ISSN - 2175-6058
DOI: https://doi.org/10.18759/rdgf.v24i3.2198

EXTERNALIDADES NEGATIVAS DA INTELIGÊNCIA ARTIFICIAL: CONFLITOS ENTRE LIMITES DA TÉCNICA E DOS DIREITOS HUMANOS

NEGATIVE EXTERNALITIES OF ARTIFICIAL INTELLIGENCE: CONFLICTS BETWEEN TECHNICAL LIMITS AND HUMAN RIGHTS

Dora Kaufman
Tainá Junquilho
Priscila Reis

RESUMO

A técnica de aprendizado de máquina utilizada na maioria das implementações atuais de inteligência artificial, é um modelo estatístico de probabilidade. Além da variável de incerteza intrínseca a modelos probabilísticos, a maneira como os algoritmos correlacionam dados, as bases de dados enviesadas, e a subjetividade humana nas decisões, engendra potenciais danos que ameaçam direitos humanos fundamentais. O propósito, com este artigo, é analisar essas externalidades, identificando as ameaças a três direitos humanos fundamentais: o direito à explicabilidade; privacidade; e não discriminação. A metodologia levou à investigação das garantias legais *versus* os limites da técnica, evidenciado que esses limites se constituem em barreiras às conformidades legal e regulatória.

Palavras-chave: Inteligência Artificial. Direitos Humanos. Danos.

ABSTRACT

The machine learning technique that permeates most current artificial intelligence implementations is a statistical probability model. In addition to the uncertainty variable intrinsic to probabilistic models, the way algorithms establish correlations in the data, the biased databases, and the human subjectivity in decisions,

engender negative externalities that threaten fundamental human rights. The purpose of the article is to reflect on these externalities identifying threats to three fundamental human rights: the right to explainability, privacy, and non-discrimination. The methodology investigated the legal guarantees *versus* the limits of the technique itself, showing that these limits constitute barriers to legal and regulatory compliance.

Keywords: Artificial Intelligence. Human Rights. Externalities.

INTRODUÇÃO

Consideradas pioneiras nos estudos sobre sistemas de computação tendenciosos, Bayat Friedman e Helen Nissenbaum, em meados da década de 1990, publicaram dois artigos alertando para o potencial impacto desses sistemas na sociedade, dado o custo relativamente baixo de sua produção e disseminação. No primeiro artigo, Friedman e Nissenbaum (1996), com base em análise de casos reais, identificam três categorias de viés: preexistente, técnica, e emergente. A categoria preexistente tem suas raízes nas instituições, em práticas e atitudes sociais; a técnica surge de restrições técnicas dos sistemas; e a emergente aparece em um contexto de uso. As autoras ponderam que, embora outros tenham apontado o problema dos vieses, são desconhecidos estudos que examinem o fenômeno de forma abrangente e ofereçam uma estrutura para compreendê-los e remediá-los, e concluem sugerindo que os critérios de confiabilidade, precisão e eficiência sejam incluídos na avaliação de qualidade dos sistemas.

No segundo artigo, Friedman e Nissenbaum (1997) enfatizam que, em breve, esses sistemas se tornarão onipresentes, e exigirão, portanto, o desenvolvimento de medidas abrangentes de avaliação para entender como o usuário pode ser beneficiado ou prejudicado pelo *design* dos modelos. Esses alertas revelam que a preocupação com o viés nos sistemas maquínicos é anterior ao recente avanço da inteligência artificial (IA) e que, como será tratado neste artigo, as três categorias de viés propostas por Friedman e Nissenbaum têm aderência aos sistemas de IA.

Em 28 de junho de 2015, um domingo à noite, o haitiano-americano Jacky Alciné recebeu uma notificação do recém-criado serviço de armazenamento e classificação de fotos Google Photos, em que a tecnologia atribui, automaticamente, uma legenda temática. Sua amiga, igualmente de pele escura, tinha compartilhado um álbum com cerca de 70

fotos dos dois juntos, legendadas pelo sistema como "Gorilas", inclusive a foto da capa do álbum.

Alciné, profissional de tecnologia, ao perceber o problema, publicou no Twitter: "Minha amiga não é um Gorila". Duas horas depois, o à época arquiteto-chefe do Google+, Yonatan Zunger, manifestou-se reconhecendo a gravidade do erro e, como solução, sua equipe removeu da base de dados a categoria rotulada como "Gorila".

A revista Wired constatou, em 2018, que o rótulo "Gorila" permanecia desativado no *Google Photos*, não classificando como gorilas nem mesmo os próprios gorilas. Noticiado amplamente pela mídia, o fato foi tratado como um problema de "algoritmo racista", quando a origem do viés estava na base de dados tendenciosa usada no treinamento dos algoritmos, predominantemente composta de homens de pele clara. O episódio é considerado o primeiro caso detectado e/ou denunciado de discriminação de sistemas de IA (Christian, 2020).

A técnica de aprendizado de máquina, subárea do campo da IA, que permeia a maior parte das implementações atuais de modelos de IA, denominada de redes neurais de aprendizado profundo (deep learning neural networks - DLNNs ou simplesmente *deep learning)* pela inspiração no cérebro biológico, é um modelo estatístico de probabilidade com duas categorias: IA preditiva e IA generativa. A técnica está em seus primórdios, pois foi reconhecida pela academia e pelo mercado em 2012, e ainda possui inúmeras limitações (Dignum, 2019; Kaufman, 2022).

Além da "evidência inconclusiva" dos resultados, função da variável de incerteza intrínseca a todo modelo estatístico de probabilidade – produz conhecimento provável, mas inevitavelmente incerto –, a técnica é sujeita às interferências de bases de dados enviesadas e da subjetividade humana presente no processo de elaboração, aplicação, visualização e interpretação dos resultados (Wachter; Mittelstadt; Floridi, 2016). Adicionalmente, a complexidade das correlações estabelecidas pelos algoritmos de IA nos dados transcende a capacidade cognitiva dos seres humanos, configurando o "problema da interpretabilidade" (ou opacidade, ou *black-box*). Essas características, e imperfeições, geram potenciais danos, que ameaçam direitos humanos fundamentais, como o direito à explicabilidade, o direito à privacidade, e o direito à não discriminação (Kaufman, 2019; 2022).

Dado que os sistemas baseados nessa técnica se tornaram fator estratégico de processos decisórios automatizados, pela capacidade de gerar resultados preditivos com taxas relativamente altas de acurácia,

na última década emergiram diversas reações da sociedade visando mitigar os potenciais danos, representadas por abordagens como: AI for Social Good; AI Ethics; AI for Good, AI Human Centric (Dignum, 2019).

A primeira iniciativa conhecida são os Asilomar Principles, originados na Conference on Beneficial AI, realizada em 2017, por iniciativa do Future of Life Institute. No evento, foi lançado um conjunto de 23 princípios gerais para assegurar o desenvolvimento de tecnologias de IA benéficas à sociedade, subdividido em três categorias – pesquisa; ética e valores; e questões de longo prazo. Esses princípios estão 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 Now Institute, da Universidade de Nova York (2018); o Human-Centered Artificial Intelligence (HAI), da Universidade de Stanford; e o Leverhulme Center for the Future of Intelligence, da Universidade de Cambridge –, e de iniciativas de organizações multilaterais e europeias como o AI4People, organizado, em 2018, pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE), primeiro fórum europeu sobre os impactos sociais da IA.

Os princípios gerais, contudo, mostraram-se de pouca aplicabilidade prática. Além da natureza abstrata e não universal, o que impossibilitou, inclusive, traduzi-los em linguagem matemática, não é suficiente aplicá-los (supondo ser viável) na etapa de desenvolvimento e implementação dos sistemas de IA. Se na partida estiverem em observâncias ética e legal, a tendência é desalinhar com a entrada de novos dados, o que requereria um monitoramento contínuo, condição não viável (Kaufman, 2021a).

Em paralelo à autorregulamentação, igualmente com baixa eficácia, afloram iniciativas de regulamentação pelo poder público, como a proposta de regulamentação da Comissão Europeia, Artificial Intelligence Act (AIA), cuja primeira versão foi disponibilizada para consulta pública em 21 de abril de 2021, e a revisão votada preliminarmente no Parlamento Europeu em 14 de junho de 2023. No Brasil, o Projeto de Lei 2.338/2023, proposto pelo senador Rodrigo Pacheco, e em tramitação no Senado Federal, é o ponto de partida no estabelecimento de um marco regulatório da IA (Kaufman; Coelho, 2023).

O propósito deste artigo é promover a reflexão sobre os principais desafios dos sistemas atuais de IA – o problema da interpretabilidade, ou não explicabilidade; a discriminação nos resultados enviesados; e a ameaça à privacidade – a partir do entendimento de sua lógica e seu

funcionamento. Busca-se identificar os conflitos e as ameaças aos direitos humanos fundamentais, *vis-à-vis* os limites da própria técnica, para agir em conformidades ética e legal. Os potenciais danos, vale ressaltar, não têm o mesmo grau de problematização, visto que a intensidade e extensão dos danos ao usuário afetado variam conforme o domínio da aplicação: o grau de risco de uma recomendação equivocada em uma plataforma de *streaming*, por exemplo, é significativamente distinto de uma recomendação equivocada de um procedimento médico.

LIMITES ÉTICOS DO AGENCIAMENTO DOS ALGORITMOS DE IA

Em "Ética a Nicômaco", Aristóteles (2003) concebe a ideia de ação como originada na noção de agente: alguém, ou algo, que tem controle sobre o que está fazendo e é responsável pelas consequências dos efeitos causados por essa ação. A atribuição aristotélica de responsabilidade distingue duas condições sobre a ação: o controle do agente e o conhecimento do agente. Com base nessas premissas, Mark Coeckelbergh (2020) argumenta que as tecnologias de IA podem ser agentes, mas não atendem aos critérios de Agente Moral (AM); ser agente, aqui, significando atuação no mundo, intervenção no ambiente e sobre os demais agentes.

Apesar de reconhecer que a IA pode agir e de que esse agir não é moralmente neutro, isto é, tem consequências morais – como as decisões automatizadas, por exemplo, de concessão de crédito; do recrutamento de Recursos Humanos (RH); dos procedimentos médicos; da definição de pena de um condenado por crime –, Coeckelbergh (2020) não atribui agenciamento moral à IA, ou seja, a responsabilidade por suas ações deve permanecer na esfera dos agentes humanos, desenvolvedores e usuários da tecnologia; interpretação endossada, em geral, pelos sistemas jurídicos mundo afora.

O debate sobre agenciamento e agenciamento moral é compartilhado por vários autores, com distintas visões. Floridi e Sanders (2004) advogam por uma moralidade diversa dos padrões morais humanos, logo, passível de ser atribuída às tecnologias de IA. Como AMs, os sistemas artificiais não necessariamente precisam exibir livre-arbítrio, estados mentais, ou responsabilidades, mas devem possuir os atributos

de interatividade (resposta ao estímulo por mudança de estado); autonomia (capacidade de mudar o estado sem estímulo); e adaptabilidade (capacidade de mudar as regras de transição pelas quais o estado é alterado).

Wendell Wallach e Colin Allen (2009), preocupados com o futuro da IA – quando esses sistemas se tornarem mais complexos e aptos a tomar decisões autônomas, ou seja, independentemente da supervisão humana –, propõem expandir o círculo de AMs dos seres humanos para agentes inteligentes artificiais, ou Artificial Moral Agents (AMAs). Wallach e Allen sugerem uma moralidade intermediária, denominada por eles de "moralidade funcional".

Nick Bostrom e Eliezer Yudkowsky (2011) recusam-se a conferir aos atuais sistemas de IA, ainda de competência restrita a um único domínio, o *status* moral, mesmo que seus algoritmos exerçam funções cognitivas próprias dos humanos. Para os autores, a não atribuição do *status* moral a esses sistemas está consubstanciada na inexistência de dois atributos: senciência, a capacidade para promover a experiência fenomenal, ou *qualia*, como a capacidade de sentir dor e sofrer; e sapiência, conjunto de capacidades associadas à autoconsciência e racionalidade responsável.

David Gunkel (2012), em suas reflexões sobre AM, recorre ao conceito de "animal-máquina" (*bête-machine*) de Descartes que, defendendo a segregação dos modos humano e animal e relacionando os animais aos seres autônomos, considerava o ser humano como a única criatura capaz de pensar racionalmente, identificando uma similitude entre os animais e as máquinas. Gunkel sugere, com base nessa visão, pensar as inovações da filosofia dos direitos dos animais como modelo a ser estendido à moral das máquinas inteligentes.

Refletindo sobre o futuro da IA, Liao (2020) identifica uma correlação positiva entre o aumento de capacidades e o *status* moral, prevendo a possibilidade de os sistemas de IA tornarem-se conscientes e sencientes. Liao, inclusive, não descarta a possibilidade da IA, no futuro, ter um *status* moral superior ao humano, em virtude de conter mais propriedades intrínsecas (independentemente das propriedades extrínsecas, de relacionamento com outras entidades).

Com base nas diversas perspectivas apresentadas e considerando o estágio atual de desenvolvimento da IA, entende-se que esses sistemas são agentes, mas não AMs, portanto, a responsabilidade sobre seus efeitos na sociedade deve circunscrever-se à esfera humana, na

observância de padrões éticos, em todas as etapas de desenvolvimento e uso da IA.

GARANTIAS LEGAIS E DIREITOS HUMANOS FUNDAMENTAIS AFETADOS PELAS EXTERNALIDADES NEGATIVAS DOS SISTEMAS DE IA

Toda tecnologia é social e humana. Seus efeitos dependem do que os seres humanos fazem com ela e como a inserem nos ambientes técnico-sociais. Cabe à sociedade humana deliberar, entre inúmeras questões, sobre se a IA deve ser aplicada em todos os domínios e para executar todas as tarefas, e se o uso da IA em aplicações de alto risco se justifica. O desafio é buscar o equilíbrio entre mitigar (ou eliminar) os riscos e preservar o ambiente de inovação, sem supervalorizar nem demonizar a IA (Coeckelbergh, 2020; Kaufman, 2022).

A ética é um atributo da sociedade humana. Os humanos é que definem como criar e usar a IA, observando as especificidades da tecnologia e as especificidades dos domínios de aplicação. A tecnologia não é determinista, mas molda e é moldada pelas escolhas humanas (Kaufman, 2022).

Na condição de agente, contudo, a IA tem prerrogativas que ameaçam os direitos humanos fundamentais, especialmente o direito à explicabilidade; o direito à não discriminação; e o direito à privacidade de dados, todos historicamente garantidos por diversas legislações.

GARANTIAS LEGAIS AO DIREITO À EXPLICABILIDADE

Na concepção do Estado Democrático de Direito está contida a ideia de que a legitimidade das decisões tomadas pelo Estado sobre os cidadãos, quando interferem diretamente em direitos humanos, só se concretiza se forem devidamente motivadas. Essa noção deriva do direito fundamental ao devido processo legal, garantido pela Constituição Federal (CF) de 1988, no artigo 5º, inciso LIV, dos quais emanam os princípios do contraditório e da ampla defesa. Com base nesses princípios constitucionais, a todo cidadão é garantido o poder de contestar

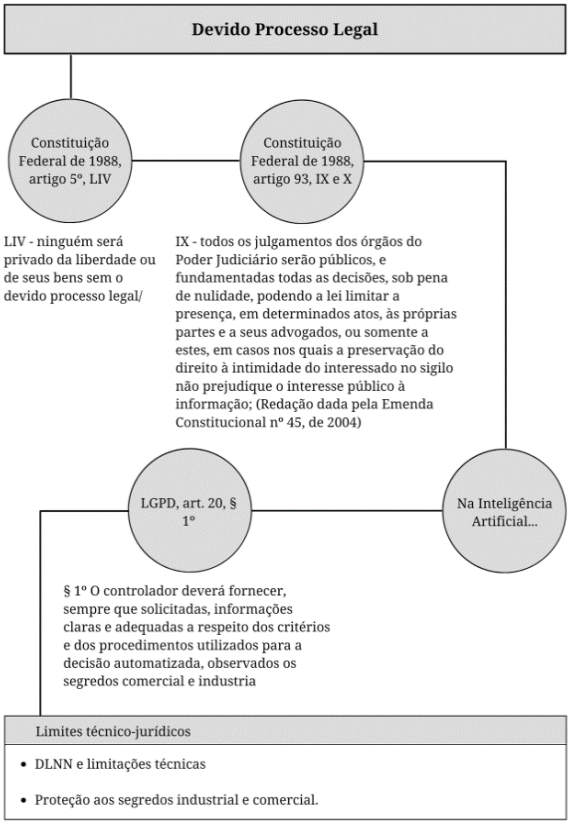
decisões arbitrárias e imotivadas, que impactem seus direitos. O direito à explicabilidade deriva, nesse sentido, do direito fundamental ao devido processo legal (Nunes; Marques, 2018).

Essa noção constitucional foi reafirmada pelo Supremo Tribunal Federal (STF), em julgamentos como o do Agravo em Recurso Extraordinário (ARE) n. 779.543/2013, e o do Mandado de Segurança (MS) n. 25.747/2012, de Santa Catarina/SC, que entenderam ser, a fundamentação das decisões judiciais e administrativas, um direito fundamental previsto no artigo 93, incisos IX e X da CF/88. Esse direito constitucional dá suporte ao direito à explicabilidade, no caso de decisões automatizadas, com base em sistemas de IA.

Normas infraconstitucionais, nacionais ou estrangeiras, igualmente, prescrevem a exigência de explicação. A Resolução do CNJ n. 332/2020, por exemplo, ao tratar do uso de sistemas de IA pelo Poder Judiciário requer que seja fornecida explicação satisfatória e passível de auditoria humana, em caso de proposta de decisão automatizada, especialmente quando se tratar de decisão judicial. A mesma resolução considera que decisões baseadas em IA devem preservar o julgamento justo, o qual não existe se não for respeitado o princípio da transparência das decisões, além do devido processo legal já mencionado.

O direito à explicabilidade também consta em leis de proteção de dados, que afetam indiretamente os sistemas de IA quando fazem uso de dados pessoais. A Lei Geral de Proteção de Dados (LGPD) explicita o direito à revisão das decisões baseadas unicamente em tratamento automatizado de dados pessoais, mas não garante, de forma irrestrita, o direito à explicação (Fig. 1). Embora trate do direito à explicabilidade, quando aponta a necessidade de explicação, pelo controlador dos dados, a respeito dos critérios e procedimentos utilizados para a decisão automatizada, na forma do artigo 20, §1º, condiciona esse dever/direito de explicação à observância dos segredos comercial e industrial, privilegiando, de certa forma, o segredo de negócio ante qualquer direito à explicação.

Figura 1 – A "explicabilidade" no direito brasileiro



Fonte: Elaboração dos autores.

No âmbito internacional, o Regulamento Geral de Proteção de Dados (RGPD), lei de proteção de dados europeia, por sua vez, embora não preveja expressamente o direito à explicabilidade no corpo da norma, traz, em seu Considerando n. 71, a necessidade de os processos automatizados de decisão sobre os aspectos pessoais do titular de dados conterem informações que expliquem a citada decisão, para permitir que o titular a conteste. A norma europeia, diferentemente da brasileira, não sobrepõe o direito à proteção do segredo de negócio ao direito à explicabilidade, dando mais importância à transparência algorítmica do que aos segredos comerciais e industriais.

O Artificial Intelligence Act, cuja proposta original fora emendada em 14/06/2023, passando a incorporar o novo artigo 4.a, traz, como

princípio geral aplicável aos sistemas de IA, a transparência, que, por sua vez, implica na explicabilidade. De forma similar, o Projeto de Lei brasileiro 2.338/2023, em seu artigo 3º, VI, eleva a explicabilidade a princípio, que deve ser observado nos casos de uso, desenvolvimento e implementação de sistemas de IA. E ainda traz em seu artigo 5º, II, que pessoas afetadas por sistemas de IA têm "direito à explicação sobre a decisão, recomendação ou previsão tomada por sistemas de inteligência artificial".

Ainda que legislações atuais e interpretações judiciais busquem garantir, em maior ou menor grau, o direito à explicabilidade, principalmente quando a própria ordem constitucional o contempla como um direito fundamental, tecnicamente, nem sempre será possível atender a tal direito.

LIMITES DA TÉCNICA AO DIREITO À EXPLICABILIDADE

Na técnica de aprendizado de máquina *(machine learning)* redes neurais de aprendizado profundo (deep learning neural networks - DLNNs), como mencionado, o processo como os algoritmos correlacionam os dados e definem os parâmetros (pesos relativos) é de tal complexidade que transcende a capacidade cognitiva humana, ou seja, os seres humanos não são capazes de compreender. Estabelece-se uma incompatibilidade entre a otimização matemática de alta dimensionalidade e o raciocínio e a interpretação semântica do ser humano, fenômeno denominado, pelos cientistas do campo, de "problema da interpretabilidade" (Goodefellow; Bengio; Courville, 2016).

Essa limitação decorre do desconhecimento de como os chamados "dados de entrada" (*inputs*) geraram os dados de saída (*output*). As arquiteturas das DLNNs são formadas por várias camadas (*layers*), e 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 constitui um dos elementos no cerne do problema da não explicabilidade, ou seja, da dificuldade de concretização, na sua plenitude, do direito à explicação. Os parâmetros (pesos relativos) que correlacionam os *pixels* no reconhecimento de imagem, por exemplo, são definidos pelos algoritmos de

IA, portanto, são variáveis não controláveis (grau de relevância de cada *pixel* para atingir o objetivo final/*output*) (Goodfellow; Bengio; Courville, 2016).

Os especialistas estão empenhados em reduzir a opacidade desses sistemas, também conhecida como *black-box*, termo que é uma metáfora com significado dual: se refere tanto a um dispositivo de gravação, como os sistemas de monitoramento de dados em aviões, quanto a um sistema de funcionamento não transparente de IA (Pasquale, 2015). Entretanto, com cada vez mais recursos computacionais e grandes conjuntos de dados (*big data*), o nível de complexidade tende a ser crescente.

Coeckelbergh (2020) aborda o problema da explicabilidade e transparência dos sistemas de IA pela perspectiva de atribuição de responsabilidade: a dos agentes (usuários da tecnologia) decorre da expectativa de os destinatários (outro lado da relação) saberem explicar as razões da decisão. Na saúde, por exemplo, o pressuposto é que o médico controle o procedimento e seja capaz de explicá-lo ao paciente. A responsabilidade é tratada como prestação de contas: para agir com responsabilidade (nesse caso, o agente precisa saber o que está fazendo, e justificar sua ação) e para explicar as razões aos afetados pela ação (pacientes), que podem e devem exigir e merecer respostas sobre o que e como foi decidido.

A visão da explicabilidade como demanda dos afetados faz sentido do ponto de vista ético, mas confronta-se com as limitações tecnológicas intrínsecas. O fato de se tratar de explicações de "humanos para humanos" não atenua o problema da explicabilidade: se a opacidade da técnica transcende a capacidade cognitiva dos seres humanos, não há como supor que os agentes humanos serão apoiados por sistemas técnicos suficientemente transparentes, como alega Coeckelberg (2022). O grau de acurácia dos sistemas de IA, inclusive, quando aplicados a tarefas em distintos domínios, gera uma tensão entre a necessidade de explicação e a eficiência, caracterizando um *tradeoff* entre precisão e transparência: quanto maior a precisão, menor a transparência (Villani, 2018).

Estão em curso esforços científicos de gerar interpretações amigáveis do funcionamento desses sistemas, o que seria a "IA Explicável" (Explanable AI – XAI) (Barredo Arrieta *et al*., 2020). Os resultados desses esforços, entretanto, ainda não são efetivos. A dificuldade aumenta quando os sistemas são produzidos por empresas privadas, em parceria

ou não com o Estado, juridicamente protegidos pelo segredo comercial, impossibilitando o questionamento do cidadão afetado pela decisão automatizada (Frazão; Goettenauer, 2021).

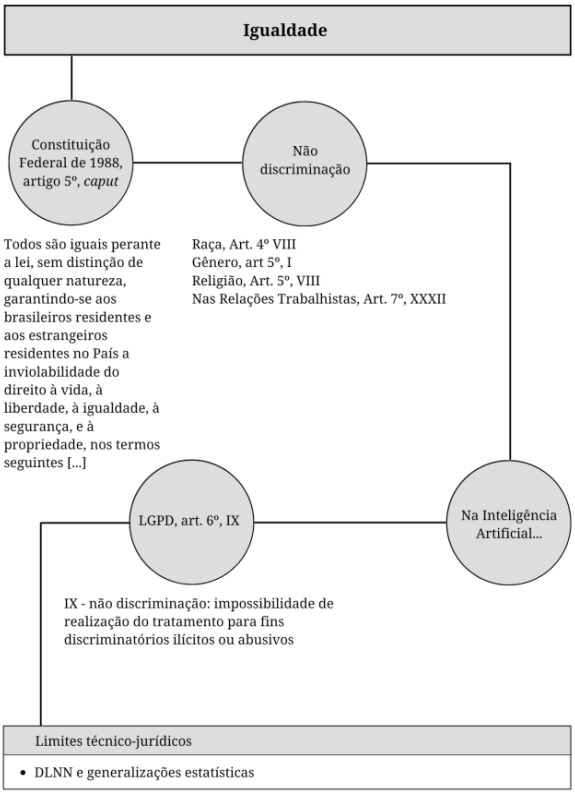
GARANTIAS LEGAIS AO DIREITO À NÃO DISCRIMINAÇÃO

Ao longo da história, muitas foram as justificativas para ações geradoras de condutas discriminatórias contra as minorias, todas elas baseadas na concepção de inferioridade de alguns grupos e/ou indivíduos em relação a outros, ou da atribuição de privilégios a uma "casta". A perspectiva da igualdade perante a lei tem origem liberal, caracterizada pela ideia de isonomia que não visava à redução das desigualdades, mas uma igualdade formal de tratamento e submissão aos regimes legais impostos pelo Estado (Da Silva, 2020).

A compreensão de que os seres humanos são iguais perante a lei (igualdade formal), com o passar do tempo e, principalmente, após os eventos da Segunda Guerra Mundial, evoluiu para abranger a ideia de igualdade material, que justifica discriminações sempre que essas servirem para reduzir as desigualdades entre grupos. Partindo dessa noção de igualdade, Bandeira de Mello (2019) explica o conteúdo que compõe o conceito jurídico de igualdade no Direito brasileiro. A importância da igualdade no Estado Democrático brasileiro expressa-se na abertura do artigo 5º e está concretizada no *caput* desse artigo na CF/88.

Diversos outros dispositivos constitucionais visam a impedir discriminações contra certos grupos: proíbem-se preconceitos de raça (art. 4º, inciso VIII); gênero (art. 5º, inciso I); religião (art. 5º, inciso VIII); ou em relações trabalhistas (art. 70, inciso XXXII). (Fig. 2). Tais previsões são corroboradas por leis infraconstitucionais, como a Lei n. 7.716/1989, que define os crimes resultantes de preconceito de raça, ou cor. A LGPD coloca como princípio previsto no artigo 6º, inciso IX, a vedação à discriminação ilícita, ou abusiva, compreendidos, esses conceitos, respectivamente, como contrariedade à lei e preconceito injustificado (Mendes; Mattiuzzo; Fujimoto, 2021).

Figura 2 – «Não discriminação» no direito brasileiro



Fonte: Elaboração dos autores.

Internacionalmente, a proteção à igualdade e à não discriminação está consagrada no artigo VII da Declaração Universal dos Direitos Humanos. Além disso, existem declarações específicas, como a Declaração das Nações Unidas sobre a Eliminação de Todas as Formas de Discriminação Racial; a Convenção sobre a Eliminação de todas as Formas de Discriminação contra a Mulher (1979); e a Convenção n. 111, da Organização Internacional do Trabalho (OIT), que dispõem sobre discriminação em matérias de emprego e ocupação.

A igualdade abrange, portanto, a não discriminação, que protege o cidadão de preconceitos cujo critério de *discrimen* não se justifique. É possível notar, entretanto, que os atuais resultados gerados por sistemas de IA têm reproduzido e, por vezes, reforçado as discriminações e os preconceitos da sociedade.

LIMITES DA TÉCNICA AO DIREITO À NÃO DISCRIMINAÇÃO

Considera-se que existe um enviesamento no resultado, quando o sistema exibe um erro sistemático ("enviesamento estatístico" ou "discriminação algorítmica"). Como todo modelo estatístico, as redes neurais profundas são projetadas para generalizar a partir de uma amostra (nesse caso, amostras formadas por grandes conjuntos de dados, *big data*). O viés refere-se ao erro que pode ocorrer nesse processo de generalização (Cozman; Kaufman, 2022).

Karen Hao (2019) adverte que, para detectar o viés, é imprescindível compreender suas origens, reconhecendo que a tendência é atribuir o enviesamento exclusivamente aos dados de treinamento tendenciosos, quando pode surgir nas várias etapas do processo, particularmente: a) no enquadramento do problema, quando o desenvolvedor traduz o objetivo a ser alcançado em linguagem computável; b) na coleta dos dados, no caso de a base não ser representativa da realidade ou refletir os preconceitos existentes na sociedade; e c) na preparação das bases de dados. Mesmo quando detectado, contudo, é difícil corrigir o viés, particularmente no caso de a detecção ocorrer em sistemas em pleno uso, o que explica a solução do Google Photos em eliminar da base de dados a categoria legendada como " Gorilas" (Hao, 2019; Kaufman, 2022).

Uma potencial discriminação, não contemplada por Hao (2019), é a originada na produção de dados, tanto na predominância de usuários dos países desenvolvidos com mais acesso às tecnologias e às redes sociais, o que engendra uma base de dados imagética enviesada pelo biotipo racial de pele clara, quanto na não desagregação dos dados por gênero e/ou o tratamento dado ao homem como "humano padrão". Caroline Perez-Criado (2021), por meio de extenso levantamento histórico da "invisibilidade" feminina, constata que a prática de não coletar dados desagregados por gênero, tratando os homens como neutros e/ou "padrão humano", distorce a suposta objetividade e a acurácia dos resultados dos modelos preditivos habilitados por IA.

Na etapa de desenvolvimento de um sistema de redes neurais, depois da determinação do objetivo, cabe ao cientista da computação traduzi-lo em variáveis que possam ser computadas, os chamados hiperparâmetros. Os desenvolvedores é que definem a arquitetura a ser utilizada; os termos de pesquisa para coletar os dados; e selecionam as

bases de dados. Identificar a influência da subjetividade humana não é trivial, além de não ser possível eliminá-la, mesmo se identificada (Hao, 2019). Equipes inter e multidisciplinares podem atenuar os efeitos discriminatórios, mas sua eficácia depende de construir "pontes" entre os campos de conhecimento (Kaufman, 2021b).

Na base de dados, o viés ocorre se sua composição for menos diversificada demograficamente do que a população-alvo, ou seja, se a base de dados não reproduzir a proporcionalidade, entre outros, de etnias e gêneros encontrados no universo objeto da ação. A diferença entre ambientes controlados (laboratórios) e ambientes não controlados (mundo real), igualmente, tem o potencial de gerar resultados tendenciosos; nas ruas, por exemplo, as câmaras captam imagens em baixa resolução; o ângulo e a luminosidade podem dificultar a extração de características faciais, ou mesmo distorcer provocando falso positivo (erro no reconhecimento facial) (Learned-Miller *et al*., 2020).

Outra possibilidade é a variância, que expressa a sensibilidade de um algoritmo às diferenças nos dados de treinamento, ou seja, não captura todas as características contidas nos dados. O processo de rotulagem dos dados, igualmente, tem o potencial de originar resultados tendenciosos. Nesse caso, o desafio é representar a complexidade do mundo em taxonomias para rotular os dados, precondição do aprendizado supervisionado – tipo de aprendizado de máquina utilizado no reconhecimento de imagem e som/voz, em que o desenvolvedor do sistema define a meta (*output);* por exemplo, identificar a imagem de um cachorro, e rotula os dados de entrada (milhares ou milhões de imagens de cachorro). O nível de complexidade aumenta, no caso de taxonomias de imagens de seres humanos, por exemplo, para classificar com precisão o gênero ou a etnia de uma pessoa, ou mesmo reconhecer a profissão a partir da fotografia (Crawford, 2021).

É recente a sensibilidade de pesquisadores, e da sociedade em geral, para o problema do viés nos dados, consequentemente, durante anos, diversas bases de dados tendenciosas foram utilizadas para desenvolver e treinar os algoritmos de IA (em parte, continuam sendo). O ImageNet, por exemplo, demorou uma década (2009 a 2019) para reconhecer o viés na rotulagem de suas imagens, mesmo assim por iniciativa do artista americano Trevor Paglen, dedicado ao tema da vigilância em massa, e do especialista em tecnologias Leif Ryge. Os dois desenvolveram o aplicativo ImageNet Roulette como parte de uma exposição de arte sobre os sistemas de reconhecimento de imagem no museu Fon-

dazione Prada, em Milão/Espanha, em cocuradoria de Kate Crawford, intitulada "Training Humans": quando o usuário efetua o *upload* de sua foto, o aplicativo retorna a imagem com o rótulo atribuído automaticamente pelo algoritmo[1].

O Labeled Faces in the Wild (LFW) é um banco de dados, organizado em 2007 com base em artigos de notícias *on-line,* e rotulado por uma equipe da "Umass Amherst". Em 2014, Hu Han e Anil Jain, da "Michigan Sates", identificaram que mais de 77% das imagens eram de homens e mais de 83% de homens de pele clara, e o então presidente dos Estados Unidos da América (EUA), George W. Bush, tinha 530 imagens exclusivas, mais do que o dobro do conjunto de imagens de todas as mulheres de pele escura. Cinco anos depois, e 12 da data de constituição do LFW, seus gestores postaram um aviso de isenção de responsabilidade, alertando que muitos grupos não estão bem representados (Christian, 2020).

O United States Office of the Director of National Intelligence – supervisor da implementação do Programa de Inteligência Nacional, principal assessor do Presidente, do Conselho de Segurança Nacional e do Conselho de Segurança Interna para assuntos de inteligência relacionados à segurança nacional –, em 2015, lançou um banco de dados de imagens faciais denominado IJB-A, supostamente, contemplando a diversidade da população americana. Estudo de Timnit Gebru e Buolamwini, contudo, constatou que 75% eram imagens de homens e 80% de homens de pele clara, e apenas 4,4% do conjunto de dados era de mulheres de pele escura (Christian,2020).

Dada a crescente consciência dos efeitos danosos do viés discriminatório contido nas decisões automatizadas por IA, particularmente as aplicações em campos sensíveis, como saúde, segurança e educação, especialistas acadêmicos e não acadêmicos estão empenhados em encontrar abordagens para detectar e remover, ou ao menos mitigar, esse viés dos sistemas de IA.

GARANTIAS LEGAIS AO DIREITO À PRIVACIDADE

A garantia legal do direito à privacidade está prevista, inicialmente, no artigo 5º, inciso X da CF/88, que positiva a privacidade e a intimidade no contexto da vida particular em família, protegendo também a correspondência e as comunicações. Marcel Leonardi (2011) argumenta que existem diversos conceitos unitários, por trás do direito à privacidade,

trazidos por diferentes normas, na doutrina e na evolução da jurisprudência. Leonardi (2011) compreende que a privacidade tem o condão de transformar a sociedade, evitando intromissão na vida dos cidadãos e prevenindo danos, além de permitir que o indivíduo continue a contribuir com a comunidade.

Tem-se compreendido que, para que a pessoa humana possa desenvolver livremente sua personalidade, bem como exercer sua autodeterminação informacional e livre consentimento, é necessário que sua privacidade esteja resguardada (Sarlet, 2021). Observa-se uma recente "ampliação normativa" da compreensão da privacidade, abrangendo não só a intimidade da vida privada, mas também a proteção aos dados que podem identificar seu titular (Bioni, 2021). Nesse sentido, o STF, em decisão histórica, fundamentada no direito à privacidade, entendeu que nesse direito está incluída a proteção de dados pessoais e o direito à autodeterminação informativa[2] .

Com a coevolução das sociedades e tecnologias, particularmente das tecnologias digitais como a IA, fez-se necessário ampliar a interpretação do direito à privacidade para que englobasse também direitos conexos, como o direito à proteção de dados pessoais, na tentativa de evitar ameaças ao cidadão às quais não estava exposto anteriormente. Isto é, a necessidade de proteger a privacidade, não só como um direito fundamental, mas como princípio constitucional, potencializa-se, considerando o alcance de tecnologias como a IA.

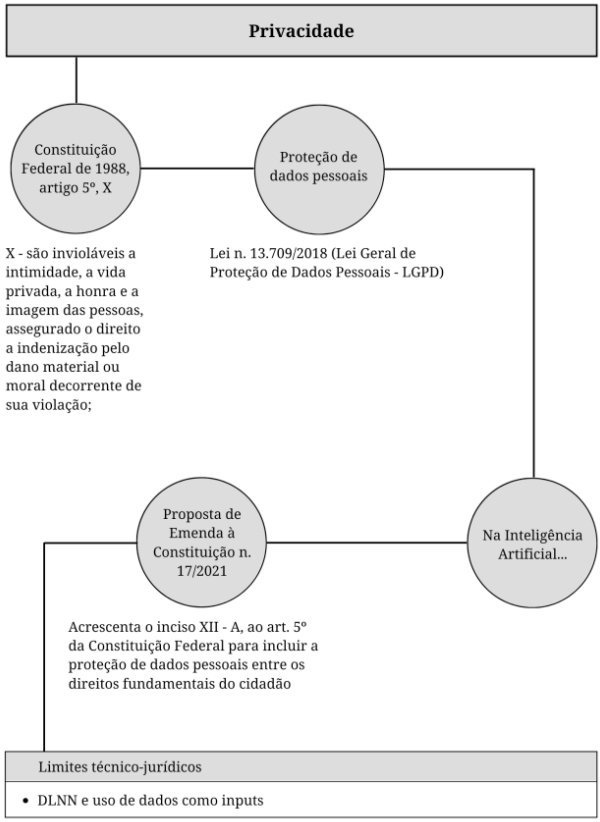
A privacidade é direito fundamental, que vem sendo relativizado, principalmente diante da sofisticação de programas de hipervigilância criados e justificados em nome da segurança, em especial, após os ataques terroristas ocorridos em 11 de setembro de 2001 (Teixeira; Datysgeld, 2016). Um exemplo ilustrativo é o programa Stellar Wind, sistema de vigilância iniciado em 2003 pelo Presidente George Bush e executado pela Agency National of Security (ANS), em conflito com os termos da 4ª Emenda da Constituição dos Estados Unidos – direito à privacidade na residência; direito de impedir que agentes do governo entrassem nas residências, e levassem os pertences, sem aprovação do morador e/ou sem mandado judicial. Com base nesse programa, firmaram-se acordos com empresas de telecomunicações, em troca de dados. No governo do Presidente Obama, o Congresso renovou a lei de vigilância, ampliando os poderes de fiscalização.

Diante dos inúmeros escândalos, como o relatado, de fiscalização e hipervigilância estatal e empresarial e dos danos a eles associados, observou-se a necessidade de ampliação legal da proteção da priva-

cidade, a qual inspirou a Lei Geral de Proteção de Dados brasileira e, mais recentemente, a Proposta de Emenda à Constituição n. 17/2021, já aprovada, que "acrescenta o inciso XII-A, ao art. 5º da Constituição Federal para incluir a proteção de dados pessoais entre os direitos fundamentais do cidadão" (Brasil, 2021).

Ingo Wolfgang Sarlet (2021) denomina tal fenômeno de "digitalização dos direitos fundamentais ou dimensão digital dos direitos fundamentais", o reconhecimento, pelas legislações protetivas de direitos humanos, de que tais direitos podem ser impactados por sistemas de IA, principalmente num contexto em que dados pessoais são processados em grande volume e velocidade e que, portanto, é fundamental reconfigurar sua proteção. (Fig. 3).

Figura 3 – "Proteção de dados" no direito brasileiro



Fonte: Elaboração dos autores.

O direito à privacidade também está previsto em normas internacionais, como a Convenção Americana sobre os Direitos Humanos – que tem o Brasil como um dos países signatários –; a Declaração Universal dos Direitos Humanos; além da Convenção Europeia dos Direitos Humanos. No cenário mundial, verifica-se que, embora o direito à proteção de dados pessoais não seja novo, sua tutela tem sido ampliada diante de novas formas e meios de disponibilização, armazenamento e captura desses dados. Como concluiu o Conselho da Europa, no Relatório "Inteligência Artificial e Proteção de Dados: Desafios e Possíveis Soluções" (2019), os dados pessoais são a base dos modelos de IA adotados em larga escala, o que torna a proteção à privacidade um potencial direito fundamental violado por essa tecnologia.

Nesse sentido, surgem regramentos, como o Regulamento Geral Europeu de Proteção de Dados, em 2016, cuja vigência, a partir de 2018, inspirou documentos como o California Consumer Privacy Act (CCPA), em 2020, e a Lei de Proteção à Informação Pessoal, da China, de 2021.

LIMITES DA TÉCNICA AO DIREITO À PRIVACIDADE

A privacidade é ameaçada pela IA em distintos momentos e usabilidades, desde as bases de dados pessoais utilizadas no desenvolvimento e treinamento de seus algoritmos, particularmente em sistemas de previsão de comportamento futuro de usuários, clientes ou consumidores, até o aperfeiçoamento de sistemas de vigilância, tais como o sistema de vigilância doméstico da ANS americana, o Sistema de Crédito Pessoal chinês (Social Credit System), ou o sistema de câmeras de vigilância de Londres/Inglaterra (em janeiro de 2020, incorporou as tecnologias de reconhecimento facial/IA)[3].

A Internet of Bodies, outro exemplo, ao ter acesso e controle de funções vitais do corpo humano, gera um conjunto de dados sensíveis que são utilizados para desenvolver e treinar os algoritmos de dispositivos de IA nas categorias de *wearable technology*, associada à vida saudável (*smartwatches*, *fitness trackers*), nos *microchips* para fins de identificação biométrica e/ou concessão de autorização, e nos dispositivos de saúde – marcapasso cardíaco monitorado remotamente; pâncreas artificial que monitora a glicose no sangue e fornece insulina; implantes cerebrais para tratar os sintomas de Parkinson e Alzheimer; próteses com *software* conectado aos ossos; *smart pills* (cápsulas com sensores

que percorrem o organismo em busca de sinais fora do padrão – anomalias, o PillCam Colon).

Os "dispositivos inteligentes", com o uso de tecnologias de IA, suscitam inéditos desafios para a governança de dados não apenas em relação às ameaças à privacidade, como também em relação ao *biohacking*. Como reconhece Andrea M. Matwyshyn (2018), são extraordinários os benefícios desses dispositivos, mas, igualmente, são potencialmente extraordinários os riscos associados à segurança física, autonomia e ao bem-estar de seus usuários.

O Fórum Econômico Mundial (WEF), em julho de 2020, publicou o relatório de autoria de Xiao Liu e Jeff Merritt (2020) "Shaping the Future of the Internet of Bodies (IoB): New challenges of technology governance", versando sobre as implicações desses dispositivos para a privacidade e equidade. Liu e Merritt examinam a governança de dados *IoB* nos EUA, comparativamente à regulamentação da União Europeia, salientando que, em ambas as regiões, existem lacunas entre as leis antidiscriminação e o inédito risco de discriminação, em função de inferências, perfis e agrupamento de dados originados nos dispositivos inteligentes. A escala de geração, armazenamento e mineração de dados é um dos elementos-chave da mudança de natureza dos atuais mecanismos de persuasão, caracterizando uma ruptura[4] nas estruturas lógicas em relação aos mecanismos anteriores (Alpaydin, 2016; Agrawal; Gans; Goldfarb, 2018; Lee, 2018; Mayer-Schönberger; Cukier, 2013; Pasquale, 2015).

A ameaça à privacidade, distinta das outras duas ameaças objeto deste artigo, está menos relacionada às especificidades do funcionamento da técnica e mais à origem, diversidade, qualidade dos dados utilizados no desenvolvimento e treinamento dos algoritmos de IA.

CONSIDERAÇÕES FINAIS

As externalidades negativas abordadas neste artigo configuram ameaças a três direitos humanos fundamentais: direito à explicabilidade; direito à não discriminação; e direito à privacidade. Para cada uma dessas ameaças, foram identificadas referências legais confirmatórias do reconhecimento do respectivo direito, indicando como, independentemente da intencionalidade de desenvolvedores e usuários, os fundamentos e a estrutura da técnica de redes neurais profundas (DLNNs) configuram barreiras para a plena observância da lei.

As barreiras não impedem, contudo, que os riscos associados ao uso dessa técnica sejam mitigados em cada uma das etapas de desenvolvimento do modelo, através de: seleção das variáveis iniciais (hiperparâmetros); escolha da base de dados de treinamento dos algoritmos do sistema; visualização e interpretação dos resultados; e posteriores auditorias para identificar vieses discriminatórios e/ou inapropriações. Em relação aos usuários, particularmente os gestores, nota-se a necessidade de adquirir conhecimento básico que permita compreender a lógica e o funcionamento dos sistemas de IA, e se capacitar para interagir com essa inédita interface homem-máquina.

O futuro da IA será função do empenho de pesquisadores acadêmicos e não acadêmicos em equacionar, ao menos, parte das limitações atuais da técnica, e da gradativa tomada de consciência da sociedade sobre seus impactos éticos e sociais, identificando os melhores caminhos para a efetivação da AI for Good, respeitando os direitos humanos fundamentais.

NOTAS

[1] O interessante resultado pode ser conferido também *on-line*. (Disponível em: http://digicult.it/slider/training-humans-an-exhibition-by-kate-crawford-and-trevor-paglen/. Acesso em: 15 set. 2022.)

[2] Por meio do julgamento de cinco ADIns conjuntas, que tratavam sobre a (in)constitucionalidade da Medida Provisória n. 954/2020, que obrigava o "compartilhamento de dados por empresas de telecomunicações prestadoras de Serviço Telefônico Fixo Comutado e de Serviço Móvel Pessoal com a Fundação Instituto Brasileiro de Geografia e Estatística, para fins de suporte à produção estatística oficial durante a situação de emergência pública de importância internacional decorrente do coronavírus (covid-19)", o STF reconheceu a existência do direito fundamental autônomo à proteção de dados pessoais. (Disponível em: https://www.conjur.com.br/2020-nov-25/lucia-ferreira-stf-direito-protecao-dados-pessoais. Acesso em: 19 set. 2022.)

[3] Pequim tem menos câmaras de vigilância proporcionalmente à sua população do que a cidade europeia: são 1,15 milhões de câmaras para uma população de 20 milhões, em Pequim (56,20 câmaras/1.000 habitantes), contra 627.727 câmaras para uma população de 9,3 milhões em Londres (67,47 câmaras/1.000 habitantes). O londrino médio é filmado 300 vezes por dia, o que atribui à cidade o título de "Capital Mundial da CCTV" (*closed-circuit television*)". O Estado controla uma parte menor do sistema de vigilância: estudo da Bristish Security Association indicou uma proporção entre câmaras privadas e públicas de 70 para 1. Os números tendem a ser maiores porque as câmaras domésticas não precisam de registro, só as câmaras de empresas precisam de registro no Information Commissioner's Office (ICO). Em média, as gravações são armazenadas por duas semanas. (Disponível em: https://www.comparitech.com/vpn-privacy/the-worlds-mostsurveilled-cities. Acesso em: 19 set. 2022.)

[4] O conceito de ruptura, no âmbito da ciência, é associado a Thomas Kuhn (1970), para quem o termo "paradigma" significa um modelo formado de métodos, tecnologias, formas de observação e experimentação, conceitos relacionados a um determinado fenômeno. O autor reconhece

a dificuldade de explicitar os elementos formadores de um paradigma por envolver, em parte, conhecimento tácito, apreendido na própria interação com o paradigma. Em geral, um paradigma representa um "mapa" a ser utilizado em relação ao objeto de estudo, e é útil enquanto não surgirem constrangimentos entre a experiência empírica e a teoria. Na visão de Kuhn (1970), em situações de ruptura, há uma perda de confiança nos paradigmas vigentes e, consequentemente, emergem novos paradigmas mais adequados aos desafios da atual realidade.

REFERÊNCIAS

AGRAWAL, A.; GANS, J.; GOLDFARB, A. **Prediction Machines**: the simple economics of artificial intelligence. Massachusetts: Harvard Press, 2018.

ALPAYDIN, Ethem. **Machine learning**. Cambridge, Massachusets: MIT Press, 2016.

ARIKAN, Cenck Lacin; EL-KHOURY, Moufid. From the internet of things toward the internet of bodies: ethical and legal considerations. **Strategic Change, Special Issue**: artificial intelligence in finance. v. 30, n. 3, maio 2021, p. 307-314.

ARISTÓTELES. **Ética a Nicômaco**. Trad. de Pietro Nassetti. São Paulo: Martin Claret, 2003.

ASSEMBLEIA GERAL DAS NAÇÕES UNIDAS. **Declaração universal dos direitos humanos**. 1948. Disponível em: https://brasil.un.org/pt-br/91601--declaracao-universal-dos- direitos-humanos. Acesso em: 1º jun. 2021.

BANDEIRA DE MELLO, Celso Antônio. **O conteúdo jurídico do princípio da igualdade**. 4. ed., São Paulo: Juspodvm, 2019.

BARREDO ARRIETA, A. *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, december 2019, p. 82-115, 2020.

BIONI, Bruno. **Proteção de dados pessoais**: a função e os limites do consentimento; 2 ed. Forense: São Paulo, 2021.

BOSTROM, Nick; YUDKOWSKY, Eliezer. **The ethics of artificial intelligence. draft for cambridge handbook of artificial intelligence**. Co-

ord. William Ramsey and Keith Frankish. Cambridge: Cambridge University Press, 2011.

BRASIL. [Constituição (1988)]. Constituição da República Federativa do Brasil: promulgada em 5 de outubro de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 25 set. 2021.

BRASIL. **Projeto de lei n. 21/20 cria o marco legal do desenvolvimento e uso da Inteligência Artificial (IA)**. Disponível em: https://www.camara.leg.br/propostas- legislativas/2236340. Acesso em: 25 ago. 2021.

BRASIL. **Projeto de lei n. 2.338/2023 dispõe sobre o uso da Inteligência Artificial**. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 8 out. 2023.

CHRISTIAN, Brian. **The alignment problem**: machine learning and human values. Nova York: W. W. Norton & Company, 2020.

COECKELBERGH, Mark. Artificial intelligence, responsibility attribution, and a relational justification of explainability. **Science and Engineering Ethics**, v. 26, 2020, p. 2051-2068.

CONSELHO DA EUROPA. **Artificial intelligence and data protection**: challenges and possible remedies. 25 jan. 2019. Disponível em: https://rm.coe.int/artificial-intelligence- and-data-protection-challenges-and--possible-re/168091f8a6. Acesso em: 3 jul. 2021.

CONSELHO DA EUROPA. **Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts**. 2021. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Acesso em: 30 jun. 2021.

CONSELHO DA EUROPA. **Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmoni-**

sed rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 14 jun. 2023. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html. Acesso em: 8 out. 2023.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Resolução n. 332, de 21 ago. 2020**. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429. Acesso em: 20 set. 2021.

COZMAN, Fabio Gabliardi; KAUFMAN, Dora. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. **Revista USP Qualis A1**, n. 135, p.195-210. Out./nov./dez. 2022. ISSN 0103-9989. DOI: https://doi.org/10.11606/issn.2316-9036.i135p195-210.

CRAWFORD, Kate. **Atlas of AI. New haven and London**. Yale University Press: Yale, 2021.

CYBERSPACE ADMINISTRATION OF CHINA. **Projeto de regulamentação de algoritmos de IA**. Disponível em: http://www.cac.gov.cn/2021-08/25/c_1631480920680924.htm. Acesso em: 28 ago. 2021.

DA SILVA, José Antônio. **Curso de direito constitucional positivo**. 43 ed. rev. atual. São Paulo: Juspodivm, 2020.

DIGNUM, V. **Responsible artificial intelligence**: howtodevelop and use ai in a responsible way. Cham: Springer Netherlands, 2019.

FERRARI, Isabela; BECKER, Daniel. Direito à explicação e decisões automatizadas: reflexões sobre o princípio do contraditório. *In*: NUNES, Dierle; LUCON, Paulo Henrique dos Santos; WOLKART, Erick Navarro (Coords.). **Inteligência artificial e direito processual**: os impactos da virada tecnológica no direito processual. Salvador: Juspodvim, 2021, p. 277-303.

FLORIDI, L. *et al*. How to design ai for social good: seven essential factors. **Science and Engineering Ethics**, v. 26, n. 3, p. 1771-1796, 2020.
FLORIDI, L., SANDERS, J. W. On the morality of artificial agents. **Minds and Machines**, v. 14, n. 3, 2004, p. 349–379.

FOOD AND DRUG ADMINISTRATION. **Artificial intelligence and machine learning (AI/ML) software as a medical device action plan**. Disponível em: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence- and-machine-learning-software-medical-device. Acesso em: 21 set. 2021.

FRAZÃO, Ana; GOTTENAUER, Carlos. Black box face à opacidade algorítmica. *In*: BARBOSA, Mafalda Miranda *et. al*. (coord.). **Direito digital e inteligência artificial**: diálogos entre Brasil e Europa. São Paulo: Foco, 2021, p. 27-42.

FRIEDMAN, Batya; NISSENBAUM, Helen. Bias in computer systems. **ACM Transactions on Information Systems**. v. 14, n. 3, jul.,1996, p. 330-334.

FRIEDMAN, Batya; NISSENBAUM, Helen. Software agents and user autonomy. **AGENTS '97**: Proceedings of the first international conference on autonomous agents. v. 14, n. 3, fev. 1997, p. 466-469.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. Cambridge: MIT Press, 2016.

GUNKEL, D. **The machine question**: critical perspectives on AI, Robots, and Ethics. Cambridge/Massachutts: MIT Press, 2012.

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-reallyhappensand- why-its-so-hard-to-fix/. Acesso em: 7 ago. 2021.

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

KAUFMAN, Dora. Inteligência artificial e os desafios éticos: a restrita aplicabilidade dos princípios gerais para nortear o ecossistema de IA. **Paulus**: Revista de Comunicação da Fapcom. São Paulo, v. 5, n. 9, jan./jul. 2021a.

KAUFMAN, Dora. Equipes interdisciplinares: não basta "juntar campos", tem que "construir pontes". **Revista Época Negócios**, 2021. Disponível em: https:// epocanegocios.globo.com/colunas/IAgora/ noticia/2021/09/equipesinterdisciplinares- nao-basta- juntar-campos-tem-que-construir-pontes.html. Acesso em: 17 set. 2021b.

KAUFMAN, Dora. **Desmistificando a inteligência artificial**. BH: Autêntica, 2022.

KAUFMAN, Dora; COELHO, Alexandre. Regular a IA, mas sem precipitação. **Valor Econômico**, 2023. Disponível em: https://valor.globo.com/opiniao/coluna/regular-a-ia-mas-sem-precipitacao.ghtml. Acesso em: 7 out. 2023.

KUHN, T. S. **The structure of scientific revolutions**. Chicago: Chicago University Press,1970.

LEARNED-MILLER *et al*. **Facial recognition technologies in the wild**: a call for a federal office. White Paper, 2020. Disponível em: https:// www.ajl.org/federal-office-call/. Acesso em: 15 set. 2021.

LEE, Kai-Fu. **AI superpowers**: China, Silicon Vallley, and the New World Order. NY: Hougton Mifflin Harcourt, 2018.

LEONARDI, Marcel. **Tutela da privacidade na internet**. São Paulo: Saraiva, 2011.

LIAO, S. Matthew. **Ethics of artificial intelligence**. New York: Oxford University Press, 2020.

MAGRANI, E. **A internet das coisas**. Rio de Janeiro: FGV Editora, 2018.
MATWYSHYN, Andrea M. **The 'internet of bodies' is here**. Are courts and regulators ready? A network of smart devices attached to or implanted in bodies raises a host of legal and policy questions, 2018. Disponível em: https://www.wsj.com/articles/the-internet-of- bodies-is-here-are-courts-and-regulators-ready-1542039566. Acesso em: 20 ago. 2021.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data**: a revolution that will transform how we live, work, and think. Nova York: Houghton Mifflin Harcourt, 2013.

MENDES, Laura Schertel; MATTIUZZO, Marcela; FUJIMOTO, Mônica Tiemy. Discriminação algorítmica à luz da lei geral de proteção de dados. *In*: DONEDA, Danilo *et.al*. **Tratado de proteção de dados pessoais**. 2. reimp. Rio de Janeiro: Forense, 2021, p. 421-446.

NEFF, G.; NAGY, P. **Talking to bots**: symbiotic agency and the case of tay. v. 10, p. 4.915-4.931, 2016.

NUNES, D.; MARQUES, A. L. P. C. Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. **Revista de Processo**, v. 285, p. 421-447, 2018.

PASQUALE, F. **The black box society**: the secret algorithms that control money and information. Cambridge: Harvard University Press, 2015.

PEREZ-CRIADO, Caroline. **Invisible women**: data bias in a world designed for men. USA: Abrams Press, 2021.

SARLET, Ingo Wolfgang. Fundamentos constitucionais: o direito fundamental à proteção de dados. *In*: DONEDA, Danilo *et. al*. **Tratado de proteção de dados pessoais**. 2. reimp. Rio de Janeiro: Forense, 2021, p. 21-60.

SOARES, Flaviana Rampazzo. Levando os algoritmos a sério. *In*: BARBOSA, Mafalda Miranda *et. al*. **Direito digital e inteligência artificial**: diálogos entre Brasil e Europa. São Paulo: Foco, 2021, p. 43-64.

SUPREMO TRIBUNAL FEDERAL. **Agravo em Recurso Extraordinário (ARE) n. 779.543**. Disponível em: https://portal.stf.jus.br/processos/detalhe.asp?incidente=4483367. Acesso em: 25 ago. 2021.

SUPREMO TRIBUNAL FEDERAL. **Mandado de Segurança (MS) n. 25747/Santa Catarina (SC)**. Disponível em: https://portal.stf.jus.br/processos/detalhe.asp?incidente=2344488. Acesso em: 25 ago. 2021.

TEIXEIRA, Carlos Gustavo Poggio; DATYSGELD, Mark William. Os clientes diplomáticos e econômicos da espionagem digital estadunidense: análise das ações contra o Conselho de Segurança da ONU e a Petrobras, estudos internacionais. **Revista de Estudos Internacionais**, Belo Horizonte, v. 4, n.1, nov. 2016, p.71-87.

VILLANI, Cédric. **For a meaningful artificial intelligence**: towards a french and european strategy. aiforhumanity.fr, 2018. Disponível em: https://www.ai4eu.eu/news/mea- ningful-artificial-intelligencetowards--french-artificial-and-european-strategy. Acesso em: 10 abr. 2021.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. **Why a right to explanation of automated decision-making does not exist in the general data protection regulation**. v. *7,* n. 2, 2016, p. 76-99.

WALLACH, Wendel; ALLEN, Colin. **Moral machines**: teaching robots right from wrong. Oxford: Oxford University Press, 2009.

WANG, Y.; KOSINSKI, M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. **Journal of Personality and Social Psychology**, v. 114, n. 2, 2018, p. 246–257.

ZUBOFF, S. **The age of surveillance capitalism**. The fight for a human future at the new frontier of power. New York: Public Affairs, 2019.

Recebido em: 9-10-2022
Aprovado em: 10-10-2023

Dora Kaufman
Pós-doutorados, pela COPPE-UFRJ e pelo TIDD PUCSP. Doutora em mídias digitais pela ECA-USP com período na Université Paris – Sorbonne IV. Professora do Programa de Tecnologias da Inteligência e Design Digital da Faculdade de Ciências Exatas e Tecnologia da PUC SP. Autora de vários livros, entre eles "A inteligência artificial irá suplantar a inteligência humana?" e "Desmistificando a Inteligência Artificial". Colunista da Época Negócios. E-mail: kaufman1955@gmail.com

Tainá Junquilho
Doutora em Direito pela UnB. Mestre em direito pela UFES. Professora de direito, inovação e tecnologia IDP. Advogada. Pesquisadora ITsrio. E--mail: taina.aguiarj@gmail.com

Priscila Reis
Mestre em Tecnologias da Inteligência e Design Digital pelo TIDD-PUC/SP. Advogada especialista em Direito Digital. E-mail: priscila.reis9@gmail.com

Pontifícia Universidade Católica de São Paulo - PUC-SP
Rua Monte Alegre, 984,
Perdizes - São Paulo - SP
05014-901