funcionamento gera uma rede apta a classificar com sucesso novos dados que não faziam parte do conjunto inicial de seu treinamento.

Apesar dos temores provocados pela não transparência, muitos cientistas da computação afirmam que os esforços para criar uma IA transparente são complementares ao aperfeiçoamento das redes neurais, não seu substituto. Simplesmente por razões de eficácia e precisão de seus resultados, como mostram os impactos positivos nas áreas de saúde, educação, meio ambiente, energia e no conjunto da economia. A dose de autonomia e as dificuldades de refazer, por engenharia reversa, os caminhos percorridos pelas redes neurais incomodam na utilização dessas técnicas e desafiam a ciência a quebrar sua opacidade.

Enquanto essas respostas não chegam, o risco e a incerteza que envolvem seu desempenho continuam gerando em todas as sociedades questões éticas importantes, porque as técnicas de DL desenvolvem-se como parte integrante de um ambiente mais amplo, muitas vezes chamado de sistemas sociotécnicos, compostos de instituições, organizações e pessoas que atuam nos mais distintos domínios, que vão dos desenvolvedores aos fabricantes, dos usuários aos gestores públicos.

Isso significa que as referências, os princípios, os protocolos e os códigos voltados para garantir a ética e a responsabilidade não podem ter os algoritmos como seu alvo exclusivo. São os componentes sociais que devem ser o alvo prioritário das recomendações éticas para que a IA seja confiável. Em outras palavras, o tratamento ético terá sentido apenas se ensejar um comportamento responsável, transparente e *accountable* de pessoas e instituições que produzem e reproduzem a ML, o que está muito vinculado ao tipo de técnica que formata o algoritmo. Por exemplo, técnicas baseadas em Árvores de Decisão ou Redes Bayesianas são muito mais transparentes e rastreáveis do que as de alta complexidade, como as Redes Neurais ou os Algoritmos Genéticos. Como vimos, algoritmos com essa complexidade mostram-se impenetráveis ao escrutínio humano, o que aumenta as preocupações com o que é aceitável (ou não) como padrão ético.

ILUMINAR A CAIXA-PRETA

Com base no que foi exposto, é possível afirmar que a maior parte das aplicações que atualmente oferece resultados positivos está dentro da valise ML-DL. Não são, portanto, sistemas ou máquinas que raciocinam ou que dispõem de consciência. A recomendação de uma dieta alimentar por um algoritmo de DL não tem o mesmo sentido daquela oferecida por um médico que realizou estudos comparativos

sobre os níveis de insulina e de açúcar e que conhece seu paciente e suas circunstâncias.[9] Para identificar um cão, por exemplo, um sistema de ML precisa do apoio de milhares de imagens e fotos, assim como de um *hardware* poderoso para processá-las. Não se trata de raciocínio, mas de uma sofisticada operação estatística voltada para identificar padrões e decidir que esses padrões representam um cão (Pasquale, 2015; Robbins, 2019).

[9] Exemplo adaptado de Topol (2019).

Como formulou Jordan (2019, p. 7), a IA de hoje se assemelha a uma "inteligência reciclada", não a uma verdadeira inteligência, e por isso é grande o risco de confiar nas máquinas, dado que não raramente fornecem respostas equivocadas. Na realidade, sistemas autônomos de ML e DL operam eticamente com reduzida previsibilidade, seja porque não foram concebidos ou não são adequados para envolver representações de raciocínio moral, seja porque os valores desses sistemas não foram devidamente sintonizados com os padrões éticos que regem as respectivas sociedades. Por isso, mostra-se inconsistente a visão de vários pesquisadores de que o aumento da autonomia dos sistemas isentaria os *designers* da sua responsabilidade. O movimento, de fato, ocorre no sentido inverso, pois, quanto maior a autonomia dos algoritmos, maior será a responsabilidade dos seus criadores. É o que toda sociedade espera para consolidar sua confiança nessas tecnologias.

É possível avançar um pouco mais para reconhecer que as técnicas de DL, ao serem alimentadas por dados, conseguem recriar novos padrões, aptos a reconhecer a representação de novos cães, por exemplo, sem a interferência do *designer*. Trata-se de um processo que gera modelos que podem ser utilizados para identificar padrões em *inputs* futuros. O conceito de DL, assim, é construído com base em sua capacidade de definir e de modificar as regras de tomada de decisão de forma autônoma. O trabalho do algoritmo de DL de incorporar novos *inputs* nos modelos pode, dessa forma, interferir nos sistemas de classificação originalmente criados. Esses *inputs* podem ser rotulados previamente (supervisionados por humanos) ou podem ser definidos pelo próprio algoritmo, ao operarem sem supervisão (Van Otterlo, 2013, pp. 60-4).

Para este artigo, interessa-nos realçar que nas duas modalidades — o aprendizado supervisionado e o não supervisionado — o algoritmo define as regras que manuseiam os novos *inputs*. Ou seja, quando alimentados e treinados por novos dados, os algoritmos realizam operações de processamento e classificação de modo automático, sem participação do operador, o que sugere uma lógica que não é transparente em todos os seus procedimentos. Essa não transparência levou a DL ao debate sobre a noção de *black box*. E, por causa desse procedimento, as noções de transparência e de explicabilidade consolidaram-se como preocupações éticas essenciais da IA atual.

A nuvem de incerteza presente nas operações de DL dificulta a identificação e a correção de desafios éticos, seja no *design*, seja na operação de algoritmos (Mittelstadt et al., 2016). A demanda por transparência e por explicação dos resultados ganhou, assim, uma dimensão muito superior à vaga ideia de clareza da informação, para se posicionar no coração das relações entre humanos e os processos de DL.

LIMITES

Não é suficiente, porém, reconhecer que a falta de transparência marca a DL. É certo que há problemas de viés ligados à seleção e ao preparo dos dados. Mas o processo aqui realçado é outro e manifesta-se já nos primeiros passos da criação de um algoritmo, quando os programadores encontram dificuldades para definir seu alcance.

Por exemplo, no setor financeiro, com a propagação de aplicativos de análise e liberação de crédito, as operações enfrentam obstáculos ao fixar o conceito de *credibilidade* para a liberação (ou não) dos créditos solicitados. Além de escorregadio, esse parâmetro ainda pode combinar-se com outros critérios, como a margem de lucro esperada, a taxa de risco aceitável ou o número de parcelas economicamente viável. Ou seja, referências de mercado, comerciais e dados pessoais (como idade, renda, gênero ou grau de escolaridade) também podem influenciar o resultado dos algoritmos, inclusive com a transposição da discriminação inerente aos dados, por causa da transferência de preconceitos existentes na sociedade. O alerta é trivial:

> *A seleção dos dados usados para construir os modelos — dados de treinamento — é uma fonte importante de potencial viés. Amostras não representativas da população geralmente levam os modelos a errar sistematicamente. Esses preconceitos de amostragem nem sempre são percebidos e, muitas vezes, impossíveis de ser totalmente reconhecidos. Mais ainda, métodos padronizados de validação, que dependem de dados extraídos da mesma amostra, também apresentarão falhas. Nem mesmo amostras representativas — como bancos que abrangem o conjunto da população alvo — conseguem garantir que os modelos tenham um desempenho igualmente positivo para diferentes segmentos da população*. (Barocas et al., 2017, p. 3)

Como se pode ver, o poder decisório que está nas mãos do programador é grande e nem sempre foi previsto, o que aumenta o grau de subjetividade inscrito no algoritmo e a incerteza sobre seu percurso e seus resultados.

Problema semelhante se apresenta também aos aplicativos na área da saúde. Embora os algoritmos de hoje sejam mais potentes e muito diferentes dos antigos, que davam apenas respostas mecânicas ou

predeterminadas a questões de saúde, a ausência de clareza sobre a escolha dos critérios que guiam os modelos e a difícil interpretação do resultado final são obstáculos à sua difusão e aceitação tanto por médicos como por pacientes, principalmente diante da potencial adoção de terapias invasivas e de alto risco (Penel, 2019).

Se for adicionado a essas dificuldades o reconhecimento de que a precisão dos algoritmos também depende do tipo de metodologia e das técnicas utilizadas, pode-se compreender por que conceitos como *explicabilidade* passaram a posicionar-se, com a transparência, no centro das preocupações de DL. Ainda mais com as pesquisas que indicam que o viés humano reproduzido nos dados pode ser amplificado ao longo do processo de aprendizagem dos algoritmos — o que torna o mundo real ainda mais desequilibrado.

Não foi à toa que pesquisas da Yale Law School, conscientes das distorções da DL, recomendaram com sabedoria: "Não precisamos trazer as desigualdades estruturais do passado para o futuro que estamos criando" (YLS, 2017).

A *explicabilidade*, em contraste com a metáfora da caixa-preta, orienta o funcionamento dos algoritmos para a transparência de seus procedimentos, desde sua concepção até a operação final com o usuário, tornando rastreável o percurso do raciocínio. Com a *auditabilidade*, o usuário ou os agentes públicos podem revisar os processos decisórios dos algoritmos, testá-los e corrigi-los quando necessário.

Esses recursos são fundamentais para bem posicionar os desenvolvedores no debate público sobre as consequências sociais geradas pela IA. Tenderão a perder credibilidade e confiabilidade os algoritmos com alto impacto social que não oferecem informações claras sobre seu funcionamento interno, sobre os padrões que orientam seu processo de aprendizagem e sobre como chegaram aos resultados finais.

Por isso, o desafio da transparência e da *explicabilidade* é enorme e premente, ainda que do ponto de vista científico tenha de enfrentar um dilema flagrante: a mesma complexidade que permite o desenvolvimento da DL com toda sua precisão e capacidade preditiva veda a transparência aos usuários e aos próprios criadores. A comparação com o viés, a confusão e o erro humano podem ser fonte de consolo para os *designers* de DL. Afinal, ao cotejar falhas humanas e dos algoritmos, é possível que em vários domínios a vantagem fique com os sistemas de IA. Contudo, em áreas de sensibilidade elevada, como na saúde, a exigência de supervisão humana é praticamente mandatória — e, mesmo assim, a dúvida é sufocante.

O esforço em equacionar esse dilema acompanha as pesquisas que buscam iluminar a caixa-preta da ML-DL. E inspirou a formulação do European Union General Data Protection Regulation (GDPR, fonte de referência para a legislação brasileira), que consagrou o direito de

cada usuário a pedir informações sobre a lógica envolvida na tomada de decisão por algoritmos — o que foi refletido também na Lei Geral de Proteção de Dados brasileira (art. 20).

As questões de fundo ligadas à *explicabilidade* permaneceram, no entanto, em aberto e suscitam dúvidas tão constantes quanto contundentes: será que o grau de acerto e de precisão da ML e da DL será suficiente para compensar a ausência de transparência desses instrumentos?[10]

[10] Para uma ponderação mais qualificada, ver Holm (2019).

ÉTICA, A NOVA FRONTEIRA

Sem avanços no campo da ética, capazes de iluminar os procedimentos dos algoritmos, a IA, em suas diferentes modalidades, poderá sofrer um processo de desgaste e corrosão da confiança das sociedades.

Essa tensão vem de longe. Ainda nos anos 1960, Norbert Wiener acendia uma luz de alerta e registrava na revista *Science*: "Minha tese é de que as máquinas podem transcender algumas limitações de seus *designers* e, ao fazê-lo, podem ser eficazes e perigosas" (1960, p. 2). O alerta apontava para a responsabilidade de programadores e para a afirmação da autodeterminação humana diante da autonomia das máquinas, em uma época em que a pesquisa ainda se restringia a pequenas comunidades e não recebia as pressões de hoje.

Os problemas éticos, no entanto, agravaram-se ao longo do tempo. Grandes corporações cresceram, criaram e controlaram gigantescos bancos de dados que condicionam as operações e as pesquisas avançadas em IA em um grande oligopólio, tanto nos Estados Unidos como na China. Isso ocorre no Ocidente, mas também no Oriente. Na verdade, um pequeno grupo de países,[11] e entre eles um pequeno grupo de empresas, domina as tecnologias de IA e tem capacidade de expandir suas fronteiras. Poucos, capitalizados e tecnologizados, não hesitam em exibir uma força inédita, que altera hábitos e influencia diretamente a elaboração de sistemas regulatórios. É fato que as autoridades nem sempre deram a atenção devida a essas corporações. Apenas nos últimos anos a gravidade da situação começou a aflorar, mas sempre *post factum* e com alto custo social, a começar pelo desgaste da democracia:

[11] Com destaque para Estados Unidos e China e, atrás deles, Reino Unido, Canadá, Alemanha, Japão, Coreia e Israel.

> *Aplicações que identificam o perfil de pessoas para direcionar a publicidade (serviço oferecido on-line e também em campanhas políticas, como revelado no caso da Cambridge Analytica), são exemplos da capacidade da IA de capturar as preferências e características de usuários e, assim, moldar seus objetivos, influenciar seu comportamento e debilitar sua autodeterminação*. (Taddeo; Floridi, 2018a, p. 2)

O aumento da capacidade de resolução de problemas complexos, com alto grau de acurácia e baixo custo, age como um forte apelo para a disseminação crescente do uso da DL. Os usuários, no entanto, ao transferirem a responsabilidade para sistemas autônomos, encontram-se precariamente protegidos pela legislação em quase todo o mundo — realidade que a GDPR da União Europeia pretende mudar.

A recorrência de sistemas de IA que tomaram decisões e discriminaram negros, hispânicos, pobres e mulheres, para citar alguns segmentos (Danks; London, 2017; O'Neal, 2016; Eubanks, 2018), não permite considerar esses resultados como equívocos menores ou ligados a um suposto custo a ser pago pela marcha da ciência.

Como ignorar que senadores negros nos Estados Unidos, que viviam em Washington, foram apontados como criminosos no sistema de reconhecimento facial da cidade de San Francisco, na Califórnia? Como justificar que sistemas públicos de avaliação de risco, baseados em IA, mantiveram ou liberaram prisioneiros por boa ou má conduta, seja por falta de supervisão humana, seja pelo *design* ou pelos dados (Ferguson, 2017)? Há aplicativos de IA que operam hoje sem a devida curadoria humana (Buranyi, 2018).

É evidente que esses desacertos não são positivos para a construção de uma sociedade mais tolerante e civilizada. E enquanto esses problemas persistirem, seja por falta de transparência, seja pelo uso de dados impregnados de preconceitos, as empresas e seus negócios vão enfrentar momentos de tensão, pois quem busca sistemas eficientes não pode conviver tranquilamente com soluções dessa natureza.

A reflexão ética sobre os algoritmos de ML faz parte do *kit* de sobrevivência da IA como a conhecemos hoje.

Mesmo com essa advertência, empresas de todo porte, mas de modo especial as gigantes de tecnologia, difundem suas soluções sem que tenham alcançado pleno sucesso nos processos de explicabilidade e na aplicação de uma IA ética, transparente e confiável.[12]

É sempre oportuno lembrar que a confiabilidade adere às máquinas apenas como metáfora, pois somente os humanos são confiáveis. Ou não são.

Após uma série de equívocos, San Francisco decretou moratória para o uso público de tecnologias de reconhecimento facial.[13] Hoje são vários os países a externar preocupação com o potencial impacto negativo das *deepfakes* para a democracia, mas a velocidade de sua sofisticação é maior do que a das técnicas de sua detecção.

Anima, porém, saber que o novo ciclo da IA está apenas no início e que o debate sobre ética cresce a olhos vistos. Mas sempre vale o alerta: os mecanismos de contenção da IA conforme os padrões éticos aceitáveis não podem se distanciar das dimensões sociais. Foi essa postura que impulsionou a formação do AI 4 People e que move o IEEE

[12] Mesmo com a multiplicação de iniciativas nessa direção, os problemas permanecem. Ver: <https://cloud.google.com/explainable-ai>; <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability>; <https://venturebeat.com/2019/05/10/microsoft-open-sources-Interpretml-for-explaining-black-box-ai>. Acesso em: 23/06/2020.

[13] San Francisco foi a primeira cidade dos Estados Unidos a proibir o uso, pelo setor público, de sistemas de reconhecimento facial, como na área de transporte e segurança, por exemplo. O volume de erros das tecnologias de IA foi preocupante, em especial quando os alvos eram mulheres, negros ou pobres. O Conselho da cidade afirmou que a tecnologia não é confiável, a legislação é imprecisa e a sociedade ainda não está pronta (*BBC News*, 15/05/2019). Outras cidades americanas aprovaram leis semelhantes às de São Francisco, como Berkeley (CA), Cambridge (MA), Sommerville (MA) e Oakland (CA). Vários estados norte-americanos (como Michigan, Nova York, Califórnia, Oregon e Washington) discutem a proibição do uso de técnicas de reconhecimento facial pelo setor público, até que amadureçam e se mostrem confiáveis (Nesta, 2020).

Global Initiative on Ethics of Autonomous and Intelligent Systems, o Berkman Klein Center, o ETH Zurique, a Partnership on AI.[14] O tipo de abordagem mais ampla e centrada no humano leva os sistemas regulatórios e de governança da IA a ir além das preocupações éticas. Enquanto as leis introduzem mecanismos de pesos e contrapesos e aceitam de alguma forma a participação democrática, a ética da IA pode, na ausência de atenção social, ser decidida por desenvolvedores, por comunidades de cientistas ou pesquisadores em laboratórios ou corporações, nem sempre envolvidos com o debate público (Bryson, 2018). Por isso é fundamental a formulação de padrões e a construção de um marco regulatório adequado.

[14] Disponíveis em: ‹https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html›; ‹digitalethicslab.oii.ox.ac.uk/ai4people›; ‹cyber.harvard.edu›; ‹ethz.ch/en.html›; ‹https://www.partnershiponai.org/partners›. Acesso em: 23/06/2020.

AVANÇAR NO DEBATE PARA ALÉM DE CÓDIGOS E CONSELHOS DE ÉTICA

Avanços importantes nos últimos cinco anos indicam o início de um consenso em torno de cinco princípios de alto nível que ajudam a nortear a construção de uma agenda de pesquisa sobre ética e IA: *transparency, justice and fairness, non-maleficence, responsibility and privacy.*[15]

Mas, embora a produção de guias e códigos de conduta tenha aumentado recentemente, assim como a proliferação de comitês e Conselhos de Ética nas empresas, os problemas avolumam-se a cada dia, o que reforça a necessidade de ir além de formulações abstratas de códigos, princípios e recomendações, principalmente porque, na maior parte dos casos, não há mecanismos de *enforcement* e nem sempre se consegue identificar a natureza real dos problemas.

A sintonia entre valores humanos e os processos de ML exige articulação entre as ações técnicas e o sistema legal-regulatório. É preciso não esquecer que aplicações de ML que apresentam falhas técnicas de *design* diminuem a eficiência de salvaguardas éticas bem estruturadas. E vice-versa.

Por isso mesmo, é necessário avançar na criação de padrões de precisão e de leis com foco mais apurado. Empresas e universidades precisam ampliar seus comitês de especialistas, dotados de legitimidade para acompanhar, avaliar e mesmo interromper projetos que se distanciam das referências éticas anunciadas publicamente. A responsabilidade institucional por pesquisas e desenvolvimento de aplicações deve estar no centro de toda atividade de ML.[16] No mesmo sentido, é preciso que os profissionais que participam de projetos de IA, de desenvolvedores a *policy makers*, sejam capazes de tratar das implicações sociais de suas criações — para isso, precisam ser qualificados e avaliados em sua formação e consistência ética. Essa capacitação é fundamental para que termos como *fairness* ou explicabilidade não se esgotem em si mesmos.

[15] *"Transparency*: is the most prevalent principle in the current literature. References to transparency comprise efforts to increase explainability, interpretability or other acts of communication and disclosure. *Justice* is mainly expressed in terms of fairness, and of prevention, monitoring or mitigation of unwanted bias and discrimination. *Non-maleficence*: harm-prevention guidelines focus primarily on technical measures and governance strategies, ranging from interventions at the level of AI research, design, technology development and/or deployment to lateral and continuous approaches. *Responsibility* and accountability: specific recommendations include acting with integrity and clarifying the attribution of responsibility and legal liability if possible upfront. *Privacy* is seen as a value to uphold and as a right to be protected. Privacy is often presented in relation to data protection and data security" (Jobin; Ienca; Vayena, 2019).

[16] Para a reflexão, ver: https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent.

Esse é o ponto relevante da agenda para uma IA ética (Theodorou; Dignum, 2020), pois os princípios não são imediata e automaticamente aplicáveis, nem há receita pronta para a solução de discordâncias normativas sobre os princípios éticos mais correntes.

Como a busca de uma IA ética não se identifica com um processo de busca de uma solução tecnologicamente ética, os princípios acordados entre instituições e países frequentemente extrapolam definições técnicas. Quando esses conceitos estão imersos na sociedade, definições superficiais esbarram na diversidade, na criatividade e na temporalidade humana. Não foi por acaso que o debate sobre esses termos povoou a trajetória da filosofia e da política ao longo dos séculos. A ingenuidade (ou prepotência) tendem a levar à simplificação desses conceitos, com base na crença de que princípios éticos podem ter sua representação simplificada e fixada em algoritmos.

Do prisma da sociedade, a questão de fundo é que a diversidade é positiva e não deve ser tomada como um obstáculo a ser superado e liquidado. Há escolhas a serem feitas no *design* dos algoritmos. E os pesquisadores de IA não devem ter medo de enfrentar a multiplicidade de interpretações desses conceitos, que ganham mais sentido quando estão sintonizados com os valores e as recomendações éticas de cada sociedade.

Contrariamente ao senso comum, as diretrizes que orientam o desempenho dos algoritmos funcionam como as leis para as atividades humanas. As técnicas de ML e DL têm na sua base concepções que definem ou condicionam a correspondência entre pessoas, comportamentos, instituições e objetos do mundo real (Selena; Kenney, 2019). Por isso, apesar de sua aura, os algoritmos não são neutros na identificação de padrões e nas previsões que fazem a partir do mergulho no mar de dados que os alimentam.

Princípios éticos variam no tempo e no espaço. E a trajetória dos últimos anos mostra que, no mínimo, é tão difícil regular algoritmos quanto seres humanos. Um caminho fértil que vem sendo trilhado, inclusive para tirar do campo da abstração esse debate sobre a ética, é o que procura confrontar as técnicas de ML e DL com benefícios e prejuízos que eventualmente causam nos direitos humanos.

Para avançar nessa direção, é fundamental que as autoridades públicas responsáveis por impulsionar uma IA ética construam mecanismos apropriados para acompanhar, estimular e avaliar sua evolução, a fim de conceber os contornos de um ecossistema próprio, que facilite a articulação institucional, a legislação, os incentivos e a pesquisa. No interior de um ambiente amigável para o desenvolvimento da IA, as políticas públicas teriam grandes ganhos de eficiência. O estabelecimento de um diálogo permanente entre governos, pesquisadores, legisladores, organizações da sociedade

civil, universidades e representantes empresariais ajudaria a qualificar o debate sobre transparência e *accountability* no universo da IA. A *Declaração de Toronto* (2018)[17] tornou-se um dos corpos de referência obrigatória no plano internacional, dada sua consistência e legitimidade. Suas diretrizes permitiriam abordar pelo menos quatro dimensões importantes das atividades de IA, com potencial de nortear a elaboração e a execução de políticas públicas voltadas para:

(i) Criar espaços formais e informais de diálogo, em nível nacional e estadual, para desenhar os mecanismos de governança a serem estabelecidos com a participação de governos, setor privado, academia e sociedade civil.
(ii) Registrar, documentar e debater experiências de práticas éticas saudáveis no campo da IA.
(iii) Avançar na discussão sobre procedimentos e protocolos voltados para mitigar riscos éticos na produção de sistemas de IA.
(iv) Incentivar a formação e qualificação ética de novas gerações de pesquisadores de IA, em todos os níveis educacionais, seja no universo público ou privado.

Um esforço conjunto entre a sociedade civil e os governos pode impulsionar este necessário debate. Esforços nessa direção serão sempre um estímulo para que as atividades de pesquisa avancem na criação de uma IA mais transparente. Ou para que, caso não consigam isso, os pesquisadores persigam os caminhos de superação de suas formas atuais, por mais avançadas que sejam.

[17] *The Toronto Declaration. Protecting the right to equality and non-discrimination in machine learning systems* (2018). Disponível em: https://www.torontodeclaration.org/declaration-text/english.

Recebido para publicação em 10 de março de 2020.
Aprovado para publicação em 13 de junho de 2020.
NOVOS ESTUDOS
CEBRAP
117, mai.–ago. 2020
pp. 395-413

GLAUCO ARBIX [http://orcid.org/0000-0002-7078-4328] é professor titular de sociologia da Universidade de São Paulo (USP). Responsável pela área de IA e Sociedade do Center for Artificial Intelligence (C4AI), coordenador do Observatório da Inovação do Instituto de Estudos Avançados (IEA-USP). Ex-presidente da Financiadora de Estudos e Projetos (Finep) e do Instituto de Pesquisa Econômica Aplicada (Ipea).

REFERÊNCIAS BIBLIOGRÁFICAS

Agrawal, Ajay; Gans, J.; Goldfarb, A. "Economic Policy for Artificial Intelligence". *Working Paper n. 24.690*, National Bureau of Economic Research (NBER), jun. 2018.
Barocas, Solon; Bradley, E.; Honavar, V.; Provost, F. "Big Data, Data Science, and Civil Rights", 2017. arXiv:1706.03102.
Bryson, Joanna. "Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics". *Ethics and Information Technology,* fev. 2018. DOI: 10.1007/s10676-018-9448-6.
Buranyi, Stephen. "Dehumanising, Impenetrable, Frustrating: The Grim Reality of Job Hunting in the Age of AI". *The Guardian*, 04/03/2018.

Burrell, Jenna. "How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms". *Big Data & Society*, 06/01/2016. DOI: 10.1177/2053951715622512.

Candès, Emmanuel; Duchi, John; Sabatti, Chiara. "On AI: The Revolution Hasn't Happened Yet". *Harvard Data Science Review*, v. 1, n. 1, mar. 2019. DOI: 10.1162/99608f92.f06c6e61.

Castelvecchi, Davide. "Can We Open the Black Box of AI?" *Nature*, v. 538, 06/10/2016.

Cockburn, Iain; Henderson, R.; Stern, S. "The Impact of Artificial Intelligence on Innovation". In: Agrawal, Gans; Goldfarb (orgs.). *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press, 2018.

Corea, F. "AI Knowledge Map: How to Classify AI Technologies, a Sketch of a New AI Technology Landscape". *Medium-AI*. Disponível em: https://medium.com/@Francesco__AI/aiknowledge-map-how-to-classify-ai-technologies-6c073b969020. Acesso em: 24/06/2020.

Danks, David; London, A. "Algorithmic Bias in Autonomous Systems". International Joint Conference on Artificial Intelligence (ICJAI), 2017. Disponível em: www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf. Acesso em: 25/06/2020.

Domingos, Pedro. *The Master Algorithm*. Nova York: Basic Books, 2015.

Donoho, David. "50 Years of Data Science". *Journal of Computational and Graphical Statistics*, v. 26, n. 4, 2017. DOI: 10.1080/10618600.2017.1384734.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Nova York: St. Martin's Press, 2018.

Feldman, Ronen. "Techniques and Applications for Sentiment Analysis". *Communications of the ACM*, v. 56, n. 4, pp. 82-9.

Floridi, Luciano. "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical". *Philosophy & Technology*, v. 32, n. 2, 2019. DOI: https://doi.org/10.1007/s13347-019-00354-x.

___; Cowls, Josh. "A Unified Framework of Five Principles for AI in Society". *Harvard Data Science Review*, 1º/07/2019. DOI: 10.1162/99608f92.8cd550d1.

Hill, Robin. "What an Algorithm Is". *Philosophy & Technology*, v. 29, n. 1, 2016.

Holm, Elizabeth. "In Defense of the Black Box". *Science*, v. 364, n. 6435, 05/04/2019.

Jobin, Anna; Ienca, M.; Vayena, E. "Artificial Intelligence: The Global Landscape of Ethics Guidelines". Preprint version, 2019. Disponível em: https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf. Acesso em: 25/06/2020.

Jordan, Michael. "AI: The Revolution Hasn't Happened Yet". *Harvard Data Science Review*, 1º/07/2019. DOI: 10.1162/99608f92.f06c6e61.

Liu, Bing. "Sentiment Analysis and Opinion Mining". *Synthesis Lectures on Human Language Technologies*, v. 5, 2012.

Luccioni, Alexandra; Bengio, Yoshua. "On the Morality of Artificial Intelligence", 2019. Disponível em: https://arxiv.org/abs/1912.11945. Acesso em: 25/06/2020.

McCarthy, John. "What is Artificial Intelligence?". John McCarthy's Homepage, 2007. Disponível em: http://www-formal.stanford.edu/jmc/whatisai.pdf; https://perma.cc/U3RT-Q7JK. Acesso em: 25/06/2020.

Miailhe, Nicolas; Hodes, C. "The Third Age of Artificial Intelligence". *Field Actions Science Reports*, ed. esp. 17, 2017.

Mittelstadt, Daniel; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. "The Ethics of Algorithms: Mapping the Debate". *Big Data & Society*, 2016. Disponível em: https://perma.cc/U3LV-6USL. Acesso em: 25/06/2020.

Nesta. "AI Governance Database". Fev./2020 Disponível em: https://www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database.

Ng, Andrew. "What Artificial Intelligence Can and Can't Do Right Now". *Harvard Business Review*, 09/11/2016.

Nilsson, Nils. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge: Cambridge University Press, 2010.

O'Neal, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Nova York: Crown, 2016.

Penel, Oliver. "X-AI, Black Boxes and Crystal Balls". *Towards Data Science: Medium*, 17/04/2019.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press, 2015.

Robbins, Scott. "Misdirected Principle with a Catch: Explicability for AI". *Minds & Machines,* n. 29, 2019, pp. 495-514. DOI: 10.1007/s11023-019-09509-3.

Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach.* 3. ed. San Francisco: Prentice Hall, 2010.

Selena, Silva; Kenney, Martin. "Algorithms, Platforms, and Ethnic Bias: A Diagnostic Model". *Communications of the Association of Computing Machinery*, out. 2019. DOI: https://doi.org/10.1145/3318157.

Taddeo, Mariarosaria; Floridi, Luciano. "Regulate Artificial Intelligence to Avert Cyber Arms Race". *Nature*, v. 556, 19/04/2018(a). DOI: 10.1038/d41586-018-04602-6.

____. "How AI Can Be a Force for Good". *Science*, v. 361, n. 6.404, ago. 2018(b). DOI: 10.1126/science.aat5991.

Theodorou, Andreas; Dignum, Virginia. "Towards Ethical and Socio-Legal Governance in AI". *Nature Machine Intelligence*, jan./2020. DOI: 10.1038/s42256-019-0136-y.

Topol, Eric. "The AI Diet". *The New York Times*, 02/03/2019.

Trajtenberg, Manuel. "AI as the next GPT: A Political Economy Perspective". In: Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi (orgs.). *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press, 2018.

Turing, Alan. "Computing Machinery and Intelligence". *Mind*, v. 59, n. 236, out./1950. DOI: 10.1093/mind/lix.236.433.

Otterlo, Martijn van. "A Machine Learning View on Profiling". In: Hildebrandt, M.; De Vries, K. (orgs.). *Privacy Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*. Abingdon: Routledge, 2013, pp. 41-64.

Wiener, Norbert. "Some Moral and Technical Consequences of Automation". *Science*, 06/05/1960.

YLS — Yale Law School Information Society Project. *Governing Machine Learning: Exploring the Intersection Between Machine Learning, Law, and Regulation*, set./2017. Disponível em: https://law.yale.edu/sites/default/files/area/center/isp/documents/governing__machine__learning__-__final.pdf. Acesso em: 25/06/2020.