"Falhas de IA" e a Intervenção Humana em Decisões Automatizadas: Parâmetros para a Legitimação pela Humanização

"AI Failures" and Human Intervention in Automated Decision-Making: Parameters for Legitimation through Humanization

MIRIAM WIMMER[1]
Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP/DF).

DANILO DONEDA[2]
Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP/SP).

RESUMO: Em um contexto em que diferentes países vêm reconhecendo um direito à intervenção humana face a decisões automatizadas, o artigo tem por objetivo investigar os elementos que podem atrair a necessidade de introdução de parâmetros humanos em processos de decisão impulsionados por sistemas de inteligência artificial. Assim, com base no método hipotético-dedutivo e a partir de pesquisa bibliográfica e documental, o artigo explora diferentes categorias de problemas que podem advir de decisões tomadas por sistemas de IA, concluindo que, em determinados casos, a necessidade de intervenção humana pode ser identificada não apenas com base em critérios de eficiência, mas também pode se constituir em um componente ético em si mesmo. Por outro lado, argumenta-se que determinados parâmetros de sistemas de IA, como o seu nível de transparência e auditabilidade, a explicabilidade das decisões, o seu baixo impacto potencial a direitos e garantias fundamentais e o grau de participação do próprio usuário do sistema na sua configuração e utilização, poderiam mitigar os riscos associados ao "déficit de humanidade" e assim proporcionar que a intervenção humana seja modulada em diferentes níveis de intensidade, mantendo-se o atendimento aos requisitos éticos de decisões legítimas, confiáveis, justas e cognoscíveis, por seres humanos, em seus principais elementos.

PALAVRAS-CHAVE: Inteligência artificial; decisões automatizadas; intervenção humana; ética.

ABSTRACT: In a context in which different countries have come to recognize a right to human intervention on automated decision-making, this paper aims to investigate the elements that may

1 Orcid: https://orcid.org/0000-0001-9210-6651.
2 Orcid: https://orcid.org/0000-0001-9535-3586.

attract the need for human parameters in decision-making supported by artificial intelligence. Thus, based on the hypothetical-deductive method and on bibliographic and documentary research, the article explores different categories of problems that may result from AI decision-making processes, concluding that the need for human intervention can be identified not only based on efficiency criteria, but may also, in certain cases, be considered an ethical component in itself. On the other hand, the paper argues that certain parameters of AI systems, such as their level of transparency, auditability and explainability, their potential low impact for fundamental rights and guarantees as well as the degree of user participation in their configuration and use, could mitigate the risks associated with the "lack of humanity". These parameters could therefore enable the modulation of human participation at different levels of intensity, while meeting the ethical requirements of legitimate, trustworthy and fair decisions, that can be understood by humans in their main elements.

KEYWORDS: Artificial intelligence; automated decisions; human intervention; ethics.

INTRODUÇÃO

Em outubro de 2020, o time escocês de futebol *Scottish Inverness Caledonian Thistle* anunciou, por meio de sua página na internet, que passaria a utilizar, em seu estádio, um novo sistema de câmeras com tecnologia de reconhecimento de imagens baseada em inteligência artificial – IA. Diante das restrições sanitárias que impediam torcedores de comparecer ao estádio, o sistema de filmagem com a tecnologia *ball tracking* permitiria que os torcedores acompanhassem as partidas em tempo real, por meio de uma plataforma de *streaming*, e visualizassem os melhores ângulos das jogadas. Para a frustração dos torcedores, entretanto, a novidade revelou-se um fiasco: as "câmeras inteligentes" reiteradamente confundiram a bola de futebol com a cabeça do árbitro, um senhor completamente calvo (ICTFC Media, 2020).

Tal episódio ilustra, de um lado, a crescente ubiquidade da inteligência artificial, presente em domínios cada vez mais amplos da vida cotidiana; de outro lado, chama atenção para o fato de que tais sistemas frequentemente geram resultados que ensejam preocupações, seja em razão do cometimento de erros objetivos, seja pela existência de fatores que tornem estes resultados questionáveis por apresentar parcialidade, viés, opacidade, tendências discriminatórias, dentre outros aspectos problemáticos. Uma vasta literatura acadêmica explora as consequências jurídicas e as implica-

ções éticas destes fatores no processo decisório de sistemas de inteligência artificial (Pasquale, 2015; O'Neil, 2017; Birhane, 2021).

De fato, ao atribuir a entes automatizados a realização de tarefas que tenham como consequência a participação, total ou parcial, em um processo decisório com impactos relevantes, surge uma demanda clara pela possibilidade de escrutínio dos diversos elementos deste processo decisório. Ainda que sistemas de inteligência artificial possam apresentar enormes vantagens de escala ou mesmo precisão em relação a decisões humanas em diversas searas, sua legitimidade não deriva meramente destas métricas – que podem, inclusive, ser ajustadas mediante critérios utilitaristas ou mesmo instrumentalizadas para determinadas finalidades. Assim, a demanda por uma espécie de instância na qual estas possíveis "falhas" possam ser verificadas e avaliadas acaba, dada a natureza destas decisões, por ser componente ínsito à sua própria natureza e fator imprescindível para a legitimação de seu uso.

Ao mesmo tempo em que se observa a ampliação dos espaços de interação e de decisão mediados por sistemas de IA, é possível verificar, internacionalmente, uma tendência a enfrentar e mitigar os riscos associados a falhas de sistemas algorítmicos. Tal tendência é bastante visível na afluência da pesquisa sobre o tema e no desenvolvimento de parâmetros éticos e normas deontológicas em geral sobre o emprego de sistemas de IA (Baxter, 2021), bem como em diversas estratégias para o uso de IA desenvolvidas por países e organizações, alguns chegando mesmo a flertar com a regulação da matéria[3].

Há, no entanto, uma estratégia observável em normas de proteção de dados pessoais de diferentes países, que é a do estabelecimento de um direito pelo qual indivíduos não estejam sujeitos a decisões baseadas unicamente em processos automatizados, invocando-se, em muitos casos, um direito à intervenção humana, que pode ser materializado, dentre outros mecanismos, por um direito à revisão humana de decisões automatizadas ou por um direito a uma explicação, compreensível por seres humanos, quanto aos seus principais critérios e parâmetros. Apesar das nuances e variações na enunciação desses direitos em diferentes ordenamentos jurídicos, é possível compreender que seu reconhecimento decorre da constatação de

3 No Brasil, no momento da conclusão deste artigo, tramita no Congresso Nacional o Projeto de Lei nº 21/2020, que estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil.

que decisões baseadas total ou predominantemente em sistemas automatizados podem ser consideradas como falhas, necessitando de um tratamento normativo que, antes mesmo de que abordagens específicas sobre IA sejam definidas no patamar regulatório, já estão presentes e, em alguns casos, sedimentadas dentro de modelos regulatórios de proteção de dados.

Diante de tal cenário, o presente artigo tem por objetivo, com base no método hipotético-dedutivo e a partir de pesquisa bibliográfica e documental, investigar os tipos de falhas em que podem incorrer sistemas de IA, com vistas a sugerir critérios que podem atrair, especificamente, a necessidade de introdução de parâmetros humanos em processos de decisão realizados por tais sistemas, assim como parâmetros que podem mitigar os riscos advindos do "déficit de humanidade".

Para os fins deste estudo, adota-se o recorte proposto por Mittelstadt *et al.* (2016) e utilizado também por Tsamados *et al.* (2020), cuja análise é voltada para os algoritmos não apenas como construções matemáticas abstratas, porém por meio de uma perspectiva funcional, sobretudo em vista da forma de sua implementação (artefatos, tecnologias, programas) e configuração (aplicações). Assim, a análise que será empreendida sobre algoritmos e sistemas de IA os considerará enquanto instrumentos usados para converter dados em evidências quanto a um determinado resultado/cenário, que, subsequentemente, é usado para desencadear e motivar uma ação que pode ter consequências que ensejem uma determinada apreciação ética. Será dedicada maior atenção àqueles sistemas que podem tomar ou recomendar decisões, dando apoio à análise humana ou substituindo-a em determinados casos, com base em processos complexos que desafiam a compreensão humana – especialmente os sistemas de inteligência artificial baseados em aprendizado por máquina.

Tem-se como hipótese que, uma vez que as falhas de sistemas de IA podem decorrer não apenas de deficiências formais (desde meros erros de programação ou a utilização de bases de dados de treinamento inadequadas), mas também da impropriedade na tomada de decisões que dependem de percepções, valores ou comportamentos que são, a princípio, incognoscíveis ou impossíveis de serem metrificados e trabalhados por máquinas, a "humanização" de uma decisão pode tornar-se necessária como um componente ético em si mesmo, quando não também sob uma perspectiva de eficiência, baseada nas taxas de erros e acertos de determinado sistema.

Assim, sugere-se que, para além dos parâmetros para atrair a intervenção humana já previstos em normas e declarações internacionais – formulados de maneira ainda muito vaga, como se verá –, devem ser considerados elementos como (i) os riscos e consequências atuais e futuros gerados para os indivíduos e grupos afetados, abrangendo elementos como impactos sobre direitos fundamentais, riscos de discriminação e possibilidade de reversão dos efeitos da decisão, e (ii) a natureza da decisão, em particular no que se refere a decisões em que os juízos de "certo" e "errado" são subjetivos ou em que a decisão deve depender de percepções/valores a princípio incognoscíveis por máquinas. Desta forma, em decisões automatizadas, tanto o seu resultado (as consequências da decisão) quanto seus aspectos procedimentais (a sua natureza) podem ensejar a necessidade do elemento remedial que é a intervenção humana.

Necessário mencionar que, ainda que o direito à intervenção humana (aqui abordado a partir da discussão sobre os direitos à revisão e à explicação de decisões automatizadas) normalmente seja invocado em momento posterior à produção dos efeitos de um sistema automatizado, este pode se materializar por meio da participação de agentes humanos nos processos de tomada de decisão algorítmica de distintas formas e em diferentes momentos do ciclo de vida do produto ou da aplicação em questão. Assim, este artigo não se propõe a identificar, de maneira conclusiva, em qual determinado momento ou de que maneira específica a intervenção humana deva ocorrer para que seja considerada significativa e relevante.

Por outro lado, argumenta-se que determinados parâmetros de sistemas de IA – tais como o grau de transparência, a auditabilidade do sistema, a explicabilidade das decisões tomadas, o seu baixo impacto potencial a direitos e garantias fundamentais, o grau de participação humana exigido ou facultado na sua utilização, entre outros – poderiam mitigar os riscos associados ao "déficit de humanidade". Sugere-se, assim, que a intervenção humana pode ser modulada em diferentes níveis de intensidade, assumindo formas mais brandas, desde que outros mecanismos assegurem o atendimento aos requisitos éticos de decisões legítimas, confiáveis, justas e cognoscíveis em seus elementos, ainda que não necessariamente tomadas por seres humanos.

O artigo está estruturado da forma a seguir descrita. Inicialmente, serão apresentados elementos que permitem compreender as diferentes categorias de problemas associados a decisões por sistemas de IA, abrangendo tanto aspectos relacionados à forma e ao processo de tomada de decisão

por tais sistemas como também aspectos relacionados aos problemas éticos que resultam do uso de tais "decisões" algorítmicas como apoio para decisões humanas ou em substituição a elas. Passa-se, em seguida, a examinar a forma pela qual diferentes jurisdições e organizações internacionais têm estabelecido direitos à intervenção humana em decisões algorítmicas, em particular por meio do direito à revisão de decisões automatizadas e direitos à explicação. Para concluir, com base na discussão apresentada, o artigo sugere parâmetros que podem ser utilizados para identificar a necessidade de intervenção humana, assim como elementos que podem viabilizar a mitigação de tal necessidade ou a modulação de sua intensidade.

1 CATEGORIAS DE PROBLEMAS EM DECISÕES DE IA

As técnicas de decisão algorítmica, cuja relevância, utilidade e ubiquidade no mundo contemporâneo são sempre mais claras, têm sido alvo de escrutínio por diversas razões. Assim, diversos esforços têm sido empreendidos para categorizar os tipos de preocupações éticas associadas a decisões algorítmicas.

Mittelstadt e outros (2016), por exemplo, propõem um mapa conceitual baseado em seis categorias de preocupações éticas ligadas a algoritmos, sendo três de natureza *epistêmica* e duas de natureza *normativa*. No primeiro grupo de preocupações, associado à *qualidade das evidências* produzidas por decisões algorítmicas, encontram-se as seguintes categorias: (i) evidências inconclusivas, que conduzem a problemas de apofenia (*i.e.*, os algoritmos indicam correlações e conduzem à identificação de padrões onde estes não existem verdadeiramente, em fenômeno por vezes também nominado de "correlação espúria"); (ii) evidências inescrutáveis (*i.e.*, opacas ou ininteligíveis); e (iii) evidências mal-orientadas (resultantes da baixa qualidade dos dados de entrada). Quanto às preocupações de natureza *normativa e ética*, os autores indicam os seguintes problemas: (iv) resultados injustos ou discriminatórios e (v) efeitos transformativos, associados à forma como os algoritmos afetam o modo de compreender o mundo, assim como sua organização social e política, com impactos sobre a autonomia humana. Por fim, agravado pelos problemas anteriormente citados, os autores apontam, ainda, para o problema da (vi) dificuldade de *rastreabilidade* das

causas de eventual dano provocado e, consequentemente, de responsabilização dos indivíduos ou organizações envolvidos[4].

Zarsky (2016), por sua vez, em abordagem um pouco mais genérica, identifica duas principais categorias de problemas, ambas relacionadas à natureza opaca e automatizada dos algoritmos, nem sempre passíveis de mitigação por meio de meras medidas de transparência: (i) problemas associados a *ineficiências*, que podem resultar tanto de imprecisões nas bases de dados como também de erros na predição de comportamentos individuais, dada a imprevisibilidade do comportamento humano; e (ii) problemas associados a *injustiças* (transferência injusta de riqueza entre grupos sociais, tratamento discriminatório ou violações à autonomia individual).

Apesar das diferenças nas categorizações acima descritas, é possível visualizar que, em ambas as propostas, há a identificação de duas questões de fundo. A primeira reside na circunstância de que sistemas de IA são intrinsecamente propensos a falhas, particularmente por tomarem decisões com base em métodos e técnicas que não são precisos e nem neutros[5], o que pode levar a resultados inaceitáveis a ponto de serem reconhecidos como falhas (problemas de ineficiência/problemas epistêmicos). Já a segunda questão consiste no fato de que a utilização de "decisões" de IA como apoio para decisões humanas ou em substituição a decisões humanas pode suscitar importantes questionamentos relacionados ao campo da ética, da justiça e da autonomia humana, ensejando igualmente algum tipo de intervenção.

Para os fins da discussão proposta neste artigo – a avaliação de critérios que podem atrair a necessidade de introdução de intervenção humana em processos de decisão automatizados –, entende-se relevante explorar, em maior profundidade, essas duas dimensões.

No que tange à primeira questão, é importante frisar que, ainda que complexos, são já conhecidos e amplamente explorados pela literatura os problemas associados à incorporação de vieses culturais e preconceitos raciais, de gênero e outros em sistemas de aprendizado por máquinas (*machine learning*), que levam a situações em que pessoas integrantes de determi-

4 Para um aprofundamento da discussão com base no mapa conceitual proposto por Mittelstadt *et al.* (2016), v. Tsamados *et al.* (2020).

5 Ao mesmo tempo em que legislações de proteção de dados começavam a se desenvolver, na década de 1970, existia também uma forte crença na objetividade, neutralidade intrínseca e eficiência de sistemas decisionais automatizados, particularmente nos Estados Unidos (Jones, 2017).

nados grupos sociais e étnicos sejam sistematicamente prejudicados por sistemas automatizados de decisão (Pasquale, 2015; O'Neil, 2017). Por serem desenvolvidos com base em dados históricos, que refletem de maneira desigual a diversidade da população ou que incorporam determinados vieses já presentes na sociedade, sistemas desse tipo frequentemente acabam por reproduzir padrões sociais e históricos de injustiça ou discriminação (Edwards; Veale, 2018); assim, as tentativas de agrupar, classificar e predizer o comportamento humano com base nessas técnicas têm se revelado, em alguns casos, bastante problemáticas. Para Birhane (2021), por exemplo, ao tentar impor ordem e identificar padrões no comportamento humano, ferramentas de IA acabam "forçando a determinabilidade, limitando possibilidades e, dessa forma, criando um mundo que se assemelha ao passado", reforçando problemas de discriminação e de injustiça, não raro com consequências particularmente cruéis para grupos marginalizados.

As diferentes abordagens apresentadas para lidar com esses tipos de problemas têm, em muitos casos, focado em ideias como transparência e explicabilidade, de modo a permitir maior visibilidade sobre os critérios que orientam decisões algorítmicas e, assim, viabilizar o seu controle e a sua correção. Tais abordagens partem, em muitos casos, do pressuposto de que é possível aprimorar e aperfeiçoar os sistemas de IA existentes de modo a diminuir as taxas de erros e eliminar eventuais vieses. Com efeito, para determinada corrente de pensamento que, de certa forma, parece carregar o legado das abordagens que propugnavam pela eficácia e objetividade das decisões tomadas por máquina, problemas dessa natureza poderiam ser facilmente corrigidos, caracterizando-se, de maneira explícita, os comportamentos e os resultados aceitáveis e viabilizando-se, assim, que sistemas de aprendizado por máquinas pudessem aprender a desconsiderar fatores discriminatórios de maneira mais efetiva do que humanos. Segundo esse raciocínio, as decisões algorítmicas tenderiam efetivamente a ter maior taxa de acertos do que os próprios seres humanos, particularmente em razão da eliminação de vieses cognitivos de tomadores de decisão humanos (Sunstein, 2018; Kahneman; Sibony; Sunstein, 2021).

De todo modo, a despeito do debate sobre a sua eficácia ou objetividade, é preciso reconhecer que o fato de que as máquinas sejam hoje capazes de realizar tarefas que normalmente são associadas a elevados níveis de discernimento e compreensão humana não significa que os computadores efetivamente possuam discernimento ou compreensão ao realizá-las (Russel; Norvig, 2010, p. 1022) ou, mais ainda, que o façam sob a pers-

pectiva de uma atuação que se possa dizer consciente, conforme veremos. A título de exemplo, muito embora sistemas de IA já sejam hoje capazes, por exemplo, de compor músicas a partir do aprendizado sobre elementos obtidos de obras consagradas, o processo de composição definitivamente é distinto daquele percorrido por um ser humano, que despeja em tal tarefa suas emoções, seu espírito criativo e sua sensibilidade artística – sua humanidade, em síntese.

Assim é que muitas das falhas em que incorrem sistemas de IA – como os sistemas de recomendação que sugerem conteúdos inapropriados, *chatbots* que fornecem respostas absurdas a perguntas formuladas pelo usuário, traduções automatizadas desprovidas de qualquer sentido e sistemas de reconhecimento facial que confundem pessoas com animais – decorrem da ausência de habilidades humanas básicas de percepção do contexto e da cultura no qual estão inseridos[6]. Nessa linha, a história narrada no início deste artigo, sobre um sistema de IA incapaz de diferenciar uma bola de futebol de uma cabeça calva, é um exemplo do chamado Paradoxo de Moravec (Moravec, 1990), segundo o qual habilidades cognitivas que requerem raciocínio lógico, que, no imaginário, costumam ser associadas a um nível elevado de inteligência, são mais facilmente simuladas em um computador do que habilidades simples, como a percepção ou a mobilidade. Desse modo, é mais fácil ensinar um computador a jogar xadrez ou a demonstrar teoremas matemáticos do que ensiná-lo a reconhecer nuances em um tom de voz ou a manipular objetos.

Seguindo essa linha de raciocínio, têm sido levantadas, no campo da filosofia da informação, uma série de objeções à ideia de que sistemas de IA seriam, algum dia, capazes de raciocinar nos mesmos moldes que um ser humano, debate frequentemente apresentado como uma oposição entre a hipótese da IA Fraca, que assevera que as máquinas são capazes apenas de *simular* o pensamento humano, ou seja, agir como se fossem inteligentes, e a hipótese da IA Forte, que afirma que as máquinas efetivamente seriam ca-

6 Para uma interessante descrição de casos históricos de falhas de inteligência artificial, categorizando-as conforme suas consequências, sua intencionalidade, sua evitabilidade e seu estágio de introdução no ciclo de vida do produto, sugere-se a leitura de Scott e Yampolski (2019). Por outro lado, um relato dos sucessos de IA pode ser encontrado em Ganascia (2019), que argumenta que as críticas à IA resultam, em muitos casos, da incompreensão quanto à natureza e aos objetivos da tecnologia, que efetivamente não serve para reproduzir a consciência humana. Para o autor, a IA se traduz em uma disciplina científica que estuda as formas pelas quais a inteligência pode ser decomposta de modo a reproduzir seus diferentes aspectos em computadores. Assim, segundo Ganascia, os falhas de IA resultam, em muitos casos, não de problemas técnicos nos programas de IA, mas residem na sua inadequação social, ou seja, na sua incapacidade de responder às exigências do ambiente social no qual são utilizados.

pazes de pensar e ter autoconsciência, da mesma forma que seres humanos (Russel; Norvig, 2010, p. 1020). As diversas objeções filosóficas apresentadas quanto à ideia de que a existência de sistemas de IA Forte seria possível (Fjelland, 2020) repousam sobre conceitos complexos como consciência, intencionalidade, compreensão e cognição, que não são passíveis de serem explorados em profundidade no contexto deste artigo[7].

Tal discussão, entretanto, permite compreender que uma segunda categoria de problemas associados a decisões de IA, que incluímos no espectro de "falhas", está relacionada não a questões epistêmicas, como deficiências decorrentes de problemas de programação ou a insuficiências nas bases de dados de treinamento, mas sim aos problemas oriundos da delegação a um sistema automatizado da responsabilidade pela tomada de decisões que devem depender de percepções, valores ou comportamentos a princípio não passíveis de conhecimento por máquinas e que consistem basicamente na pletora de emoções e estados de espírito, como bondade, compaixão ou senso de humor, bem como as decisões e tomadas de posição em que os juízos de "certo" e "errado" são subjetivos. Conforme relatam Awad *et al.* (2018), tomando-se como exemplo o conhecido cenário do carro autônomo que, na iminência de um acidente fatal, se vê diante de diferentes possibilidades de decisão – que resultariam na morte do ocupante do carro, de uma criança ou de uma pessoa idosa –, é possível observar que a resposta considerada aceitável pode ter significativas variações entre países, com forte dependência de cultura e de religião – evidenciando a dificuldade de trabalharmos com soluções consideradas mais eficientes, ainda que de ponto de vista utilitarista, que sejam legítimas a despeito da diversidade dos valores culturais envolvidos.

Trata-se de ponto enfatizado, sob outro prisma, também por Birhane (2021), que, ao salientar a "incapacidade de automatizar a ambiguidade", chama atenção para o fato de que seres humanos são sistemas adaptativos complexos, dotados de indeterminabilidade e de uma inerente imprevisi-

7 A título de ilustração, e sem pretensão de esgotar a temática, vale notar que importante campo atual de pesquisa em IA diz respeito à ideia de cognição incorporada, que refuta a ideia de que a cognição poderia se dar por agentes lógicos desconectados de um corpo físico. Segundo essa linha, a cognição é um processo que não funciona de maneira separada de sentidos e estados corporais, que se dão sempre dentro de determinado contexto. Assim, fazendo referência aos trabalhos de Dreyfus (1986) e Clark (1998), Russel & Norvig (2010:1026) apontam que um agente cujo entendimento do termo "cachorro" provém somente de um conjunto limitado de enunciados lógicos (como, por exemplo, "cachorro (x) = mamífero (x)") nitidamente tem uma compreensão mais limitada do animal do que aquela de um agente humano que já observou um cachorro, brincou com o animal e foi lambido por ele.

bilidade, o que contrasta fortemente com a lógica que costuma orientar os sistemas de aprendizado por máquina. Para a autora, as pessoas não podem ser compreendidas fora de seu ambiente (e das normas sociais e das assimetrias de poder ali existentes), de suas trajetórias históricas e de seus valores morais e políticos, que representam elementos cruciais para a sua própria identidade.

Assim é que há domínios que suscitam outro tipo de questionamento: ainda que determinado sistema autônomo atinja um patamar considerado aceitável de taxas de erros e acertos, seria legítimo, eticamente, delegar certos tipos de decisão integralmente a sistemas automatizados, sem intervenção humana relevante?

No campo dos conflitos armados internacionais, por exemplo, sujeitos a regras de direito internacional humanitário concebidas de maneira marcadamente antropocêntrica, debate-se, à luz da chamada "Cláusula de Martens"[8], a própria aceitabilidade de que uma decisão de vida ou morte seja completamente delegada a máquinas, incapazes de mostrar compaixão, de sentir empatia ou de reconhecer a dignidade humana (Wimmer, 2021). Nesse cenário, o que está em discussão, conforme elucida Asaro (2012), não é apenas investigar se um computador, uma máquina ou um processo automatizado são capazes de tomar decisões de vida ou morte e atingir um nível de desempenho considerado aceitável à luz dos preceitos do Direito Internacional Humanitário; mas, sim, se é eticamente aceitável que o ser humano esteja tão distanciado do processo de identificação do alvo a ponto de delegar quase que integralmente a máquinas esse tipo de decisão. Em outras palavras: em determinadas ocasiões, seria legítimo fundamentar decisões em elementos exclusivamente utilitaristas e normativos, extirpando a participação de uma avaliação humana – ainda que esta seja em sentido contraposto?

Debates semelhantes poderiam ser travados com relação a sistemas automatizados operantes em outras searas em que podem existir significativos impactos sobre direitos fundamentais, como na definição de tratamentos médicos ou alocação de leitos em hospitais, no estabelecimento de penas

8 A chamada Cláusula de Martens, também conhecida como o Princípio da Humanidade, encontra-se presente em tratados internacionais e também no Protocolo Adicional às Convenções de Genebra, que estabelece, em seu art. 1º, que as pessoas civis e os combatentes permanecem sob a proteção e o domínio dos princípios do Direito Internacional derivado dos costumes estabelecidos, dos princípios de humanidade e dos ditames da consciência pública.

privativas de liberdade, em decisões críticas de segurança em veículos autônomos e até mesmo, em certas circunstâncias, nas decisões algorítmicas relacionadas à moderação de conteúdos em redes sociais.

É importante ainda registrar que o debate, nestes casos difíceis, não gira apenas em torno da pergunta se, em que circunstâncias e em quais momentos deve ou não existir (ou haver a possibilidade de requerer) a intervenção humana; trata-se, também, de definir que tipo de intervenção humana seria qualitativamente apta a suprir o "déficit de humanidade" em tomadas de decisões sobre as quais recai uma elevada carga moral ou que produzem um significativo impacto sobre direitos fundamentais.

Embora diferentes estratégias jurídicas e regulatórias venham sendo debatidas para lidar com tais desafios, é possível identificar, conforme mencionado anteriormente, uma tendência, em instrumentos internacionais e em legislações de proteção de dados pessoais de variados países, de previsão de direitos associados à intervenção humana em determinados tipos de decisões automatizadas. Diferentes modalidades de intervenção humana previstas em outras jurisdições e os critérios usados para desencadear o exercício de tais direitos serão, a seguir, brevemente examinados.

2 MODALIDADES DE INTERVENÇÃO HUMANA EM DIFERENTES JURISDIÇÕES

Pelo fato de que as propostas de regulamentação abrangente da inteligência artificial são ainda embrionárias, é no campo da proteção de dados pessoais que se pode observar, mais claramente, tentativas de promover a introdução de elementos "humanos" em decisões tomadas automaticamente. A ideia de que, pelo menos em alguns casos, indivíduos não devem estar sujeitos a decisões tomadas unicamente com base em algoritmos tem sido endereçada por meio de diferentes estratégias jurídicas, com destaque para o reconhecimento de direitos (i) à explicação sobre as características, critérios e consequências de decisões automatizadas e (ii) à revisão de tais decisões.

Os direitos à explicação e à revisão são, aqui, tomados em um vetor genérico. Há, por exemplo, tanto leituras como proposições do primeiro que ora se assemelham mais a um direito à explicação, ora mais a um direito à informação. E, ainda, há uma interessante ressonância entre ambos: muito embora o direito à explicação não pressuponha necessariamente a participação humana, trata-se de ferramenta que permite que um ser humano compreenda e exerça controle sobre os principais aspectos relacionados

a decisões automatizadas e pode também representar elemento essencial para o próprio exercício do direito de revisão. Da mesma forma, o direito à revisão, tomado em toda a sua tessitura, implica o reconhecimento de um inafastável componente informativo ligado ao próprio direito à explicação, visto que uma revisão somente se legitima quando é capaz de explicitar os critérios e vetores que a inspiraram – o que nada mais é que um elemento de "explicação" sobre a sua *ratio*.

Verifica-se, ademais, o surgimento de crescente consenso acerca do tema em organismos internacionais. Observe-se, por exemplo, que, muito embora as Diretrizes da OCDE sobre privacidade (2013) não tenham abordado o tema da participação humana em decisões automatizadas, essa ideia é claramente enunciada na Recomendação do Conselho da OCDE sobre Inteligência Artificial (2019), que estabelece a necessidade de que organizações e indivíduos envolvidos com o desenvolvimento e a utilização de sistemas de IA implementem mecanismos e salvaguardas, como a capacidade para a determinação humana, que sejam apropriados ao contexto e consistentes com o estado da arte, de modo a assegurar o respeito à lei, aos direitos humanos e aos valores democráticos[9]. Ao abordar o tema da transparência e da explicabilidade, a Recomendação indica ainda a necessidade de que sejam fornecidas informações relevantes, apropriadas ao contexto e consistentes com o estado da arte, que permitam que aqueles afetados por um sistema de IA possam compreender o resultado e contestá-lo, com base em informações simples e facilmente compreensíveis sobre os fatores e a lógica que serviram de base para a predição, recomendação ou decisão.

Também a versão modernizada da Convenção para a Proteção das Pessoas relativamente ao Tratamento Automatizado de Dados de Caráter Pessoal, conhecida como Convenção 108+, explicitamente prevê direitos relacionados à participação humana em decisões automatizadas[10].

Mais recentemente, em novembro de 2021, a Conferência Geral da Unesco aprovou a Recomendação sobre a Ética da Inteligência Artificial, na qual se enuncia que, em cenários em que as decisões de sistemas de IA pos-

9 Os valores e direitos citados são a liberdade, a dignidade, a autonomia, a privacidade e a proteção de dados, a não discriminação, a igualdade, a diversidade, a equidade, a justiça social e os direitos trabalhistas.

10 "*Article 9 – Rights of the data subject. 1. Every individual shall have a right: a. not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration; [...] 2. Paragraph 1.a shall not apply if the decision is authorised by a law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights, freedoms and legitimate interests*."

sam produzir impactos irreversíveis ou de difícil reversão, ou que envolvam decisões de vida ou morte, deve existir determinação humana final. Ademais, a Recomendação indica que a supervisão humana compreende não apenas a supervisão humana individual, mas também a supervisão pública, conforme apropriado. A Recomendação afirma, por fim, que um sistema de IA nunca pode substituir a responsabilidade e *accountability* humanos em última instância (Unesco, 2021).

Não há dúvidas, entretanto, de que os direitos que decorrem da afirmação da necessidade de intervenção humana carecem de maior densidade conceitual, de modo que existe ainda ampla margem para disputas interpretativas sobre seu conteúdo e sobre a sua concreta forma de fruição em diferentes jurisdições (Jones, 2017). Merece destaque, em particular, a controvérsia acerca da própria viabilidade de exercício de um direito à explicação face à complexidade e opacidade dos modelos usados por sistemas de IA para chegarem a determinadas decisões.

De modo a exemplificar duas abordagens relevantes para a presente discussão, passa-se a examinar como o tema tem sido abordado na Europa e no Brasil.

2.1 Europa

O Regulamento Europeu de Proteção de Dados Pessoais – RGPD[11] estabelece, em seu art. 22, que:

> O titular dos dados tem o direito de não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, que produza efeitos na sua esfera jurídica ou que o afete significativamente de forma similar.

Observa-se, portanto, que o RGPD tem como ponto de partida uma vedação à tomada de decisões exclusivamente automatizadas com efeitos jurídicos ou similarmente significativos. As exceções a tal regra são definidas também pelo art. 22, admitindo-se as decisões exclusivamente automatizadas nos seguintes casos: (i) quando o tratamento for necessário para a execução ou celebração de um contrato; (ii) quando o tratamento for autorizado pelo direito da União ou do Estado-Membro, sendo necessário,

11 Regulamento (UE) nº 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva nº 95/46/CE (Regulamento Geral sobre a Proteção de Dados).

entretanto, que a legislação preveja medidas adequadas para salvaguardar os direitos e liberdades e os legítimos interesses do titular de dado; ou (iii) quando o tratamento for baseado no consentimento explícito do titular de dados.

Ademais, ainda que o titular de dados consinta com o tratamento automatizado ou que este seja necessário para a celebração de um contrato, conforme hipóteses (i) e (iii) *supra*, o Regulamento determina a necessidade de estabelecimento, pelo controlador, de "medidas adequadas para salvaguardar os direitos e liberdades e legítimos interesses do titular dos dados, designadamente o direito de, pelo menos, *obter intervenção humana por parte do responsável*, manifestar o seu ponto de vista e contestar a decisão".

O art. 22 do RGPD tem sido objeto de intensa discussão, notadamente no que diz respeito à existência e à extensão de um possível "direito à explicação" que dele decorreria. Conforme relatam Souza, Perrone e Magrani (2021), a controvérsia resulta do fato de que, muito embora o Considerando 71 do RGPD mencione o direito do titular de dados de obter uma explicação sobre a decisão automatizada e contestá-la[12], tal previsão acabou por não ser explicitada no corpo do art. 22 do Regulamento. Ao mesmo tempo, o art. 15 do RGPD estabelece que, no caso de decisões automatizadas, o titular tem o direito de obter informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento. Tais circunstâncias conduziram a uma ampla gama de interpretações quanto ao escopo do dispositivo, como a posição de Wachter, Mittelstadt e Floridi (2017) no sentido de que o GDPR não estabelece um amplo direito à explicação, mas um limitado direito de ser informado quanto às funcionalidades de um sistema de decisão automatizado; e o posicionamento no sentido contrário, pela concretude do direito à explicação no ordenamento europeu, esposada por Selbst e Powles (2017), que derivam este direito diretamente da previsão do RGPD quanto ao direito a uma informação significativa (*meaningful information*) acerca da lógica utilizada na decisão e suas consequências.

Edwards e Veale (2018) descrevem outras polêmicas que têm cercado a interpretação do art. 22 do RGPD. O que seria uma "decisão" para fins

12 Confira-se: "Em qualquer dos casos, tal tratamento deverá ser acompanhado das garantias adequadas, que deverão incluir a informação específica ao titular dos dados e o direito de obter a intervenção humana, de manifestar o seu ponto de vista, *de obter uma explicação* sobre a decisão tomada na sequência dessa avaliação e de contestar a decisão".

do RGPD? Esta incluiria, por exemplo, o envio de publicidade dirigida, que poderia ser facilmente ignorada pelo titular? No que consistiriam os citados "efeitos na esfera jurídica" ou outros efeitos que afetem significativamente o titular? E, por fim: o que seria uma decisão tomada "exclusivamente" com base no tratamento automatizado, e que nível de participação humana afastaria a incidência da norma?

Ao analisar o art. 22 do RGPD para fins de estabelecer orientações sobre decisões automatizadas e a definição de perfis, o Grupo de Trabalho do Artigo 29 para a Proteção de Dados (WP29, 2018) acabou por fixar uma interpretação ampla quanto ao seu escopo, indicando que decisões tomadas "exclusivamente com base" no tratamento automatizado são aquelas em que não há uma supervisão humana relevante. Assim, segundo a interpretação do grupo, o controlador não pode se eximir da incidência do artigo fabricando uma intervenção humana meramente simbólica; esta deve ser realizada "por alguém com autoridade e competência para alterar a decisão e que, no âmbito da análise, deverá tomar em consideração todos os dados pertinentes"[13].

Ao tempo em que reconheceu que o art. 22 abrange apenas as situações em que há impactos graves para o titular de dados, o Grupo de Trabalho do Artigo 29 também interpretou, de maneira abrangente, a expressão "que produza efeitos na sua esfera jurídica ou que o afete significativamente de forma similar". Para o colegiado, quando o RGPD faz referência a "efeitos na esfera jurídica", é preciso que a decisão em questão afete concretamente direitos de alguém[14]. Por outro lado, segundo o grupo de trabalho, quando a norma introduz a expressão "ou que o afete significativamente de forma similar", há uma abertura do âmbito de incidência da norma, que passa a ser aplicável quando os efeitos de uma decisão automatizada sejam suficientemente grandes ou importantes para merecerem atenção – vale dizer, quando a decisão puder: (i) afetar, de maneira significativa, as circunstâncias, o comportamento ou as escolhas das pessoas em causa;

13 Na mesma linha, a autoridade de proteção de dados britânica indica que a regra em questão se refere a situações em que não há influência humana sobre o processo decisório. Assim, um processo pode ser considerado totalmente automatizado se um ser humano apenas inserir os dados a serem processados, sendo a decisão tomada por uma máquina. Um processo não será considerado totalmente automatizado se um ser humano avaliar e interpretar os resultados de uma decisão automatizada antes de aplicá-la a um indivíduo (ICO, 2018).

14 Por meio, por exemplo, da limitação à liberdade de associação, ao direito de voto ou à possibilidade de mover ações judiciais; ou quando forem afetados o estatuto jurídico de uma pessoa ou os seus direitos no âmbito de um contrato.

(ii) provocar um impacto prolongado ou permanente no titular dos dados; ou (iii) ensejar a exclusão ou discriminação das pessoas.

O art. 22.2 (b) do RGPD indica a possibilidade de que os Estados-Membros definam, por lei, hipóteses autorizativas de decisões exclusivamente automatizadas, sendo necessário, entretanto, estabelecer medidas adequadas para salvaguardar os direitos e liberdades e os legítimos interesses dos titulares de dados. Como nota Malgieri (2019), é possível visualizar abordagens bastante heterogêneas entre os países europeus na implementação da norma; para os fins deste artigo, interessa chamar atenção para duas das quatro abordagens descritas pelo autor. A primeira, caracterizada como "procedimental", foi adotada pelo Reino Unido, Irlanda e Eslovênia, e inclui a definição de procedimentos a serem adotados por controladores de dados quando houver tomadas de decisão automatizadas, como a obrigação de notificação ao titular sobre a decisão automatizada e o estabelecimento de procedimentos para exercício de um direito de revisão. A segunda abordagem, chamada de "proativa", inclui o estabelecimento de salvaguardas novas e mais detalhadas, como o direito de conhecer os métodos e critérios usados em sistemas específicos ou o direito de receber informações específicas sobre a implementação do sistema de decisão algorítmica – segundo Malgieri, é esse o caso da França e da Hungria[15].

Por fim, vale notar que, especificamente no campo da prevenção, investigação, detecção ou repressão de infrações penais ou execução de sanções penais, a Diretiva (UE) nº 2016/680 estabelece, em seu art. 11, a proibição de decisões tomadas exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, que produzam efeitos adversos na esfera jurídica do titular dos dados ou que o afetem de forma significativa. Tal regra é excepcionada unicamente em casos que sejam autorizados pelo direito da União ou do Estado-Membro ao qual o controlador está sujeito, desde que a legislação preveja garantias adequadas dos direitos e

15 No caso da França, é importante também chamar atenção para a decisão de 2018 do Conselho Constitucional acerca da constitucionalidade da lei de proteção de dados francesa, que aborda o tema da transparência algorítmica e da explicação de decisões automatizadas. V. Conseil Constitutionnel, Décision nº 2018-765 DC du 12 juin 2018, §71: "*le responsable du traitement doit s'assurer de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard. Il en résulte que ne peuvent être utilisés*, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement".

liberdades do titular dos dados, pelo menos o direito de obter a intervenção humana do responsável pelo tratamento.

2.2 Brasil

Também no Brasil, a Lei Geral de Proteção de Dados Pessoais – LGPD[16] introduziu previsões semelhantes, embora formuladas de maneira ainda mais aberta do que se verificou na Europa e com ênfase menos direta na participação humana. Com efeito, o art. 20 da LGPD estabelece que "[o] titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses [...]". O § 1º do mesmo artigo, por sua vez, determina que o controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial[17].

Observa-se, de imediato, que, apesar das supramencionadas críticas quanto à sua vagueza, o RGPD europeu apresenta parâmetros mais detalhados do que aqueles trazidos pela LGPD no que tange ao exercício dos direitos associados à intervenção humana em decisões algorítmicas. Para além das perguntas já enunciadas com relação ao RGPD – o que é uma decisão e no que consiste uma decisão tomada unicamente com base em tratamento automatizado –, a LGPD enseja ainda outra: a que tipo de "interesses" afetados a lei estaria a se referir no art. 20, que justificariam a incidência da norma? Quaisquer interesses afetados por uma decisão automatizada, por mais triviais que fossem, seriam aptos a ensejar um direito à revisão?[18]

É interessante notar que, enquanto a discussão na União Europeia claramente tem sido pautada pela defesa da participação humana nos processos de decisão capazes de produzir efeitos jurídicos ou de afetar significativamente os interesses do titular de dados, boa parte do debate no Brasil

16 Lei nº 13.709, de 14 de agosto de 2018.

17 Merece destaque a posição de Renato Leite Monteiro (2018) no sentido de que a LGPD teria expandido o conceito de um direito à explicação em comparação com o RGPD, trazendo um rol de proteções mais amplo que aquelas previstas na regulação europeia. Tal interpretação decorre, sobretudo, do entendimento do autor de que, diferentemente do que ocorre no regulamento europeu, a LGPD teria previsto a possibilidade de que, caso o processo automatizado tenha por finalidade formar perfis comportamentais ou se valha de um perfil comportamental para tomar uma decisão subsequente, haveria também a possibilidade de o titular ter acesso aos dados anonimizados utilizados para enriquecer tais perfis. Outra diferença significativa, segundo o autor, residiria no fato de que, diferentemente da LGPD, o RGPD veio a limitar o direito de oposição do titular nos casos em que a base legal para o tratamento é a execução de um contrato ou o consentimento.

18 Levantando questionamentos similares, v. Mulholland e Frajhof (2019) e Frazão (2018).

ainda gira em torno do questionamento se o citado direito à revisão acarretaria sempre e necessariamente uma revisão por um ser humano, ou se uma revisão automatizada de uma decisão automatizada seria então admissível perante a lei. A controvérsia decorre das diversas alterações ao dispositivo durante a sua tramitação pelo Congresso Nacional, que resultaram na eliminação da expressão "por pessoa natural", que integrava a versão original do art. 20[19].

Conforme delimitado anteriormente, o objetivo deste estudo não é de propor uma interpretação definitiva dos referidos artigos, que são trazidos a lume apenas para exemplificar a complexidade do debate e evidenciar um ponto a ser abordado na próxima seção: a carência de critérios claros que possam ser considerados para atrair a necessidade de intervenção humana no caso de decisões tomadas por sistemas automatizados, em especial no caso de utilização de sistemas de inteligência artificial.

3 PARÂMETROS PARA ATRAIR OU ATENUAR A NECESSIDADE DE INTERVENÇÃO HUMANA

Da discussão precedente, é possível compreender que o principal critério eleito tanto pelo RGPD como pela LGPD para viabilizar a contestação de decisões automatizadas diz respeito, essencialmente, aos seus *efeitos*. O RGPD estabelece, como critério para a intervenção humana face a decisões automatizadas, a produção de *efeitos na esfera jurídica* do titular ou outros que o afetem *significativamente* de forma similar; já a LGPD, ao tratar da revisão de decisões automatizadas (ainda que sem explicitar a participação humana) e do direito do titular de obter informações acerca dos critérios e procedimentos utilizados, menciona, de maneira ainda mais genérica, as decisões "que *afetem seus interesses*", inclusive aquelas destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade[20]. Conforme visto anteriormente, a imprecisão da lingua-

19 A redação original da LGPD foi alterada pela Medida Provisória nº 869, de 2018, suprimindo-se a referência à pessoa natural. Na tramitação da medida provisória pelo Congresso, foi novamente incluída menção à revisão por pessoa natural, por meio da inserção de um parágrafo ao dispositivo. Esse parágrafo foi, por fim, objeto de veto presidencial, baseado no seguinte argumento: "A propositura legislativa, ao dispor que toda e qualquer decisão baseada unicamente no tratamento automatizado seja suscetível de revisão humana, contraria o interesse público, tendo em vista que tal exigência inviabilizará os modelos atuais de planos de negócios de muitas empresas, notadamente das *startups*, bem como impacta na análise de risco de crédito e de novos modelos de negócios de instituições financeiras, gerando efeito negativo na oferta de crédito aos consumidores, tanto no que diz respeito à qualidade das garantias, ao volume de crédito contratado e à composição de preços, com reflexos, ainda, nos índices de inflação e na condução da política monetária".

20 Interessante notar que, no projeto de lei de conversão da MP 869, chegou a ser aprovada a seguinte redação pelo Congresso Nacional, posteriormente vetada pelo Presidente da República, para definir a necessidade de revisão humana: "Art. 20. [...] § 3º A revisão de que trata o caput deste artigo deverá ser realizada por

gem jurídica impõe o estabelecimento de critérios interpretativos adicionais que permitam avaliar se, quando e em que circunstâncias a intervenção humana seria necessária para legitimar a decisão adotada.

Nesse sentido, convém recordar que existe um amplo espectro de aplicações de IA e uma diversidade quase infindável de contextos em que tais tecnologias podem ser utilizadas, desde tradutores de textos usados em ambientes profissionais a armamentos letais usados em conflitos armados. Assim, embora seja possível visualizar, nas extremidades de tal espectro, circunstâncias em que seja mais evidente a necessidade de maior ou menor participação humana no processo decisório, há um sem-número de situações intermediárias que podem suscitar fundadas dúvidas.

Outro aspecto a se considerar é que também a intervenção humana pode assumir diferentes matizes. Ainda que se afaste a validade de uma intervenção humana meramente simbólica, conforme preconizado pelo Grupo de Trabalho do Artigo 29 para a Proteção de Dados (WP29, 2018), permanecem diversas questões relacionadas à definição das características para que uma participação humana seja significativa e adequada a cada contexto.

Assim, cabe, em primeiro lugar, considerar que a intensidade da intervenção humana pode variar bastante em função do grau de delegação de decisões à máquina. Para ilustrar o raciocínio, tome-se como exemplo uma aeronave completamente autônoma, em que há apenas indicação da origem e do destino do voo, sendo todos os demais parâmetros definidos de maneira automatizada, sem qualquer intervenção humana. Agora, imagine-se, como segundo exemplo, que essa mesma aeronave seja capaz de fazer a viagem autonomamente, mas permanece sob supervisão de um piloto humano ao longo de todo o trajeto. Por fim, considere-se, como terceiro cenário, a situação de uma aeronave em que o sistema autônomo é responsável apenas pelas etapas de pouso e decolagem, ficando todas as demais definições e ações a cargo do piloto. Embora em todos os casos descritos o voo tenha sido viabilizado por um sistema automatizado, é evidente que a

pessoa natural, conforme previsto em regulamentação da autoridade nacional, que levará em consideração *a natureza* e o *porte da entidade ou o volume de operações de tratamento de dados*". Como se pode observar, a redação vetada previa, como critérios para a revisão humana, elementos essencialmente empresariais e econômicos, afastando-se, assim, da discussão que correlaciona o direito à revisão a um mecanismo de proteção da autonomia individual.

intensidade da participação humana foi radicalmente distinta em cada um dos cenários.

Ainda quanto às formas de intervenção humana, um segundo ponto a ser considerado diz respeito à efetiva capacidade humana (intelectual, emocional, motora) de (re)avaliação substantiva de decisões sugeridas por sistemas de IA. Em muitos casos, a própria opacidade do processo decisório de sistemas de IA dificulta a identificação de erros que possam ter sido cometidos. Em outras circunstâncias, o volume de dados tratados por meio de decisões automatizadas é tão elevado que não há condições para que uma avaliação caso a caso seja sequer realizável por um humano. Por fim, é preciso ainda considerar a amplamente difundida visão de que as decisões automatizadas são sempre mais precisas e mais corretas do que as decisões humanas, o que acaba gerando, para o decisor humano, um ônus argumentativo por vezes excessivo ou mesmo intransponível para não adotar a solução sugerida pelo sistema automatizado.

Um terceiro ponto a ser considerado diz respeito ao momento da intervenção humana. Vale destacar que a questão não se limita a saber se determinada intervenção humana deva ocorrer antes ou após a produção de efeitos pela decisão automatizada, mas requer, inclusive, uma avaliação quanto ao grau de afastamento temporal entre a decisão humana e a decisão de IA que faz tal decisão humana ainda ser relevante. No caso de utilização de armas autônomas letais, por exemplo, pode-se compreender que quanto maior o intervalo de tempo entre o momento em que uma arma autônoma é ativada por um operador humano e o momento em que a arma seleciona e ataca um alvo, maior o risco de que as premissas que embasaram a decisão humana não mais sejam válidas, especialmente quando seu uso se dá em ambientes dinâmicos ou densamente povoados (Lawand, 2020).

À luz do exposto, pode-se compreender que os critérios comumente previstos em normas e em documentos internacionais para atrair direitos associados à intervenção humana, anteriormente apresentados, são ainda muito vagos e pouco sistematizados para que se possa, em casos concretos, definir, de maneira precisa, em quais circunstâncias e de que maneira a intervenção humana pode ser exigível.

Decerto, a dificuldade de estabelecimento de critérios rígidos para tal avaliação decorre também do fato de que a análise deve necessariamente levar em consideração o contexto da decisão automatizada. Ainda assim, entende-se que é possível, com base nas reflexões anteriores, apontar ele-

mentos que podem trazer maior densidade para tais critérios, a partir da consideração dos diferentes tipos de "falhas" que podem ser cometidas em decisões automatizadas com apoio em sistemas de IA – passíveis, como se viu, de cometerem erros não apenas em razão de deficiências de programação ou falhas nas bases de dados de treinamento, mas também em razão de sua incapacidade de tomada de decisões em que os juízos de "certo" e "errado" são subjetivos ou em que a decisão deve depender de percepções, valores ou comportamentos a princípio não passíveis de conhecimento por máquinas.

Assim, sugere-se que a avaliação quanto à necessidade, forma e momento de intervenção humana, baseada na avaliação dos efeitos da decisão automatizada, deve incluir, ao menos, um juízo quanto: (i) aos riscos e consequências atuais e futuros gerados para os indivíduos e grupos afetados, abrangendo elementos como impactos sobre direitos fundamentais, riscos de discriminação e possibilidade de reversão dos efeitos da decisão; e (ii) à natureza da decisão, em particular no que se refere a decisões em que os juízos de "certo" e "errado" são subjetivos ou em que a decisão deve depender de percepções/valores a princípio incognoscíveis por máquinas. Desta forma, em decisões automatizadas, tanto o seu resultado (as consequências da decisão) quanto seus aspectos procedimentais (a sua natureza) podem ensejar a necessidade do elemento remedial que é a intervenção humana.

É válido apontar que esse tipo de avaliação pode ser concretizada pela realização de avaliações de impacto que, conforme elucida Mantelero (2018), devem ser pautadas pelos direitos e valores em jogo e, consequentemente, podem ter abordagens específicas para o contexto (por exemplo, avaliações de risco no campo da saúde podem levar em consideração elementos distintos daquelas empreendidas no campo da segurança pública). Para o autor, a adoção de uma abordagem orientada a valores impõe, adicionalmente, um foco sobre o impacto social do uso de dados, abrangendo potenciais resultados negativos para direitos e princípios fundamentais e levando em conta, também, as consequências éticas e sociais do tratamento de dados[21]. Nessa linha, Mantelero argumenta que modelos de avaliação de

21 É importante observar, entretanto, que existem determinados tipos de riscos que podem ser identificados em avaliações desse tipo que não serão necessariamente solucionados por meio da intervenção humana – mencione-se, a título exemplificativo, riscos associados à coleta excessiva de dados pessoais ou riscos associados à formação de perfis por meio de inferências.

impacto já existentes ou em discussão – como os já conhecidos relatórios de impacto à proteção de dados, os relatórios de impacto a direitos humanos, ou, ainda, os relatórios de impactos éticos – poderiam evoluir para um modelo mais completo, abrangendo a avaliação de impactos quanto a direitos humanos, ética e sociedade (*Human Rights, Ethical and Social Impact Assessment – HRESIA*), incluindo, em alguns casos, a previsão de consultas a comitês independentes de especialistas para apoiar avaliações éticas mais complexas.

A avaliação dos riscos e dos possíveis efeitos, atuais e futuros, de uma decisão automatizada movida por um sistema de IA pode também ser útil para que se permita concluir em qual momento uma participação humana relevante é necessária e de que maneira ela deve ocorrer, caso ela de fato seja imprescindível – ou seja, se, em determinado caso, um direito à intervenção humana *a posteriori* seria mecanismo adequado e suficiente para viabilizar a proteção dos direitos dos indivíduos afetados; se haveria necessidade de participação humana *a priori*, ou seja, antes que a decisão automatizada produzisse efeitos concretos; ou, ainda, se a supervisão humana significativa seria necessária ao longo de todo o ciclo de vida do sistema (incluindo etapas como pesquisa, desenvolvimento, utilização, manutenção, operação, monitoramento, avaliação etc.). Nesse sentido, a irreversibilidade dos efeitos da decisão automatizada certamente é elemento central a ser considerado, pois, embora um direito à explicação pudesse eventualmente apoiar demandas de reparação por danos experimentados, pouco sentido haveria em prever o direito à revisão de uma decisão cujos efeitos são irreversíveis.

Por outro lado, é possível, sem pretensão de exaustividade, vislumbrar a existência de outros parâmetros de sistemas de IA que, ao conferirem maior legitimidade às decisões tomadas, têm o condão de suprir, em certa medida, o "déficit de humanidade" e assim proporcionar que sejam consideradas formas mais brandas de participação humana, com a condição de que sejam atendidos plenamente os requisitos éticos de decisões legítimas, confiáveis, justas e cognoscíveis, por seres humanos, em seus principais elementos.

Nesse sentido, tomando-se como ponto de partida a constatação de que a transparência é, efetivamente, um princípio fundamental da proteção de dados pessoais tanto no Brasil como em outros países, um dos aspectos centrais a ser considerado diz respeito justamente ao grau de transparência/

opacidade do sistema em questão[22] e à compreensibilidade do processo decisional. Os casos de algoritmos inteiramente determinísticos, em que os resultados são sempre previsíveis e passíveis de compreensão, suscitam questões éticas e jurídicas de muito mais fácil resolução do que aquelas geradas por sistemas do tipo "caixa preta forte" (Bathaee, 2018), caracterizados por processos de tomada de decisão em que não há forma (a) de se determinar como a IA chegou a uma decisão ou previsão, (b) de saber qual informação foi determinante para o resultado alcançado pela IA, ou (c) de obter um *ranking* das variáveis processadas pela IA em sua ordem de importância.

Nesse sentido, tomando o requisito da transparência pelo seu viés da "explicabilidade" e abordando a questão dos modelos que, por sua complexidade, geram decisões que não são intrinsecamente interpretáveis, Maranhão, Cozman e Almada (2021) esclarecem que técnicas de Explainable AI, ou XAI, têm sido utilizadas para viabilizar a geração de explicações quanto à forma pela qual decisões foram tomadas, permitindo, assim, que tais decisões sejam simuladas, contrastadas com outras alternativas plausíveis e, eventualmente, questionadas pelos indivíduos afetados quanto a possíveis efeitos antijurídicos, por meios judiciais ou extrajudiciais.

Um segundo parâmetro a ser considerado diz respeito ao impacto da decisão automatizada sobre direitos fundamentais. Neste sentido, pode-se trabalhar com uma matriz de risco na qual as decisões automatizadas podem ser classificadas em torno dos seus impactos potenciais aos direitos fundamentais, pela qual aquelas com menor potencial de impacto podem, em conjunto com outros fatores, justificar modelos mais genéricos de intervenção humana ou mesmo permitir que outros elementos, tal como a transparência, uma vez verificados a contento, possam mesmo eximir a necessidade de que esta intervenção seja direta. A partir desta métrica, o aumento potencial dos impactos aos direitos fundamentais presentes em decisões automatizadas pode ensejar a necessidade de que formas mais específicas de intervenção humana sejam observadas, compreendendo, inclusive, hipóteses nas quais o recurso às decisões automatizadas seja desaconselhado de todo.

22 Nesse sentido, veja-se a proposta de classificação de Burrell (2016), que identifica três diferentes formas de opacidade algorítmica: (i) a opacidade intencional, como mecanismo corporativo ou institucional de autoproteção e ocultação; (ii) a opacidade decorrente do fato de que escrever e ler código computacional é uma habilidade limitada a especialistas; e (iii) a opacidade que resulta do descasamento entre os procedimentos matemáticos de algoritmos capazes de aprendizado e os estilos humanos de interpretação semântica.

Por fim, um terceiro aspecto a ser considerado para modular a necessidade ou intensidade de intervenção humana nos processos de tomada de decisão automatizados diz respeito às possibilidades de participação do próprio indivíduo afetado pela decisão na configuração e nos resultados do sistema – que, a depender de sua efetividade, poderiam mesmo descaracterizar o conceito de decisão tomada unicamente com base no tratamento automatizado de dados pessoais. De fato, a ideia de *design* centrado no usuário tem ganhado força nos debates éticos sobre sistemas de IA, particularmente no contexto de sistemas de recomendação, levando à sugestão de que alguns problemas éticos decorrentes de exposição a conteúdo inapropriado, por exemplo, podem ser endereçados, inclusive, por meio do estabelecimento de filtros especificados pelo próprio usuário (Milano *et al.*, 2020). Assim, as possibilidades e o efetivo grau de participação do indivíduo afetado pela decisão automatizada na sua configuração – por exemplo, permitindo que ele afaste ou calibre a relevância de determinados resultados ou critérios de decisão – podem ajudar a orientar uma decisão quanto à forma e ao momento em que o direito à intervenção humana (por meio de um direito à explicação ou à revisão, por exemplo) pode ser exercido, ou mesmo afastar a sua exigibilidade[23].

CONCLUSÃO

Considerações sobre a necessidade e a modulação de intervenções humanas em procedimentos que incluem decisões automatizadas são, de certa forma, variações de um dos eixos clássicos dos estudos – e de divagações! – sobre inteligência artificial, que é a própria natureza da atividade realizada pelos sistemas e artefatos que incluem essa técnica: eles realmente "pensam" como os humanos? Desta atividade pode decorrer uma legitimidade que permita considerar tais decisões em paralelo com aquelas realizadas por humanos?

A efetiva utilização e implementação de sistemas decisionais baseados em inteligência artificial trouxe, necessariamente, uma boa dose de pragmatismo para este debate. Ainda que diversos de seus elementos ontologicamente mais relevantes continuem demandando intensa reflexão,

23 Por outro lado, é importante não perder de vista que as soluções centralizadas no usuário possuem, também, determinadas limitações, que são expostas em detalhes no estudo de Milano *et al.* (2020). Para os autores, um ponto importante a ser considerado diz respeito ao fato de que, embora tal abordagem possa estimular a transparência, ela tem também o efeito colocar integralmente sobre os ombros do usuário a responsabilidade pela proteção de direitos e pela utilidade da aplicação.

surgiu a necessidade de respostas imediatas para problemas e dilemas que se colocam crescentemente diante de nós a partir deste crescente recurso a estes sistemas. A partir destas demandas mais concretas e imediatas, outras abordagens se impõem. Por exemplo, saber se sistemas de IA podem ou não "pensar" como os humanos acaba sendo, nestas dinâmicas, uma discussão secundária – algo equivalente a questionar se submarinos sabem nadar (Russel; Norvig, 2010, p. 1021), por exemplo, diante da constatação de que tais sistemas, na prática, já fazem, efetivamente, escolhas e deliberam acerca de questões que impactam, de maneira profunda, a vida das pessoas.

Sendo, portanto, fato que o impacto destes sistemas decisionais automatizados já se percebe em inúmeras circunstâncias, também é necessário que as ferramentas aplicadas para equalizar seus efeitos levem em conta os óbices em se contar exclusivamente com ações específicas dos cidadãos, titulares de dados, para legitimá-las ou não. A depender de como sejam empregados, estes mecanismos – seja a revisão, explicação ou outras formas de intervenção humana – acabam colocando um fardo pesado sobre os ombros do titular de dados[24].

A esse argumento se somam outros, como o de que eventualmente a revisão ou a explicação *a posteriori* de decisões automatizadas podem vir tarde demais, a depender dos efeitos já produzidos, como nos casos, por vezes limítrofes porém concretos, de danos referentes à liberdade ou à integridade física por conta destas decisões.

Assim, a discussão sobre intervenção humana deve, necessariamente, contemplar outras questões relacionadas a estruturas de governança e supervisão de decisões automatizadas de forma mais ampla e matizada conforme as circunstâncias específicas e os riscos específicos aos direitos e garantias das pessoas envolvidas.

Conclui-se, assim, que os mencionados direitos à revisão ou à explicação têm fundamento, sobretudo, na introdução de um componente humano no processo decisório automatizado, em particular, no caso de sis-

24 "*The right to an explanation is only one tool for scrutinizing, challenging, and restraining algorithmic decision making. While it has rhetorical strength in demanding transparency to enable user challenge, it has serious practical and conceptual flaws [...] In short, a legal right to an explanation may be a good place to start, but it is by no means the end of the story. Rights become dangerous things if they are unreasonably hard to exercise or ineffective in results, because they give the illusion that something has been done while in fact things are no better. It is instructive here to compare the history of consent to sharing of data, which has moved in the online world from a real bulwark of privacy to something most often described as meaningless or illusory*." (Edwards; Veale, 2018)

temas de IA, com o objetivo de viabilizar que a decisão seja submetida a um controle de legitimidade que envolve critérios e parâmetros compreensíveis e contestáveis por seres humanos. Constatando-se que os direitos em questão, portanto, são essencialmente instrumentos que servem para endereçar os problemas da confiança na máquina, pode-se compreender que outros parâmetros de legitimação são também importantes e podem relativizar a forma e a intensidade da necessidade de participação humana em processos decisórios automatizados.

Justamente a conjugação de diferentes instrumentos de "humanização" permite compreender que a revisão humana não é elemento imprescindível em todos os casos, por ser instrumental para o elemento ético de ter decisões legítimas, confiáveis, justas e cognoscíveis em seus elementos, ainda que não necessariamente revistas por humanos.

REFERÊNCIAS

ASARO, P. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94 (886), p. 687-709.

AWAD, E.; DSOUZA, S.; KIM, R.; SCHULZ, J.; HENRICH, J.; SHARIFF, A.; BONNEFON, J; RAHWAN, I. The moral machine experiment. *Nature*, v. 563, p. 59-64, 2018.

BATHAEE, Yavar. The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, Vol. 31, N. 2, p. 890-938, Spring 2018.

BAXTER, Kathy. Ethical AI frameworks, tool kits, principles, and certifications – Oh my! 2021. Disponível em: https://blog.einstein.ai/frameworks-tool-kits-principles-and-oaths-oh-my/. Acesso em: dez. 2021.

BIRHANE, A. The impossibility of automating ambiguity. *Artif Life (2021)*, 27 (1): 44-61. Disponível em: https://doi.org/10.1162/artl_a_00336. Acesso em: ago. 2021.

BURREL, J. How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data & Society*, January-June 2016. Disponível em: journals.sagepub.com/doi/pdf/10.1177/2053951715622512. Acesso em: 26 dez. 2018.

DONEDA, D.; MENDES, L.; DE SOUZA, C.; ANDRADE, N. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. *Pensar*, v. 23, n. 4, 2018. Disponível em: https://periodicos.unifor.br/rpen/article/view/8257. Acesso em: dez. 2021.

EDWARDS, L.; VEALE, M. Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy*, 16(3), 46-54, 2018, doi:10.1109/MSP.2018.2701152.

FJELLAND, R. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7, 10 (2020), https://doi.org/10.1057/s41599-020-0494-4. Disponível em: https://www.nature.com/articles/s41599-020-0494-4. Acesso em: dez. 2021.

FRAZÃO, A. Controvérsias sobre direito à explicação e à oposição diante de decisões automatizadas. *Revista Jota*, 12 dez. 2018. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/controversias-sobre-direito-a-explicacao-e-a-oposicao-diante-de-decisoes-automatizadas-12122018. Acesso em: nov. 2021.

GANASCIA, J. G. Epistemology of AI revisited in the light of the Philosophy of Information. *Knowledge, Technology & Policy*, v. 23, p. 57-73, 2010.

ICO – INFORMATION COMMISSIONER'S OFFICE. *Detailed Guidance on Automated Decision-Making and Profiling* (2018). Disponível em: https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling-1-1.pdf. Acesso em: nov. 2021.

ICTFC MEDIA. ICTTV Live Streaming From Caledonian Stadium. *Inverness Caledonian Thistle Football Club*, 16 out. 2020. Disponível em: https://ictfc.com/icttv-live-streaming-from-caledonian-stadium. Acesso em: set. 2021.

JONES, M. L. Right to a Human in the loop: political constructions of computer automation & personhood from data banks to algorithms. *Soc. Stud. of Sci.*, v. 47, 2017.

KAHNEMAN, D.; SIBONY, O.; SUNSTEIN, C. R. *Noise*. A flaw in human judgement. London: William Collins, 2021.

LAWAND, K. International law, incluing IHL, on LAWS: is there a need for a new protocol? In: *Rio Seminar on Autonomous Weapons Systems*, Rio de Janeiro, Naval War College. Brasília: Funag, 2020.

MALGIERI, G. Automated decision-making in the EU Member States: the right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, 35, 2019. Disponível em: https://www.researchgate.net/publication/334359463_Automated_decision-making_in_the_EU_Member_States_The_right_to_explanation_and_other_suitable_safeguards_in_the_national_legislations. Acesso em: set. 2021.

MANTELERO, A. AI and Big Data: a blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, Vol. 34, Issue 4, p. 754-772, August 2018.

MARANHÃO, J.; COZMAN, F. G.; ALMADA, M. Concepções de explicação e do direito à explicação de decisões automatizadas. In: VAINZOF, R.; GUTIERREZ, A. (Org.). *Inteligência artificial*: sociedade, economia e Estado. São Paulo: Thomson Reuters, 2021.

MILANO, S., TADDEO, M.; FLORIDI, L. Recommender systems and their ethical challenges. *AI and Society*, 2020, 35, 957-967. Disponível em: https://link.springer.com/article/10.1007/s00146-020-00950-y. Acesso em: set. 2021.

MITTELSTADT, B. D.; ALLO, P.; TADDEO, M.; WACHTER, S.; FLORIDI, L. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society 3*, 2016, https://doi.org/10.1177/2053951716679679.

MONTEIRO, Renato Leite. Existe um direito à explicação na Lei Geral de Proteção de Dados Pessoais do Brasil? *Artigo Estratégico 24*. Rio de Janeiro: Instituto Igarapé, dez. 2018. p. 9-10. Disponível em: https://igarape.org.br/existe-um-direito-a-explicacao-na-lei-geral-de-protecao-de-dados-no-brasil/. Acesso em: set. 2021.

MORAVEC, H. *Mind children*. The future of robot and human intelligence. Cambridge/London: Harvard University Press, 1990.

MULHOLLAND, C.; FRAJHOF, I. Z. Inteligência artificial e a Lei Geral de Proteção de Dados Pessoais: breves anotações sobre o direito à explicação perante a tomada de decisões por meio de machine learning. In: FRAZÃO, A.; MULHOLLAND, C. (Org.). *Inteligência artificial e o Direito*: ética, regulação e responsabilidade. São Paulo: Thomson Reuters, p. 272-290, 2019.

O'NEIL, Cathy. *Weapons of math destruction*: how Big Data increases inequality and threatens democracy. New York: Random House Audio, 2017.

OECD – ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. *The OECD Privacy Framework*. Paris, 2013. Disponível em: https://www.oecd.org/digital/ieconomy/privacy-guidelines.htm. Acesso em: set. 2021.

_____. *Recommendation of the council on artificial intelligence*. Paris, 2019. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Acesso em: set. 2021.

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

RUSSEL, Stuart J.; NORVIG, Peter. *Artificial intelligence*. A modern approach. New Jersey: Pearson Education Inc., 2010.

SCOTT, P. J.; YAMPOLSKI, R. V. Classification Schemas for artificial intelligence failures. *Delphi – Interdisciplinary Review of Emerging Technologies*, Vol. 2, Issue 4, Pages 186 – 199, 2019. Disponível em: https://delphi.lexxion.eu/article/DELPHI/2019/4/8. Acesso em: ago. 2021.

SELBST, A. D.; POWLES, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, vol. 7, n. 4, p. 233-242. Disponível em: https://ssrn.com/abstract=3039125. Acesso em: dez. 2021.

SOUZA, C. A.; PERRONE, C.; MAGRANI, E. O direito à explicação entre a experiência europeia e a sua positivação na LGPD. In: BIONI, B. R.; DONEDA, D.; SARLET, I. W.; MENDES, L. S.; RODRIGUES JR, O. L. (Org.). *Tratado de Proteção de Dados Pessoais*. São Paulo: Forense, p. 243-270, 2021.

SUNSTEIN, Cass. Algorithms, correcting biases (December 12, 2018). No prelo. *Social Research*. Disponível em: https://ssrn.com/abstract=3300171. Acesso em: set. 2021.

TSAMADOS, A.; AGGARWAL, N.; COWLS, J.; MORLEY, J.; ROBERTS, H.; TADDEO, M.; FLORIDI, L. *The ethics of algorithms*: key problems and solutions (28 de julho de 2020). Disponível em: https://ssrn.com/abstract=3662302. Acesso em: set. 2021.

UNESCO – UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION. *Recommendation on the Ethics of Artificial Intelligence*. Adotada em 24 de novembro de 2021 pela Conferência Geral da Unesco. Disponível em: https://en.unesco.org/artificial-intelligence/ethics#recommendation. Acesso em: nov. 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. (December 28, 2016). *International Data Privacy Law* (2017). Disponível em: https://ssrn.com/abstract=2903469. Acesso em: set. 2021.

WIMMER, M. Inteligência artificial e conflitos armados internacionais: o problema das armas autônomas letais. In: VAINZOF, R.; GUTIERREZ, A. (Org.). *Inteligência artificial* – Sociedade, economia e Estado. 1. ed. São Paulo: Revista dos Tribunais, 2021.

WP29 – Grupo de Trabalho do Artigo 29 para a Proteção de Dados. *Orientações sobre as decisões individuais automatizadas e a definição de perfis para efeitos do Regulamento (UE) 2016/679*. Adotadas em 3 de outubro de 2017, com a última redação revista e adotada em 6 de fevereiro de 2018. Disponível em: https://ec.europa.eu/info/law/law-topic/data-protection_pt. Acesso em: set. 2021.

ZARSKY, T. Z. (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology & Human Values*, 41(1), p. 118-132, 2016. Disponível em: https://law.haifa.ac.il/images/documents/0162243915605575.pdf. Acesso em: set. 2021.

Sobre a autora e sobre o autor:

Miriam Wimmer | *E-mail:* miriam.wimmer@yahoo.com.br

Doutora em Políticas de Comunicação e Cultura pela Faculdade de Comunicação da Universidade de Brasília (UnB). Mestre em Direito Público e Graduada em Direito pela Universidade do Estado do Rio de Janeiro – UERJ. Ex-Bolsista do Programa Internacional da Universidade de Waseda, com Distinção Acadêmica. Certificada como Especialista em Proteção de Dados Pessoais (Europa) pela International Association of Privacy Professionals (CIPP/E). Professora do Corpo Permanente do Mestrado Profissional em Direito do IDP-Brasília e Professora Convidada em diversas instituições de ensino de nível superior. Servidora Pública desde 2007, integrante da carreira de Especialista em Regulação de Serviços Públicos de Telecomunicações da Anatel. Ocupou diferentes cargos de direção no Ministério das Comunicações – MC e no Ministério de Ciência, Tecnologia, Inovações e Comunicações – MCTIC, onde coordenou a elaboração da Estratégia Brasileira para a Transformação Digital e atuou na propositura da Estratégia Brasileira de IA. É, atualmente, Diretora da Autoridade Nacional de Proteção de Dados – ANPD.

Danilo Doneda | *E-mail:* danilo.doneda@idp.edu.br

Doutor em Direito Civil pela Universidade do Estado do Rio de Janeiro (UERJ). Professor no Instituto Brasiliense de Direito Público (IDP). Membro indicado pela Câmara dos Deputados para o Conselho Nacional de Proteção de Dados e Privacidade. Diretor do CEDIS/IDP (Centro de Estudos de Internet e Sociedade). Membro do Conselho Diretor da IAPP (International Association of Privacy Professionals). Foi Coordenador-Geral na Secretaria Nacional do Consumidor do Ministério da Justiça, onde coordenou a redação do Anteprojeto de Lei de Proteção de Dados, a base da Lei Geral de Proteção de Dados. Membro da Comissão de Juristas da Câmara dos Deputados para redação de Projeto de Lei sobre Proteção de Dados nos setores de segurança pública e investigação criminal. Membro do Grupo de Trabalho sobre proteção de dados e informações judiciais do Conselho Nacional de Justiça. Membro dos Conselhos Consultivos do Projeto Global Pulse (ONU) e do Projeto Criança e Consumo (Instituto Alana). Foi Pesquisador Visitante na Autoridade Garante para a Proteção de Dados em Roma (Roma, Itália), na Università degli Studi di Camerino (Camerino, Itália), e no Instituto Max Planck para Direito Privado Comparado e Internacional (Hamburgo, Alemanha). Parte do seu trabalho está disponível em: www.doneda.net.

Data de submissão: 30 de setembro de 2021.
Data de aceite: 15 de dezembro de 2021.