Pontifícia Universidade Católica
do Rio de Janeiro

Isabella Zalcberg Frajhof

Direito à explicação e proteção de dados pessoais nas decisões por algoritmos de inteligência artificial

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do grau de Doutora pelo programa de Pós-Graduação em Direito do Departamento de Direito da PUC-Rio.

Orientadora: Profª. Caitlin Sampaio Mulholland

Rio de Janeiro
Julho de 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Isabella Zalcberg Frajhof

Direito à explicação e proteção de dados pessoais nas decisões por algoritmos de inteligência artificial

Tese apresentada como requisito parcial para obtenção do grau de Doutora pelo Programa de Pós-Graduação em Direito do Departamento de Direito da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profª. Caitlin Sampaio Mulholland
Orientadora
Departamento de Direito – PUC-Rio

Profª. Ana Frazão
Universidade de Brasília (UnB)

Profª. Miriam Wimmer
Autoridade Nacional de Proteção de Dados

Prof. Illié Antonio Pele
Departamento de Direito – PUC-Rio

Prof. Sérgio Marcos Carvalho de Ávila Negri
Universidade Federal de Juiz de Fora (UFJF)

Rio de Janeiro, 06 de Julho de 2022.

Isabella Zalcberg Frajhof

Mestre em Teoria do Estado e Direito Constitucional pela Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio (2018). Graduada em Direito pela PUC-Rio (2015), com intercâmbio acadêmico na American University (Washington D.C). Foi pesquisadora bolsista do Programa de Iniciação Cientifica, na área de Liberdade de Expressão (2010). Atualmente é pesquisadora do Legalite (PUC-Rio), realizando pesquisas sobre a relação entre Direito e Novas Tecnologias.

Ficha Catalográfica

Frajhof, Isabella Zalcberg

Direito à explicação e proteção de dados pessoais nas decisões por algoritmos de inteligência artificial / Isabella Zalcberg Frajhof ; orientadora: Caitlin Sampaio Mulholland. – 2022.
224 f. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Direito, 2022.
Inclui bibliografia

1. Direito – Teses. 2. Direito à explicação. 3. Decisão algorítmica. 4. Inteligência artificial. 5. Proteção de dados pessoais. 6. Direitos fundamentais. I. Mulholland, Caitlin Sampaio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Direito. III. Título.

CDD: 340

Agradecimentos

À minha orientadora, Caitlin Sampaio Mulholland, pela orientação atenta, inteligente, divertida e generosa. Muitas das ideias presentes neste trabalho são frutos das nossas inúmeras trocas. Obrigada por ser uma grande parceira e por todas as oportunidades que você possibilitou. Você é uma inspiração em todos os domínios da vida.

Ao Núcleo Legalite: aos pesquisadores e às pesquisadoras do grupo de pesquisa Legalite, pelos frutíferos debates e por ser um laboratório para testar e melhorar muitos dos argumentos que se encontram nesta tese. Um agradecimento especial à Ana Lara Mangeth, Bianca Kremer, Mariana Palmeira, e Priscilla Silva Laterça Monteiro, pelo acolhimento, pela escuta atenta e pelas inúmeras conversas que me ajudaram atravessar este trabalho; à Maria Regina Rigolon Korkmaz, pelo tempo compartilhado para discutirmos nossas dúvidas e incertezas sobre o instigante tema do direito à explicação; ao Gustavo Robichez e Rafael Nasser, pelo apoio e confiança ao longo de todos estes anos.

Ao João Vitor Barcellos, pela ajuda na pesquisa, fundamental para este trabalho.

Ao Carlos Affonso Souza, por me apresentar ao tema desta tese, pela disponibilidade em compartilhar comigo os seus brilhantes insights, além das indicações de leituras, que foram estruturais para este trabalho.

À Heloísa Carpena, pelos empréstimos de livros, indicações de textos e pela escuta sempre atenta, obrigada por me guiar desde o início.

Ao Programa de Pós-Graduação em Direito da PUC-Rio: aos professores e às professoras do Programa de Pós-Graduação em Direito da PUC-Rio, pelo estímulo ao pensamento crítico e pelos aprendizados, que extrapolam as salas de aula; aos companheiros de Doutorado, em especial, ao Reinaldo Cintra, pelo carinho e incentivo em toda nossa longa jornada na PUC-Rio; à Carmen e ao Anderson, pela disponibilidade e auxílio com todas as demandas e dúvidas.

Ao CNPQ e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Ao meu amor, João Arthur, o responsável por tornar meus dias mais felizes, leves e me acalmar quando tudo parece estar confuso. Obrigada pelo apoio e o

incentivo incondicional em todas as minhas escolhas profissionais e acadêmicas, sem qualquer tipo de cobrança ao longo desses últimos longos anos.

À minha família: à minha mãe, Ilana, por ser a "culpada" por tudo isso, minha maior admiração e inspiração mora em você; ao Alfredo, a quem não consigo agradecer o suficiente pelo tempo dedicado a ler, revisar, pensar e debater cada palavra neste trabalho. Sem vocês esta tese estaria inacabada. Ao meu pai, Leonardo, pelo entusiasmo em me acompanhar ao longo da tese, com a indicação de leituras e a troca de ideias. Aos meus irmãos e irmãs, Lucas, Nicolas, Laura e Chloé, por entenderem minha ausência e também por me tirarem dela. O que seria da vida sem vocês? À minha avó Margot, quem me ensinou que o conhecimento é a única coisa que não nos podem tirar, um ensinamento que reflete e reforça todas as minhas escolhas.

Às minhas amigas e aos meus amigos, indispensáveis na minha vida, fontes inesgotáveis de alegria, cumplicidade, acolhimento e diversão. Agradeço: Anna Bentes, parceira desse e de todos os outros momentos da minha vida, com quem vivi intensamente esses anos de tese, compartilhando ideias e traçando caminhos; à Silvia Bailly, Alice Gelli, Julia Barros, Bruna Aragão e Nathalia Guinet, por trazerem a festa, dança e música à minha vida, me distraindo, compreendendo e incentivando sempre; à Paula Rabacov e Mariah Cretton, minhas confidentes e maiores entusiastas, obrigada pela escuta e por me darem confiança; aos amigos do João, agora meus amigos, em especial, ao Pedro Henrique Miranda, Guilherme Libman, e Dmitri Saramago, pelas nossas longas conversas, confidências e pelo apoio caloroso, mas principalmente pelas muitas risadas, vocês me divertem sempre; ao William Paulo Ducca Fernandes, pela parceria e revisão atenta, obrigada por ser meu especialista técnico em IA e responder sempre prontamente às minhas dúvidas; ao Paulo Henrique Alves, por me aturar todos os dias, você é um grande parceiro e amigo, me apoiando, incentivando e animando sempre que precisei.

Resumo

Frajhof, Isabella Zalcberg; Mulholland, Caitlin Sampaio (Orientadora). **Direito à explicação e proteção de dados pessoais por algoritmos de inteligência artificial.** Rio de Janeiro, 2022. 222p. Tese de Doutorado – Departamento de Direito. Pontifícia Universidade Católica do Rio de Janeiro.

Em um mundo mediado por algoritmos, em que espaços de tomada de decisão antes destinados a humanos passam a ser dominados por estes artefatos, surge uma demanda para que estas decisões algorítmicas sejam explicáveis. Este desafio ganha uma camada de complexidade quando há o uso de técnicas de inteligência artificial, em especial, a aplicação de modelos de aprendizado de máquina, diante da opacidade e inescrutabilidade do modo de funcionamento e dos resultados gerados de alguns tipos destes algoritmos. Neste sentido, esta tese tem início com a apresentação do conceito e dos desafios da inteligência artificial e do aprendizado de máquina para o Direito, particularmente para direitos fundamentais (i.e. proteção de dados pessoais, privacidade, liberdade, autonomia e igualdade). Em seguida, é compartilhada a discussão envolvendo o direito à explicação quando do seu surgimento, e como a sua previsão na LGPD poderá ser interpretada à luz dos aprendizados e interpretações já colhidos no âmbito do GDPR. Ainda, serão analisados como os principais desafios para os direitos fundamentais que são colocados por tais algoritmos de tomada de decisão podem ser resumidos sob os princípios de transparência, prestação de contas e responsabilização e justiça/igualdade. É proposta uma abordagem multifacetada e multidisciplinar, a ser aplicada em diferentes momentos, para assegurar a observância de tais princípios no desenvolvimento e uso de algoritmos de tomada de decisão de aprendizado de máquina. Por fim, propõe-se que a garantia de um direito à explicação, atualmente inserido em uma discussão mais ampla de prestação de contas e responsabilização, deve atender a uma perspectiva de mérito e de procedimento. São identificados os diferentes tipos de conteúdos que têm sido mapeados como passíveis de serem exigidos a título de explicação, e os valores e direitos que um direito à explicação visa proteger, demonstrado, ao final, a importância de que este conteúdo possa estar sujeito a algum tipo de escrutínio público.

Palavras-chave

Direito à explicação; decisão algorítmica; inteligência artificial; proteção de dados pessoais; direitos fundamentais.

Abstract

Frajhof, Isabella Zalcberg; Mulholland, Caitlin Sampaio (Advisor). **Right to an explanation and data protection in decisions by artificial intelligence algorithms**. Rio de Janeiro, 2012. 222p. Tese de Doutorado – Departamento de Direito. Pontifícia Universidade Católica do Rio de Janeiro.

In a world mediated by algorithms, in which decision-making spaces previously destined for humans are now dominated by these artifacts, urges a demand for these algorithmic decisions to be explainable. This challenge gains a layer of complexity when artificial intelligence techniques are used, in particular, the application of machine learning models, given the opacity and inscrutability of the operating mode and the results generated by some types of these algorithms. In this sense, this thesis begins with the presentation of the concept and challenges of artificial intelligence and machine learning for the area of Law, particularly for fundamental rights (i.e. data protection, privacy, freedom, autonomy and equality). Then, the discussion involving the arise of a right to explanation is presented, and how its provision in the LGPD can be interpreted in the light of the lessons learned and interpretations already gathered under the GDPR. Furthermore, it will be analyzed how the main challenges for fundamental rights that are posed by such decision-making algorithms can be summarized under the principles of transparency, accountability and justice/equality. A multifaceted and multidisciplinary approach is proposed, to be applied at different moments in time, to ensure that such principles are incorporated during the development and use of machine learning decision-making algorithms. Finally, this thesis proposed that guaranteeing a right to explanation, which is currently allocated in a broader discussion involving accountability, must take into account a perspective of merit and procedure. The different types of content that have been mapped as likely to be required as an explanation are identified, as well as the values and rights that a right to explanation aims to protect, demonstrating, finally, the importance that such content be subject to public scrutiny.

Keywords

Right to explanation; algorithm decision; artificial intelligence; data protection; fundamental rights.

Sumário

Lista de Tabelas

Lista de Figuras

Introdução

Quando um evento inesperado ocorre, como, quando uma pessoa conhecida falece, é comum que a primeira pergunta que seja feita é: o que aconteceu? Espera-se que seja dada uma explicação simples e clara, envolvendo uma ou duas palavras: câncer; ataque cardíaco; acidente de carro; acidente vascular (Klein, 2018, p. 83). No entanto, se essa pergunta fosse feita entre profissionais da área médica, a explicação se daria em outros termos: se procuraria saber a origem, as causas e quais foram as complicações que levaram à morte do paciente. Por exemplo, em caso de uma infecção pulmonar (pneumonia), o médico plantonista explicaria ao médico da paciente qual foi o vírus ou a bactéria responsável pela infecção, a resposta da paciente ao tratamento e por que o tratamento não foi eficiente, levando à sua morte. Percebe-se, portanto, que os tipos de explicações apresentadas variam de acordo com o contexto, assim como em relação a quem fornece, e a quem solicita a informação. Elas podem ser informações completas, ou apenas parciais, mais técnicas, ou menos específicas; podem ser feitas analogias ou usos de exemplos para facilitar a compreensão; elas dependem da interação entre o locutor e o destinatário da explicação. O objetivo, ao final, é permitir que se compreenda os motivos que justificaram um determinado acontecimento, e que a explicação oferecida seja confiável e fundada em razões sólidas, a fim de convencer a sua correção e legitimidade.

Neste cenário, estamos diante de um evento onde se espera que um humano seja responsável por apresentar uma explicação. No entanto, em um mundo cada vez mais conectado e mediado pela tecnologia, os algoritmos[1] tornaram-se um "fato da vida" (Doneda; Almeida, 2018, p. 142). Eles executam tarefas, tomam decisões, realizam análises e predições, substituindo – algumas vezes, ainda que parcialmente, – espaços decisórios antes dominados por humanos. A abrangência dos seus usos faz com o que os mesmos exerçam um papel importante na regulação das nossas vidas (Mulholland; Frajhof, 2019, p. 268). Tanto é assim que se tornou corriqueiro o uso de algoritmos para apoiar a tomada de decisão de um magistrado para

[1] Os algoritmos podem ser compreendidos como "um conjunto de instruções para realizar uma tarefa, produzindo um resultado final a partir de um algum ponto de partida" (Doneda; Almeida, 2018, p. 142). É comum que o termo seja utilizado ele próprio ou como sinônimo para se referir a computadores, máquina, código, software, entre outros (Doneda; Almeida, 2018, p. 142).

sentenciar ou avaliar a probabilidade de uma pessoa ser reincidente[2], para decidir se uma investigação de fraude contra benefícios sociais será iniciada pelo Estado[3], para conceder um empréstimo financeiro, ou para selecionar pessoas para vagas de trabalho[4].

A evolução destes algoritmos e a possibilidade de delegar a eles a tomada de decisão (parcial ou total) em contextos complexos foi possível diante da disponibilidade massiva de dados e sua facilidade de acesso, confiabilidade e baixo custo de dispositivos de armazenamento (*hardware)*, além do atual avanço de poder computacional (Agrawal et al., 2017; Stone et al., 2016, p. 51). Isto está fortemente relacionado com o atual desenvolvimento e interesse, tanto da indústria, quanto acadêmica, pela inteligência artificial (IA)[5]. Em especial, uma técnica dentro da área de estudo da IA tem chamado atenção: o aprendizado de máquina (*machine learning* – ML). O entusiasmo em torno desta última se refere basicamente à capacidade que algoritmos e modelos de ML possuem de produzir resultados mais acurados do que técnicas preditivas padrões (Ohm; Lehr, 2017, p. 710), melhorando a tomada de decisão baseada em dados em uma escala e por um custo que seria impensável se realizado por humanos (Casey et al., 2019, p. 149).

A presença constante destes algoritmos de ML, e outras técnicas de IA, fazem surgir a necessidade de assegurar uma regulação jurídica deste fenômeno, especialmente pelo seu potencial de afetar direitos fundamentais. Uma das maiores preocupações sobre o tema se refere à opacidade relacionada a como certos

[2] Tal como o software COMPAS, utilizado pode poder judiciário norte americano, para avaliar o potencial risco que um sujeito possui de cometer um crime. Inclusive, este software ganhou notoriedade pela publicação de um artigo que demonstrava que a sua predição afetava desproporcionalmente pessoas pretas em relação a pessoas brancas. Ver em: ANGWIN, Julia, LARSON, Jeff, MATTU, Surya; KIRCHNER, Lauren. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 de maio de 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acessado em 16.04.2021.

[3] Tal como o algoritmo aplicado pelo governo holandês, que será melhor abordado no capítulo 3 deste trabalho. Ver em: SIMONITE, Tom. Europe Limits Government by Algorithm. The US, Not So Much. Wired, 02 de julho de 2020. Disponível em: https://www.wired.com/story/europe-limits-government-algorithm-us-not-much/. Acessado em 13.02.2021.

[4] Tal como o algoritmo implementado pela Amazon para realizar um processo seletivo, que evidenciou um forte viés de gênero, favorecendo candidatos masculinos. Ver em: Ver em: DASTIN, Jeffrey. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 10 de out. de 2018. Disponível em: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Acessado em 15.05.2021.

[5] O termo "inteligência artificial" é um termo guarda-chuva que abriga diversas técnicas computacionais, assim como várias outras áreas do conhecimento que exigem das máquinas certa inteligência (Raso et al., 2018, p. 10).

algoritmos funcionam,[6] diante da dificuldade em compreender o seu processo de trabalho interno e as razões que motivaram seus resultados. Tendo isto em vista, é comum que a literatura se refira a estes algoritmos como "caixas pretas" (Pasquale, 2015; Selbst; Barocas, 2018, p. 1.085), justamente pela ausência de transparência, previsibilidade e inteligibilidade dos seus resultados e modo de funcionamento. Além da questão técnica, estes algoritmos também são considerados confidenciais, em razão da proteção jurídica assegurada ao segredo empresarial e ao direito à propriedade intelectual, sendo uma defesa comumente utilizada para impedir qualquer acesso ao seu código para avaliação e investigação. É neste contexto que demandas por transparência, prestação de contas destas decisões automatizadas, e a possibilidade de pedir a sua revisão e contestação, vêm surgindo como formas de viabilizar meios de compreender estas predições, classificações, resultados e análises feitas por algoritmos.

Assim, a garantia de direitos e de novas maneiras de proteger as pessoas envolvidas nestes processos algorítmicos são fundamentais. Um exemplo disto é o chamado direito à explicação, previsto no Novo Regulamento Geral Europeu de Proteção de Dados Pessoais (*General Data Protection Regulation* – GDPR, em inglês)[7]. Embora na antiga Diretiva 95/46/EC (substituída pelo GDPR) já houvesse previsões regulando este fenômeno[8], o GDPR foi responsável por sistematizar, detalhar e trazer maior robustez para os direitos e salvaguardas relacionados a processos de tomada de decisão algorítmica. O direito à explicação, tal como previsto no GDPR[9], despertou interesse e motivou debates em torno do tema. Quando a norma foi publicada, iniciou-se um debate acadêmico que discutia no que exatamente consistia ser este direito, e como o mesmo deveria ser interpretado, chegando ao ponto de ter sido questionado se ele sequer existiria no GDPR.

[6] Não são todos os algoritmos que são considerados "caixas pretas", no sentido de não ser possível compreender o seu método de trabalho. Os algoritmos que este trabalho apresentará como problemáticos e controversos são aqueles em que não é possível averiguar seus processos internos, e justamente geram esses questionamentos sobre a ausência de transparência e prestação de contas.

[7] Neste trabalho será utilizada a abreviação em inglês do novo Regulamento Europeu de Proteção de Dados Pessoais para se referir a esta norma, tendo em vista a rápida assimilação que os leitores fazem entre a sigla e a norma europeia, diante do uso reiterado que acadêmicos, e a própria mídia, têm feito para se referir à regulação.

[8] A antiga Diretiva 96/46/EC, em seu art. 15, já tratava sobre decisões tomadas por meios totalmente automatizados.

[9] Pode-se dizer que o art. 17, do GDPR, que trata sobre o direito ao apagamento dos dados, ou, o direito a ser esquecido, é uma outra previsão que entrou para o debate público.

Seguindo o exemplo da União Europeia, a Lei Geral de Proteção de Dados Pessoais brasileira (Lei 13.709/2018 – LGPD), promulgada em agosto de 2018, e inspirada na referida norma europeia, também previu, em seu art. 20, um direito à explicação e à revisão de decisões tomadas unicamente com base em tratamento de dados automatizado. Apesar das diferenças entre ambas as normas, a análise comparativa do GDPR e da LGPD torna-se absolutamente necessária, para que se possa avaliar as possíveis interpretações e inspirações que poderão – ou não – serem adotadas pelo direito brasileiro.

Em relação ao GDPR, a controvérsia sobre a existência de um direito à explicação foi iniciada[10] por Sandra Wachter, Brent Mittelstadt e Luciano Floridi (Wachter et al., 2017). Por meio de uma análise linguística da regulamentação, os referidos autores questionaram os tipos de explicações que deveriam ser fornecidas quando uma pessoa estivesse sujeita a uma decisão tomada por meios exclusivamente automatizados, ou seja, por uma decisão algorítmica. O foco da discussão consistiu em analisar o conteúdo que deveria ser apresentado enquanto explicação, além de demarcar em que momento esta explicação poderia ser exigida (Wacther et al., 2017). A partir de uma análise gramatical da norma e uma análise comparativa entre versões anteriores do GDPR, Wachter et al. (2017) argumentam que tal direito não teria sido adotado intencionalmente pelos legisladores[11]. Ao realizarem uma leitura conjunta dos artigos 13, 14 e 15 e 22, do GDPR, apontam que o que poderia ser derivado daquela norma é o que os autores denominaram como um "direito a ser informado". Diferentemente do direito à explicação, o direito a ser informado teria um significado estreito e uma aplicação mais limitada.

Por sua vez, contestando esta posição, Selbst e Powles (2017) sustentam que um "direito à explicação" poderia sim ser extraído do GDPR (Selbst; Powles,

[10] Como se verá nos capítulos subsequentes deste trabalho, a origem da discussão pode ser remetida ao artigo publicado por Goodman e Flaxman (2016). No entanto, o artigo não se debruça em relação às previsões legais deste direito, tampouco apresenta um debate mais profundo sobre o tema (Casey et al., 2019).

[11] Segundo os autores, alguns argumentos sustentam esta posição. Um deles é a ausência do termo "direito à explicação" nas previsões do GDPR que fazem referência a decisões totalmente automatizadas (art. 13, 14, 15 e 22). Esta terminologia consta apenas no Considerando 71 que, na visão de Wachter et al. (2017), não teria força vinculante e, por isso, não poderia ser um dever imposto aos controladores. Além disso, os autores apontam que os legisladores teriam intencionalmente adotado redações diferentes entre o Considerando 71 e o art. 22, do GDPR, apesar de ambas previsões apresentarem similaridade. Neste mesmo sentido, os autores indicam que a redação do art. 15, da antiga Diretiva, que tratava de decisões automatizadas, e do novo art. 22, do GDPR sobre o tema, continuam semelhantes, apesar das diferenças, e na primeira norma não havia qualquer referência a um direito à explicação.

2017). Os autores defendem a necessidade de uma interpretação funcional e flexível que permita aos indivíduos exercerem seus direitos à luz desta legislação, bem como da normativa de direitos humanos. De acordo com os autores, este direito não se encontra diretamente positivado em um artigo, mas ele decorre de uma leitura sistemática e conjunta das previsões enfrentadas por Wachter et al. (2017) (Mulholland; Frajhof, 2019).

Assim, em um primeiro momento, o debate sobre o direito à explicação no âmbito do GDPR teve como foco a discussão sobre o tipo e o conteúdo da explicação de decisões totalmente automatizadas no regulamento europeu. Embora esta discussão ainda não esteja resolvida, atualmente, o direito à explicação encontra-se inserido em uma discussão mais ampla sobre a prestação de contas e responsabilização (*accountability*) de algoritmos[12] (Doshi-Velez; Mason, 2017, p. 2), em especial, em relação aos aspectos da "explicabilidade" e "interpretabilidade" de algoritmos de IA. Além disso, há uma percepção que o tema contribui para questões que transcendem a discussão sobre proteção de dados, tal como a moderação de conteúdo das plataformas de aplicação, e no debate sobre o desenvolvimento de uma IA ética (Frajhof, 2021, p. 470).

A discussão sobre os desafios colocados pelos algoritmos de tomada de decisão e sistemas de IA traz, portanto, reflexões de natureza diversa, sendo a um só tempo técnicas – referente à sua opacidade – e jurídica – referente ao segredo atribuído ao seu código –. Por isso, qualquer solução que busque apresentar meios de garantir explicações sobre estes algoritmos de ML destinados à tomada de decisão deverá implicar em uma abordagem que seja multifacetada e multidisciplinar, dependendo de diferentes técnicas e métodos e pessoas com diferentes expertises (Burrell, 2016, p. 10).

À luz dessas considerações, esta tese tem por objetivo discutir o direito à explicação que está inserido dentro da discussão mais ampla sobre prestação de contas e responsabilização algorítmica de sistemas de IA, inspirada na proposta apresentada por Margot Kaminski, chamada de governança binária dos algoritmos (Kaminski, 2019). A autora identifica três principais razões que justificam a

[12] Neste sentido, como indicado por Doshi-Velez e Mason: "While there are many tools to increasing accountability in AI systems, we shall focus on one in this report: explanation. (We briefly discuss alternatives in Section 7.) By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute" (Doshi-Velez; Mason, 2017, p. 2).

regulação de algoritmos de tomada de decisão: (i) proteger a dignidade da pessoa humana e a sua autonomia nestes contextos, (ii) assegurar a legitimidade das razões e da tomada de decisão algorítmica, e (iii) garantir meios para fiscalizar e corrigir os erros, vieses e discriminações que podem advir de tais sistemas. Para tanto, ela propõe um desenho regulatório de governança binária que possui uma abordagem de direitos individuais (itens i e ii), e sistêmica (item iii), propondo mecanismos de prestação de contas que visem assegurar transparência e legitimidade ao processo de elaboração e de uso de algoritmos. Tais mecanismos devem ser implementados antes do desenvolvimento de algoritmos de tomada de decisão (*ex ante*), e depois que uma decisão foi tomada (*ex post*). Esta perspectiva individual e sistêmica, em que uma depende da outra, destaca a importância de assegurar meios que viabilizem uma prestação de contas focada tanto no sujeito impactado por uma decisão algorítmica, quanto em atender a interesses coletivo e público.

Ademais, é preciso que o direito à explicação também atenda ao ponto discutido inicialmente, qual seja, o tipo de informação que deverá ser apresentada quando uma explicação sobre uma decisão algorítmica é exigida. Afinal, qual é a explicação que uma pessoa impactada por uma decisão automatizada deseja receber? Quais são os objetivos que uma explicação visa alcançar? Quais são os interesses e os direitos que esta explicação e sua revisão visam proteger? Ainda sob a perspectiva de direitos individuais, o direito à explicação deve ser compreendido sob um aspecto de mérito (referente ao conteúdo envolvido na explicação sobre a decisão algorítmica) e procedimental (relacionado aos meios e salvaguardas assegurados para o seu pleno exercício). Neste sentido, a discussão sobre os diferentes tipos de explicações proposta por Wachter et al. (2017) veio se desenvolvendo no sentido de detalhar ainda mais o conteúdo que pode ser apresentado, e em que contexto e de que forma isto deve ocorrer.

Desta forma, a presente tese encontra-se estruturada da seguinte maneira. O primeiro capítulo tem por objetivo apresentar o presente objeto de estudo: o conceito e determinadas ponderações técnicas e legais envolvendo a inteligência artificial e a técnica de aprendizado de máquina. São introduzidos conceitos e definições sobre estas áreas de estudo, apontando quais são os desafios que os seus usos (como a aplicação da técnica de perfilamento) têm trazido para o Direito, em especial, aos direitos fundamentais à liberdade (autonomia), proteção de dados e privacidade, que é o foco deste trabalho. Apresenta-se, então, diferentes propostas

sobre como regular a tecnologia sob o conceito de governança algorítmica, introduzindo o debate que irá ser desenvolvido no capítulo subsequente. Em seguida, é introduzida a discussão que se iniciou com o GDPR sobre o direito à explicação, analisando os artigos que tratam sobre o tema, além de realizar uma análise minuciosa dos artigos deste direito previsto no GDPR, bem como na própria LGPD.

Já o segundo capítulo organiza quais são os três princípios básicos que devem ser atendidos no desenvolvimento de tecnologias que se valem de IA: justiça/igualdade (*fairness),* prestação de contas/responsabilização *(accountability)* e transparência *(transparency)*, representados pelo acrônimo FAT[13]. Procura-se identificar como que os algoritmos de tomada de decisão criam desafios para estes três princípios, e aponta para meios que buscam viabilizar e operacionalizar o direito à explicação a partir destes três princípios, que são os pilares para proporcionar e assegurar o direito à explicação. Neste capítulo é introduzido o conceito de governança binária algorítmica, indicando quais são os documentos e técnicas que devem ser implementados a fim de atender o direito à explicação, e em que momento eles devem ser produzidos.

Por fim, o terceiro capítulo se volta para a perspectiva de mérito e de procedimento do direito à explicação e de revisão, a fim de definir o que seria uma explicação adequada para o sujeito impactado por uma decisão automatizada, bem como para o poder público e para a sociedade civil, quando necessário. São apresentadas diferentes taxonomias de explicações, e os respectivos conteúdos que devem ser compartilhados (i.e. explicações contrafactuais, de dia a dia, estatísticas, científicas, etc), além da organização do aspecto procedimental do direito à explicação, que envolveria uma espécie de exercício do devido processo algorítmico. Além disso, são apresentados os objetivos que o direito à explicação visa atingir, e os valores e direitos que ele visa proteger. Ainda, são analisados alguns projetos de leis que tangenciam o tema do direito à explicação, evidenciando como o tema não se limita à discussão de normas de proteção de dados pessoais. Ao final, a análise da LGPD é retomada, com o objetivo de propor definições sobre

[13] O assunto vem sendo debatido por alguns anos na academia. A criação da *Association for Computing Machinery Fairness, Accountability and Transparency* (ACM FAT) demonstra este interesse. Há, também, uma grande produção de artigos sobre o assunto, como demonstra Linadartos et al., 2021.

o direito à explicação e à revisão de decisões automatizadas, e demonstrar como a proposta do direito à explicação ora apresentada pode ser extraída da referida norma.

Em suma, o que esta tese visa propor são maneiras de viabilizar o direito à explicação, à luz de considerações de diversas naturezas, diante dos desafios técnicos e jurídicos existentes. A elaboração de documentos, aplicação de técnicas, e a implementação de salvaguardas, são formas de assegurar a efetividade e garantia de uma explicação adequada, e permitir uma revisão e contestação de uma decisão automatizada. A proposta visa se esquivar daqueles desafios, de maneira a proteger adequadamente o bem jurídico ora em discussão. Como se verá, não se defende uma abertura indistinta e indiscriminada de toda a sorte de informações sobre o algoritmo e os sistemas de IA, mas formas de exercer uma transparência qualificada (Pasquale, 2015). Esta deve levar em consideração o agente que solicita, e o agente que fornece uma explicação, a fim de permitir um escrutínio, avaliação e julgamento da legitimidade e legalidade de sistemas e decisões algorítmicas.

1.
O que é a Inteligência Artificial?

Existe uma dificuldade em definir o que é inteligência artificial[14] (IA). De maneira geral, o termo "inteligência artificial" é um termo guarda-chuva que abriga diversas técnicas computacionais, assim como de várias outras áreas do conhecimento[15], que requerem das máquinas certa inteligência (Raso et al., 2018, p. 10). A dúvida sobre a sua definição reside principalmente na ambiguidade conceitual da palavra "inteligência", tendo em vista que definições sobre o que seria a IA estão intrinsecamente relacionadas ao que é considerado inteligência[16]. No contexto da informática, Ben Coppin aponta para dois conceitos sobre IA, um mais simples e outro mais complexo. O mais simples permite defini-la como o "estudo de sistemas que agem de uma forma que para qualquer observador aparentaria ser inteligente"[17], enquanto a definição mais complexa entende que a IA "envolve o uso de métodos baseados no comportamento inteligente de humanos e outros animais para resolver problemas complexos"[18] (Coppin, 2004, p. 4).

Há, ainda, uma divisão na literatura técnica sobre o quão similar é a inteligência das máquinas em relação a dos humanos. Essa similitude é um dos critérios que divide acadêmicos entre aqueles que defendem a existência de uma IA forte e de uma IA fraca[19]. Pesquisadores que acreditam na primeira supõem que os

[14] Inclusive, a dificuldade em atribuir um conceito mais estreito à inteligência artificial dificulta as diferentes tentativas de regulação e definição de políticas públicas sobre o assunto (Calo, 2017, p. 407).

[15] As áreas do conhecimento que fundam a IA são: a filosofia, matemática, economia, neurociência, psicologia, engenharia da computação, teoria do controle e cibernética, linguística (Russell; Norvig; 2009, p. 10-16).

[16] De acordo com a definição do Aurélio, o termo inteligência possui as seguintes definições: "1. Faculdade de aprender, apreender ou compreender; percepção, apreensão, intelecto, intelectualidade. 2. Qualidade ou capacidade de compreender e adaptar-se facilmente; capacidade de penetração, agudeza, perspicácia" (....). FERREIRA, Aurélio Buarque de Holanda. *1*. 2ª Ed., rev., e aum.. São Paulo: Editora Nova Friburgo, 1995.

[17] Tradução livre de: "Artificial intelligence is the study of systems that act in a way that to any observer would appear to be intelligent".

[18] Tradução livre de: "Artificial Intelligence involves using methods based on the intelligent behavior of humans and other animals to solve complex problems".

[19] Esta diferenciação não deve ser confunda com os métodos fracos e fortes da IA, como aponta Ben Coppin. O autor explica que, enquanto os métodos fracos estão preocupados em resolver problemas a partir de determinados métodos (como o uso de lógica) a uma ampla gama de assuntos, não há uma apreensão do conhecimento para executar esta tarefa. Por isso, este método estaria mais relacionado à metodologia aplicada para resolver um determinado problema. Por sua vez, os métodos fortes pressupõem um conhecimento sobre o mundo e possíveis problemas que existem e podem ter que ser enfrentados. Contudo, os métodos fortes dependem do método fraco, pois "a system with knowledge is useless without some methodology for handling that knowledge" (Coppin, 2004, p. 6).

computadores irão desenvolver tal poder de processamento e capacidade de aprendizado que eles serão capazes de pensar e ter consciência. Essa posição é contestada por aqueles que acreditam ser impossível criar um artefato dotado de emoção e consciência, como os humanos. Além disso, aqueles que defendem uma IA fraca sustentam que a máquina tem um comportamento considerado inteligente quando o mesmo é modelado para resolver problemas complexos, mesmo que essa capacidade não implique na máquina ser inteligente como uma pessoa (Coppin, 2004, p. 5).

Apesar dessas visões distintas, pode-se afirmar que, o parâmetro para medir se o comportamento de um dispositivo ou de um sistema é "inteligente" é a medida "ser humano", seja porque a avaliação de inteligência será feita por um observador humano, seja pela percepção de que o artefato realiza uma tarefa que, supostamente, apenas humanos poderiam realizar, e por isso, reproduz a cognição deste ser. Embora existam outras maneiras de definir e conceituar o que constitui ser IA[20], a mesma pode ser compreendida como a pretensão de que os sistemas ou dispositivos embarcados com técnicas de inteligência artificial sejam tão inteligentes quanto um humano.

Esse elemento "humano" é tão intrínseco à definição do que seja IA, que conceitos trazidos por autores do âmbito jurídico também fazem esta associação. Ryan Calo aponta que a inteligência artificial pode ser compreendida como o uso de técnicas que tentam aproximar aspectos da cognição humana a máquinas (Calo, 2017, p. 404). Harry Surden indica que tais técnicas automatizam determinadas tarefas que normalmente estão associadas à inteligência humana, tais como jogar xadrez, traduzir palavras, reconhecer padrões ou dirigir carros (Surden, 2019, p. 1307). Matthew Schrerer define a IA como máquinas que realizam tarefas que, se feitas por humanos, teria sido requerido certa inteligência (Scherer, 2016, p. 362).

No entanto, vale a ressalva de que, o fato de a IA executar determinados tipos de tarefas ou rotinas de maneira bem-sucedida, atendendo à finalidade para a qual foi construída (como, vencer uma partida de xadrez, imitar uma conversa, realizar um diagnóstico médico ou apoiar um juiz em uma decisão), não significa que o seu processo cognitivo seja semelhante ao do humano. Essa afirmação, que

[20] Além dessa medida humana para averiguar a inteligência de uma máquina, é possível definir a inteligência artificial a partir do que os pesquisadores andam estudando, como por exemplo, o atual interesse da comunidade em software, e não tanto em hardware (Stone et al., 2016, p. 13)

se aproxima da perspectiva da IA fraca, é interessante, pois chama atenção para as diferentes maneiras que os problemas são resolvidos por uma máquina e por uma pessoa (Surden, 2014, p. 87). Enquanto este último é capaz de pensar abstratamente e compreender a sintaxe da linguagem, determinados tipos de sistemas de IA pensam de maneira mais estreita, a partir do que é certo ou errado, identificando padrões que muitas vezes não são perceptíveis ao humano, e em uma velocidade inalcançável[21].

Stuart Russell e Peter Norvig, autores de um dos livros mais tradicionais e utilizados na área de ciência da computação, "Artificial Intelligence: A Modern Approach", respondem à pergunta sobre "O que é IA", indicando que esta é direcionada a alcançar determinados objetivos, e que estes são atendidos a partir de uma ação e/ou um pensamento racional. Para os autores, um artefato, seja um dispositivo que possui materialidade física, ou um sistema computacional, para ser considerado inteligente deve agir ou pensar de maneira racional, como um ser humano supostamente assim o faria[22].

Tal artefato estaria representado a partir do conceito de "agente inteligente" dos autores, compreendido como coisas que interagem com o ambiente no qual estão inseridas, que adquirem conhecimento sobre o mesmo, e agem de acordo com a finalidade para a qual foram construídas. Ou seja: o agente inteligente é "qualquer coisa que pode ser vista como percebendo o seu ambiente por meio de sensores e atuando nesse ambiente por meio de atuadores"[23] (Russell; Norvig, 2009, 34). Assim como os seres humanos possuem sensores – olhos e ouvidos –, e atuadores – pernas e os braços – os agentes inteligentes, tal qual um robô, possuem diversos sensores e câmeras que o guiam, cujos motores o permitem agir no mundo. Esses agentes podem agir sob a forma física de um objeto, como o aspirador de pó

[21] Para uma visão mais abrangente e técnica das definições de inteligência artificial, ver: <https://medium.com/@jackkrupansky/untangling-the-definitions-of-artificial-intelligence-machine-intelligence-and-machine-learning-7244882f04c7/>. Acessado em 03.01.2020.

[22] Os autores apresentam trazem oito definições do que é a inteligência artificial, as organizando em quatro principais categorias. São elas: (i) sistemas que agem como humanos, (ii) sistemas que pensam como humanos, (iii) sistemas que pensam racionalmente e (iv) sistemas que agem racionalmente (Russell; Norvig; 2009, p. 3). Os sistemas que agem como humanos são programados para imitar uma reação humana; os sistemas que pensam como humanos simulam o funcionamento da mente e seus processos cognitivos; os sistemas que pensam racionalmente são baseados em regras lógicas que se valem do silogismo para funcionarem; por fim, os sistemas que agem racionalmente buscam alcançar a melhor resposta possível para uma determinada ação, por meio da sua percepção do mundo e uma ação apoiada por essa percepção.

[23] Tradução livre de: "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators".

inteligente, o Roomba[24], que aprende a mapear e se adaptar à uma casa, ou por meio de uma interface digital, sem aderir a um formato material no mundo físico, tal como o provedor de pesquisa Google ou qualquer outro sistema corporativo ou aplicativo.

Neste sentido, alinhado a estes elementos dos agentes inteligentes, o grupo especializado em inteligência artificial formado pela Comissão Europeia aborda os aspectos mais relevantes da IA, propondo a seguinte definição sobre tais sistemas:

> Sistemas de Inteligência Artificial (IA) são softwares (e possivelmente hardwares) desenhados por humanos que, dado um objetivo complexo, atuam na dimensão física ou digital percebendo o seu ambiente por meio da aquisição de dados, interpretando os dados estruturados e não estruturados coletados, raciocinando sobre o conhecimento ou processando a informação derivada desse dado e decidindo a(s) melhor(es) ação(ões) para alcançar aquele objetivo. Sistemas de IA podem usar regras simbólicas ou aprenderem com modelos numéricos, e também podem adaptar seu comportamento analisando como o ambiente é afetado por suas ações pretéritas (...) (Comissão Europeia, 2019b, p. 6)[25].

Como indicado, as regras implementadas podem ser simbólicas, representadas por comandos que pessoas são capazes de compreender, com conhecimento de sentenças declarativas e métodos de raciocínio lógico, ou não-simbólicas, em que o conhecimento não é representado explicitamente por símbolos, e é construído por adaptação ou inferência[26], tal qual ocorre com as redes neurais[27]. Neste último caso, não será possível compreender a forma de trabalho do artefato com regras não-

[24] Ver em: https://www.irobot.com.br/roomba. Acessado em 16.02.2020.

[25] Tradução livre de: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions".

[26] Informações obtidas na aula "Inteligência Artificial: Introdução", proferida pelo Professor Augusto Baffa, no Programa de Inovação Tecnológica (PIT) da PUC-Rio, em 11 de outubro de 2018.

[27] Rede neural é um modelo computacional que simula os neurônios cerebrais para emular o aprendizado. Conforme explicam Russel e Norvig: "The first work that is now generally recognized as AI was done by Warren McCulloch and Walter Pitts (1943). They drew on three sources: knowledge of the basic physiology and function of neurons in the brain; a formal analysis of propositional logic due to Russell and Whitehead; and Turing's theory of computation. They proposed a model of artificial neurons in which each neuron is characterized as being "on" or "off," with a switch to "on" occurring in response to stimulation by a sufficient number of neighboring neurons (…). They showed, for example, that any computable function could be computed by some network of connected neurons, and that all the logical connectives (and, or, not, etc.) could be implemented by simple net structures. McCulloch and Pitts also suggested that suitably defined networks could learn" (Russell; Norvig, 2009, 16).

simbólicas, e como que os resultados finais foram alcançados, justamente pela maneira que o conhecimento é gerado. Isto cria um entrave relevante para o direito à explicação.

Estes sistemas e artefatos são desenvolvidos por algoritmos que, sob uma perspectiva computacional, podem ser compreendidos como um conjunto de etapas que, quando ordenados de maneira correta, vão processar computacionalmente instruções e/ou dados (*input*) para alcançar o resultado (*output)* desejado (Kitchin, 2014, p. 3). Pode-se dizer que um algoritmo nada mais é do que um conjunto de regras que tem como objetivo realizar uma determinada tarefa. O exemplo mais corriqueiro é a sua comparação a uma receita de bolo[28] que, uma pessoa ao seguir as instruções dos ingredientes necessários, a ordem que eles devem ser misturados, e o tempo de espera para que o bolo asse, o objetivo final será bem sucedido. Ademais, é possível compreender os algoritmos, em especial, um tipo específico de algoritmo destinado à tomada de decisão, a partir de um conceito que vem sendo difundido pela opinião pública. Isto é, a ideia de algoritmos geralmente são confiáveis, e desenvolvidos "baseados em regras complexas que desafiam ou frustram a capacidade humana para ação e compreensão" (Mittelstadt et al., 2016, p. 3).[29]

Este é um dos motivos que tem chamado a atenção da doutrina especializada na relação entre direito e tecnologia, tendo sido considerada como a tecnologia desde a internet que traz os maiores desafios jurídicos e sociais (Calo, 2015; Balkin, 2015). Neste sentido, na tentativa de compreender tais reveses, Ryan Calo (2015) elenca três qualidades essenciais que a IA e a robótica[30] possuem,[31] e que motivam novos e diferentes desafios regulatórios, tal como a criação de uma nova área de estudo no Direito para dar conta da sua excepcionalidade (Calo, 2015). As três

[28] Uma explicação simples sobre o que é um algoritmo pode ser acessada aqui: https://www.youtube.com/watch?time_continue=1&v=eVhz6nytufY.

[29] Tradução livre de: "based upon complex rules that challenge or confound human capacities for action and comprehension".

[30] Embora Calo já tenha aventado a possibilidade de separar a IA que não incorpora uma presença física para fins regulatórios (Calo, 2017, p. 407), é verdade que a robótica é apenas uma das facetas das dificuldades que a IA apresenta ao Direito, e alguns dos problemas que surgem com ela, assim como algumas das suas características, também são pertinentes ao contexto dos algoritmos de aprendizado de máquina.

[31] Algo que é criticado por Jack Balkin (2015), em seu texto em resposta a Ryan Calo. Segundo Balkin, não se deve pensar em "qualidades essenciais" de uma tecnologia, pois estas qualidades não são independentes por si só, elas se manifestam de acordo com o uso que é feito delas a partir da interação humana, e se desenvolvem ao longo do tempo.

qualidades da IA e da robótica seriam: a materialidade (*embodiment*), o comportamento emergente (*emergence*) e o valor social (*social valence*) [32]. Essas são apontadas como as principais razões pelas quais a robótica e a IA devem ser consideradas como tecnologias transformativas, em que essas características, em conjunto ou individualmente, irão impactar o Direito e suas instituições de formas diferentes. A materialidade se refere à presença física e capacidade de agir no mundo dos agentes inteligentes que incorporam um formato físico; o comportamento emergente diz respeito à aptidão dos agentes inteligentes de aprenderem e se adaptarem a novas circunstâncias, que ocorre de maneira imprevisível e ininteligível, atribuindo autonomia/agência ao seu comportamento[33]; e, por fim, seu valor social se refere à forma como os humanos interagem e enxergam os robôs, muitas vezes como se fosse um animal, um objeto ou até mesmo como uma pessoa, impactando como o Direito vai atribuir responsabilidades.

1.1 As diferentes maneiras de regular a Inteligência Artificial

Quando se considera a tecnologia como objeto de regulação, ou da tecnologia como agente regulador do comportamento humano (Hildebrandt, 2020, p. 251)[34], é importante compreender como que as suas características vão afetar direitos, em especial, direitos fundamentais. Lições sobre a regulação da internet ainda continuam pertinentes para a IA. Em ambos os cenários, a regulação jurídica, especificamente, a lei, deve preservar direitos fundamentais, bem como assegurar "que o desenvolvimento tecnológico se torne um elemento que aprimore o desenvolvimento da personalidade e as condições econômicas e sociais dos indivíduos e coletividades, e não o contrário" (Souza; Lemos, 2016, p. 16).

Um dos desafios na regulação destas tecnologias é compreender qual é o meio mais adequado e eficaz para constranger e delimitar comportamentos, especialmente para assegurar direitos e delimitar deveres. Há, portanto, diferentes propostas sobre como esta regulação deverá ocorrer, para além do Direito. Há a

[32] Conforme traduzido por Doneda et al., 2019, p. 7.
[33] Calo indica que o uso do termo "*emergence*" no lugar de autonomia se justifica, pois comportamento emergente aponta para a ideia de que as máquinas têm a intenção de agir de uma ou outra forma. Aqui fica evidente o objetivo do autor em atribuir agência a estes artefatos para permitir a atribuição de responsabilidade às danosidades causadas pelos mesmos.
[34] Neste sentido, Heidelbrandt aponta que: "So, technology can be either the object or the subject of regulation (and maybe both), whereas law is usually only seen as a subject of regulation (that which regulates)" (Hildebrandt, 2020, p. 251)

tecno-regulação, representado pela teoria de Lawrence Lessig em sua famosa afirmação de que o "código é lei" (*code is law*); a autorregulação, que ocorre pelos próprios agentes do mercado que exploram a tecnologia, sendo um exemplo disso os diversos documentos éticos que têm sido produzidos, sujeito a críticas pelo fato de o agente regulado, implicado na atividade, estar liderando a sua própria regulação[35], além de ser considerada como uma *soft law*[36], carente, portanto, de força vinculante; e a regulação da própria tecnologia, que pode se dá por meio de uma governança algorítmica[37], e assumir diferentes formatos (Correa, 2021, p. 201).

Para Lessig, existiriam diferentes ferramentas/instrumentos que seriam capazes de constranger e regular comportamentos. São elas: (i) o direito que, através do seu corpo normativo e instituições, visa oferecer enunciados de como as pessoas devem se comportar, ameaçando com sanções em caso de inobservância dessas regras; (ii) as normas sociais, que constrangem comportamentos, e ameaçam com sanções *ex post* de maneira descentralizada, mas que são impostas por toda a sociedade; (iii) o mercado, que regula por meio de preços (o mercado limita o comportamento pelo preço, porque ele também está sendo regulado pelo Direito e por normas sociais), e (iv) a arquitetura do mundo real como, por exemplo, uma rodovia que divide dois bairros, mudando a sua extensão e influenciando nos limites

[35] Neste sentido que Yonchai Benkler escreveu um curto, mas forte, artigo de posição defendendo que o financiamento da tecnologia não deve ser feito pela indústria, mas pelo Governo para evitar enviesamentos de seus interesses. BENKLER, Yochai. Don't let the industry write the rules for AI. *Nature*. World View, v. 569, 9 mai 2019, p. 161.

[36] A definição de *soft law* tem origem no Direito Internacional Público, na discussão sobre as fontes de direito, e a natureza das obrigações jurídicas que existem nesta área de estudo. A *soft law* é considerada uma fonte de direito flexível, e se diferencia das normas de Direito Internacional Público de natureza *jus cogens,* uma categoria de normas de natureza imperativa, que não podem ser derrogadas, salvo se por outra posterior e de mesma natureza. Conforme conceitua Mazuolli, a *soft law* pode ser compreendida como: "todas aquelas regras cujo valor normativo é menos constringente que o das normas jurídicas tradicionais, seja porque os instrumentos que as abrigam não detêm o status de "normas jurídicas", seja porque os seus dispositivos, ainda que insertos no quadro de instrumentos vinculantes, não criam obrigações de direito positivo aos Estados, ou não criam senão obrigações pouco constringentes. Portanto, um dos maiores problemas desse tipo de norma se encontra na falta de elementos que garantam a sua efetiva aplicação" (Mazuolli, 2020, p. 207).

[37] Governança de algoritmos, ou governança algorítmica, é uma maneira de governar o poder exercido pelo algoritmo em suas interações, podendo assumir diferentes formatos, com múltiplos agentes. Saurwein et al. (2015), por exemplo, propõem uma governança de algoritmos que é baseada em riscos a partir de uma taxonomia que identifica nove tipos de riscos que os algoritmos podem causar. De acordo com os autores, sob uma perspectiva institucional, a governança algorítmica pode ser analisada em um *continuum*, estando em um extremo a regulação pelo mercado, e no outro a regulação estatal. Entre estes dois extremos podem existir diferentes modelos de governança, categorizados da seguinte maneira: auto-organização de uma empresa, auto-regulação setorial, co-regulação entre autoridades estatais e a indústria (Saurwein et al., 2015, p. 37). A escolha de uma ou outra forma dependerá dos riscos que se visa evitar ou mitigar.

e no direito da cidade (Lessig, 1999, p. 506). Estas forças atuam conjuntamente, de maneira cooperativa ou competitiva. Para avaliar se uma regulação foi bem-sucedida é necessário observar a interação dessas quatro forças regulatórias em um determinado cenário. A contribuição de Lessig para esse debate foi reconhecer que na internet, ou como diria o autor, no ciberespaço, o código representa esta arquitetura. Consequentemente, isto traz reveses para a privacidade e transparência de ambientes regulados por código.

Virgílio Almeida e Danilo Doneda destacam que a governança algorítmica pode assumir diferentes naturezas, que variam desde uma perspectiva jurídica e regulatória, até um ponto de vista técnico. A escolha de qual governança será a mais adequada irá variar de acordo com a natureza, o contexto e os riscos que o algoritmo em questão poderá gerar[38] e os danos que poderá causar. A governança visa, de forma geral, "priorizar a responsabilização, transparência e as garantias técnicas" de maneira que os resultados danosos, discriminatórios e os erros causados sejam eliminados ou ao menos mitigados (Almeida; Doneda, 2018, p. 145), podendo assumir diferentes frentes.

O Direito pode constranger e limitar os usos de IA por meio de diferentes etapas na cadeia de desenvolvimento da IA. Uma delas ocorre pela regulação dos bancos de dados que são utilizados pelos algoritmos[39], bem como da proteção garantida aos usos de dados pessoais, a fim de assegurar que os dados sejam legitimamente obtidos para tratamento, sejam corretos, atualizados e exatos. Inclusive, muitos dos princípios previstos pelas normas de proteção de dados pessoais se equivalem a exigências normativas no contexto de IA.

Já as garantias técnicas podem ser capazes de dar uma resposta para como evitar resultados enviesados, discriminatórios e danosos dos algoritmos, tal como pelo estabelecimento de padrões de concepção, taxas de desempenho (*performance)* ou estabelecimento de responsabilização.

As diferentes abordagens que a governança algorítmica pode assumir também refletem nas diversas maneiras de supervisionar esta governança, podendo ocorrer por agências reguladoras, órgãos internos[40] ou comitês multissetoriais (Almeida; Doneda,

[38] Tal qual proposta por Saurwein et al. (2015).
[39] No Brasil o art. 43 do Código de Defesa do Consumidor também traz previsões específicas sobre o tema.
[40] Doneda et al. (2018) propõem o estabelecimento de quadros éticos corporativos, que entendem exercer um papel relevante, tendo "a oportunidade de maximizar os benefícios que essas tecnologias

2018, p. 146-147). Percebe-se que não há uma maneira correta ou previamente indicada para estabelecer esta governança, mas diversas formas de implementá-la, devendo ser uma abordagem multifacetada e multidisciplinar.

Em relação à regulação pela ética, considerado um mecanismo de *soft law*,[41] diversas entidades têm proposto códigos de conduta e códigos éticos, apresentando um núcleo de princípios que devem ser observados ao longo de toda a cadeia de desenvolvimento da IA. As entidades envolvem: organismos internacionais (OCDE[42] e *Access Now*[43]), instituições privadas (IBM[44] e Microsoft[45]) e instituições de pesquisa (*Association for Computer Machinery* -- ACM[46] e *The Institute of Electrical and Electronics Engineers* -- IEEE[47]). Estes guias deontológicos têm sido frequentemente citados como limites e balizas no desenvolvimento e implementação da IA, ante à ausência de uma regulação jurídica sobre o tema. Pode-se dizer, de forma geral, que estes documentos partem de um ponto em comum, que é: o fomento e desenvolvimento de uma que IA que deve ser confiável[48], auditável, e seus processos conhecidos e controlados – ou controláveis – por humanos (Mulholland; Frajhof, 2021,

podem trazer, garantindo que danos e resultados negativos sejam evitados, destaca a necessidade da ética como uma estrutura analítica e operacional destinada a orientar a estratégia dos atores corporativos e a moldar as suas práticas nesse domínio" (Doneda et al., 2018, p.13).

[41] Julia Black e Andrew Murray criticam a abordagem de regulação da IA com códigos de ética, pois os autores acreditam que isto pode marginalizar a regulação jurídica e normativa sobre o tema, em que esta última seria considerada como preferencial para regular a tecnologia. Conforme escrevem os autores: "Our experience though is that the lure of soft regulation through ethical codes of practice were a crutch for governments who did not want to set hard standards. Eventually though with contractualised regulation replacing ethics the folly of that error would become apparent" (Black; Murray, 2019, p. 9). Na concepção dos autores, seria necessário adotar uma perspectiva policêntrica da regulação, com a atuação dos diferentes agentes envolvidos no ecossistema de IA, sendo certo a definição das pautas do que e como a IA deve ser regulada não deveria estar sendo liderada pelos agentes econômicos interessados.

[42] O Guia da OCDE sobre recomendações na regulação da Inteligência Artificial por meio de princípios éticos, encontra-se disponível em: https://www.oecd.org/going-digital/ai/principles/. Os 36 países membros da OCDE aderiram aos Princípios da Inteligência Artificial da OCDE. Embora o Brasil não seja membro da OCDE, o país também aderiu ao documento, em maio de 2019.

[43] Disponível em: https://www.accessnow.org/artificial-intelligence-we-just-became-a-member-of-the-partnership-on-ai/#:~:text=Access%20Now%20recognises%20that%20ethical,only%20be%20a%20first%20step. Acessado em 21.08.2021.

[44] Disponível em: https://www.ibm.com/blogs/policy/trust-principles/. Acessado em 21.08.2021.

[45] Disponível em: https://www.microsoft.com/en-us/ai/responsible-ai. Acessado em 21.08.2021.

[46] Disponível em: https://ethics.acm.org/code-of-ethics/. Acessado em 21.08.2021

[47] Disponível em: https://www.ieee.org/about/corporate/governance/p7-8.html. Acessado em 21.08.2021.

[48] Para ser considerada como uma IA confiável, três princípios devem estar presentes: o princípio da justiça (*fairness*), o princípio da acurácia (*accuracy*) e o princípio da inteligibilidade (*intelligibility*), que exigem adoção de medidas que: (i) impeçam sistemas de IA de violar o princípio da igualdade de tratamento; (ii) os insumos e resultados da IA devem ser precisos; (iii) permitam que o humano possa conhecer os processos humanos (Mulholland; Frajhof, 2021, p. 73).

p. 72). Logo, a governança da IA deve ser pautada por princípios éticos que viabilizem estes objetivos e, consequentemente, deve adotar medidas que façam valer estes princípios.

O que se percebe nesta descrição de governança algorítmica é a proposta e o reconhecimento de que os desafios possuem ordem e natureza diferentes. Por isso, será necessário se valer de diferentes instrumentos regulatórios para proteger direitos.

A discussão sobre como regular a tecnologia frequentemente invoca dúvidas e incertezas sobre o fino balanço entre proteger os direitos fundamentais das pessoas, como a privacidade e intimidade e a proteção de dados pessoais (art. 5º, inciso X e LXXIX da CF)[49/50], e a livre iniciativa dos agentes econômicos responsáveis por desenvolver e implementar estas tecnologias, princípio que é fundamento do Estado Democrático de Direito e que rege a ordem econômica brasileira (art. 1º, inciso IV c/c art. 170, *caput,* da CF)[51]. É comum que haja um conflito entre tais direitos e princípio. Contudo, a ponderação nestes casos deve levar em consideração que a livre iniciativa, que também incentiva a inovação e o empreendedorismo, não pode se dar às custas de uma violação ou uma restrição a tais direitos fundamentais que, muitas das vezes, são violados e ameaçados sem que o próprio titular tenha conhecimento.

Um exemplo deste conflito é quando a finalidade para a qual uma determinada tecnologia foi desenvolvida é posteriormente utilizada para outros e novos fins, descontextualizados e não adequados à finalidade inicialmente pretendida. Isto também ocorre com o tratamento de dados pessoais, especialmente quando há um novo agente econômico detentor desta tecnologia. Um caso recente disto é o escândalo de espionagem digital pelo *spyware* "Pegasus", desenvolvido pela empresa israelense

[49] Art. 5º, inciso X e LXXIX: "Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes: (…) X - são invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação; LXXIX - é assegurado, nos termos da lei, o direito à proteção dos dados pessoais, inclusive nos meios digitais"

[50] A recente aprovação da Proposta de Emenda à Constituição 17/2019 incluiu o inciso LXXIX, no art. 5º da Constituição Federal, que possui a seguinte redação: "é assegurado, nos termos da lei, o direito à proteção dos dados pessoais, inclusive nos meios digitais". Além disso, foi incluído o inciso XXVI ao art. 21, e o inciso XXX ao art. 22, para fixar a competência da União Federal para legislar sobre o tema.

[51] Art. 1º, inciso IV, da CF: "A República Federativa do Brasil, formada pela união indissolúvel dos Estados e Municípios e do Distrito Federal, constitui-se em Estado Democrático de Direito e tem como fundamentos: IV - os valores sociais do trabalho e da livre iniciativa (…)" Art. 170, *caput,* da CF: "A ordem econômica, fundada na valorização do trabalho humano e na livre iniciativa, tem por fim assegurar a todos existência digna, conforme os ditames da justiça social, observados os seguintes princípios (…)"

de cibersegurança NGO Group. Este software foi desenvolvido para invadir telefones celulares e espionar seus usuários e, supostamente, seria comercializado apenas para agências governamentais com bom histórico de direitos humanos[52], e para combater crimes que causam grande impacto na sociedade, tais como terrorismo e pornografia infantil.[53] No entanto, o que vem sendo amplamente noticiado é o seu uso por governos autoritários para monitorar ativistas de direitos humanos e jornalistas.

Isto demonstra como a tecnologia, bem como a maneira de regulá-la, deve sempre ter como norte o desenvolvimento de políticas, práticas, técnicas e normas jurídicas que respeitem e priorizem a proteção a direitos fundamentais, em especial, aqueles que protegem as pessoas, diante dos potenciais abusos que podem ocorrer. A governança algorítmica é uma proposta que reconhece que os desafios da tecnologia, especialmente com a inteligência artificial, possuem ordem e natureza diferentes. Por isso, será necessário se valer de diferentes instrumentos regulatórios para proteger direitos.

Há uma área de estudo da IA que tem chamado a atenção da doutrina especializada em direito e tecnologia: o aprendizado de máquina (*machine learning* -- ML). Cada vez mais o uso de algoritmos de tomada de decisão de aprendizado de máquina, ou apenas, algoritmos de tomada de decisão, tem sido um foco de preocupação, principalmente pela imprevisibilidade e ininteligibilidade do seu comportamento. Assim, para delimitar e compreender o que são estes algoritmos de tomada de decisão, neste trabalho será utilizada a definição ampla proposta por Rovatsos et al. (2019, p. 18) de algoritmos de tomada de decisão, que podem ser compreendidos como:

> uma gama de ferramentas, que incluem sistemas que podem propor recomendações para uma tomada de decisão humana (assim como muitas ferramentas de avaliação de impacto), até sistemas que podem tomar uma decisão quase sem nenhum ou totalmente sem nenhum *input* ou supervisão humana (...)[54].

[52] Conforme noticiado pela mídia. Ver em: BBC. Pegasus: o que é o sistema que espionou jornalistas, ativistas e advogados. *BBC*, 19 de jul. de 2020. Disponível em: https://www.bbc.com/portuguese/internacional-57885795. Acessado em 30.01.2022.

[53] Conforme noticiado pela mídia. Ver em: CABRAL, Carlos. Embargo dos EUA contra o software espião Pegasus não torna ambiente cibernético mais seguro. *El País Brasil*, 12 de dez., de 2021. Acessado em 30.01.2022.

[54] Tradução livre de: "range of tools, ranging from systems which might provide advice to a human decision-making (as with many risk assessment tools), through to systems which might make a decision with almost no human input or oversight".

De forma mais específica, mas não necessariamente precisa, este trabalho se alinha à perspectiva de Mittelstadt et al. (2016, p. 3) em relação ao tipo de algoritmo que é objeto de estudo, que seriam aqueles responsáveis por tomarem decisões que são consideradas confiáveis (sob um ponto de vista subjetivo, e não de acurácia) com base em regras complexas que dificultam a compreensão e ação humanas. Nas palavras dos autores: "estamos interessados em algoritmos cujas ações são difíceis para os humanos predizer ou que a lógica de tomada de decisão é difícil de explicar depois do fato".[55] Estes tipos de algoritmos de tomada de decisão tipicamente envolvem a adoção de técnicas de aprendizado de máquina, razão pela qual é importante compreender o que é esta técnica e os seus efeitos.

Assim, será importante delimitar quando um sistema ou uma ferramenta será considerado como um algoritmo de tomada de decisão, e a maneira como ele é aplicado, pois a sua caracterização poderá implicar na incidência de normas jurídicas, como é caso do direito à explicação.

Em relação à técnica de aprendizado de máquina, esta tem testemunhado um aumento do seu uso "diante da sua habilidade de radicalmente melhorar a tomada de decisão baseada em dados, por um custo e em uma escala incomparáveis aos humanos" (Casey et al., 2019, p. 7).[56] Apesar de existir uma série de técnicas de IA[57], a área de aprendizado de máquina tem apresentado resultados frutíferos e bem-sucedidos de automação de tarefas que supostamente necessitariam de cognição humana (Surden, 2014, p. 88). A abrangência do uso desses sistemas para a tomada de decisão faz com que os mesmos exerçam um papel importante na regulação das nossas vidas. Por este motivo, cabe ser feita uma breve apresentação do que seja ML.

1.2 Breves notas sobre a área de Aprendizado de Máquina (*machine learning* – ML)

O aprendizado de máquina é uma sub-área da Ciência da Computação, que tem por objetivo criar algoritmos que evoluem seu comportamento e, portanto, o deu

[55] Tradução livre de: "(…) we are interested in algorithms whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact. Algorithms that automate mundane tasks, for instance in manufacturing, are not our concern"
[56] Tradução livre de: "due to their ability to radically improve data-driven decision-making at a cost and scale incomparable to that of humans".
[57] Tal como Processamento de Linguagem Natural, Visão Computacional, Mineração de Dados, entre outras.

desempenho, a medida que aprendem com os dados que recebem[58]. Seu objetivo é delegar à máquina a identificação e o reconhecimento de padrões nos dados analisados, a fim de que esta possa, ao final, automatizar determinada tarefa ou realizar predições e conhecimento sobre determinados cenários. A inteligência atribuída à máquina está relacionada à sua capacidade de reconhecimento de padrões – até então uma tarefa comumente humana – com um alto nível de qualidade (Surden, 2014, p. 90).

A área de aprendizado de máquina é subdividida em três classes: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

A técnica de aprendizado supervisionado pode ocorrer de diferentes maneiras, com diferentes técnicas[59] que, basicamente consistem em:

> Classificar *inputs* de acordo com um conjunto de classificações finitas (ou às vezes, infinitas). Tipicamente, é fornecido ao sistema de aprendizado um conjunto de dados de treinamento, que foram classificados a mão. O sistema em seguida tenta aprender a partir desse dado treinado como classificar o mesmo dado (comumente uma tarefa fácil) e também classificar novos dados que ainda não foram vistos (Coppin, 2004, p. 268)[60].

Por sua vez, a técnica de aprendizado não supervisionado encontra padrões ou reúne determinados grupos de dados que sejam semelhantes entre si, formando heurísticas a partir desses conjuntos/grupos de dados analisados sem que uma pré-categorização tenha sido feita por um humano[61]. Uma técnica comum nesse tipo de aprendizado é a clusterização, que é a reunião de dados em diferentes grupos (*clusters*) de acordo com a presença de elementos de similaridade entre eles.

O aprendizado de máquina por reforço é utilizado quando não há uma categorização prévia, tal como ocorre com o aprendizado supervisionado, havendo a necessidade de apontar para a máquina quando uma determinada tarefa foi

---

[58] Esse aprendizado difere do tipo de aprendizado cognitivo humano, o que Surden chama de "aprendizado funcional", ocorrendo a partir da experiência adquirida pelo sistema com a interação com o ambiente externo, fazendo com que o seu comportamento seja alterado com o intuito de melhorar a sua atuação (Surden, 2014, p. 89).

[59] Como: o *rote learning, concept learning, general-to-specifics ordering*, *simple algorithm, version spaces, candidate elimination, inductive bias, induction of decision trees* (Coppin, 2004, p. 268-286).

[60] Tradução livre de: "to classify inputs according to a finite (or sometimes infinite) set of classifications. Typically, a learning system is provided with a set of training data, which have been classified by hand. The system then attempts to learn from these training data how to classify the same data (usually a relatively easy task) and also how to classify new data that it has not seen.

[61] Conforme definição apresentada no site Math Works em <https://www.mathworks.com/discovery/machine-learning.html>. Acessado em 03.03.2020.

executada de maneira bem-sucedida ou não, penalizando e recompensando o aprendizado. Esse tipo de técnica é utilizado para programar uma máquina para jogar xadrez, ensinando as movimentações do tabuleiro e para que seja possível antecipar por probabilidade qual será a próxima jogada do seu oponente[62]. No entanto, sem um retorno (*feedback)* sobre se uma jogada é boa ou ruim – como se houve um xeque-mate seu ou do seu oponente – a máquina terá dificuldade em saber se o seu movimento foi positivo ou negativo, e qual deverá ser o próximo movimento que ela deverá tomar. Esse *feedback* é exatamente o reforço ou a recompensa do aprendizado, que pode ocorrer ao longo da execução da tarefa, ou ao final (Russell; Norvig, 2009, p. 830-831).

Uma das finalidades para as quais algoritmos de tomada de decisão que se valem de ML possui é a de fazer predições sobre os mais variados assuntos, desde predizer o tempo, ou recomendar livros, músicas, vídeos ou qualquer outro produto em plataformas digitais. A acurácia e eficiência dos algoritmos de ML, aliadas à maior facilidade de acesso a *hardware*, e à disponibilidade massiva de dados, acabaram barateando a capacidade de realizar estas previsões e predições, ampliando o seu uso aos mais variados contextos [63/64] (Agrawal et al., 2017). Contudo, apesar dos evidentes benefícios gerados pela tecnologia, a expansão do seu uso não vem sem riscos.

Um dos maiores desafios, se não o maior, relacionado ao desenvolvimento e uso de algoritmos de tomada de decisão de ML se refere à sua dificuldade em permitir que seus resultados sejam interpretados por e explicados para um humano. Isto porque as predições são baseadas em regras e heurísticas inferidas a partir dos padrões encontrados nos dados, podendo ocasionar correlações entre eventos que não possuem qualquer causalidade entre eles, gerando um conhecimento não

[62] O Deep Blue, criado pela IBM, foi o primeiro programa de IA a derrotar o campeão de xadrez mundial, Garry Kasparov, em 1997, com a utilização desse tipo de aprendizado. Como se verá mais a frente, em 2019 a Google desenvolveu um novo programa de xadrez, AlphaZero, aplica uma outra forma de aprendizado.

[63] O argumento dos autores pode ser representado na seguinte afirmação: "cheaper prediction brings more prediction. Simple economics" [63] (Agrawal et al., 2017)

[64] Como aponta o relatório de Stanford sobre os cem anos da IA: "Technological progress had also made the task of building systems driven by real-world data more feasible. Cheaper and more reliable hardware for sensing and actuation made robots easier to build. Further, the Internet's capacity for gathering large amounts of data, and the availability of computing power and storage to process that data, enabled statistical techniques that, by design, derive solutions from data. These developments have allowed AI to emerge in the past two decades as a profound influence on our daily lives (…)" (Stone et al., 2016, p. 51)

genuíno[65] (Mittelstadt et al., 2016, p. 5) ou até mesmo associações evidentemente incorretas (Selbst; Barocas, 2018, 1.123)[66]/[67]. Neste sentido, Selbst e Barrocas descrevem o que seria ML, e sintetizam os problemas advindos do mesmo da seguinte maneira:

> Ao invés de programar computadores manualmente com regras explícitas, o aprendizado de máquina se baseia em algoritmos de reconhecimento de padrões e uma série de exemplos para descobrir relações entre dados que podem servir de insumos confiáveis para a tomada de decisões. O poder do aprendizado de máquina se fundamenta não apenas na sua habilidade de retirar dos programadores a difícil tarefa de produzir instruções explícitas para os computadores, mas na sua capacidade de aprender exemplos sutis entre dados que podem passar despercebidos para os humanos ou até mesmo não serem reconhecidos. Esse poder pode tornar os modelos desenvolvidos com aprendizado de máquina extremamente complexos e, portanto, impossíveis de serem analisados por um ser humano[68] (Selbst; Barocas, 2018, 1.094).

Além da impossibilidade de se compreender a forma de trabalho e como esses algoritmos de ML alcançaram certos resultados ou fizeram certas predições, há outros desafios que surgem com a aplicação de algoritmos de ML, conforme se verificará a seguir.

[65] Como demonstrado no site "*Spurious correlations",* que evidencia a possibilidade de relacionar e associar eventos que ocorreram de maneira totalmente aleatória, sendo possível afirmar que há correlação entre eles, mas não causalidade. Ou seja, não é possível dizer que o evento X foi motivado, causado em razão da ocorrência do evento Y. Por exemplo, o site mostra que existiria correlação entre o número de pessoas que se afogaram em uma piscina entre os anos 1999 e 2009, e o número de filmes que o ator americano, Nicolas Cage, aparece naqueles anos. Ver em: < https://www.tylervigen.com/spurious-correlations> Acessado em 03.01.2020.

[66] Selbst e Barrocas apresentam um estudo feito por Rich Caruana sobre o uso de um modelo para prever complicações de pacientes com pneumonia. O modelo associou que uma pessoa com asma seria considerada como alguém com baixo risco de morte caso desenvolvesse um quadro de pneumonia, o que era algo claramente errado para especialistas. O motivo que causou esta associação incorreta eram os dados que foram utilizados pelo modelo: "The model was trained on clinical data from past pneumonia patients, and it turns out that patients who suffer from asthma truly did end up with better outcomes. What the model missed was that these patients regularly monitored their breathing, causing them to go to the hospital earlier". De acordo com Caruana, após o ocorrido, modelos mais simples e passíveis de serem inspecionados deveriam ser utilizados (Selbst; Barocas, 2018, p. 1.123).

[67] Por tal motivo que aplicações desses tipos de algoritmos têm levantado questões éticas relacionadas à epistemologia. Nesse sentido: BOYD, Danah; CRAWFORD, Kate. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, & Society* 15:5, 2012, p. 662-679.

[68] Tradução livre de: "Rather than programming computers by hand with explicit rules, machine learning relies on pattern-recognition algorithms and a large set of examples to uncover relationships in the data that might serve as a reliable basis for decision-making.[51] The power of machine learning lies not only in its ability to relieve programmers of the difficult task of producing explicit instructions for computers, but in its capacity to learn subtle relationships in data that humans might overlook or cannot recognize. This power can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse"

1.2.1
Desafios dos usos de algoritmos de ML

Como se viu, algoritmos de aprendizado de máquina podem ser aplicados a vários contextos, tendo como insumo o uso de dados de diversas naturezas (como o uso de dados financeiros na bolsa de valores, ou de dados meteorológicos para a predição do clima). Este trabalho, no entanto, tem por objeto de estudo os sistemas de IA e algoritmos de tomada de decisão que fazem uso de dados pessoais para viabilizar o seu treinamento e dependem destes dados para implementar uma ação. Assim, o foco das análises feitas ao longo deste e dos próximos capítulos é sobre o uso de dados pessoais por estes artefatos, e os impactos dos seus usos, especialmente aos direitos fundamentais à proteção de dados pessoais, privacidade e intimidade da pessoa impactada e da coletividade.

Assim, o principal insumo para o desenvolvimento dos sistemas de IA são os dados coletados e utilizados para treinar o modelo. Sua capacidade de aprendizado, sua autonomia e o seu comportamento estão intrinsecamente relacionadas à abundância e à qualidade dos dados tratados, acessados, analisados e utilizados para a sua aprendizagem. A máxima do aprendizado de máquina é que um modelo de ML é tão bom quanto os dados que o alimentam[69]. Atualmente, como vivemos uma "datificação" (*datafication*)[70] (Mayer-Schönberger; Cukier, 2013, p. 97), onde tudo é capaz de ser transformado e quantificado em dados, aumenta-se o espectro de temas submetidos a análises probabilísticas feitas por algoritmos, e a possibilidade de associações com novos tipos de informações.

A coleção destes dados no contexto do *big data*[71] – literalmente uma enorme quantidade de dados coletados e armazenados em bancos de dados – tem por objetivo realizar análises, predições e correlações, se valendo de algoritmos e

[69] Ver: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Acessado em 26.06.2020.

[70] De acordo com os autores, no contexto da datificação, qualquer informação pode ser capturada e transformada em dado legível por máquinas, para posterior análise. Como consequência, cria-se a quantificação de quase tudo que fazemos: nossos hábitos, gostos musicais, livros, filmes, sentimentos, batimentos cardíacos, ciclo menstrual, entre outros. Além disso, a novidade da datificação no contexto do *big data* é que os dados coletados para uma determinada finalidade muitas vezes são utilizados para um novo fim, totalmente diferente, e neste uso secundário, inédito e inesperado, as análises realizadas se mostram bastante efetivas.

[71] O potencial do *big data* pode ser compreendido a partir do que Doug Laney cunhou como "3 V": volume, velocidade e variedade de informações. Volume se refere à grande quantidade de dados coletados; velocidade é a rápida capacidade de coleta e transmissão destes dados, que pode ocorrer em tempo real; e, variedade se refere aos diferentes tipos de conteúdo e fontes de onde são extraídos os dados (Gomes, 2019, p. 25-26).

modelos de aprendizado de máquina, para apoiar ou automatizar uma tomada de decisão. Assim, o que antes se atribuía genericamente ao contexto de *big data*, atualmente pode ser atribuído mais especificamente às técnicas de aprendizado de máquina, embora não se ignore as diferenças entre elas[72].

Os dados utilizados para o aprendizado dos modelos de ML influenciam diretamente na qualidade dos seus resultados. Um exemplo disso foi apresentado pela pesquisadora do laboratório da universidade americana MIT, Joy Buolamwini, e a cientista de dados, Timnit Gebru, que à época era pesquisadora do Google. As autoras identificaram que os softwares para reconhecimento facial desenvolvidos por três grandes empresas (Microsoft, IBM e Face+++) não eram capazes de reconhecer usuários de peles pretas. Os sistemas de reconhecimento facial quando testados em pessoas com tons de pele escuros eram mais propensos a erro, porque tais sistemas não teriam sido treinados com imagens diversificadas, o que afetava principalmente o reconhecimento de rostos de mulheres negras.

Para melhorar estes resultados, as autoras construíram seu próprio banco de dados, adotando os seguintes cuidados para evitar discriminações no seu resultado: equilíbrio na presença de mulheres e homens nos dados de treinamento, se valer de imagens de pessoas com diferentes tons de pele, além de equilibrar a qualidade entre as fotografias em termos de iluminação, *pixels*, e posição da pessoa retratada. O cuidado das pesquisadoras não se dirigiu apenas em diversificar seu conjunto de dados (*dataset)*, mas em nivelar a qualidade destes dados (iluminação e *pixels*) para evitar ruídos e diferenças no aprendizado. Em suas conclusões, as autoras apontaram para a necessidade de que sejam divulgadas as métricas dos testes aplicados nestes tipos de softwares, a fim de que possam ser avaliadas as condições e os contextos em que se deu o treinamento desses modelos, além de chamar a atenção dos responsáveis pelo desenvolvimento sobre os riscos discriminatórios dos resultados quando estes tipos de cuidados não são adotados.

Outra dificuldade envolvendo elementos relacionados aos dados utilizados para treinar algoritmos de ML, consiste em verificar se estes são capazes de medir o que se busca com o modelo de aprendizado de máquina (Lehr; Ohm, 2017, p. 679). Ou seja, é necessário verificar se os exemplos analisados permitem que o

[72] Mayer-Schönberger e Cukier diferenciam o termo *big data* das técnicas de aprendizado de máquina pois, segundo os autores, o *big data* não busca ensinar à máquina a pensar como um humano, que visa a mera aplicação de fórmulas matemática a quantidades enormes de dados.

modelo de aprendizado de máquina aprenda e tenha boa capacidade de generalização quando analisar novos dados, e que os dados sejam representativos do mundo real que o conjunto de dados visa retratar (Lehr; Ohm, 2017, p. 680). Parece algo banal, mas a sua inobservância pode ter um alto impacto nos resultados gerados. Qualquer modelo preditivo será sempre uma mera simplificação da realidade, o que implica em reconhecer que informações serão deixadas de fora, levando à apresentação de erros ou caracterizações incorretas do que se visa representar (O'Neil, 2016, p. 5).

Por exemplo, uma instituição financeira desenvolveu um algoritmo de aprendizado de máquina para avaliar quais clientes que pediram empréstimos possuem a maior probabilidade de não cumprir com o pagamento e, com base nesta análise, deferir ou não deferir o pedido de empréstimo. Caso os dados utilizados para o treinamento do algoritmo se refiram apenas a dados históricos de pessoas que solicitaram e tiveram concedidos o empréstimo, o algoritmo terá problemas em analisar casos de pessoas que teriam seu pedido de empréstimo rejeitado pelo banco, pois o modelo nunca teria visto exemplos como esses (Lehr; Ohm, 2017, p. 680).

Estes são alguns dos diversos exemplos[73] que existem e apontam para como a qualidade dos dados reflete no resultado (*output*) do modelo, mostrando um impacto real que a interação com a tecnologia pode ter em diferentes aspectos na vida de uma pessoa, desde a permissão de acesso ao crédito, definição de um diagnóstico médico, ou a experiência com um aplicativo. Contudo, os dados não são os únicos responsáveis pelos problemas advindos dos algoritmos de aprendizado de máquina.

David Lehr e Paul Ohm (2017) apontam que existem diversas etapas, antes e depois da coleta e uso de dados, que podem causar percalços e erros nos resultados de modelos e algoritmos de ML. Um deles, que é especialmente complexo, diz respeito à dificuldade em representar um determinado objetivo que o algoritmo visa atender em uma medida que seja quantificável (Lehr, Ohm, 2017, p. 675).

[73] Neste sentido, o relatório produzido em abril de 2019 pela *AI Now Institute*, um hub interdisciplinar que realiza pesquisa sobre as implicações sociais que a IA têm gerado, foram reunidos diversos casos que revelam a discriminação, especialmente de gênero e de raça, de produtos e programas de IA, que foram motivados pelos dados utilizados como *input*. Ver em: WEST, S.M., WHITTAKER, M; CRAWFORD, K. *Discriminating Systems: Gender, Race and Power in AI.* AI Now Institute, 2019. Disponível em: <https://ainowinstitute.org/ discriminatingsystems.html> Acessado em 07.11.2020.

Por exemplo, modelos preditivos desenvolvidos para os times de beisebol, como narrado no livro de Michael Lewis, "Moneyball: The Art of Winning an Unfair Game" (2004), que deu origem ao filme "Moneyball", pode ser considerado um exemplo "inofensivo" de uso de algoritmos de ML. Os dados utilizados são públicos (disponibilizados até mesmo em cartões de coleção de jogadores), as métricas são objetivas (tempo do *home run*[74], velocidade da bola, vezes de acerto) e o objetivo é claro, direto e transparente: selecionar os melhores jogadores, colocá-los nas posições que eles mais se destacam e melhor pontuam, com o intuito de vencer a temporada.

Em contrapartida, um modelo preditivo mais complexo seria a avaliação de professores. Como definir o que é um(a) bom(a) professor(a)? Quais dados seriam utilizados como parâmetros? Poderiam ser: a média de aprovação de uma determinada turma, avaliação dos pais, do(a)s próprio(a)s aluno(a)s ou dos pares do(a) professor(a)? Como narrado por O'Neil em um caso real[75], estas escolhas, que foram incorporadas a um algoritmo, nem sempre vão ser compreendidas por ou explicitadas para aqueles afetados diretamente pelo modelo, no caso, o(a) professor(a). A ausência de transparência em relação a estas escolhas afeta o reconhecimento da legitimidade do algoritmo de tomada de decisão, pois a pessoa impactada por ele não terá a possibilidade de conhecer os motivos que resultaram em sua avaliação. Este resultado, por exemplo, poderá ser utilizado pela escola tanto para justificar a sua demissão, quanto o atraso na sua progressão de carreira. Em ambos os casos a compreensão dos critérios, dos dados utilizados, e dos motivos que ocasionaram estes resultados devem ser compartilhados para que seja passível de verificação se as regras que autorizam um ou outro caso foram atendidas.

Além disso, o parâmetro utilizado para verificar o comportamento de um algoritmo de ML, e o que foi definido como a taxa de sucesso, que seria a avaliação sobre a capacidade do algoritmo em concretizar o objetivo final para o qual ele foi desenvolvido, também pode levantar dúvidas e questionamentos sobre os seus resultados. Como aponta Cathy O'Neil, a taxa de sucesso é capaz de influenciar diretamente nos valores que serão observados ou negligenciados pelo algoritmo

[74] *Home run* é quando o rebatedor lança a bola para fora do campo de beisebol, e em seguida deve completar a volta no campo e nas bases para pontuar para o seu time.
[75] Cathy O'Neil narra o caso de Sarah Wysocki, uma professora nos EUA que foi demitida por conta de uma avaliação realizada por um algoritmo, apesar de Wysocki ser uma professora adorada pelos alunos e diretor da escola (O'Neil, 2016).

(O'Neil, 2016, p. 6). Um exemplo disto, que levanta dilemas éticos, diz respeito à programação de carros autônomos. Dependendo das variáveis imputadas ao código o carro poderá ter uma direção utilitarista ou deontológica, ora valorizando a maior efetividade da direção, ora sendo mais cauteloso. Isso implica em uma escolha do que será definido como sucesso.[76] Outro exemplo, no cenário de pontuação de crédito (*credit scoring*), seria o caso de que uma instituição financeira que baliza o seu modelo para ter como objetivo autorizar mais empréstimos em cenários de maior risco de inadimplência, pois o que se pretende é incentivar economicamente uma determinada área, onde tradicionalmente os moradores possuem dificuldades financeiras (Selbst, Barrocas, 2018, p. 1.130).

Em suma, alguns dos desafios apontados neste subcapítulo apontam para como o amplo uso de algoritmos de ML desconstroem a crença de que eles seriam neutros, objetivos e racionais.[77] Isto porque o desenvolvimento de algoritmos de ML implica necessariamente em escolhas subjetivas por parte dos seus criadores (Selbst, Barrocas, 2018, p. 1.130), seja na seleção dos parâmetros, na taxa de sucesso ou no banco de dados que será utilizado. Isto acaba refletindo nos valores que são incorporados ao código, uma ideia sintetizada por Cathy O'Neil ao afirmar que os "algoritmos seriam opiniões incorporadas em código".[78] Isto, por si só, revela a importância de assegurar transparência e estruturar meios que possibilitem uma prestação de contas sobre o processo de desenvolvimento de algoritmos e dos seus resultados.

Como se pode notar, os exemplos apresentados até este momento, em sua grande maioria, se valem do uso de dados pessoais para treinar algoritmos de tomada de decisão. Estes algoritmos são treinados para reconhecer padrões de comportamento a partir da análise de dados pessoais coletados, e categorizar pessoas dentro de determinados perfis para obter ou predizer certo

[76] A Universidade americana MIT desenvolveu a plataforma "Moral Machine", que tem por objetivo observar a "perspectiva humana em relação às decisões morais feitas pelas inteligências das máquinas, como em carros autônomos". A plataforma apresenta dilemas morais envolvendo carros autônomos, e os participantes têm que indicar suas ações de comportamento para, ao final, a plataforma comparar a sua resposta com a de outros participantes, e apontar para o seu posicionamento moral. Ver em: https://www.moralmachine.net/hl/pt. Acessado em 05.10.2021.

[77] O'Neil, 2016; Kitchin, 2014; Ananny; Crawford, 2018; Boyd; Crawford, 2012; Mittelstadt et al.; 2016.

[78] Tradução livre de: "*algorithms are opinions embedded in code*". O'Neil, Cathy. The era of blind faith in big data must end. TedTalk, 2017. Disponível em: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end?language=en. Acessado em 09.06.2021.

comportamento [79]. É neste sentido que Rodotà alerta para a necessidade de reconhecer que as pessoas deixaram de ser meros corpos físicos, para se tornarem corpos eletrônicos (Rodotà, 2008).

A construção destes corpos eletrônicos, e estas identidades digitais, foi facilitada pelo atual contexto tecnológico e de datificação. As pessoas são classificadas e categorizadas a partir de inferências realizadas sobre aspectos da sua personalidade, para que uma ação seja tomada a partir da predição de determinados gostos, preferências e comportamentos. Esta prática, conhecida como perfilamento *(profiling*) ou perfilamento automatizado (*automated profiling*), é um dos principais pressupostos para o uso de algoritmos de tomada de decisão e de ML, conforme se verá a seguir.

1.3
A aplicação de algoritmos de tomada de decisão e aprendizado de máquina para a construção de perfis e realização de inferências

A atual capacidade tecnológica de gerar e processar dados sobre tudo que pode ser conectado a dispositivos e à internet ampliou e diversificou as possibilidades de análises e combinações que podem ser feitas com estes dados. Por exemplo, por meio de um relógio inteligente é possível registrar com precisão o batimento cardíaco, os níveis de oxigênio no sangue[80] e a qualidade do sono do usuário; a partir de análises sobre estas condições de saúde, são feitas sugestões para que a pessoa dê alguns passos para movimentar o seu corpo ou que beba água; o relógio é capaz de chamar até mesmo o serviço de emergência caso identifique que a pessoa teve uma queda brusca, avisar ao usuário que o seu corpo não se encontra mais em movimento, ou que o seu batimento cardíaco está diferente.[81] Estes dados sensíveis são gerados indiretamente a partir da interação das pessoas com estes dispositivos. Ou seja, tais dados não são fornecidos diretamente, tal como

[79] Como descreve Kitchin, os algoritmos de tomada de decisão de ML estão aptos a pesquisar, colar, organizar categorizar, agrupar, juntar, analisar, criar perfis, modelar, simular, visualizar e regular pessoas, processos e lugares, moldando a maneira como as pessoas veem o mundo (Kitchin, 2014, p. 11), e como os outros veem as pessoas.

[80] Tal qual o Apple Watch Series 6. Ver em: https://www.apple.com/br/apple-watch-series-6/. Acessado em 07.07.2021.

[81] O Apple Watch detectou a queda do penhasco de um homem, acionando o serviço de emergência americano 911, que conseguiu localizá-lo pelo GPS do aparelho, o que salvou a sua vida. Ver em: https://www.techtudo.com.br/noticias/2019/10/apple-watch-chama-emergencia-e-salva-vida-de-homem-que-caiu-de-penhasco.ghtml. Acessado em 09.07.2021.

ocorre quando um cadastro ou formulário é preenchido, mas são produzidos a partir do monitoramento deste contato constante.

Esta coleta massificada de dados, possibilitada pela datificação, desenvolvimento de tecnologias de baixo custo, e aplicação de técnicas inovadoras, viabilizaram novos modelos de negócio, assim como novos tipos de tratamento de dados pessoais. Esta variedade da natureza e dos tipos de dados que podem ser coletados permite que estas informações inicialmente pensadas para uma certa finalidade (por exemplo, avaliar o comportamento do condutor no volante para atribuir e tornar dinâmico o valor do seguro do seu automóvel), possam ser utilizadas para outros fins (tal como: o perfil do condutor também poderá ser um bom fator preditivo sobre os riscos que ele estaria disposto assumir, e que podem ser relevantes para definir o valor do seu seguro de vida). São justamente estes novos usos, aplicados em um contexto diferente daquele em que o dado foi inicialmente coletado, que geram novas e inesperadas informações sobre um grupo de pessoas ou uma pessoa. Consequentemente, os mesmos ameaçam direitos fundamentais, em especial, a liberdade, autonomia, igualdade, privacidade e proteção de dados pessoais.

A elaboração de perfis de comportamento de uma pessoa é uma das técnicas aplicadas no contexto do *big data* e da inteligência artificial, que apresentam riscos a tais direitos fundamentais, assim como para os fundamentos do Estado Democrático de Direito. Estas ameaças foram reconhecidas e destacadas nos votos dos Ministros do Supremo Tribunal Federal (STF)[82], no caso da Ação Direta de Inconstitucionalidade (ADI) nº 6.387/DF[83] que, antes da aprovação da PEC n. 17/2019[84] que incluiu a proteção de dados pessoais no rol de direitos fundamentais, em um julgamento histórico, reconheceu implicitamente na Constituição Federal de 1988 a existência de um direito fundamental à proteção de dados pessoais. A relatora da ação, Ministra Rosa Weber, sustentou este argumento no direito à

[82] Neste sentido, destacamos um trecho do voto do Ministro Alexandre de Moraes que ressalta esta preocupação: "Pela primeira vez, no inciso XII, nosso sistema constitucional expressamente previu a proteção constitucional ao sigilo de dados, mostrando, dessa forma, a importância, na visão ocidental de democracia, da interligação entre democracia e Estado de Direito e democracia e limitação do exercício do Poder - ambos indissoluvelmente combinados -, sendo imprescindível a observância dos direitos e garantias fundamentais. São comandos proibitórios expressos dirigidos ao Estado não violar a intimidade, a vida privada e o sigilo de dados".

[83] STF, ADIs ns. 6.387, 6.389, 6.390 e 6.393, sob a relatoria da Ministra Rosa Weber, j. em 07/05/2020, publicado em 14/11/2020.

[84] Ver nota de rodapé n. 37.

autodeterminação informativa, constituído a partir da liberdade individual, da privacidade, do livre desenvolvimento da personalidade e da inviolabilidade do sigilo das comunicações (art. 5º, *caput,* e incisos X e XII, da CF).

O julgamento da referida ADI analisou a constitucionalidade da Medida Provisória (MP) n. 954/2020, que autorizava que as empresas de telefonia pudessem compartilhar com o Instituto Brasileiro de Geografia e Estatística (IBGE) dados pessoais dos seus usuários para fins estatísticos durante a pandemia do coronavírus. O voto da relatora, ao discutir os riscos do compartilhamento de dados pessoais, como nome, telefone e endereço, reconheceu os perigos, no atual contexto tecnológico, dos usos dinâmicos que podem ser aplicados a estes dados, citando expressamente que um deles seria a formação de perfis extremamente detalhados.[85] Ao final, a Corte, por maioria, referendou a medida cautelar deferida pela Ministra Relator Rosa Weber, para suspender a eficácia da referida MP.

A compreensão do direito à proteção de dados pessoais enquanto direito fundamental já vinha sendo ventilada pela doutrina (Doneda 2011; Mendes; 2014; Mulholland, 2018), a partir de uma interpretação conjunta do art. 5º, X, da CF, da garantia de *habeas data,* e do princípio da dignidade da pessoa humana (Mendes, 2014, p. 172), diante dos riscos causados aos direitos da personalidade, à luz dos princípios constitucionais da igualdade e liberdade (Doneda, 2011, p. 103). Este entendimento foi selado na Constituição Federal de 1988, por meio da aprovação da Proposta de Emenda à Constituição (PEC) 17/2019, culminando na promulgação da Emenda Constitucional nº 115, de 2002, que prevê a proteção de dados pessoais no art. 5º, LXXIX, da CF, e acrescenta o inciso XXX, ao art. 22, e o inciso XXVI, ao art. 21, da CF, para fixar competência privativa da União para legislar sobre o tema.

O reconhecimento da proteção de dados enquanto direito fundamental ganha um âmbito de proteção duplo: ao mesmo tempo que visa proteger o indivíduo em relação aos riscos à sua personalidade em razão da coleta, tratamento, uso e

[85] Nos termos do voto na Ministra Relatora: "Certamente há quem ainda se lembre de que há poucas décadas, antes da ubiquidade da telefonia móvel, era comum a edição de listas telefônicas impressas contendo nomes, telefones e endereços dos assinantes residenciais e comerciais dos serviços de telefonia em uma dada localidade. Além de ser facultado aos usuários dos serviços de telefonia optarem pela exclusão dos próprios dados dessas listas, é crucial ter presente que o que podia ser feitos a partir da publicização de tais dados pessoais não se compara ao que pode ser feito no patamar tecnológico atual, em que poderosas tecnologias de processamento, cruzamento e filtragem de dados permitem a formação de perfis individuais extremamente detalhados".

circulação de dados pessoais, também busca garantir ao indivíduo que tenha a garantia de controlar o fluxo de seus dados pessoais (Mendes, 2014, p. 176). Quanto a este último ponto, a ideia de controle sobre os dados pessoais está intimamente relacionada ao direito à autodeterminação informativa, firmado em decisão histórica pela Corte Alemã de 1983[86], que reconheceu o direito subjetivo de um indivíduo conhecer e controlar o fluxo de suas informações. Este direito também foi expressamente previsto e reconhecido pelo STF na mencionada ADI 6.387/DF.

No contexto de constituição destes perfis, o direito à autodeterminação informativa,[87]/[88] definido por Rodotà como "o direito de manter o controle sobre suas próprias informações e de determinar a maneira de construir sua própria esfera particular" (Rodotà, 2008, p. 15), possui um grande potencial de ser ameaçado ou até mesmo não observado, diante da ausência de transparência, conhecimento e, portanto, de controle, do sujeito em relação às informações que são geradas sobre si e aos seus usos futuros. Isto porque quando há o desenvolvimento destes perfis, gera-se uma distância entre o motivo que justificou a coleta de dados pessoais, as informações geradas a partir destes dados, e os usos atribuídos a estas informações. Este é um dos exemplos[89] que geram um questionamento sobre a efetividade da autodeterminação informativa, visto que a dinâmica e a opacidade inerente a tais técnicas acabam criando verdadeiros óbices para gerenciar este controle.

[86] Esta expressão teve origem no julgamento sobre a constitucionalidade do censo populacional previsto na "Lei do Rencenseamento de População, Profissão, Moradia e Trabalho, de 25/03/1983, pela Corte Constitucional alemã (*Bundesverfassungsgericht,* 15 dezembro 1983, *Neue Juristische Wochenschrift*, 1983, p. 419). Neste julgamento, decidiu-se pela inconstitucionalidade parcial da referida lei, sob o argumento de que a mesma possuía dispositivos que ameaçavam o direito fundamental ao livre desenvolvimento da personalidade dos indivíduos, visto que estes perdiam a possibilidade de controlar o fluxo das suas informações pessoais coletadas pelo Estado. Conforme aponta Laura Schertel Mendes, "A sentença da Corte Constitucional, na sua formulação de um direito à autodeterminação da informação, criou um marco para a teoria da proteção de dados pessoais e para as subsequentes normas nacionais e europeias sobre o tema, ao reconhecer um direito subjetivo fundamental e alçar o indivíduo a protagonista no processo de tratamento de seus dados" (Mendes, 2014, p. 31)

[87] Tal direito é um desdobramento do conceito de privacidade formulado por Warren e Brandeis em 1890, formulado como o direito a ser deixado só (*the right to be left alone*), que compreendia a privacidade apenas sob sua dimensão negativa, de não invasão de terceiros na esfera individual e na vida privada. A autodeterminação informativa representa um avança do conceito de privacidade, que passa a ser lida sob uma dimensão positiva, exigindo de terceiros, tanto o Estado, quanto agentes privados, que estabeleçam garantias que protejam os dados pessoais dos cidadãos (Moraes, 2008, p. 7).

[88] Tal direito também encontra respaldo na Lei Geral de Proteção de Dados Pessoais Brasileira (Lei n. 13.709/2018 -- LGPD), constituindo-se como um dos seus fundamentos: "Art. 2º A disciplina da proteção de dados pessoais tem como fundamentos: (…) II - a autodeterminação informativa (…)"

[89] Outro exemplo seria o consentimento solicitado aos usuários de aplicativos e plataformas para o tratamento de dados, em que é preciso "ceder" à coleta de todos os dados solicitados, sob pena de criar-se um impedimento para o uso e acesso aos serviços e bens ofertados.

Em relação à formação destes perfis comportamentais, estes podem ser formados tanto pelos dados que uma pessoa fornece ativamente (i.e. gênero, idade, endereço) ou que são captados a partir da sua interação com uma plataforma ou dispositivo (i.e. tal qual ocorre com o relógio inteligente, como frequência cardíaco, número de passos, etc). A construção destes perfis tem como pressuposto o tratamento de dados, em especial, dados pessoais, com o uso de tecnologias (Hildebrandt, 2008, p. 17), e técnicas, tais como métodos estatísticos e técnicas de inteligência artificial, capaz de sintetizar "hábitos, preferências pessoais e outros registros da vida desta pessoa" (Doneda, 2019, p. 151).

O perfilamento (*profiling)* ou perfilamento automatizado (*automated profiling)*, pode ser aplicado a indivíduos ou grupos, e visa avaliar propensões individuais e/ou coletivas, com o intuito de predizer tendências futuras, comportamentos ou o destino de uma pessoa ou um grupo de pessoas classificadas dentro de um dado perfil (Doneda, 2019, p. 151). Isto é feito com o objetivo de apoiar uma decisão, que pode se dar sem qualquer intervenção humana (Hildebrandt, 2008, p. 18). A formação de perfis tem cada vez mais variadas finalidades, e é aplicada nos mais diversos contextos, como: recomendação de livros, roupas ou músicas, definição de qual tipo de propaganda será recebida[90], realocação de pessoas no mercado de trabalho[91], investigação para apurar potenciais fraudes contra o Estado[92], definição de um diagnóstico médico, ou identificação de padrões de comportamento de uma "típica família brasileira de classe média".[93] Percebe-se a variedade de contextos e finalidades para as quais tais

[90] Ciente dos potenciais discriminatórios de suas práticas de campanhas de anúncios pagos em sua rede social (*targeting)*, o Facebook anunciou que iria alterar a maneira como estas campanhas são realizadas, removendo o uso de certos dados em determinadas campanhas envolvendo oportunidades de moradia, emprego e questões relacionadas a crédito. Ver em: IVES, Nat. Facebook Axes Age, Gender and Other Targeting for Some Sensitive Ads. *The Wall Street Journal,* 19 de março de 2019. Disponível em: https://www.wsj.com/articles/facebook-axes-age-gender-and-other-targeting-for-some-sensitive-ads-11553018450. Acessado em 01.12.2021.

[91] NIKLAS, Jędrzej. *Can an algorithm hurt? Polish experiences with profiling of the unemployed*. Centre for Internet and Humans Right, março de 2017. Disponível em: https://cihr.eu/can-an-algorithm-hurt/. Acessado em 01.12.2021.

[92] NAÇÕES UNIDAS. Landmark ruling by Dutch court stops government attempts to spy on the poor – UN expert. Escritório do Alto Comissariado das Nações Unidas para os Direitos Humanos, 05 de fev. de 2020. Disponível em: shorturl.at/arxFN. Acessado em 13.02.2021.

[93] Conforme noticiado pelo site Intercept Brasil, a companhia telefônica Vivo compartilhou dados considerados "anonimizados" com o governo do Espírito Santo para que fosse realizado um estudo sobre as tendências de comportamento da família brasileira de classe média. A reportagem pode ser acessada em: https://theintercept.com/2020/04/13/vivo-venda-localizacao-anonima/, e o estudo do governo aqui: https://observatoriodoturismo.es.gov.br/Media/observatorio/Pesquisas/Telefonia%20M%C3%B3vel/Descritivo%20Metodol%C3%B3gico.pdf. Ambos acessados em 15.07.2021.

técnicas são aplicadas, que podem ser empregadas pela iniciativa privada, visando categorizar (potenciais) clientes para apresentar ofertas personalizadas, ou pelo Estado para fins de investigação criminal ou atender a políticas sociais.

A datificação e a era do *big data* permitiram, também, o uso de outra técnica, comumente utilizada para a construção de perfis, chamada de mineração de dados[94] (*data mining*), que busca fazer "correlações, recorrências, formas, tendências e padrões significativos a partir de quantidades muito grandes de dados, com o auxílio de instrumentos estatísticos e matemáticos" (Doneda, 2019, p. 154). A mineração de dados realiza um tipo de predição baseada em dados históricos, em que decisões são tomadas com base naquelas, representando a probabilidade de que as coisas deverão se repetir daquela exata maneira no futuro (Hildebrandt, 2008, p. 18). As motivações do porquê uma predição ocorre, ou seja, quais seriam os eventos, dados ou fatos que influenciaram a predição não são reveladas, tampouco são exigidas por quem aplica estas técnicas. As causas ou razões são menos importantes diante de uma predição que viabilize uma tomada de decisão confiável.[95]

As técnicas de perfilamento buscam gerar correlações entre os dados (Hildebrandt, 2008, p. 20). As categorias dos perfis podem ser criadas a partir de correlações entre determinados eventos, em razão de seus atributos, estabelecendo um grupo e um conjunto de propriedades que pertencem ao perfil das pessoas que são alocadas juntas (por exemplo: pessoas que se interessam por livros de proteção de dados pessoais, seriam potenciais interessadas em livros sobre inteligência artificial). Ou, as correlações são feitas a partir de características comuns a um grupo de pessoas que já constituem uma comunidade (por exemplo: alunos e alunas de pós-graduação em direito), e que estão sujeitas a atributos comuns a elas, como, interesses por certos autores e autoras e hábitos de estudo. Portanto, técnicas de perfilamento podem formar grupos de perfis a partir de uma certa comunidade ou

[94] Apesar de a mineração de dados ser considerada semelhante à técnica de aprendizado de máquina, elas diferem entre si. Esta última tem por pressuposto o processo de aprendizado, e a adaptação e melhoria do comportamento do algoritmo para casos futuros e semelhantes. Como reiteradamente destacado, a técnica de aprendizado de máquina vem sendo aplicada a diversos contextos para a tomada de decisão, em que a construção de perfis é imprescindível e, por vezes, um pressuposto para a sua aplicação.

[95] Hildebrandt também aponta uma outra maneira de articular o tipo de conhecimento produzido pelo perfilamento, que seria a análise de perfis como se fossem hipóteses que não são desenvolvidas previamente à análise de dados, mas que surgem ao longo do processo de mineração de dados. Esta é uma abordagem também denominada como *discovery-driven approach*, que se opõe a *assumption-driven approach* (Hildebrandt, 2008, p. 18).

grupos já existentes (alunos e alunas de pós-graduação em direito), ou, a partir de traços comuns, formando um conjunto que não existia previamente (pessoas interessadas em livros de proteção de dados se interessam por livros de inteligência artificial). Em ambos os casos, o perfil é constituído por meio de inferências de dados coletados e, a partir da sua formação, as pessoas podem ser alocadas e categorizadas dentro de tais perfis, de acordo com o seu comportamento *online* (Hildebrandt, 2019, p. 20-21), ou ações e eventos que são datificados.

Os atributos podem ser preenchidos por todos que fazem parte do perfil do grupo (perfil distributivo), ou apenas por certos integrantes do grupo (perfil não distributivo). Neste último caso, isto pode ter consequências jurídicas e sociais relevantes, uma vez que o reconhecimento de uma pessoa dentro de um determinado perfil pode ensejar consequências mesmo que ela não possua as características necessárias (Hildebrandt, 2008, p. 21).

As correlações responsáveis por construir estes perfis não fornecem um motivo e/ou uma justificativa para explicar o atributo, a categorização e a designação de um grupo ou indivíduo dentro deste perfil, o que tem sido fonte de críticas[96], motivando muitos autores a defenderem que, em certos contextos, modelos correlacionais não devam ser aplicados (Zarsky, 2016; Selbst; Barrocas, 2018). Além do problema das "correlações espúrias" (*spurious correlations*), que ocorrem quando determinados eventos são altamente correlacionados, mas não possuem qualquer relação ou nexo causal entre eles,[97] resultados baseados em correlações podem desafiar a capacidade humana de compreender a relevância de certos critérios de decisão.

É neste contexto que explicações têm sido exigidas como uma maneira de avaliar a legitimidade e legalidade das motivações que ocasionaram certa tomada de decisão algorítmica. No clássico exemplo do uso de algoritmos para avaliação de pontuação de crédito, uma explicação sobre a decisão tomada pelo algoritmo poderia incluir o conhecimento do perfil que a pessoa está caracterizada, quais dados sobre ele ou ela foram utilizados pelo modelo estatístico para constituição de

[96] Mayer-Schönberger; Cukier; Annany; Crawford, 2018; Wachter; Mittelstadt, 2019; Hildebrandt; 2008; Zarsky, 2016; Selbst; Barrocas, 2018.

[97] Como o número de pessoas que se afogaram em piscinas nos EUA, e o número de filmes que o Nicolas Cage aparece, ou consumo per capita de queijo mozarela e doutorados concluídos em engenharia civil. Ver em: https://www.tylervigen.com/spurious-correlations. Acessado em 09.07.2021. O criador do site, Tyler Vigen, também é o autor do livro que leva o mesmo nome: "Spurious Correlations".

cada perfil, e os benefícios/prejuízos e consequências delegadas a cada perfil. Devem ser fornecidas informações suficientes para que a pessoa compreenda como os seus dados contribuíram para que ela fosse classificada dentro de um determinado grupo, como este foi constituído, e as consequências que podem decorrer para cada tipo de perfil.

As vantagens no desenvolvimento e uso de perfis é inegável. Eles geram eficiência e confiança na tomada de ação, tanto para entes públicos como privados, em que as generalizações de atributos e agrupamento de casos semelhantes facilita a aplicação de decisões idênticas a situações semelhantes (Hidelbrandt, 2008, p. 24),[98] gerando previsibilidade e segurança. Entretanto, o perfilamento também tem o revés de limitar a liberdade do indivíduo e o seu livre desenvolvimento (Doneda, 2019, p. 152), pois este acaba sendo "visto" e "tratado" a partir de gostos, preferências e inclinações pessoais de um certo recorte no tempo, havendo a anulação da "capacidade de perceber as nuances sutis, os gostos não habituais", com a imobilização de "perfis historicamente determinados" (Rodotà, 2008, p. 83). Consequentemente, há o risco de penalização daqueles que não se enquadram dentro de uma determinada categoria, tal como ocorre no caso do perfil não distributivo, em que muitas vezes a formação e uso destes grupos são realizados sem o conhecimento do indivíduo. Não por outro motivo, a decisão do STF na ADI nº 6.387/2020 apontou a necessidade de que o tratamento de dados pessoais deva assegurar a "proteção das cláusulas constitucionais assecuratórias da liberdade individual (art. 5º *caput*, da CF), da privacidade e do livre desenvolvimento da personalidade (art. 5º, X e XII, da CF), sob pena de lesão desses direitos". Por isso, o direito de conhecer estes dados e os seus usos torna-se essencial como uma maneira de proteger estes direitos fundamentais.

Como a maior parte dos insumos para a construção destes perfis envolve o tratamento de dados pessoais coletados em plataformas digitais e objetos inteligentes conectados à internet[99], isto acaba atraindo, quando vigentes nas respectivas jurisdições, a incidência de normas de proteção de dados pessoais. É o

[98] A autora parte das posições de Frederick Schauer, em "*Profiles, Probabilities and Stereotypes"* (2003).

[99] A expressão IoT (*Internet of Things,* e, em português Internet das Coisas) é utilizada para retratar esta realidade, e busca "designar a conectividade e interação entre vários tipos de objetos do dia a dia, sensíveis à internet" e se refere "a um mundo onde objetos e pessoas, assim como dados e ambientes virtuais, interagem uns com os outros no espaço e no tempo" (Magrani, 2018, p. 44).

caso, por exemplo, da aplicação do Regulamento Geral de Proteção de Dados Pessoais da União Europeia[100] (*General Data Protection Regulation* – GDPR[101]), relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados, que traz diversas previsões tratando da técnica de perfilamento. A referida norma, que entrou em vigor em maio de 2018, revogando a antiga Diretiva 95/46/CE do Parlamento Europeu e do Conselho que tratava sobre este assunto, é um dos mais importantes marcos normativos sobre o tema.

A influência do GDPR inspirou e incentivou a aprovação de diversos projetos de lei sobre a proteção de dados pessoais no mundo, tal como aqui no Brasil, culminando na promulgação, em 14 de agosto de 2018, da Lei Geral de Proteção de Dados brasileira (Lei n. 13.709/2018 – LGPD). O GDPR, portanto, é uma fonte de interpretação importante para a LGPD, assim como são as decisões tomadas pelas autoridades de proteção de dados pessoais Europeia, e de outros órgãos que atuam nesta área, como o antigo Grupo de Trabalho do Artigo 29 (GTA29),[102] atualmente sob a forma do Comitê Europeu para a Proteção de Dados (CEPD), responsável por prover guias interpretativos sobre os artigos do GDPR.

Neste contexto, como a atividade de perfilamento pode envolver o uso de dados pessoais para a formação de um perfil sobre um indivíduo (Bioni, 2019, p. 91), o GDPR traz uma série de previsões sobre o assunto, garantindo direitos aos titulares de dados, e impondo deveres aos agentes de tratamento. Assim, para entender como que esta legislação se aplica a este contexto, é preciso compreender dois personagens principais previstos nesta lei: o titular de dados e o controlador de dados. O primeiro é a pessoa natural a qual o perfilamento se refere e se dirige, e o segundo é a pessoa natural ou jurídica, de direito público ou privado, que realiza esta prática.

[100] Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016.
[101] Neste trabalho vamos nos valer do uso da abreviação em inglês do novo Regulamento Europeu de Proteção de Dados Pessoais, GDPR, para se referir a esta norma, tendo em vista a rápida assimilação que os leitores fazem entre a sigla e a norma europeia, diante do uso reiterado que acadêmicos, e a própria mídia, têm feito para se referir à regulação.
[102] O *Article 29 Working Party* era um órgão consultivo criado pela antiga Diretiva 95/46 composto por representantes de cada Estado Membro da União Europeia, e responsável por dar suporte técnico, recomendações sobre a referida Diretiva e emitir opiniões sobre práticas que pudessem afetar o direito à proteção de dados pessoais de cidadãos europeus. Com a entrada em vigor do GDPR, em 25 de maio de 2018, o referido grupo de trabalho foi substituído pelo Comitê Europeu para a Proteção de Dados (CEPD) (*European Data Protection Board* - EDPB). Ver em: <https://edpb.europa.eu/about-edpb/about-edpb_en> Acessado em 12.12.2019.

Em sua definição de dados pessoais, o GDPR indica que o titular de dados é a pessoa natural a quem se referem os dados pessoais, que são informações aptas a tornar o titular identificado ou identificável, de forma direta ou indireta por um identificador (nome, registro de identidade, dados de localização geográfica, identificador digital, ou outros elementos referentes à sua identificação física, fisiológica, genética, mental, econômica, cultural ou social) (art. 1(1), do GDPR[103]). Por sua vez, o controlador é a pessoa natural ou jurídica, pública ou privada, que em conjunto, ou individualmente, determina a finalidade e os meios do tratamento de dados pessoais (art. 1(7), do GDPR[104]).

Em seu art. 4(4), o GDPR define o que seria perfilamento, e ao longo do seu texto normativo limita diversos tipos de tratamento de dados pessoais que se valem desta técnica. De acordo com o GDPR, esta última seria "qualquer forma de tratamento automatizado de dados pessoais que consista em utilizar esses dados para avaliar certos aspectos de uma pessoa singular, nomeadamente para analisar ou prever aspectos relacionados com o seu desempenho profissional, a sua situação econômica, saúde, preferências pessoais, interesses, fiabilidade, comportamento, localização ou deslocações". Desta forma, a técnica de perfilamento prescinde (i) de um tratamento automatizado, (ii) no uso de dados pessoais para avaliação de aspectos pessoais de uma pessoa natural, e (iii) técnicas que analisam ou preveem aspectos pessoais.

A definição de perfilamento prevista no GDPR é mais restrita do que aquela proposta por Hildebrandt, fato que é reconhecido pela autora, pois a norma atrela o seu conceito à avaliação de aspectos pessoais de pessoas humanas, identificados ou identificáveis. Por sua vez, para a autora, os titulares de dados (*data subject)* podem ser uma pessoa humana ou não, sob uma perspectiva individual ou de grupo. A definição simples de perfilamento, a partir do seu processo, se propõe a ser uma

[103] Art. 1(1), do GDPR: "«Dados pessoais», informação relativa a uma pessoa singular identificada ou identificável («titular dos dados»); é considerada identificável uma pessoa singular que possa ser identificada, direta ou indiretamente, em especial por referência a um identificador, como por exemplo um nome, um número de identificação, dados de localização, identificadores por via eletrónica ou a um ou mais elementos específicos da identidade física, fisiológica, genética, mental, económica, cultural ou social dessa pessoa singular"

[104] Art. 1(7), do GDPR "«Responsável pelo tratamento», a pessoa singular ou coletiva, a autoridade pública, a agência ou outro organismo que, individualmente ou em conjunto com outras, determina as finalidades e os meios de tratamento de dados pessoais; sempre que as finalidades e os meios desse tratamento sejam determinados pelo direito da União ou de um Estado-Membro, o responsável pelo tratamento ou os critérios específicos aplicáveis à sua nomeação podem ser previstos pelo direito da União ou de um Estado-Membro".

atividade que busca descobrir correlações que podem identificar e representar um objeto humano ou não humano (individual ou coletivo). Na perspectiva do uso do perfilamento, a autora aponta que este visa individualizar e representar uma pessoa ou classificá-la dentro de uma categoria ou grupo, com o objetivo de avaliar riscos e oportunidades para o controlador em relação a ela (Hildebrandt, 2008, p. 19-20). Apesar desta perspectiva não antropocêntrica e individualista da autora, é possível se valer do seu conceito, sem que haja inconsistências ou contradições com a proposta deste trabalho, embora este seja mais abrangente do que a definição prevista no GDPR.

De acordo com o parecer do antigo Grupo de Trabalho do Artigo 29 (GTA29),[105] a atividade de perfilamento pode envolver a reunião de aspectos para mera análise, e não necessariamente deve implicar em uma predição. Logo, a constituição de um perfil poderia significar reunir informação sobre um indivíduo, ou um grupo de indivíduos, para analisar suas características ou padrões de comportamento com o objetivo de categorizá-lo em um grupo, ou fazer predições ou análises sobre o mesmo, tal como seu comportamento, sua habilidade em realizar uma tarefa ou identificar um interesse (GTA29, 2016, p. 7).

A LGPD, inspirada pelo GDPR, também traz em seu art. 5º uma série de definições e conceitos importantes. Assim como o GDPR, a LGPD traz em seu art. 5º, V, a definição de titular de dados, que é a "pessoa natural a quem se referem os dados pessoais que são objeto de tratamento", e o art. 5º, VI, define o controlador como a "pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais". Por sua vez, a LGPD não apresenta o conceito da técnica de perfilamento, mas o termo "perfil" aparece na norma em duas ocasiões. A primeira no artigo 12, § 2º, da LGPD, ao tratar dos

[105] O *Article 29 Working Party* era um órgão consultivo criado pela antiga Diretiva 95/46 composto por representantes de cada Estado Membro da União Europeia, e responsável por dar suporte técnico, recomendações sobre a referida Diretiva e emitir opiniões sobre práticas que pudessem afetar o direito à proteção de dados pessoais de cidadãos europeus. Com a entrada em vigor do GDPR, em 25 de maio de 2018, o referido grupo de trabalho foi substituído pelo Comitê Europeu para a Proteção de Dados (CEPD) (*European Data Protection Board* - EDPB). Ver em: <https://edpb.europa.eu/about-edpb/about-edpb_en> Acessado em 12.12.2019.

dados anonimizados[106], abrindo uma exceção à regra geral do art. 12, *caput,*[107] da LGPD, para indicar que os dados pessoais anonimizados quando utilizados para fins de formação de perfil comportamental de pessoa natural, se identificada, serão considerados como dados pessoais. O intuito desta previsão é proteger os titulares de dados das consequências que o tratamento de dados pessoais pode ter em sua esfera individual e em seu livre desenvolvimento (Bioni, 2019, p. 80).

Embora a LGPD use as expressões "determinada pessoa" e "identificada" em seu art. 12, § 2º, elas devem ser compreendidas em relação aos desdobramentos do tratamento para o indivíduo, e não se o perfil comportamental pode ser atribuído a uma pessoa em especial. Isto estaria justificada em uma interpretação sistemática no próprio conceito expansionista de dados pessoais, ao usar os termos pessoa identificada ou identificável, bem como em um dos fundamentos da lei, que visa promover o livre desenvolvimento da personalidade (art. 1º e 2º, VII, da LGPD) (Bioni, 2019, p. 80-81).

A segunda vez que o termo "perfil" é citado na LGPD é no artigo 20, que trata sobre o direito à revisão e explicação de decisões automatizadas (que será explorado no próximo subitem deste trabalho). O art. 20, da LGPD, incide quando decisões tomadas unicamente com base em tratamento automatizado afetarem os interesses dos titulares de dados, "incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade". Percebe-se, assim, certa semelhança entre a LGPD e o GDPR, ao associar à formação de perfis aspectos pessoais do titular de dados. No entanto, há uma diferença relevante, visto que a LGPD não atrela a técnica de perfilamento necessariamente a uma atividade automatizada, tampouco se refere à finalidade para a qual estes perfis poderão ser constituídos, se para mera análise ou para realizar predições.

Em relação ao uso de técnicas e tecnologias no tratamento de dados pessoais, vale ressaltar, ainda, que ambas as normas preveem uma série de direitos

[106] O art. 5º, inciso III, da LGPD define dados anonimizados como o: "dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento". Ainda, a técnica de anonimização é conceituada pelo art. 5o, inciso XI, da LGPD, como a: "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo".
[107] Art. 12, *caput*, da LGPD: "Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei, salvo quando o processo de anonimização ao qual foram submetidos for revertido, utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido".

e salvaguardas que exigem transparência em relação aos dados de entrada que vão ser utilizados em modelos estatísticos para a formação de perfis, podendo haver o uso de ML. Contudo, há pouca exigência de transparência em relação a como estas inferências são alcançadas, o que poderia revelar informações sobre quais são os dados tipicamente utilizadas para a constituição de perfis[108], a relevância confiada a cada um dos dados, e como eles influenciam na categorização de uma pessoa dentro de um ou outro perfil. Apesar de o GTA29 já ter reconhecido que as inferências e derivações feitas a partir de dados pessoais para a constituição de perfis e para decisões totalmente automatizadas seriam considerados como dados pessoais (GTA29, 2016, p. 20), [109] o mesmo grupo indicou que a maneira como essas inferências são alcançadas, comumente por meio de algoritmos e outras metodologias, estariam protegidas pelo segredo comercial e industrial dos algoritmos (GTA29, 2016b, p.10). Isto é um dos principais impedimentos para o pleno exercício do direito de ter acesso a qualquer informação relacionada ao algoritmo que dê pistas sobre como uma decisão algorítmica foi alcançada, um tema que será explorado no próximo capítulo.

Como já indicado, um aspecto importante das técnicas de construção de perfis comportamentais e mineração de dados é o distanciamento que é gerado entre a informação que é fornecida pela pessoa, seja direta ou indiretamente por meio da sua interação, e os novos usos atribuídos aos dados (Doneda, 2019, p. 157). Muitas vezes os dados que são coletados não seriam considerados como informações sensíveis, mas as finalidades e as inferências que podem ser feitas dos mesmos podem revelar aspectos íntimos e privados que não teriam sido fornecidos pelos titulares (Wachter; Mittelstadt, 2019, p. 18).

Um caso emblemático que revelou as engrenagens de como estas técnicas são aplicadas, além de trazer consequências concretas, foi o escândalo da Cambridge Analytica (CA), que veio a público em 2018, e envolveu o acesso indevido aos dados pessoais de 50 milhões de usuários do Facebook para

[108] Neste sentido, Wachter e Mittelstadt propõem que as inferências consideradas de "alto risco" devam ser consideradas como dados pessoais, tais como aqueles que podem causar danos à reputação no presente ou no futuro, sejam invasivas à privacidade da pessoa, ou que não sejam passíveis de verificação e utilizadas para apoiar decisões importantes. De acordo com os autores, o benefício deste reconhecimento seria garantir que os titulares de dados tenham transparência em relação às inferências que são feitas sobre ele (Wachter e Mittelstadt, 2019, p. 22-23).

[109] "The process of profiling is often invisible to the data subject. It works by creating derived or inferred data about individuals – 'new' personal data that has not been provided directly by the data subjects themselves" (GTA29, 2016, p 20).

influenciar digitalmente as pessoas em campanhas eleitorais. Os dados pessoais eram coletados a partir do aplicativo desenvolvido pelo pesquisador Aleksandr Kogan, MyPersonality, que se valeu das premissas e resultados apresentados em um artigo científico escrito por professores de Psicologia da Universidade de Cambridge, e um pesquisador da Microsoft, que apontavam a possibilidade de predizer atributos pessoais e traços da personalidade das pessoas a partir das suas curtidas no Facebook (Kosinski et al., 2013). Aleksandr implementou o que havia sido sugerido no artigo, sendo capaz de extrair e analisar dados, estabelecendo padrões de comportamento e de personalidade, classificando os usuários dentro de determinados perfis comportamentais, em um trabalho que se valeu de conhecimento sobre psicometria e mineração de dados[110].

Em seguida, Aleksandr conseguiu expandir o número de usuários do MyPersonality por meio de uma parceria feita com o Facebook[111]. A partir das respostas, foram realizadas análises estatísticas, correlações e identificação de padrões entre as respostas dos participantes de pesquisa, os dados do Facebook do participante, como curtidas e informações do seu perfil, e dos amigos destes participantes, com o intuito de realizar correlações, inferências de preferências e inclinações políticas [112], de acordo com seus vestígios digitais. [113] Como amplamente noticiado pela mídia, a empresa Cambridge Analytica (CA) adquiriu esta base de dados com o objetivo de influenciar comportamentos no contexto eleitoral. A estratégia da CA teve vitórias relevantes, como as eleições presidenciais de Donald Trump em 2016, e a campanha a favor do Brexit, havendo notícias de

[110] CARDOSO, *Bruno. Por que fazer uma sociologia da internet*? Sobre o caso Cambridge Analytica e Facebook. Blog do Laboratório de Estudos Digitais da UFRJ, 25 de março de 2018. Disponível em: https://ledufrj.wixsite.com/ledufrj/post/2018/03/25/por-que-fazer-uma-sociologia-da-internet-sobre-o-caso-cambridge-analytica-e-facebook. Acessado em 07.07.2021.

[111] Pessoas interessadas em realizar o teste de personalidade do aplicativo tinham que baixá-lo, se cadastrar com a sua conta do Facebook, e informar o número de seu registro eleitoral americano. Participaram da pesquisa 32 mil pessoas, que foram remunerados de U$ 2 a U$ 5 para tanto.

[112] CARDOSO, *Bruno. Por que fazer uma sociologia da internet*? Sobre o caso Cambridge Analytica e Facebook. Blog do Laboratório de Estudos Digitais da UFRJ, 25 de março de 2018. Disponível em: https://ledufrj.wixsite.com/ledufrj/post/2018/03/25/por-que-fazer-uma-sociologia-da-internet-sobre-o-caso-cambridge-analytica-e-facebook. Acessado em 07.07.2021.

[113] CARDOSO, *Bruno. Por que fazer uma sociologia da internet*? Sobre o caso Cambridge Analytica e Facebook. Blog do Laboratório de Estudos Digitais da UFRJ, 25 de março de 2018. Disponível em: https://ledufrj.wixsite.com/ledufrj/post/2018/03/25/por-que-fazer-uma-sociologia-da-internet-sobre-o-caso-cambridge-analytica-e-facebook. Acessado em 07.07.2021.

que a empresa teria atuado na Nigéria, no Quênia, na República Tcheca, na Índia, na Argentina, e chegou a ter tratativas no Brasil.[114]

A verdade é que existem tantas outras empresas que realizam atividades semelhantes à da CA. Ao final, o intuito é predizer o comportamento de um grupo de pessoas e oferecer algum tipo de conteúdo, serviço ou produto considerado estatisticamente relevante para ela, a partir da análise de dados pretéritos que assim indicavam. O caso da CA é emblemático, pois revela e concretiza o tratamento abusivo dos dados pessoais, e os seus impactos, extrapolando questões existenciais da pessoa humana e atingindo diretamente no processo democrático de deliberação e eleição de representantes políticos.

Este potencial abuso não passou despercebido pelos autores do artigo científico que embasou o aplicativo MyPersonality, que estavam cientes dos riscos sobre os seus achados. Eles reconhecem que tais predições poderiam ocasionar "perigosas invasões à privacidade", e dificultaria que as pessoas pudessem controlar quais características sobre elas poderiam ser reveladas, tais como sua orientação sexual, posição política ou inteligência, sendo certo que elas não gostariam que isto acontecesse (Kosinski et al., 2013). Por exemplo, menos de 5% dos participantes classificados como homossexuais tinham ativamente revelado este atributo por meio da adesão a grupos públicos, como "Eu amo ser Gay" (*I love being Gay*) ou "Casamento Gay" (*Gay Marriage*). A predição era feita por aspectos menos informativos, mas considerados abrangentes, como curtidas nas páginas da cantora norte americana Britney Spears, ou da série de televisão americana "Desperate Housewives", considerados prognósticos moderados sobre a orientação sexual dos participantes.

Como indicado, os próprios autores reconhecem que muitas vezes as correlações que são feitas são pouco – ou nada – convincentes, e não apresentam qualquer razão justificável entre o fator preditivo e o seu resultado, como o caso das predições feitas sobre a inteligência dos participantes. Os autores apontam que curtidas em páginas sobre "Trovões" (*thunderstorm*), a revista "Science", "The

[114] LEITE, Renato M. Cambridge Analytica e a nova era Snowden na proteção de dados pessoais. *El País,* 20 de março de 2018. Disponível em: https://brasil.elpais.com/brasil/2018/03/20/tecnologia/1521582374_496225.html. Acessado em 09.07.2021

Colbert Report"[115] e "Batata Frita Enrolada" (*Curly Fries*) eram fortes indicativos de inteligência, enquanto curtidas nas páginas da loja "Sephora"[116], "Eu amo ser mãe" (*I love being a Mom*), a companhia de motocicletas "Harley Davidson" e "Lady Antebellum"[117] eram fortes fatores de predição de baixa inteligência. Eles chamam atenção para o fato de que, embora algumas curtidas evidenciem uma relação de confiança sobre a predição, outras não, como a correlação encontrada entre inteligência e curtidas na página "Batata Frita Enrolada".

1.3.1
A privacidade e a proteção de dados pessoais sob uma perspectiva coletiva

Como visto, estas inferências dependem da análise de um número considerável de dados, proveniente de diversas pessoas. É neste contexto que alguns autores têm defendido a ideia de privacidade de grupo (*group privacy*) (Mantelero, 2016; Floridi; 2014; Taylor et al., 2016) e de proteção coletiva de dados pessoais *(collective data protection*) (Mantelero, 2016). A motivação destes pesquisadores são os riscos e potenciais danos que podem ser causados pelos usos discriminatórios e indevidos de dados pessoais, que afetam uma coletividade de pessoas que encontra-se reunida sob um mesmo grupo de perfil (Mantelero, 2016, p. 17-21). O argumento para defender esta perspectiva coletiva de direitos que tradicionalmente foram pensados sob uma perspectiva individual[118] é que o tratamento dos dados pessoais não estaria especialmente preocupado com o perfil de um único usuário, mas com os agrupamentos (*clusters*) criados a partir da coleta de dados de milhões de indivíduos. Nas palavras de Taylor, Floridi e Van der Sloot:

> (...) embora um indivíduo específico possa sofrer um dano ou ser beneficiado por certos usos de dados, isto de novo é bastante incidental na era de *big data*. Políticas e decisões são tomadas com base em perfis e padrões e isto afeta negativamente ou positivamente grupos ou categorias. É por isso que foi sugerido que o foco deve

[115] *The Colbert Report* é um programa de televisão americano que é uma sátira aos noticiários americanos politicamente conservadores, tal como a Fox News. O programa era estrelado por Stephen Colbert, que criou um personagem de um âncora do jornal fictício.
[116] Uma loja multinacional francesa que vende produtos de beleza.
[117] The Lady Antebellum é uma banda country norte-americana.
[118] Stefano Rodotà reconhece a necessidade de dilatar o conceito de privacidade, antes compreendida como "o direito de ser deixado só", para uma ideia de "tutela global das escolhas da vida contra qualquer forma de controle político e de estigmatização social" com o objetivo de proteger e garantir a liberdade das escolhas existenciais e políticas (Rodotà, 2008, p. 129).

> ser em grupos de interesses: se o grupo floresce, se pode agir autonomamente, se é tratado com dignidade, etc[119] (2017, p. 15)

Esta perspectiva coletiva de direitos fundamentais não é novidade no Brasil, especialmente no que se refere à proteção de dados, em que há autores que apontam para a "coletivização da proteção de dados pessoais" (Zanatta, 2020, p. 3), atribuída à própria natureza difusa do dano causado pelo tratamento inadequado de dados pessoais (Mendes, 2014, p. 36)[120]. Neste sentido, a LGPD apresenta uma estrutura ambígua que reforça essa perspectiva: ao mesmo tempo que a lei é orientada para proteger os direitos individuais dos titulares de dados[121], ela também assegura uma proteção de direitos difusos,[122] ao indicar a possibilidade de que os titulares de dados poderão se defender em juízo coletivamente (art. 22, da LGPD), bem como perante organismos de defesa do consumidor (art. 18 § 8º, da LGPD) (Zanatta, 2020, p. 14).

Ainda, há uma conexão da LGPD com o Código de Defesa do Consumidor (CDC), e com o sistema de tutela coletiva, que viabiliza que a proteção aos dados

[119] Tradução livre de: "(…) although specific individuals may be harmed or benefited by certain data uses, this again is increasingly incidental in the big data era. Policies and decisions are made on the basis of profiles and patterns and as such negatively or positively affect groups or categories. This is why it has been suggested that the focus should be on group interests: whether the group flourishes, whether it can act autonomously, whether it is treated with dignity, etc".

[120] Neste sentido, ao tratar sobre os vazamentos de dados e a possibilidade de reparação coletiva, Heloísa Carpena reconhece a dimensão coletiva quando há o vazamento de dados pessoais, visto que isto atinge os direitos da personalidade à privacidade e autodeterminação informativa, ocasionando um dano moral coletivo (Carpena, no prelo).

[121] A norma anuncia como seu objetivo "proteger os direitos fundamentais da liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" (art. 1º); define como dado pessoal a "informação relacionada a pessoa natural identificada ou identificável" (art. 5º, inciso I), como como titular de dados "a pessoa natural a quem se referem os dados pessoais que são objeto de tratamento" (art. 5º, inciso V). É assegurada aos titulares de dados a titularidade de seus dados pessoais (art. 17), podendo exercer perante o controlador, em relação aos dados por ele tratados, uma série de direitos que podem ser exercidos individualmente por ele (correção, acesso, portabilidade, eliminação, oposição, entre outros, previstos no art. 18). Ainda, este titular de dados tem a possibilidade de peticionar contra o controlador sobre os seus dados, perante a Autoridade Nacional de Proteção de Dados Pessoais (ANPD) (art. 18 § 2º c/c art. 55-J, V). Portanto, há uma clara identificação nesta lei, assim como há no GDPR, de um caráter que visa proteger os direitos individuais" (Zanatta, 2020, pp. 13-14).

[122] A LGPD possui forte ligação com o CDC, e o sistema de tutela coletiva, razão pela qual conforme apontam os seguintes dispositivos. A LGPD aponta que a "defesa dos interesses e dos direitos dos titulares de dados poderá ser exercida em juízo, individual ou coletivamente, na forma do disposto na legislação pertinente, acerca dos instrumentos de tutela individual e coletiva" (art. 22), além de indicar que os titulares de dados poderão exercer seus direitos perante os organismos de defesa do consumidor, atraindo toda a estrutura própria, como os "Procons, Defensorias Públicas, ONGs e Ministérios Públicos (o que é chamado de "Sistema Nacional de Defesa do Consumidor"), (Zanatta, 2020, p. 14). Além disso, o art. 42 prevê que os agentes de tratamento poderão responder quando causarem a outrem dano patrimonial, moral individual coletivo, quando houver violação à legislação de proteção de dados pessoais

pessoais possa ser realizada de maneira coletiva, em conjunto com as diversas formas de assegurar a proteção individual dos direitos dos titulares de dados (Zanatta, 2019, p. 205). Pode-se dizer que esta perspectiva coletiva da privacidade e proteção de dados exige uma avaliação do impacto do tratamento de dados em relação à comunidade como um todo, com a possibilidade de que possa ser pleiteada uma reparação coletiva (Zanatta, 2019, p. 203).

Este processo de "coletivização da proteção de dados", assim cunhado por Zanatta, encontraria respaldo em quatro elementos presentes no Brasil, bem como em outros países: (i) a importância da linguagem de "direitos coletivos" e "direitos difusos", baseados em teorias como a de *group privacy* e dimensões coletivas da proteção de dados, destacando o reconhecimento da violação "aos valores da sociedade", em contrapartida aos mecanismos de reparação compensatórios individuais; (ii) o deslocamento da maneira como estes direitos são protegidos, indo de uma perspectiva liberal, em que há a proteção individual da pessoa em relação aos seus próprios direitos, para a possibilidade de que entidades civis especializadas tenham legitimidade para propor ações coletivas[123]/[124]; (iii) adoção de uma perspectiva preventiva sobre os riscos envolvidos na atividade, tal como ocorre no Direito Ambiental, que possui respaldo tanto no GDPR, quanto na LGPD, sendo que nesta última há a previsão expressa do princípio da precaução (art. 6º, VIII)[125], que justifica a criação de obrigações impostas aos controladores para mitigar potenciais riscos causados aos titulares de dados, tais como os relatórios de impacto à proteção de dados[126]; e (iv) alteração das estruturas administrativas de defesa do consumidor, além de questões relacionadas à proteção de dados pessoais, que são

[123] Conforme aponta Zanatta (2019, p. 203): "Com a coletivização da proteção de dados, essa proteção passa a ser feita, cada vez mais, por entidades civis especializadas e que pos- suem legitimidade ativa para a proposição de ações civis públicas (ou *Privacy Class Actions,* como nota- do na literatura internacional) (MANTELERO, 2016). Esse fenômeno foi chamado por Mauro Cappelletti, há mais de 30 anos, de "despublicização do direito" (CAPPELLETTI, 1985), ou seja, a ideia de que a própria sociedade civil possa tutelar seus interes- ses. Esse elemento também está ligado ao que Ada Pellegrini Grinover chamou de ampliação da "tute- labilidade" (GRINOVER, 1984, p. 284), e que foi uma das marcas dos movimentos progressistas de cria- ção da Lei da Ação Civil Pública e do Código de Defesa do Consumidor (CDC)."

[124] Além das previsões do art. 22 e 18 § 8o, da LGPD, os arts. 80 e 81 do GDPR expressamente autorizam entidades civis de ajuizarem ações perante o judiciário ou casos perante às Autoridades.

[125] Art. 6o, inciso VIII, da LGPD: "prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais".

[126] Além do relatório de impacto, há a a constituição de *boards* consultivos com ONGs, que tem por objetivo avaliar o impacto de algoritmos em direitos humanos (Zanatta; 2020, o. 19)

tratadas como um problema coletivo de "direito do consumidor"[127]. Rafael Zanatta adiciona um quinto elemento, apontado por ele como um qualificador no Brasil: o Ministério Público brasileiro. Esta instituição tem tido uma atuação intensa em ações civis públicas[128], antes mesmo da entrada em vigor da LGPD, visando resguardar o direito à privacidade, intimidade e proteção de dados pessoais em casos envolvendo desde incidentes de segurança, até eventuais abusos no tratamento de dados pessoais (Zanatta, 2019, pp. 203-204).

Embora se reconheça que a tutela coletiva possui um forte fundamento processual,[129] tema que foge do escopo deste trabalho, o que é relevante para a presente pesquisa é o reconhecimento da dimensão coletiva da proteção de dados pessoais e da privacidade. Isto reflete na garantia de direitos e na implementação de mecanismos e salvaguardas de quando há o uso de dados pessoais para a tomada de decisão automatizada com o uso de algoritmos de ML. Este reconhecimento fornece fundamentos para que sejam implementadas medidas que visem assegurar uma *accountability* algorítmica binária (Kaminski, 2019b). Esta última possui uma dimensão de direitos individuais e outra de interesses coletivos, com intuito de justificar e apresentar diferentes maneiras de implementar uma prestação de contas, especialmente em atenção ao direito à explicação de decisões algorítmicas.

No mais, o reconhecimento da proteção de dados como um direito fundamental enseja uma dimensão subjetiva (direito de defesa subjetivo do indivíduo) e outra objetiva (dever de proteção estatal) (Mendes, 2014, p. 176), que se relaciona com essa perspectiva individual e coletiva da *accountability* algorítmica binária. Tal reconhecimento garante não apenas o direito subjetivo à autodeterminação informativa ao titular de dados, mas exige a garantia concreta de

[127] Nos EUA, a Federal Trade Commission (FTC), em se dedicado a um programa de *privacy protection* investigando casos importantes, tais como o escândalo envolvendo a Cambridge Analytics, em que fechou um acordo com a empresa Facebook no valor de 5 bilhões de dólares. No Brasil. A Secretaria Nacional do Consumidor (Senacon) tem cumprido este papel, instaurando inquéritos e processos administrativos para investigar possíveis violações de direitos da personalidade de consumidores (Zanatta, 2019, p. 203).

[128] Conforme selecionado por Zanatta: Ministério Público Federal do Piauí (MPF/PI) vs. Google (JFPI, processo n. 25463-45.2016.4.01.4000); Ministério Público Federal de São Paulo (MPF/SP) vs. Microsoft (JFSP, processo n. 5009507-78.2018.4.03.610); Instituto Brasileiro de Defesa do Consumidor (Idec) vs. ViaQuatro (concessionária da Linha Amarela do metrô de São Paulo) (2018) (TJSP, processo n. 1090663.42.2018.8.26.0100); Ministério Público Federal do Distrito Federal e Territórios (TJDFT) vs. Telefonica Brasil (2019) (TJDFT, processo n. 0721735-15.2019.8.07.0001); e Defensoria de São Paulo vs. Metrô (2020) (TJSP, processo n. 1006616-14.2020.8.26.0053).

[129] Ver: Carpena, no prelo.

como esta proteção deve ocorrer, com o estabelecimento de meios e direitos que viabilizem isto.

Como aponta Laura S. Mendes, a dimensão subjetiva assegura "ao indivíduo um espaço de liberdade e privacidade, não sujeito a intervenções estatais", o que envolve o poder de ter controle sobre o fluxo das suas informações pessoais. Uma vez que haja ameaça ou intervenção a este direito, devem ser garantidos meios para protegê-lo. Assim como todo direito fundamental, esta proteção não é absoluta, e encontra limites em direitos e interesses de terceiros (Mendes, 2014, p. 177-178), com a necessidade de realizar uma ponderação sobre a necessidade e proporcionalidade[130], entre a finalidade da atividade de tratamento de dados pessoais, e o direito à proteção de dados pessoais.[131]

Por sua vez, a dimensão objetiva "representa a necessidade de concretização e delimitação desses direitos por meio de ação estatal" (Mendes, 2014, p. 176), o que exige uma ação positiva do Estado para promover mecanismos e procedimentos para o legítimo tratamento de dados pessoais, além de uma ação negativa de não interferência no âmbito de liberdade e privacidade do sujeito. Com a promulgação da LGPD, o Estado-legislador houve por bem estabelecer um arcabouço de proteção que prevê princípios e deveres que exigem a transparência da coleta e processamento de dados pessoais para viabilizar a autodeterminação informativa do titular de dados; direitos que efetivam o controle sobre os dados pessoais (como o direito de acesso, o direito de correção de dados incompletos, inexatos ou desatualizados, direito de oposição); assegura que os dados pessoais coletados não sejam utilizados em contexto diferente do qual foi coletado, devendo atender à finalidade compatível à inicial (de acordo com os princípios da finalidade e adequação); e, por fim, previu uma autoridade de controle de dados pessoais, a Autoridade Nacional de Proteção de Dados Pessoais (ANPD), que exerce uma

[130] Dois critérios sugeridos pela autora para guiarem a interpretação sobre os limites à proteção dos dados pessoais e à autodeterminação informativa: "i) necessidade de determinado processamento de dados pessoais para atender a um fim legítimo protegido pelo ordenamento jurídico ou para cumprimento de direito de terceiros; b) [sic] pertinência temática (ou de conteúdo) entre o tratamento de dados e a finalidade a ser atingida" (Mendes, 2014, p. 177).

[131] Um dos exemplos utilizados pela autora é a transferência de dados pessoais de beneficiados no programa bolsa família para outro órgão governamental como sendo necessário para atender a uma política pública e programas sociais. Seria considerado abusivo, no entanto, se a transferência destes dados fosse destinada a entes privados, visto que descontextualizaria o seu uso, com o potencial de causar discriminação e estigmatização (Mendes, 2014, pp. 177-178)

papel de fiscalização e controle da atividade de processamento de dados pessoais (Mendes, 2014, p. 180).

Assim, o reconhecimento da proteção de dados como direito fundamental traz novas exigências, dando maior robustez à sua proteção. Isto vai refletir, sem dúvidas, na maneira como o direito à explicação deverá ser interpretado.

Neste sentido, para explorar as complexidades do encontro do direito e da tecnologia, em especial, das técnicas de IA e os algoritmos de tomada de decisão de ML, e a necessidade de proteger as pessoas envolvidas nestes processos algorítmicos, o ponto de partida será a discussão iniciada na União Europeia, com a promulgação do GDPR, que prevê o direito à explicação. Este direito tem sido objeto de disputa acadêmica, onde pesquisadores têm exigido uma definição mais precisa de quais salvaguardas e garantias devem ser asseguradas, questionando-se, até mesmo, se tal direito estaria de fato previsto no GDPR.

Analisar a origem da discussão sobre o direito à explicação no GDPR é importante para delimitar e compreender quais são as salvaguardas e os direitos que devem ser assegurados, e as diferentes ferramentas e instrumentos que devem ser fornecidos para garantí-lo. A sua compreensão no âmbito da UE certamente influenciará a leitura deste direito no Brasil, que possui previsão expressa no art. 20, da LGPD. Desta forma, o próximo subcapítulo se dedicará a apresentar as disposições da LGPD e do GDPR que tratam do direito à explicação, e a discussão que se desenvolveu sobre o tema após a promulgação do GDPR.

1.4
O direito à explicação previsto no GDPR

Os artigos do GDPR que tratam da tomada de decisão algorítmica são os arts. 13 e 14, que se referem às informações que devem ser fornecidas aos titulares de dados quando seus dados pessoais forem coletados e tratados (chamado também como direito de notificação), o art. 15, que se refere ao direito de acesso a dados, e o art. 22[132], que trata especificamente de decisões individuais automatizadas,

[132] Tal Zarsky faz uma crítica ao art. 22, indicando que a sua previsão pode ameaçar ou até mesmo inviabilizar usos de dados no contexto de *big data* por três principais motivos. O primeiro é o estabelecimento da regra geral de proibição de decisões totalmente automatizadas, o que obviamente limita atividades de *big data;* o segundo seria a ameaça à eficiência, otimização e precisão de técnicas aplicadas, pois haveria a necessidade de explica-las; e, terceiro, a intervenção humana também poderia atrasar a inovação de novas tecnologias (Zarsky, 2017, p. 1017).

incluindo definição de perfil. Ainda, os Considerandos 63[133] e 71[134] também tratam do direito de acesso e das decisões tomadas exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, respectivamente. A partir da leitura conjunta e sistemática destes dispositivos pode-se extrair o que ficou conhecido

[133] Considerando 63: "Os titulares de dados deverão ter o direito de aceder aos dados pessoais recolhidos que lhes digam respeito e de exercer esse direito com facilidade e a intervalos razoáveis, a fim de conhecer e verificar a tomar conhecimento do tratamento e verificar a sua licitude. Aqui se inclui o seu direito de acederem a dados sobre a sua saúde, por exemplo os dados dos registos médicos com informações como diagnósticos, resultados de exames, avaliações dos médicos e quaisquer intervenções ou tratamentos realizados. Por conseguinte, cada titular de dados deverá ter o direito de conhecer e ser informado, nomeadamente, das finalidades para as quais os dados pessoais são tratados, quando possível do período durante o qual os dados são tratados, da identidade dos destinatários dos dados pessoais, da lógica subjacente ao eventual tratamento automático dos dados pessoais e, pelo menos quando tiver por base a definição de perfis, das suas consequências. Quando possível, o responsável pelo tratamento deverá poder facultar o acesso a um sistema seguro por via eletrônica que possibilite ao titular aceder diretamente aos seus dados pessoais. Esse direito não deverá prejudicar os direitos ou as liberdades de terceiros, incluindo o segredo comercial ou a propriedade intelectual e, particularmente, o direito de autor que protege o *software*. Todavia, essas considerações não deverão resultar na recusa de prestação de todas as informações ao titular dos dados. Quando o responsável proceder ao tratamento de grande quantidade de informação relativa ao titular dos dados, deverá poder solicitar que, antes de a informação ser fornecida, o titular especifique a que informações ou a que atividades de tratamento se refere o seu pedido".

[134] Considerando 71: "O titular dos dados deverá ter o direito de não ficar sujeito a uma decisão, que poderá incluir uma medida, que avalie aspetos pessoais que lhe digam respeito, que se baseie exclusivamente no tratamento automatizado e que produza efeitos jurídicos que lhe digam respeito ou o afetem significativamente de modo similar, como a recusa automática de um pedido de crédito por via eletrônica ou práticas de recrutamento eletrônico sem qualquer intervenção humana. Esse tratamento inclui a definição de perfis mediante qualquer forma de tratamento automatizado de dados pessoais para avaliar aspetos pessoais relativos a uma pessoa singular, em especial a análise e previsão de aspetos relacionados com o desempenho profissional, a situação económica, saúde, preferências ou interesses pessoais, fiabilidade ou comportamento, localização ou deslocações do titular dos dados, quando produza efeitos jurídicos que lhe digam respeito ou a afetem significativamente de forma similar. No entanto, a tomada de decisões com base nesse tratamento, incluindo a definição de perfis, deverá ser permitida se expressamente autorizada pelo direito da União ou dos Estados-Membros aplicável ao responsável pelo tratamento, incluindo para efeitos de controlo e prevenção de fraudes e da evasão fiscal, conduzida nos termos dos regulamentos, normas e recomendações das instituições da União ou das entidades nacionais de controlo, e para garantir a segurança e a fiabilidade do serviço prestado pelo responsável pelo tratamento, ou se for necessária para a celebração ou execução de um contrato entre o titular dos dados e o responsável pelo tratamento, ou mediante o consentimento explícito do titular. 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. Essa medida não deverá dizer respeito a uma criança.

A fim de assegurar um tratamento equitativo e transparente no que diz respeito ao titular dos dados, tendo em conta a especificidade das circunstâncias e do contexto em que os dados pessoais são tratados, o responsável pelo tratamento deverá utilizar procedimentos matemáticos e estatísticos adequados à definição de perfis, aplicar medidas técnicas e organizativas que garantam designadamente que os fatores que introduzem imprecisões nos dados pessoais são corrigidos e que o risco de erros é minimizado, e proteger os dados pessoais de modo a que sejam tidos em conta os potenciais riscos para os interesses e direitos do titular dos dados e de forma a prevenir, por exemplo, efeitos discriminatórios contra pessoas singulares em razão da sua origem racial ou étnica, opinião política, religião ou convicções, filiação sindical, estado genético ou de saúde ou orientação sexual, ou a impedir que as medidas venham a ter tais efeitos. A decisão e definição de perfis automatizada baseada em categorias especiais de dados pessoais só deverá ser permitida em condições específicas".

como o direito à explicação de decisões algorítmicas, embora ainda não esteja bem definido e delimitado como estes comandos normativos devem ser atendidos (Wachter et., 2017; Kaminski, 2019a; Selbst; Powles, 2017).

O *caput* do art. 22, do GDPR, estabelece que o titular de dados possui "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". A norma possui natureza proibitiva (Mulholland, Frajhof, 2019, p. 275; GTA29, 2017) e, via de regra, coíbe que seja tomada uma decisão exclusivamente automatizada, incluindo o tratamento destinado à formação de perfis. É importante mencionar que uma decisão totalmente automatizada nem sempre incluirá a técnica de perfilamento. Uma decisão totalmente automatizada pode se valer de qualquer tipo de dado, e não apenas dados pessoais, além de avaliar aspectos que não sejam considerados como pessoais[135]. Para que seja caracterizada como uma decisão "tomada exclusivamente com base no tratamento automatizado", não se admite o envolvimento de uma pessoa humana. Por sua vez, o perfilamento, nos termos do art. 4º, do GDPR, deve se destinar à avaliação de aspectos pessoais de um titular de dados e não necessariamente exclui o envolvimento humano no seu processo. Assim, a decisão tomada exclusivamente com base no tratamento automatizado pode ser feita sem a aplicação do perfilamento, embora se reconheça que nem sempre serão atividades separadas (GTA29, 2017).

Apesar de a regra geral ser a proibição da decisão tomada exclusivamente com base no tratamento automatizada, há exceções de quando esta poderá ocorrer, listadas pelo art. 22 (2), do GDPR. São elas: (a) necessária para a celebração ou execução de um contrato entre o titular de dados e o controlador, (b) prevista em lei pela União ou Estado-Membros, com a previsão de medidas adequadas para salvaguardar direitos, liberdades e os legítimos interesses do titular de dados, ou (c) quando baseada no consentimento explícito do titular de dados. A contrapartida para a implementação de uma decisão totalmente automatizada é que o controlador deverá empreender medidas para salvaguardar direitos, liberdades e interesses legítimos dos titulares de dados. Isto inclui o direito de, pelo menos, (i) obter intervenção humana por parte do responsável,

[135] O referidoConsiderando 71 se vale do verbo poderá, e não deverá, indicando a discricionariedade e não obrigatoriedade de que a decisão totalmente automatizada se destina a definir aspectos pessoais do titular de dados: "O titular dos dados deverá ter o direito de não ficar sujeito a uma decisão, **que poderá incluir** uma medida, que avalie aspetos pessoais que lhe digam respeito, que se baseie exclusivamente no tratamento automatizado (…)" (grifos meus).

(ii) manifestar o seu ponto de vista, (iii) contestar a decisão e (iv) obter uma explicação sobre a decisão tomada na sequência desta avaliação (art. 22 (3) c/c Considerando 71, do GDPR). É importante notar que as exceções não se aplicam ao tratamento de categorias especiais de dados (i.e. dados sensíveis), salvo quando houver o consentimento explícito, nos termos do art. 9 (2), a)[136], do GDPR, ou se houver interesse público, nos termos do art. 9(2), g)[137], do GDPR, devendo ser aplicadas medidas adequadas para salvaguardar os direitos e liberdades e os legítimos interesses do titular de dados.

Além disso, quando houver uma decisão totalmente automatizada, há a incidência dos arts. 13, 14 e 15, do GDPR, que preveem disposições que se relacionam com o art. 22, do GDPR, que exigem que o titular de dados tenha: (i) o direito a ser informado sobre decisões automatizadas, incluindo formação de perfis (arts. 13, (2), f)[138] e 14, (2), g)[139], do GDPR, e (i.a) informações úteis relativas à lógica subjacente, bem como (i.b) a importância e as consequências previstas de tal tratamento para o titular dos dados, e o (ii) o direito de acessar dados, incluindo (ii.a) o acesso às informações úteis relativas à lógica subjacente, bem como (ii.b) a importância e as

[136] Art. 9(2) a): "1. É proibido o tratamento de dados pessoais que revelem a origem racial ou étnica, as opiniões políticas, as convicções religiosas ou filosóficas, ou a filiação sindical, bem como o tratamento de dados genéticos, dados biométricos para identificar uma pessoa de forma inequívoca, dados relativos à saúde ou dados relativos à vida sexual ou orientação sexual de uma pessoa. 2. O disposto no nº 1 não se aplica se se verificar um dos seguintes casos: a) Se o titular dos dados tiver dado o seu consentimento explícito para o tratamento desses dados pessoais para uma ou mais finalidades específicas, exceto se o direito da União ou de um Estado-Membro previr que a proibição a que se refere o nº 1 não pode ser anulada pelo titular dos dados".

[137] Art. 9(2) g): "Se o tratamento for necessário por motivos de interesse público importante, com base no direito da União ou de um Estado-Membro, que deve ser proporcional ao objetivo visado, respeitar a essência do direito à proteção dos dados pessoais e prever medidas adequadas e específicas que salvaguardem os direitos fundamentais e os interesses do titular dos dados".

[138] Art. 13, (2), f) "1. Quando os dados pessoais forem recolhidos junto do titular, o responsável pelo tratamento faculta-lhe, aquando da recolha desses dados pessoais, as seguintes informações: (…) 2) Para além das informações referidas no nº 1, a quando da recolha de dados pessoais, o responsável pelo tratamento fornece ao titular as seguintes informações adicionais, necessárias para garantir um tratamento equitativo e transparente: (...) f) A existência de decisões automatizadas, incluindo a definição de perfis, referida no artigo 22º, nºs 1 e 4, e, pelo menos nesses casos, informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento para o titular dos dados".

[139] Art. 14, (2), g): "1. Quando os dados pessoais não forem recolhidos junto do titular, o responsável pelo tratamento fornece-lhe as seguintes informações: (...) 2. Para além das informações referidas no nº 1, a quando da recolha de dados pessoais, o responsável pelo tratamento fornece ao titular de dados as seguintes informações, necessárias para lhe garantir um tratamento equitativo e transparente: (…) g) A existência de decisões automatizadas, incluindo a definição de perfis referida no artigo 22 º, n os 1 e 4, e, pelo menos nesses casos, informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento para o titular dos dados."

consequências previstas de tal tratamento para o titular dos dados (art. 15, (1), h), do GDPR)[140].

Nos termos dos arts. 13 e 14, do GDPR, sob o princípio da transparência, os titulares de dados devem ser informados de maneira simples e clara sobre como funciona o processo de perfilamento e de decisão exclusivamente automatizada. Ainda, o art. 15, do GDPR, estabelece que o titular de dados tem direito a ter acesso à categoria e aos dados pessoais que foram utilizados para formar o perfil, devendo ser observado o que dispõe o Considerando 63, do GDPR, que informa que os controladores não poderão se abster de fornecer estes dados sob a justificativa da proteção ao segredo empresarial (GTA29, 2017, p. 16).

A leitura sistemática destes artigos motivou a discussão e gerou dúvidas sobre qual o tipo de informação que deve ser fornecida ao titular de dados quando este estiver sujeito a uma decisão tomada exclusivamente com o tratamento de dados automatizado. Em suma, para a interpretação destes artigos podemos pensar no seguinte esquema:

[140] Art. 15, (1), h): "1. O titular dos dados tem o direito de obter do responsável pelo tratamento a confirmação de que os dados pessoais que lhe digam respeito são ou não objeto de tratamento e, se for esse o caso, o direito de aceder aos seus dados pessoais e às seguintes informações: (…) A existência de decisões automatizadas, incluindo a definição de perfis, referida no artigo 22 º, n os 1 e 4, e, pelo menos nesses casos, informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento para o titular dos dados".

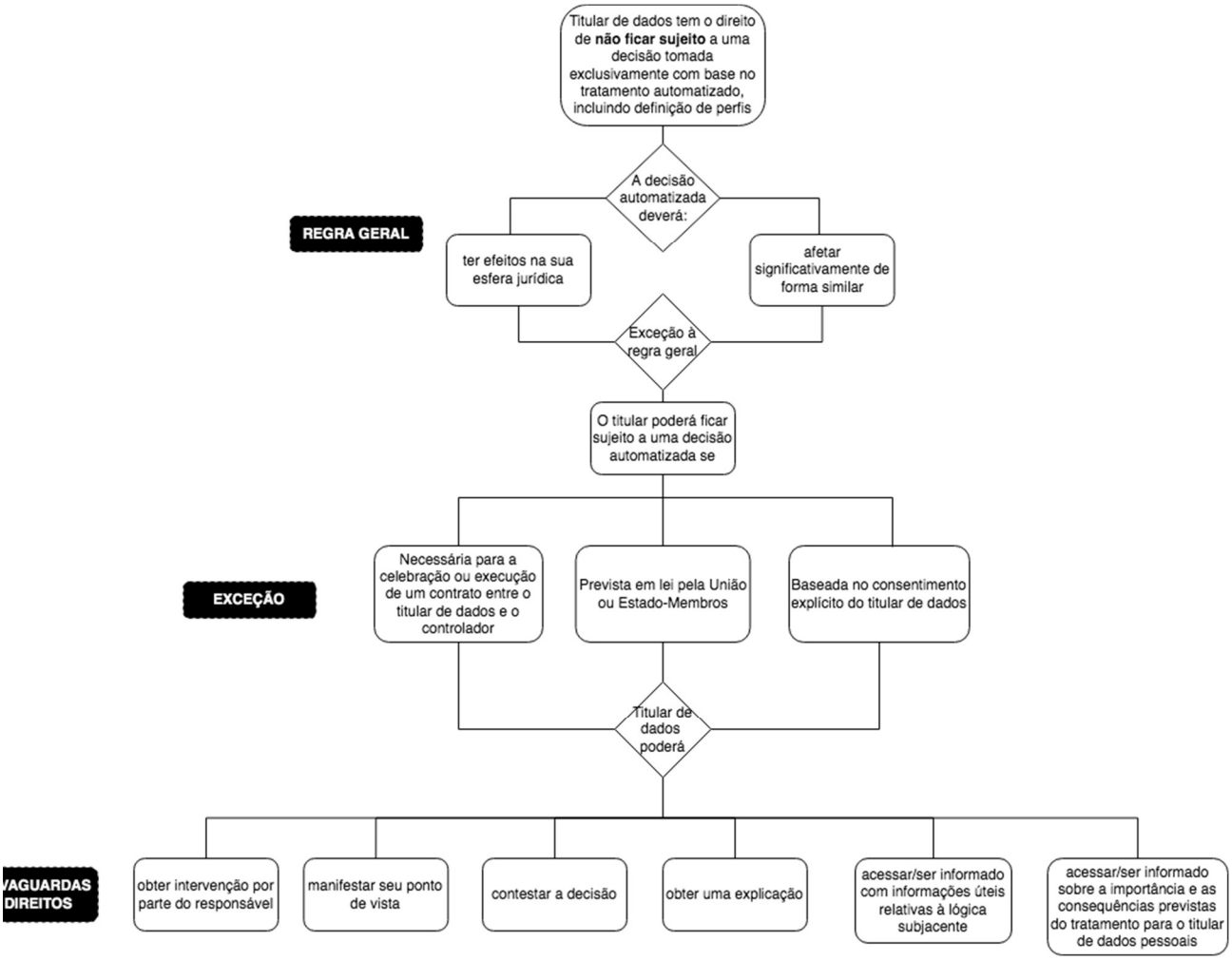

Figura 1 - Esquematização do direito à explicação no GDPR

No entanto, não é qualquer decisão automatizada que faz incidir as medidas e o próprio direito à explicação, mas apenas aquela "decisão tomada exclusivamente com base no tratamento automatizado" (*based solely on automated processing)*. O parecer do GTA29 que analisa estes artigos joga luz sobre a delimitação desta atividade. Para o antigo GTA29, uma decisão totalmente automatizada implica em uma decisão que se vale de algoritmos, utilizada ou não para o desenvolvimento de perfis, e que não possui qualquer envolvimento humano em seu resultado final. Por exemplo, se uma pessoa revisa ou leva em consideração outros fatores, como outros dados, na decisão final, esta não será considerada como exclusivamente automatizada. Um ponto de atenção ressaltado neste parecer é a preocupação de que os controladores "fabriquem" envolvimento humano para descaracterizar a decisão algorítmica. Por isso, para denotar tal envolvimento, a pessoa deverá ter autoridade e competência dentro da organização para alterar esta decisão, e deverá considerar os dados utilizados como entrada, e os resultados de saída disponíveis como parte da sua análise (GTA29, 2017, p. 9-10).

Não basta que seja uma decisão tomada exclusivamente com base no tratamento automatizado, mas o art. 22, do GDPR se refere especificamente a decisões que (i) afetem a esfera jurídica do titular de dados, ou que (ii) o afetem significativamente de forma similar. O Considerando 71, do GDPR, apresenta alguns exemplos de decisões automatizadas que seriam consideradas como capazes de afetar a esfera jurídica do titular de dados, ou de o afetar de forma semelhante. Estas podem incluir a definição de perfis, e podem ser destinadas a realizar uma avaliação de aspectos pessoais relativos a uma pessoa natural, em especial a análise e previsão de aspectos relacionados com o desempenho profissional, a situação econômica, de saúde, de preferências ou interesses pessoais, fiabilidade ou comportamento, localização ou deslocamento do titular dos dados.

Em uma análise concreta, o GTA29 avalia que uma decisão totalmente automatizada afeta a esfera jurídica do titular de dados quando há um impacto em um direito seu previsto em lei ou estabelecido em contrato, tal como seu direito à liberdade de expressão, ao devido processo legal, ou de gozar de um benefício social. A decisão algorítmica pode causar efeitos semelhantes, como a recusa a um pedido de crédito ou realizar práticas de recrutamento eletrônico sem qualquer intervenção humana (conforme exemplos do Considerando 71, do GDPR)[141]. Esta decisão automatizada deve trazer um impacto relevante, que não pode ser trivial, e deve influenciar de maneira decisiva circunstâncias, comportamentos ou escolhas das pessoas, sendo que a consequência mais grave nestes casos será considerada a exclusão ou discriminação do titular de dados (GTA29, 2017). Seria considerado pelo GTA29 como um tratamento de baixo impacto uma propaganda direcionada a um determinado tipo de perfil, mas com alto impacto a decisão que afetar minorias e pessoas vulneráveis, como o direcionamento de propaganda de jogos de aposta para adictos em jogos.

Os direitos que devem ser assegurados e as medidas e salvaguardas que devem ser implementadas envolvem uma perspectiva de direito material e procedimental. Enquanto este último tem sido considerado uma espécie de "devido processo

[141] Tal como testado pela Amazon, que desenvolveu um algoritmo de aprendizado de máquina para realizar um processo seletivo, o que acabou revelando um viés de gênero, favorecendo candidatos homens do que mulheres. Ver em: DASTIN, Jeffrey. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 10 de out. de 2018. Disponível em: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Acessado em 15.05.2021.

algorítmico" (Kaminski, 2019b), a questão substantiva (material), de mérito, envolve a discussão sobre o tipo de explicação e informação que pode ser exigida do controlador sobre o resultado e sobre o algoritmo responsável pela decisão automatizada. O "devido processo algorítmico" traz garantias para que o indivíduo sujeito à decisão algorítmica deva ser notificado e tenha conhecimento de que está sujeito a uma decisão tomada exclusivamente com base no tratamento automatizado, e possa se manifestar e ter uma intervenção humana (Kaminski, 2019a e 2019b; Citron; Pasquale, 2014). Enquanto a questão de forma possui uma conotação procedimental, a segunda tem como objeto de discussão a substância, o conteúdo que deve ser fornecido ao titular de dados em relação à decisão algorítmica, ou seja, a sua justificativa, que consistiria nas razões e fundamentos da mesma (Kaminski, 2019b, 1.529).

Tanto a perspectiva procedimental, quanto de conteúdo visam justamente garantir algum nível de transparência para que uma pessoa possa compreender o resultado de uma decisão totalmente automatizada, bem como avaliar a legitimidade e legalidade da mesma, para verificar se houve algum tipo de discriminação ou erros. As disposições que tratam do direito à explicação no GDPR estão fortemente ligadas aos princípios de transparência e de justiça e igualdade (*fairness*). Como aponta Kaminski:

> O "quem" e "porque" da transparência no GDPR ditam o que, o quando, e o como. As previsões de transparência individual, o parecer [do GTA29] deixa claro, tem a intenção de empoderar os indivíduos para invocarem seus direitos sob o GDPR. Embora o código fonte não precise ser apresentado aos indivíduos, eles precisam ter garantido muito mais do que uma visão sobre uma-sentença sobre como um sistema de tomada de decisão algorítmica funciona. Eles precisam de informações suficientes para serem capazes de compreender o que eles estão aceitando (se uma empresa estiver se baseando na exceção do consentimento explícito) para contestar uma decisão; e de achar e corrigir informações erradas, incluindo inferências (Kaminski, 2019a, p. 21) [142]

O início da discussão sobre o direito à explicação esteve muito pautado sobre o aspecto de substância/mérito do que deveria ser fornecido ao titular de dados enquanto explicação de uma decisão automatizada. Tendo em vista o amplo uso de

[142] Tradução livre de: "The "who" and "why" of transparency in the GDPR dictates the what, when, and how. Individual transparency provisions, the guidelines make clear, are intended to empower individuals to invoke their rights under the GDPR. Thus while individuals need not be provided with source code, they need to be given far more than a one-sentence overview of how an algorithmic decision-making system works. They need to be given enough information to be able to understand what they are agreeing to (if a company is relying on the explicit consent exception) to contest a decision; and to find and correct erroneous information, including inferences." (removidas as notas de rodapé).

algoritmos de ML para a tomada de decisão, a discussão se pautou sobre os desafios que surgem para o fornecimento destas explicações quando há a aplicação de IA, em especial, modelos de ML. Isto porque o uso de algoritmos e árvores de decisão simples, sistemas conversacionais e modelos de ML considerados como modelos caixa-branca (*white-box models*)[143], não encontram barreiras técnicas relevantes para a apresentação de explicações, o que não é o caso quando há o uso de ML, especialmente os modelos caixa-preta (*black-box models)*, técnicas mais avançadas, como as redes neurais e o aprendizado profundo[144] (*deep learning*). Quanto mais complexo o modelo de ML, maior será o desafio em explicar suas predições e resultados, sendo certo que muitas vezes isso sequer será viável. Portanto, a utilidade das explicações e o seu conteúdo tornaram-se o centro da discussão sobre o direito à explicação.

No contexto do uso de inteligência artificial, especialmente de algoritmos de tomada de decisão de ML, é que o direito à explicação ganha relevância, pela inescrutabilidade, imprevisibilidade, opacidade, correlação das predições e de seus resultados. Neste cenário, indaga-se: o que poderia ser explicado quando tampouco humanos são capazes de justificar o funcionamento e resultado destes sistemas?

Este é o desafio deste trabalho: compreender quais medidas e salvaguardas devem ser implementadas para permitir uma explicação de uma decisão algorítmica que se vale de ML, além de especificar o conteúdo adequado que deve ser fornecido sobre o mérito da explicação em si. Como se verá mais à frente no segundo capítulo, será necessário adotar uma série de medidas, de ordem técnica e jurídica, para dar conta dessa explicação, seja sobre os aspectos procedimentais, quanto materiais. Entretanto, a discussão sobre o "conteúdo" do que deve ser apresentado por controladores quando há o uso de uma decisão tomada exclusivamente por meios automatizados é o que motivou um debate público "explosivo" (Casey et al., 2019, p. 158) e uma disputa acadêmica inicial sobre o direito à explicação. Assim, pesquisadores têm exigido uma definição mais precisa do que pode ser pleiteado a partir deste direito, tendo sido colocado em debate, até mesmo, se um direito à explicação poderia ser derivado do GDPR, conforme se passará a expor a seguir.

[143] *White-box models* se referem a soluções para o desenvolvimento de modelos de aprendizado de máquina que sejam passíveis de terem o seu funcionamento compreendido por humanos, sendo a transparência inerente a ele (Linadartos et al., 2020, p. 17). Este tema será melhor abordado no capítulo subsequente.

[144] Aprendizado profundo é um dos métodos de aprendizado de máquina.

1.4.1
O início do debate sobre o direito à explicação no GDPR

A origem da discussão pode ser remetida ao artigo publicado por Bryce Goodman e Seth Flaxman (2016), ambos pesquisadores do *Oxford Internet Institute* (OII) à época, alertando sobre as potenciais proibições a diversos usos de algoritmos de aprendizado de máquina (tais como sistemas de recomendações, análises de riscos de crédito e seguros) que poderiam ocorrer por conta do art. 22 do GDPR. Um ponto interessante é que o referido artigo não se vale do termo "explicação". Este termo apenas é utilizado no Considerando 71, o que é objeto de críticas por Wachter et al. (2017), sendo curioso notar a associação que Goodman e Flaxman fazem da terminologia "direito à explicação" ao art. 22, do GDPR. Na perspectiva dos autores, o referido artigo cria um "direito do cidadão de receber uma explicação para decisões algorítmicas jogando luz para a importância da interpretabilidade humana no *design* do algoritmo" (Flaxman; Goodman, 2016, p. 1)[145]. O texto, contudo, não se dedica a estabelecer uma definição mais precisa sobre o que seria o "direito à explicação", ou travar um debate mais profundo sobre o que constituiria este direito (Casey et al., 2018), apontando apenas que uma explicação deveria envolver, no mínimo, como um determinado dado de entrada (*input*) se relaciona com suas predições. Em sua conclusão[146], os autores reconhecem que o GDPR cria "bons problemas", indicando que os algoritmos não poderão mais ser meramente eficientes, mas devem ser transparentes, igualitários e justos, destacando a importância da discussão ética no contexto técnico.

Este artigo motivou Sandra Wachter, Brent Mittelstadt e Luciano Floridi, também pesquisadores do OII, a questionarem se um direito à explicação de decisões exclusivamente automatizadas no GDPR poderia ser extraído desta norma (Wachter et al., 2017).[147] O que é especialmente interessante neste trabalho é a

[145] Tradução livre de: "The GDPR's policy on the right of citizens to receive an explanation for algorithmic decisions highlights the pressing importance of human interpretability in algorithm design".

[146] Além de questões relacionadas a explicações de modelos de ML, os autores mostram como modelos de ML podem apresentar predições discriminatórias mesmo que dados pessoais sensíveis (i.e. raça e etnia, por exemplo) não sejam utilizados em análises, diante das inferências que podem ser feitas na análise de outros dados (como a partir do endereço). Além disso, dependendo do modelo adotado e do objetivo que ele visa alcançar (i.e. modelo de análise de crédito que é avesso a risco), também podem ocasionar discriminações pela análise conjunta de uma série de dados que não seriam, a priori, considerados como dados sensíveis.

[147] Esta posição é feita no próprio título do artigo: *Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation.*

delimitação e a diferenciação dos autores do que o GDPR estaria exigindo em termos de explicação. Neste sentido, os autores acreditam que existiriam dois tipos de explicações de decisões tomadas exclusivamente com base em tratamento automatizado. O primeiro se refere a uma explicação sobre a funcionalidade do sistema (*system functionality*), que envolveria informações sobre a:

> lógica, significância, consequências desejadas e funcionalidades gerais de um sistema automatizado de tomada de decisões, e.g. as especificações sobre os requerimentos do sistema, a árvore de decisão, os modelos pré-definidos, critérios e classificação de estruturas"[148] (Wachter et al., 2017, p. 6).

O segundo seria uma explicação específica sobre uma determinada decisão (*specific decisions*), que diz respeito à decisão em si, especialmente ao

> "*rationale*, às razões e circunstâncias individuais de uma decisão automatizada específica, e.g. o peso atribuído às características, casos definidos por máquina e regras de decisões específicas, informações sobre grupos de referência ou perfil"[149] (Wachter et al., 2017, p. 6)[150].

Cada uma dessas explicações deve ser exigida em um determinado momento no tempo. As explicações sobre funcionalidade do sistema podem ser apresentadas antes de uma decisão algorítmica ocorrer (*ex ante*), enquanto uma explicação sobre uma decisão específica apenas pode ser exigida depois que esta aconteceu (*ex post*). No entanto, neste momento, também é possível fornecer explicações sobre as funcionalidades do sistema.

Por exemplo, caso um algoritmo seja utilizado para avaliar e decidir sobre a concessão de um empréstimo financeiro, antes de o perfil do cliente solicitante ser avaliado pelo algoritmo, apenas seria possível requerer uma explicação sobre a funcionalidade do sistema (categorias de árvore de decisão, ferramentas utilizadas no desenvolvimento do sistema, e funcionalidades do mesmo). Após o algoritmo ter analisado o pedido do cliente e tomado uma decisão sobre a concessão do empréstimo, além da explicação sobre a funcionalidade do sistema, seria possível apresentar uma explicação sobre a lógica e as circunstâncias daquela decisão

[148] Tradução livre de: "significance, envisaged consequences and general functionality of an automated decision-making system, e.g. the system's requirements specification, decision trees, pre-defined models, criteria, and classification structures".
[149] Tradução livre de: "rationale, reasons, and individual circumstances of a specific automated decision, e.g. the weighting of features, machine-defined case- specific decision rules, information about reference or profile groups".
[150] Conforme exposto em Mulholland; Frajhof, 2019, p. 279-280.

específica (Wachter et al., 2017, p. 7) (categorias de perfis, peso das variáveis e dos dados, como os perfis se relacionam com o resultado), uma vez que esta já foi tomada.

Vale notar que esta diferenciação se assemelha à perspectiva da literatura técnica de aprendizado de máquina, que distingue uma explicação global de uma explicação local. Esta última se refere a uma explicação sobre uma decisão tomada pelo modelo de ML, em relação a uma parte específica do mesmo, visando oferecer os elementos relevantes para aquele resultado em especial. Por sua vez, uma explicação global se refere a informações que dizem respeito ao modelo como um todo (Ribeiro et al., 2016). Embora o texto não mencione explicitamente esta referência, é possível que a motivação para a separação entre explicações sobre funcionalidade de um sistema e de uma decisão específica tenham se originado desta literatura.

Ademais, segundo Wachter et al., o art. 22 e o Considerando 71, os arts. 13 e 14, e os Considerandos 60-62, e o art. 15 e o Considerando 63, todos do GDPR, se referem às salvaguardas que devem ser adotadas quando há uma decisão automatizada e aos direitos de notificação e de acesso a dados de seus titulares, formando um conjunto que permitiria compreender o suposto direito à explicação. Os autores chamam a atenção para o fato de que o termo "explicação" apenas é utilizado no Considerando 71, do GDPR, e que o legislador teria optado por não repetí-lo no art. 22, da norma. Para Wachter et al., os Considerandos não possuem caráter vinculante, e exercem uma função de guia interpretativo do texto normativo para auxiliar na resolução de eventuais ambiguidades.[151]Portanto, a ausência de um direito explícito deixaria dúvidas sobre o que exatamente deve ser exigido dos controladores em termos de acesso à informação sobre a decisão automatizada.

Em suma[152], os autores defendem que não existiria algo parecido como um direito à explicação sobre uma decisão algorítmica, e sim um direito a ser informado, que estaria fundamentado no direito de acesso, e se restringiria às

[151] Kaminski discorda desta posição, indicando que: "A Recital is supposed to "cast light on the interpretation to be given to a legal rule [but] it cannot in itself constitute such a rule." This gives Recitals a liminal legal status—they are not binding hard law, but they are often cited as authoritative interpretations where the GDPR is vague" (Kaminski, 2019a, p. 8) (notas de rodapé removidas).

[152] Os autores de dedicam a uma análise mais detida dos artigos 13, 14, 15 e 22 do GDPR, e comparam a nova redação desta norma com a antiga Diretiva 95/46/EC. No entanto, nem todos os pontos argumentativos serão apresentados, tendo em vista que, para fins deste trabalho, o que é interessante é de fato a diferenciação dos autores sobre os dois tipos de explicações que poderiam existir.

explicações sobre a funcionalidade do sistema[153]. Contudo, estas estariam restritas aos interesses dos controladores (i.e. segredos de negócio, previsto no Considerando 63, do GDPR) e sujeitas a outras restrições em interpretações futuras (i.e. hipóteses que se enquadrariam dentro da definição de "decisão totalmente automatizada"). Para os autores, para que uma explicação sobre uma decisão específica possa ser exigida, seria necessária uma alteração legislativa, seja pelos Estados-Membros, seja no próprio GDPR, ou pelas definições realizadas pelo CEPD e pela própria jurisprudência.

Contudo, muitas das dúvidas apresentadas pelos autores foram respondidas pelo parecer sobre o tema do GTA29, o que permitiu maior clareza sobre interpretações normativas sobre quais são os direitos, medidas e salvaguardas que devem ser adotadas quando há uma decisão totalmente automatizada.[154]

Em resposta à Wachter et al. (2017), Julia Powles e Andrew Selbst (2017) publicaram um artigo, criticando também a publicação de Goodman e Flaxman, apontando que ambos os trabalhos acabaram influenciando o debate acadêmico sobre o tema, sem de fato tratar do que os autores consideram relevante: o direito do titular de dados de obter "informações úteis relativas à lógica subjacente" quando há a aplicação de ML e IA no tratamento de dados pessoais.

São quatro principais argumentos apresentados pelos autores rebatendo Wachter et al. (2017), no que se refere à análise dos arts. 13(2)(f), 14(2)(g), 15(1)(h), do GDPR. O primeiro é que Powles e Selbst destacam a importância de que as informações úteis exigidas devem ser significativas para um humano que não possui expertise técnica. Isto se relaciona com o segundo argumento, que é: a interpretação do que seja "informação útil" deve ser compreendida como um valor instrumental, que viabiliza o exercício de outros direitos, e não um valor intrínseco[155] à autonomia humana. Esta perspectiva dá maior concretude e facilita

[153] Isto corre principalmente pela linguagem utilizada nos arts. 13(2)(f), 14(2)(g), 15(1)(h), do GDPR, que se refere às "informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento para o titular dos dados". Na visão dos autores, isto estaria exigindo informações sobre a finalidade e objetivos que o sistema buscava alcançar, e a disponibilização de informações sobre a funcionalidade do sistema seria capaz de atender ao comando normativo.

[154] Tal qual a discussão sobre a natureza do art. 22, do GDPR se o mesmo seria um direito do titular de dados de se opor, ou uma proibição, em que o GTA29 já se manifestou sobre a natureza proibitiva do artigo (GTA29, 2017).

[155] Sob a perspectiva da explicação como um valor intrínseco poderia haver o risco de que fosse exigida uma explicação muito casuística, focada na demanda subjetiva e pessoal de cada pessoa (Selbst; Powles, 2017, p. 7).

a análise sobre o quão relevante deve ser esta informação, além de fortalecer sua defesa quando este direito entrar em conflito com outros interesses e direitos, tal qual o segredo de negócio. O terceiro ponto é que a informação deve ter funcionalidade, ou seja, deve ser útil o suficiente para permitir que o titular de dados possa exercer seus direitos assegurados pelo GDPR. Isto implica na informação ser capaz de permitir que o titular de dados possa identificar se houve ou não uma discriminação, por exemplo. Por fim, a exigência legal deve ser interpretada de maneira flexível e funcional, sem a necessidade de se estabelecer aprioristicamente qual é o tipo de explicação que deve ser fornecida. A rigidez em definir uma explicação *ex ante* ou *ex post*, sobre a funcionalidade do sistema ou sobre o *rationale* da decisão em si, pode acabar afetando usos complexos de soluções de ML, como é o caso de redes neurais, em que diferentes formatos e tipos de informações poderão ser apresentadas e servir como explicações adequadas (Selbst; Powles, 2017, p. 7-9).[156]

Em relação ao art. 22, do GDPR os autores indicam que, ao contrário do que pressuposto por Wachter et al., é este artigo e o Considerando 71, do GDPR, que sustentam a leitura dos arts. 13, 14 e 15, do GDPR, e são os fundamentos para defender a existência de um direito à explicação, e não o contrário. Chamam atenção que, embora os autores rejeitem a força vinculante do Considerando 71, do GDPR, os mesmos se valem do Considerando 47 e 63, do GDPR, para possíveis restrições que poderiam ser feitas por outros interesses e que podem entrar em conflito com o chamado direito a ser informado. Além disso, é visto como incoerente a rejeição da existência de um direito à explicação, embora Wachter et al. reconheçam a existência de um direito a ser informado, apesar de limitado a informações relacionadas às funcionalidades de sistemas.

Selbst e Powles acreditam que a maneira como modelos de ML funciona é determinística, ou seja, considerando um certo número de *inputs* utilizado pelo modelo, será apresentado o mesmo *output* se o modelo não for alterado ou não se altere. Os autores acreditam que, por tal motivo, o modelo seria previsível, de forma que, assim como uma explicação sobre a funcionalidade poderá ser fornecida, uma explicação sobre uma decisão específica também poderá ser apresentada, o que

[156] Caso se entendesse que o GDPR apenas exige explicações sobre a funcionalidade do sistema, qual seria a utilidade em explicar os resultados de uma rede neural apresentando informações sobre suas multicamadas e arquitetura complexa? (Flaxman; Goodman, 2016, p. 6-8)

desafia o argumento dos pesquisadores do OII sobre o momento que uma explicação poderia ser exigida (*ex ante* ou *ex post*). Para Selbst e Powles a distinção dos tipos de explicações não teria fundamento, pois tanto explicações sobre a funcionalidade, quanto sobre uma decisão em si, são vistas como complementares.

Esta perspectiva de os autores sobre os resultados serem determinísticos em modelos de ML não é totalmente verdadeira. Como visto, uma das características da IA, e que é especialmente relevante na área de aprendizado de máquina, é justamente a sua imprevisibilidade, diante do seu comportamento emergente. É inerente aos modelos de ML sua alteração conforme vão aprendendo e se desenvolvendo, evoluindo e se aperfeiçoando com a análise constante de novos dados. Ademais, mesmo que o resultado fosse considerado determinístico, a opacidade e complexidade de alguns modelos de ML pode não permitir que seja compreendido o que motivou certa predição e resultado. Logo, se o objetivo de um titular de dados é compreender os motivos que ocasionaram certa decisão em relação a ele ou ela, o fato de o modelo de ML ser considerado determinístico não implica em garantir que o titular de dados irá ter acesso a estes motivos, a fim de conhecer quais informações foram determinantes para àquela decisão.

Por fim, os autores discordam de Wachter et al., e reconhecem que a previsão nos arts. 13 a 15, e art. 22, do GDPR, poderia ser chamada de um direito à explicação, apontando que menos importaria o nome de tal direito, e sim a valorização e proteção que é garantida aos titulares de dados a partir do mesmo. É defendido que o direito à explicação deverá ser interpretado de maneira funcional e flexível, de forma que os titulares de dados sejam capazes de exercer seus direitos previstos no GDPR, assim como em normas de direitos humanos que estabeleçam previsões sobre o tema.

A discussão apresentada por este conjunto de artigos[157] levanta pontos interessantes para a construção de um direito que assegure às pessoas a possibilidade de: (i) saber quando estão sujeitas a uma decisão algorítmica em que não há intervenção humana, (ii) exercer uma espécie de devido processo algorítmico, com a garantia de salvaguardas que visem proteger seus direitos e

[157] Como apontam Casey et al (2019, p. 162), este conjunto de artigos motivou um debate global profundo, que discute os benefícios econômicos, viabilidade técnica e *tradeoffs* sociais sobre a importância de *accountability* algorítmica e as práticas que devem ser adotadas por governos e pelo merado.

liberdades, (iii) conhecer os motivos que embasaram esta decisão, especialmente quando diante do uso de algoritmos de tomada de decisão que se valem de ML. Como se verá no capítulo subsequente, há, ainda, uma perspectiva coletiva do direito à explicação, sob uma visão mais ampla, como sendo um meio de realizar uma prestação de contas de sistemas de IA e algoritmos de tomada de decisão de ML. Esta exige a documentação de informações relacionadas ao *design* e desenvolvimento destes artefatos como formas de viabilizar uma interpretação e explicação.

No mais, cabe destacar que as dúvidas sobre o tema no âmbito do GDPR certamente vão auxiliar no debate que deverá ser travado no Brasil sobre o direito à explicação. A seguir será apresentado como o direito à explicação foi introduzido na LGPD, e compartilhadas algumas propostas interpretativas sobre o mesmo.

1.5
O direito à explicação previsto na LGPD[158]

O direito à explicação previsto no art. 20, da LGPD, assim como no próprio GDPR, não diz respeito a um mero pedido de acesso a informações. Aquele pode ser compreendido como um conjunto de direitos e prerrogativas que o titular de dados pode exercer quando for impactado por uma decisão tomada unicamente com base em tratamento automatizado de dados pessoais, que afete "seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade". Assim como ocorre com o GDPR, o artigo se aplica a decisões automatizadas, que podem ou não estar destinadas à formação de perfil, indicando a possibilidade de que o direito à explicação possa ocorrer independentemente de ter havido ou não a formação de perfil.

A norma, de natureza permissiva (Mulholland; Frajhof, 2019, p. 275), autoriza "decisões tomadas unicamente com base em tratamento automatizado de dados pessoais" que afetem os interesses do titular de dados, conforme indicado acima. Quando esta ocorrer, o titular de dados poderá demandar uma série de medidas específicas. São elas: (i) pedido de revisão (art. 20, *caput,* da LGPD), (ii)

[158] Este subitem contém trechos que são transcrições integrais, com pequenas modificações, do artigo Frajhof, 2021.

acesso a "informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada" (art. 20, § 1º, da LGPD), e (iii) pedido por petição à Autoridade Nacional de Proteção de Dados Pessoais (ANPD) para que realize uma auditoria para verificar aspectos discriminatórios em tratamento automatizado de dados pessoais, caso o controlador se negue a fornecer estas informações com base no alegado segredo comercial e industrial (art. 20, § 2º c/c art. 18 § 1º, da LGPD), e (iv) direito de se opor a este tratamento (art. 18 § 2º, da LGPD) [159]. Este conjunto de direitos visa que ninguém possa "ficar sujeito, de forma irrestrita e sem garantias, aos julgamentos decorrentes de decisões totalmente automatizadas" (Frazão, 2018b), assegurando o que tem se denominado como uma espécie de devido processo algorítmico (Kaminski, 2019a; Citron; Pasquale, 2015). Como se verá, este conjunto também deverá englobar outros direitos.

Embora a LGPD tenha reunido em seu artigo 18 alguns dos direitos que o titular de dados poderá requerer perante o controlador[160], o legislador já vinha os enunciando desde o art. 1º da Lei, de forma que o referido artigo tem a função de sistematizar e reunir direitos anteriormente estabelecidos (Silva, 2020, p. 195). A previsão do art. 20, da LGPD, da mesma forma, também reúne e organiza uma série de prerrogativas que decorrem não apenas da autodeterminação informativa do titular de dados (art. 2º, II, da LGPD) (Frazão, 2018a), mas dos próprios princípios da Lei, que são:

> (i) livre acesso (art. 6º, IV), (ii) qualidade e a clareza dos dados (art. 6º, V), (iii) transparência dos dados, (art. 6º, VI), (iv) a prevenção de danos (art. 6º, VII), (v) não discriminação (art. 6º, IX) e (vi) responsabilização e prestação de contas (art. 6º, X) (Frazão, 2018a).

[159] Ana Frazão também aponta para a existência de um direito a se opor à decisão automatizada, com base no art. 18, § 2º, da LGPD. Contudo, ressalvo que, embora o GDPR estabeleça de maneira mais explícita que nos casos em que há o tratamento de dados pessoais automatizado o titular de dados poderá se opor ao referido tratamento, na LGPD isto apenas poderá ocorrer quando o tratamento de dados pessoais tiver como fundamento uma das hipóteses legais que não o consentimento, e quando este estiver descumprindo uma das disposições legais da LGPD.

[160] São eles: (i) confirmação da existência de tratamento; (ii) acesso aos dados; (iii) correção de dados; (iv) anonimização, bloqueio ou eliminação de dados desatualizados, excessivos ou tratados em desconformidade com a lei; (v) portabilidade de dados; (vi) eliminação dos dados pessoais tratados com o consentimento, (vii) informações sobre o compartilhamento de dados; (viii) informação sobre a possibilidade de não fornecer consentimento e sobre as consequências da negativa; (ix) revogação do consentimento. O direito à portabilidade é verdadeiramente a única novidade prevista no referido artigo.

Portanto, esta leitura do artigo viabiliza que possam ser exigidas uma série de salvaguardas e direitos para fazer valer o direito à explicação.

O art. 20, da LGPD, entretanto, sofreu um revés que fragiliza este amplo conjunto de direitos do titular de dados frente a uma decisão automatizada. A redação original da Lei aprovada pelo Congresso Nacional estabelecia que tal revisão fosse realizada por "pessoa natural", o que foi vetado pela Medida Provisória nº 869/2018, convertida na Lei nº 13.853/2019 [161], removendo a obrigação de que um humano tivesse que revisar a decisão automatizada. A ausência de uma pessoa para analisar o pedido de revisão[162] torna a sua efetividade questionável, diante do desafio técnico de implementar sistemas aptos a analisarem este requerimento e procederem com a revisão, além de limitar a autonomia do sujeito que requer a revisão, pois este pode não ter conhecimento específico para interagir com o sistema, bem como com as respostas fornecidas pelo mesmo. Ainda sobre a revisão, é importante ressaltar que o fato de a lei autorizar o pedido de revisão, não implica que, após a sua análise, o resultado final necessariamente será alterado (Mulholland; Frajhof; 2019, p. 272).

A intervenção humana em processos algorítmicos, ao longo do desenvolvimento e após a sua implementação, tem sido considerada como um princípio ético importante, e uma forma de prevenir e mitigar danos causados pela tecnologia (Comissão Europeia, 2019a, p. 16). Neste sentido, guias de recomendação para a adoção de princípios éticos no desenvolvimento e regulação da IA [163], produzidos por governos, pela iniciativa privada e por organismos

[161] As razões ao veto apresentado pelo Presidente Jair Bolsonaro foram: "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." Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Msg/VEP/VEP-288.htm>. Acessado em 20.03.2021.

[162] A classificação para avaliar a presence humana em sistemas de IA pode ser compreendida como: humanos-*dentro*-do-loop, humanos-*no*-loop ou humanos-*no*-comando.

[163] Foram analisados os princípios éticos elaborados pela OCDE, *Access Now*, IBM e Microsoft. Além disso, também foram consultados: o Relatório "European Group on Ethics in Science and New Technologies Artificial Intelligence, Robotics and 'Autonomous' Systems", a Estratégia Brasileira para a Inteligência Artificial do MCTIC, a "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems", Princípios de Asilomar, dentre outros.

internacionais, colocam a supervisão humana[164] como um elemento fundamental, tornando-se um princípio constante em todos estes documentos. Portanto, o veto à necessidade de que a revisão fosse realizada por uma pessoa humana constitui não apenas um óbice relevante para a efetividade deste direito, mas como uma inobservância de um princípio ético para a IA que se encontra mapeado e previsto em diversos documentos sobre o tema.

A importância da supervisão humana em processos algorítmicos foi reconhecida pelo Comitê de Supervisão do Facebook (*Oversight Board*), que foi criado para auxiliar a rede social em relação às decisões de moderação de conteúdo tomadas por ela, bem como pela rede social Instagram. Em janeiro de 2021, o Comitê proferiu uma decisão importante para a discussão do direito à explicação, reconhecendo a necessidade de haver uma supervisão humana em processos e decisões automatizadas envolvendo a moderação de conteúdo. Isto deixa evidente como o direito à explicação, embora tenha origem e previsão no contexto de proteção de dados pessoais, não está restrito às discussões envolvendo este tema, abrangendo diferentes contextos e afetando outros direitos.

O Comitê Supervisor tem o objetivo de auxiliar a rede social em suas decisões envolvendo a moderação de conteúdo, e visa defender a liberdade de expressão e segurança dos usuários, com base nas políticas e valores declarados pela rede social. As decisões são vinculantes, e o Facebook deve implementá-las, salvo se as mesmas violarem a lei. Apesar da interessante proposta da formação deste Comitê para fins de governança, bem como para maneiras alternativas de regulação de plataformas digitais, o que cabe ser analisado no objeto deste trabalho é a decisão do caso 2020-004-IG-UA.[165] O caso tratou de uma decisão do Facebook que, por meio de um processo automatizado, baseado em sua política de moderação de conteúdo, removeu uma publicação da plataforma Instagram por violação aos padrões da comunidade do Facebook sobre nudez adulta e atividade sexual. A publicação tratava de conteúdo de "conscientização sobre o câncer de mama", em que fotos de mamilos femininos apareciam descobertos. O Conselho considerou

[164] Como destacam Mulholland e Frajhof, isto "significa dizer que toda IA deve necessariamente ser centrada na pessoa humana, a ela direcionada e por ela supervisionada, engajando os sistemas para servir à coletividade, no sentido de amplificar não só a sua segurança, mas também garantir a autonomia e a capacidade de decidir das pessoas" (Mulholland; Frajhof, 2021, p. 73).

[165] Comitê Supervisor do Facebook. Decisão 2020-004-IG-UA. Proferida em janeiro de 2021. Disponível em: https://www.oversightboard.com/sr/decision/004/Portuguese. Acessado em 02.10.2021.

que o conteúdo removido tratava de questões importantes de direitos humanos e que deveria ser restabelecido, além de reconhecer que este tipo de conteúdo poderia ser publicado pelos usuários por ser uma das exceções previstas nas políticas do Facebook em relação a nudez adulta e atividade sexual.

O caso tratou de uma denúncia feita em outubro de 2020, por um usuário brasileiro, que denunciou a remoção de uma foto no Instagram da campanha internacional chamada "Outubro Rosa" sobre conscientização do câncer de mama. Conforme narrado pelo Comitê, eram oito fotografias compiladas em uma única imagem com mamilos que apresentavam sintomas do câncer de mama: cinco delas tinham mamilos femininos que estavam descobertos e visíveis, e três eram mamilos que não estavam enquadrados na foto ou estavam cobertos por uma mão. A publicação havia sido removida de maneira automatizada por violar as diretrizes e políticas de comunidade do Facebook e do Instagram. Quando o caso foi selecionado para análise pelo Conselho, o Facebook restaurou a publicação, alegando haver um erro da aplicação, em que a publicação removida se enquadraria em uma das exceções dos padrões da comunidade, mas que apenas foi tomada ciência por conta da escolha do caso. Contudo, o Conselho entendeu que deveria prosseguir com a sua análise, uma vez que seria importante que fosse oferecida "uma explicação completa do motivo pelo qual a publicação foi removida".

Em sua decisão, o Conselho ressaltou que a remoção do conteúdo havia sido feita por um classificador de aprendizado de máquina. Este tinha sido treinado para identificar fotos de nudez e proceder com a sua moderação, atendendo aos termos dos padrões da comunidade do Facebook, em especial, a parte que trata de nudez adulta e atividade sexual, que também é aplicável ao Instagram. A decisão do Conselho analisou a questão sob os padrões internacionais de direitos humanos, em especial, os Princípios Orientadores da ONU sobre Empresas e Direitos Humanos (UNGPs), endossados pelo Conselho de Direitos Humanos da ONU de 2011, citando uma série de direitos e os respectivos tratados internacionais de direitos humanos que embasaram a decisão [166]. Com base neste arcabouço jurídico, entendeu-se que a remoção da referida imagem não estaria aderente aos padrões de

[166] O direito à liberdade de expressão: Pacto Internacional dos Direitos Civis e Políticos (PIDCP); Pacto Internacional dos Direitos Econômicos, Sociais e Culturais (PIDESC); Convenção sobre a eliminação de todas as formas de discriminação contra as mulheres (CEDAW); Convenção internacional sobre os direitos da criança (CRC).

comunidade do Facebook e do Instagram, bem como aos padrões de direitos humanos internacional. Em especial, aquela não atenderia aos requisitos necessários à restrição da liberdade de expressão (legalidade, objetivo legítimo, necessidade e proporcionalidade) e uma perspectiva de igualdade e não discriminação.

Em relação à necessidade e proporcionalidade, o Conselho chama atenção para o fato de a remoção automatizada ter cometido um erro em sua moderação, revelando a limitação da tecnologia, e a preocupação na ausência de uma análise humana em relação ao conteúdo removido. Embora reconhecida a importância de se valer da tecnologia automatizada para a moderação de conteúdo, foi ressaltado que a mesma não consegue compreender certos contextos, e isto pode ocasionar restrições excessivas, interferindo de maneira desproporcional na expressão do usuário. A remoção automatizada não estaria aderente às regras das políticas da comunidade das plataformas, pois a publicação removida estaria autorizada visto que destinada a fins médicos e educacionais. Além disso, destacou-se a necessidade de que a rede social notificasse e explicasse ao usuário por que o seu conteúdo foi removido. Foi indicado que qualquer remoção deve estar sujeita a uma auditora interna, bem como que o recurso apresentado por um usuário a uma moderação de conteúdo realizada pelo Facebook deve ser analisado por um humano, de forma a permitir que erros na aplicação possam ser corrigidos.

Por fim, as recomendações do Conselho ao Facebook envolveram: melhorar sua detecção automática de imagens quando há a sobreposição de texto em imagens no contexto de campanhas de câncer de mama; que os usuários sejam notificados dos motivos de verificação de suas publicações, devendo ser informada a regra específica do Padrão de Comunidade do Facebook aplicada a elas; informar quando há automação aplicada a um conteúdo postado, devendo incluir descrições acessíveis sobre o que significa essa automação; garantir a possibilidade de que os usuários possam recorrer a um humano quando há a aplicação de decisões tomadas por sistemas automatizados que impliquem em remoção de conteúdo em violação às regras e políticas da rede social envolvendo nudez adulta e atividade sexual; implementar um procedimento de auditoria interna que faça análises contínuas sobre as amostras representativas e estatísticas das decisões automatizadas de remoção de conteúdo, a fim de melhorar o aprendizado e futuras moderações de conteúdo; e expandir relatórios de transparência, trazendo dados sobre as decisões

automatizadas e remoções de conteúdo de acordo com o Padrão da Comunidade, e a proporção de decisões que foram posteriormente revertidas por uma supervisão humana.

Este posicionamento é fundamental no contexto do direito à explicação e decisões totalmente automatizadas, pois revela, na prática, a importância de que haja um monitoramento e controle humano dos resultados gerados por estas aplicações, a fim de avaliar o seu impacto nos direitos das pessoas afetadas pela tecnologia e sua legitimidade e conformidade. O reconhecimento da importância da supervisão e revisão humana indicam o quão limitante será o direito à explicação na LGPD. Além disso, a transparência sobre a moderação de conteúdo automatizada e a comparação sobre as taxas das reversões, além da própria auditoria interna, também auxiliam na avaliação de erros e vieses destes algoritmos, dando a oportunidade para melhorá-los e corrigi-los. Por fim, a valorização da informação e explicação aos usuários de quando há uma verificação sendo realizada, além da possibilidade de recorrer de uma remoção automatizada, com supervisão humana, implementam uma espécie de devido processo algorítmico, e apontam para o reconhecimento da transparência e do acesso à informação como meios essenciais para o exercício de direitos fundamentais. A exigência de explicação também demonstra que não apenas uma única pessoa que teve seu conteúdo removido se beneficia da compreensão do porquê isso ocorreu, mas isto serve como precedente para todos os demais usuários da rede, evidenciando um interesse coletivo no benefício da explicação individual.

Retornando à análise do art. 20, da LGPD, esta não definiu em seu texto normativo o que seria considerada como uma decisão totalmente automatizada. Diante desta lacuna, já foi apresentado no Senado Federal o Projeto de Lei (PL) n° 4.496, de 2019, que propõe tal definição[167]. Contudo, este PL não traz ao debate qualquer questão relevante, tal como feito pelo GTA29, em relação ao grau de interferência humana necessário para descaracterizar ou caracterizar uma decisão como tomada exclusivamente por meios automatizados. O PL apenas elenca uma série de atividades que poderiam ser objeto de automação, pouco, ou em nada, auxiliando na compreensão do assunto. Na lacuna sobre este tema, o parecer do GTA29 aponta para interpretações interessantes para o que pode ser considerado

[167] Ver em: https://www25.senado.leg.br/web/atividade/materias/-/materia/138136. Acessado em 20.01.2021.

como “decisões tomadas unicamente com base em tratamento automatizado de dados pessoais”, indicando o nível de interferência humana exigido para qualificar ou desqualificar o que poderia ser considerada como uma decisão totalmente automatizada.

Ademais, não é qualquer decisão totalmente automatizada que atrairá a aplicação do art. 20, da LGPD. As hipóteses listadas no *caput* do referido artigo devem ser compreendidas como sendo exemplificativas, e não taxativas (Monteiro, 2018, p. 10), acionando uma responsabilidade para a ANPD, a doutrina e o judiciário, em fornecer exemplos e prever as situações em que o direito à explicação e à revisão poderão ser exigidos, a fim de dar maior concretude ao que constituem ser estes interesses. Tendo em vista que a LGPD não traz um requisito mais específico sobre o tipo de impacto que uma decisão automatizada deverá ter para o titular de dados, as hipóteses de incidência da norma acabam sendo muito amplas, bastando que a mesma afete seus interesses, ou se destine a formar perfis sobre aspectos pessoais, profissionais, de consumo e de crédito ou os aspectos de sua personalidade.

A categoria de “perfis pessoais” se relaciona, portanto, à pessoa, identificada ou identificável, e acaba sendo uma categoria ampla que permite abarcar diversos usos, sendo suficientes que estejam relacionados a aspectos pessoais do titular de dados. A fim de especificar o contexto no qual estes perfis são constituídos e utilizados, os exemplos trazidos pelo Considerando 71 do GDPR podem auxiliar nesta delimitação.

Como já defendido em outra oportunidade, o reconhecimento da proteção de dados como direito fundamental a dota de eficácia direta e imediata, característica inerente a este tipo, fazendo com que tal direito deva ser respeitado tanto pelo Estado, quanto nas relações travadas entre particulares (Sarmento, 2008, p. 107). Portanto, quando há a aplicação de algoritmos de tomada de decisão envolvendo o tratamento de dados pessoais, dependendo do que estiver em disputa sobre o que será decidido, o que deverá ser analisado é se o controlador está ou não negando, ou até mesmo deixando de promover, a fruição de um direito fundamental. Desta forma, a avaliação é necessária para verificar se houve ou não discriminação no tratamento de dados automatizado, bem como se o mesmo deixa de promover adequadamente o direito à proteção de dados. Além disso, é necessário avaliar o bem jurídico que está envolvido na decisão automatizada, se o bem irá realizar

funções sociais constitucionalmente asseguradas (concessão de financiamento de crédito estudantil, por exemplo) (Mulholland; Frajhof, 2019, p. 270-271).

Ademais, o art. 20, § 1º, da LGPD também estabeleceu de maneira ampla quais são as informações que, uma vez pleiteadas pelo titular de dados, devem ser disponibilizadas pelo controlador sobre os critérios e procedimentos utilizados para a decisão automatizada, resguardando o segredo comercial e industrial. Este parágrafo introduz a discussão sobre o conteúdo que deve ser disponibilizado quando ocorrer uma decisão automatizada. Ora se defende que não seja utilizada a distinção feita por Wachter et al., adotando-se o argumento de Selbst e Powles de que não seria interessante definir de maneira rígida um ou outro tipo de explicação, mas que ambos os tipos de explicações, sobre as funcionalidades e sobre as decisões em si, devem ser compreendidos como complementares. Ainda, como se verá no próximo capítulo, a explicação sobre decisões algorítmicas deve exigir a documentação de outros tipos e natureza de informações relacionadas ao desenvolvimento de sistemas de IA.

Caso o controlador tenha se negado a oferecer as informações indicadas no §1º, e a negativa tenha sido baseada na necessidade de proteger o segredo comercial e industrial do controlador, a ANPD poderá realizar uma auditoria, com a finalidade de verificar aspectos discriminatórios do sistema. Esta previsão revela a importância do direito à explicação também como uma forma de proteger e atender ao princípio da não discriminação (Silva, 2019). Algumas críticas, contudo, merecem ser feitas. Primeiro, a discricionariedade da ANPD em proceder ou não com a auditoria pode incentivar o controlador a meramente alegar a proteção do sigilo comercial ou industrial, pois sabe que a opção de iniciar uma auditoria seria optativa. Segundo, caso o controlador apresente uma motivação diferente para não fornecer estas informações, tal como dificuldades técnicas, resta saber se isto será considerado como um fato impeditivo para a ANPD proceder com a auditoria ou não (Mulholland; Frajhof, 2019, p. 271).

No mais, atender ao comando do art. 20, da LGPD irá implicar na adoção de uma série de medidas prévias e posteriores, técnicas e não-técnicas, que envolvem o desenvolvimento do sistema envolvido na tomada de decisão automatizada. Serão necessárias condutas que não sejam apenas reativas, mas

preventivas[168], gerando um "grande dever de cuidado aos agentes de processamento" (Frazão, 2018c). Embora em um primeiro momento a discussão sobre o direito à explicação no GDPR tenha caminhado para debater qual o tipo de explicação que poderia ser exigido, atualmente, o que está em andamento é uma discussão mais ampla sobre *accountability* algorítmica (Kaminski, 2019b). Neste sentido, o direito à explicação tem se colocado como uma das maneiras de implementar esta prestação de contas,[169] contribuindo para questões que transcendem a discussão sobre proteção de dados.

Portanto, a seguir serão apresentados os desafios que os algoritmos de aprendizado de máquina apresentam para o Direito, e os possíveis instrumentos que podem auxiliar na viabilização do direito à explicação. Serão explorados os instrumentos jurídicos e as técnicas que podem ser aplicadas para obter explicações sobre as decisões algorítmicas.

[168] Até mesmo em atenção ao princípio da prevenção (art. 6o, VIII) e responsabilização e prestação de contas (art. 6o, X).

[169] Como indicado por Doshi-Velez e Mason: "While there are many tools to increasing accountability in AI systems, we shall focus on one in this report: explanation. (We briefly discuss alternatives in Section 7.) By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute" (Doshi-Velez; Kortz, 2017, p. 2).

2.
Os problemas da IA e dos algoritmos de tomada de decisão de aprendizado de máquina: FAT

A imprevisibilidade e a inescrutabilidade de algoritmos de aprendizado de máquina utilizados para a tomada de decisão ofuscam e impossibilitam a capacidade de interpretação do seu processo de trabalho e, portanto, desafiam o fornecimento de uma explicação dos seus resultados. Esta opacidade é um verdadeiro entrave para avaliar a legitimidade e legalidade dos seus frutos, e afeta a confiança das pessoas em relação ao mesmo (Zarsky, 2016, p. 129; Edwards; Veale, 2017; Wachter et al., 2018; Casey et al., 2019; Doshi-Velez; Kortz, 2017). A dificuldade de compreender estes resultados, predições e o seu modo de trabalho ocorre em determinados modelos matemáticos, e em razão da técnica aplicada ao algoritmo, que não permite identificar quais dados ou qual aspecto dos dados foram mais relevantes[170] para a predição ou seu resultado, qual funcionalidade influenciou e como, quais dados foram utilizados, e quais deles tiveram mais relevância nas inferências e predições realizadas. A opacidade acaba atrapalhando a capacidade de as pessoas questionarem as saídas e o processo interno destes algoritmos, razão pela qual medidas de transparência podem ser adotadas para mitigar esta desconfiança (Zarsky, 2016, p. 130).

Como já exposto no capítulo anterior, o direito à explicação busca oferecer "uma espécie de devido processo legal para proteger os cidadãos contra a 'tirania' dos julgamentos automatizados" (Frazão, 2018a)[171]. Estas garantias e proteções visam dar transparência sobre a justificativa e as motivações que ocasionaram a decisão automatizada, para que o titular pudesse ter uma interferência humana, possa contestá-la, pedir a sua revisão, e acessar e corrigir dados dos algoritmos. A explicação sobre uma decisão automatizada, no entanto, extrapola a própria redoma individual da pessoa afetada, visto que, diante da escalabilidade da IA, tal

[170] Por exemplo, quando pensamos em aprendizado de máquina aplicado a imagem, o modelo pode apenas sopesar certos aspectos de uma imagem para classificá-la como uma bicicleta. O algoritmo pode identificar que o que é relevante para classificar um objeto como bicicleta e não uma scooter é o guidão, a estrutura mais fina do quadro e o pedal, sendo irrelevante o tamanho das rodas ou o banco com espigão, por exemplo. A questão do aprendizado de máquina diz respeito exatamente a dificuldade em conhecer o que foi considerado pelo modelo como sendo relevante para a classificação.

[171] Assim como proposto por Citron e Pasquale (2014), que defendem uma espécie de devido processo legal algorítmico, assim como tantos outros autores e autoras, como reunido por Kaminski, 2019.

explicação pode fornecer subsídios que permitam compreender e avaliar a legitimidade e adequação do sistema como um todo (Frajhof, 2021, p. 468).

É neste sentido que algum nível de transparência torna-se importante para verificar se os resultados de um algoritmo de tomada de decisão de ML violam ou não preceitos legais, principalmente se violam ou deixam de promover o gozo de direitos fundamentais, como a liberdade, igualdade, privacidade e proteção de dados pessoais (art. 5º, *caput*, incisos X, XII, LXXIX, da CF), ou até mesmo direitos sociais, como o direito à educação, à saúde, ao trabalho, à moradia e a assistência aos desamparados (art. 7º, *caput,* da CF).

Tais preocupações são especialmente relevantes quando estamos diante de algoritmos e modelos de aprendizado de máquina, que são o objeto de estudo deste trabalho. A complexidade do seu funcionamento e a proteção legal assegurada aos mesmos pelo direito de propriedade intelectual e o segredo empresarial têm sido dois fortes argumentos que são apresentados como entraves para garantir esta transparência. Tais argumentos justificam a adjetivação destes algoritmos como sendo "caixas-pretas" e "sigilosos" (Pasquale, 2015), diante da dificuldade em interpretar e explicar seus resultados, e a impossibilidade de acessar seu código para realizar esta investigação. Considerando que a natureza dos problemas tem origens distintas (técnica e jurídica), qualquer solução que busque apresentar meios de garantir explicações sobre estes algoritmos de ML destinados à tomada de decisão deverá implicar em uma abordagem que seja multifacetada e multidisciplinar, dependendo de diferentes técnicas e métodos e pessoas com diferentes expertises[172] (Burrell, 2016, p. 10).

A transparência, no entanto, não significa uma abertura indiscriminada de todo e qualquer tipo de informação, com a mera disponibilização, por exemplo, do código fonte destes algoritmos. Visualizar o código e o seu funcionamento interno não significa entendê-lo (Ananny; Crawford, 2018, p. 978). Logo, uma "simples transparência" não seria o suficiente. Isto é, a abertura de elementos técnicos destes algoritmos, tal como, os parâmetros utilizados pelo sistema, a redução e abstração dos dados, a presunção de racionalidade estatística e a autoexecutividade da decisão,

[172] Conforme indica Burrell: "Alleviating problems of black boxed classification will not be accomplished by a single tool or process, but some combination of regulations or audits (of the code itself and, more importantly, of the algorithms functioning), the use of alternatives that are more transparent (i.e. open source), education of the general public as well as the sensitization of those bestowed with the power to write such consequential code" (Burrell, 2016, p. 10).

ainda assim tornam "a avaliação da conformidade jurídica um problema" (Frazão; Gottenauer, 2020, p. 57). É necessária uma tradução da linguagem matemática para a linguagem natural, capaz de construir uma ponte entre o sistema linguístico que sustenta o direito e a linguagem técnico-matemática que fazem parte das decisões automatizadas (Frazão; Gottenauer, 2020, p. 57). Podemos arriscar afirmando que o direito à explicação tem a pretensão de servir como esta ponte conciliatória.

A transparência é, portanto, um pressuposto para a verificação destes algoritmos, a fim de avaliar a legitimidade e legalidade dos seus resultados. Assim, neste capítulo serão abordados quais são os desafios que surgem com o amplo uso de algoritmos de tomada de decisão que se valem de ML. Isto é, como o comportamento emergente tensiona a transparência, a prestação de contas/responsabilização, e a justiça/igualdade. Neste sentido, três princípios básicos[173] (Casey et al., 2019, p. 148)[174] devem lastrear o desenvolvimento de qualquer tecnologia que se vale da IA, que são: justiça/igualdade (*fairness),* prestação de contas/responsabilização *(accountability)* e transparência *(transparency)*, representados pelo acrônimo FAT[175]. O FAT busca evitar que sistemas de IA produzam resultados discriminatórios, cometam injustiças ou causem erros e danos (Frajhof, 2021, p. 481). Além disso, este capítulo irá apontar meios para viabilizar e operacionalizar o direito à explicação a partir destes três princípios. O FAT é um pilar para proporcionar o direito à explicação, sendo a transparência um pressuposto para a viabilização dos outros dois, visto que permite *insights*, explicações, contestações e a própria prestação de contas e responsabilização (Artigo 19, 2019, p. 13).

À luz dos desafios dos algoritmos de tomada de decisão colocados ao FAT, será apontado como a doutrina tem apresentado maneiras de enfrentá-los, e indicado maneiras de eficazmente atendê-los. A observância do FAT é, também, um pressuposto para viabilizar a explicação de decisões algorítmicas, de forma que

[173] Conforme já exposto no capítulo anterior.

[174] Ao se referir às discriminações reveladas nos últimos anos causadas por algoritmos de tomada de decisão, Casey et al. indicam que: "These revelations, in turn, have had a pronounced effect on scholars, policymakers, industry leaders, and society *writ large*—often serving as a rallying cry for greater efforts to promote fairness, accountability, and transparency in the design and deployment of highly automated systems".

[175] O assunto vem sendo debatido por alguns anos na academia. A criação da *Association for Computing Machinery Fairness, Accountability and Transparency* (ACM FAT) demonstra este interesse. Há, também, uma grande produção de artigos sobre o assunto, como demonstra Linadartos et al., 2021.

avaliar as soluções apresentadas pela doutrina permitirá identificar quais são os possíveis caminhos que permitam o exercício do direito à explicação. Por ser a transparência um princípio que é o pressuposto para avaliar a prestação de contas e a responsabilização, além de eventuais discriminações e erros, este será analisado primeiro.

2.1
O Princípio da Transparência (*transparency*)

"A transparência não é apenas um fim em si mesmo, mas um passo provisório para o caminho da inteligibilidade"[176]

A demanda por transparência algorítmica tem sido comum na literatura especializada, e estaria justificada como uma forma de revelar e identificar potenciais violações à privacidade, proteção de dados, liberdade, igualdade e autonomia dos sujeitos impactados pelos algoritmos de tomada de decisão de ML. A transparência possui um papel importante para conhecer os fatores usados no processo de tomada de decisão do algoritmo, por que tais fatores foram utilizados, e oferece elementos que permitem que uma pessoa possa se defender e contestar esta decisão (Zarsky, 2013, p. 17), além de verificar a legitimidade da mesma (i.e quais foram as "razões de decidir" que justificaram uma decisão) (Kaminski, 2019b, p. 1.546).

A típica demanda por "abrir a caixa-preta" tem como presunção o fato de que a transparência sobre o sistema viabilizaria a prestação de contas necessária para avaliar a legitimidade e legalidade dos seus resultados, compreender o seu funcionamento, além de ser capaz de fornecer uma explicação sobre estes *outputs*. Contudo, apenas "abrir a caixa-preta" não é suficiente, e envolve diferentes medidas, diante dos diversos tipos de barreiras que impedem o acesso a certas informações. De forma geral, pode-se dizer que transparência visa enfrentar a opacidade dos diferentes tipos de algoritmos de ML, aplicados a diversos contextos e finalidades. Contudo, esta opacidade não diz respeito apenas à "interpretabilidade"

[176] Tradução livre de: "transparency is not just an end in itself, but an interim step on the road to intelligibility" (Pasquale, 2015, p. 8).

e inteligibilidade a nível de sistema (Burrell, 2016), mas aos diferentes tipos de sigilos que permeiam o mesmo (Pasquale, 2015)[177].

A doutrina apresenta o conceito de opacidade como sendo o oposto de transparência, e a coloca como um dos principais problemas no uso de algoritmos de ML (Zarsky, 2013, p. 119)[178]. Ou seja, a transparência destes sistemas e de seus códigos seria ofuscada (Pasquale, 2015, p. 7), dificultada e impossibilitada por diferentes motivos. O reconhecimento e a identificação de cada um deles torna-se uma premissa para tatear e apresentar soluções para contorná-los.

Neste sentido, Jenna Burrell identifica três eventos que podem ocasionar a opacidade dos algoritmos de aprendizado de máquina: (i) o segredo corporativo ou de estado; (ii) a necessidade de um conhecimento técnico específico (letramento técnico) para explicá-lo, e (iii) as características inerentes aos algoritmos de ML, em que há um descompasso entre a dimensão do código e a capacidade humana de interpretação e explicação do mesmo (Burrell, 2016). Distinguir cada um deles é importante para direcionar soluções adequadas e específicas para cada tipo de opacidade.

A opacidade por questões de segredo de negócio para obter vantagem comercial e competitiva (Burrell, 2016, p. 3), também é chamada por Pasquale como "segredo corporativo" (*corporate secrecy*) (Pasquale, 2015). Pode-se dizer que este tipo de opacidade é bastante conveniente para os agentes econômicos que detêm o código, pois permite que seus interesses sejam resguardados por um sigilo que é juridicamente protegido (Frazão; Gottenauer, 2020, p. 53).

Esta proteção legal atribuída aos algoritmos sob o manto do segredo empresarial tem sido objeto de crítica contundente por pesquisadores (Watcher; Mittelstadt, 2019; Pasquale, 2015; Frazão, 2021b), visto que esta tem se constituído como uma defesa legal para evitar investigações sobre o artefato e suas razões de

[177] Frank Pasquale afirma que a opacidade dos algoritmos resulta de três tipos sigilos, que são motivados por diferentes interesses. O primeiro seria o sigilo real (*real secrecy*), que se refere a quando desejamos manter privado determinados fatos para preservar nossa privacidade e intimidade; o segundo é o sigilo legal (*legal secrecy*), de quando há um dever ou obrigação legal ou contratual de manter em sigilo determinados fatos; o terceiro seria a ofuscação (*obfuscation*), que seria um sigilo que visa dificultar a compreensão de um determinado fato e desestimular o interesse em ter acesso a certas informações (Pasquale, 2015, p. 6-7).

[178] De acordo com Zarsky seriam dois tipos de problemas que são tipicamente invocados com o uso de algoritmos de ML: automação e opacidade (Zarsky, 2013). O fato de serem considerados como problemas não necessariamente vai implicar em uma política regulatória que proíba o uso desses algoritmos, diante dos reconhecidos benefícios do seu uso, mas as diferentes questões que surgem a partir desses problemas devem ser endereçadas de formas específicas e diferenciadas.

decidir. No Brasil, o ordenamento jurídico protege os algoritmos para incentivar a criação e promoção da circulação de riquezas, e estimular a livre iniciativa (arts. 1º, inciso IV e 170, da CF), que são promovidas pelos agentes econômicos responsáveis pelo desenvolvimento e o uso do código (Fernandes; Oliveira, 2020, p. 12). Esta proteção jurídica pode ter diferentes naturezas (segredo de indústria, de comércio ou proteção autoral), e estar fundamentada em diferentes marcos normativos (Lei de Propriedade Industrial (Lei n. 9.279/1996), Lei de Direitos Autorais (Lei n. 9.610/1998), e Lei do Software (Lei n. 9.609/1998)).

Quando o algoritmo está protegido pelo segredo de negócio ou empresarial, o que se busca garantir é a competitividade do negócio em relação a outros concorrentes do mercado[179]. O segredo visa proteger o *know how*, as fórmulas, as metodologias de produção, as técnicas, os fluxos e as organizações relacionadas à atividade empresarial (Anjos, 2019). De acordo com Micaela Fernandes e Camila Oliveira, o segredo empresarial seria o gênero, e o segredo comercial e industrial espécies. O primeiro se refere a um conhecimento aplicado ao comércio ou à prestação de serviços, e o segundo à produção industrial (Fernandes; Oliveira, 2020, p. 12). Ou seja, o objetivo dessa proteção tem origem no contexto empresarial e industrial, visando proteger e incentivar a concorrência.

Contudo, esta proteção cria uma tensão com outros direitos e interesses, tal qual com a proteção de dados pessoais. É interessante notar que a proteção ao segredo comercial e industrial é mencionada ao menos treze vezes[180] ao longo da LGPD. Até mesmo o princípio da transparência (art. 6º VI, da LGPD) prevê tal ressalva, para que seja garantido aos titulares "informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento", observando-se os segredos comercial e industrial. Esta restrição indica uma forte preocupação do legislador em balancear os interesses dos agentes econômicos com os interesses individuais e coletivos dos titulares de dados.[181]

[179] Neste sentido, Lucas Anjo, pesquisador do Instituto de Referência em Internet e Sociedade (IRIS), aponta que o segredo de negócio "refere-se a uma vantagem competitiva no modelo de negócio da empresa, cujo segredo e confidencialidade é justamente aquilo que garante sua proteção contra os concorrentes" (Anjos, 2019).

[180] Art. 6º, inciso VI; art. 9, inciso II; art. 10, § 3º; art. 18, inciso V; 19, inciso II e § 3º; 20 §§ 1º e 2º; art. 38; art. 48, § 1º inciso III; art. 55-J, incisos II e X e § 5º, todos da LGPD.

[181] Um questionamento relevante sobre este tema se refere ao deslocamento e à adequação deste tipo de argumento no contexto de demandas consumeristas e de proteção de dados, que nada se relacionam com o ambiente concorrencial e empresarial em que a justificativa e proteção ao segredo corporativo e industrial tem origem. Apesar de relevância deste tema, o mesmo acaba fugindo do escopo deste trabalho (agradeço ao Sérgio Ávila Negri quanto a este ponto, manifestado em debate

Contudo, esta proteção jurídica "não pode ser utilizad[a] para a ausência de explicações" (Frazão, 2021b), tal como vem sucedendo como uma justificativa legal por agentes econômicos para se abster de compartilhar qualquer tipo de informação envolvendo o algoritmo.

Esta proteção não deve ser considerada como absoluta. Como destaca Ana Frazão, até mesmo a propriedade intelectual, seja a industrial ou a autoral, também deve ceder diante de outros interesses sociais. A autora segue apontando que:

> (...) há boas razões para não considerar o segredo de empresa como algo absolutamente intocável ou sacrossanto, de forma a se exigir que, em algumas situações, ele seja sopesado diante de relevantes interesses sociais que possam ser prejudicados em virtude do segredo (Frazão, 2021b).

O tema da opacidade, explicação e interpretação de algoritmos chamou a atenção da Justiça do Trabalho recentemente[182]. O caso envolveu um pedido do motorista de aplicativo de transporte Uber, que pleiteava por uma perícia técnica do algoritmo da empresa para demonstrar a "presença dos elementos caracterizadores da relação de emprego". Embora o pedido tenha sido deferido em primeira instância, a amplitude da perícia foi delimitada pela desembargadora relatora em segunda instância, razão pela qual a empresa interpôs um recurso para o Tribunal Superior do Trabalho (TST), requerendo uma tutela cautelar de urgência, com pedido de efeito suspensivo ao recurso, para que a perícia fosse suspensa[183]. A justificativa alegada pela Uber, e acolhida pelo TST, seriam os riscos que poderiam advir da realização da perícia, "porquanto tem ela potencial de trazer à tona informações sigilosas, aparentemente fundamentais no segmento empresarial de atuação da Requerente, baseado em tecnologia digital".

Caso semelhante a este ocorreu em Wisconsin, nos Estados Unidos da América (EUA).[184] Eric Loomis foi preso em fevereiro de 2013, acusado de dirigir um automóvel que não era de sua propriedade e que havia sido utilizado em um tiroteio. Ele admitiu que havia fugido das autoridades policiais e que estaria

conduzido no âmbito do grupo de pesquisa Legalite, coordenado pela Professora Caitlin Mulholland, quando apresentei parte da minha pesquisa aos colaboradores do grupo em 26.05.2021, com o trabalho intitulado: "Meios para viabilizar o direito à explicação a partir do FAT").

[182] TST, acórdão n. 1000825-67.2021.5.00.0000, sob a relatoria do Ministro Douglas Alencar Rodrigues, j. em 28 de maio de 2021.

[183] Ver em: https://www.migalhas.com.br/quentes/346444/tst-ministro-suspende-pericia-tecnica-no-algoritmo-da-uber. Acessado em 24.06.2021.

[184] TASHEA, Jason. *Courts are using AI to sentence criminal. That must stop now.* Wired, 17 de abril de 2017.

dirigindo o carro sem a autorização do seu proprietário. Ao ser detido, Loomis respondeu algumas perguntas que foram inseridas em uma controvertida ferramenta utilizada pelo sistema criminal de Wisconsin, chamada COMPAS[185], destinada a avaliar o risco que os acusados possuíam de cometer novos crimes. Quando ele foi sentenciado, o juiz o condenou a seis anos de prisão, sob o argumento de que ele havia sido classificado pela ferramenta como alguém de "alto risco" à sociedade. Loomis contestou a sentença, sob o argumento de que o uso do algoritmo violava seus direitos ao devido processo legal, a ter uma decisão individualizada, e a ser sentenciado com base em informações exatas e corretas. Contudo, o tribunal julgou improcedente seu recurso, tendo o caso alcançado a Suprema Corte de Wisconsin (State *vs*. Loomis, 2016), que tampouco acolheu o pedido de Loomis, alegando que o resultado do algoritmo apresentava um nível suficiente de transparência.[186]

Os julgamentos evidenciados acima, em especial o caso criminal, são exemplos da importância em assegurar a interpretação e explicação de decisões algorítmicas quando direitos fundamentais são afetados (como o devido processo legal e a ausência de uma decisão judicial fundamentada), tal como no caso Loomis, ou quando impacta relações assimétricas, tal como ocorreu na Justiça do Trabalho (Frazão, 2021b). É necessário, portanto, considerar a importância de realizar "temperamentos" no direito ao segredo empresarial, quando este restringir direitos sociais e fundamentais (Frazão, 2021b). Caso isto não seja feito, tal direito será uma defesa absoluta, sem qualquer tipo de ponderação, ou repercussão quanto às exigências de transparência e prestação de contas e responsabilização dos agentes econômicos.[187]

Uma maneira de equilibrar estes interesses, de acordo com alguns autores (Pasquale, 2015; Scherer, 2016; Zarsky, 2013), seria disponibilizar o código para "auditores confiáveis", o que não implicaria em uma abertura indiscriminada do seu

[185] Ver nota de rodapé n. 20 para informações sobre o COMPAS.
[186] Para uma leitura mais detalhada sobre o julgamento, ver em: State *v*. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. 130 *Harvard Law Review* 1530, março de 2017. Disponível em: https://harvardlawreview.org/2017/03/state-v-loomis/. Acessado 24.06.2021.
[187] Sobre este assinto, Ana Frazão sustenta que este tipo de argumento acaba afetando até mesmo a questão probatória em demandas envolvendo a abertura de algoritmos: "Com efeito, considerando todos os riscos da opacidade algorítmica, o pior cenário é aquele em que se permitiria à plataforma se beneficiar do segredo de negócios, de forma ampla e absoluta, igualmente para efeitos probatórios. Caso as cortes trabalhistas assim entendam, chancelarão, ainda que indiretamente, a possibilidade de todos os tipos de fraudes e burlas à legislação trabalhista sob a conveniente justificativa do segredo de negócios" (Frazão, 2021b).

conteúdo ao público em geral. Esta delimitação do que e para quem essas informações podem ser apresentadas é denominada por Frank Pasquale como uma transparência qualificada (*qualified transparency).* Para o autor, esta medida respeitaria os interesses das partes envolvidas, ou seja, dos agentes econômicos e daqueles impactados pelas decisões dos algoritmos (Pasquale, 2015, p. 142).

Estes auditores confiáveis podem constituir-se sob diferentes formatos, podendo ser institucionalizados em um órgão público ou uma agência reguladora[188], ou constituídos sob organizações privadas com expertise sobre o tema. O exercício desta função pressupõe um dever de confidencialidade, e as informações disponibilizadas apenas podem ser utilizadas para a finalidade específica para a qual foi destinada a investigação do algoritmo de tomada de decisão de ML.[189]

No entanto, este tipo de transparência limitada, restrita e qualificada, pode afetar a confiança no funcionamento e nos resultados destes algoritmos por parte de um público mais amplo, aqui considerados não apenas as pessoas diretamente impactadas pelas decisões algorítmicas, mas a sociedade civil, instituições científicas, tecnológicas e de inovação, governantes, e representantes de diferentes segmentos econômicos, sociais e industriais. É necessário permitir um escrutínio para um público mais amplo.[190] Por tal motivo, defende-se que seja disponibilizado um documento, como um sumário executivo, contendo as principais conclusões e recomendações alcançadas nesta auditoria.

No mais, enquanto soluções de auditoria para investigar discriminações, e apresentar interpretações e explicações sobre os seus resultados ainda não fazem

[188] Mathew Scherer defende que seja criada uma agência reguladora dedicada a regular iniciativas de inteligência artificial, em que a mesma seria responsável por criar um procedimento de certificação para entidades que desenvolvem sistemas de IA. O intuito desta certificação seria pesquisar e desenvolver estes sistemas de maneira segura. Scherer traça um caminho alternativo para regular a IA, defendendo que seja feito um gerenciamento dos riscos públicos associados à IA, sem engessar a inovação, por meio de uma órgão regulador, ao invés de regular a tecnologia por normas emitidas pelo legislativo (Scherer, 2016, p. 393).

[189] Por exemplo, o artigo 206 da Lei de Propriedade Industrial (Lei n. 9.279/1996) indica que informações confidenciais, como aquelas protegidas pelo segredo comercial e industrial, que forem compartilhadas em juízo deverão ser mantidas em sigilo por meio da determinação do segredo de justiça, vedando-se expressamente que as informações sejam utilizadas pela parte contrária para outras finalidades.

[190] Tal como ocorreu no estudo produzido pelo ProPublica sobre os vieses existentes no software COMPAS, utilizado pode poder judiciário, para avaliar o potencial risco que um sujeito possui de cometer um crime, em que a predição afetava desproporcionalmente pessoas pretas do que em relação a pessoas brancas. Ver em: ANGWIN, Julia, LARSON, Jeff, MATTU, Surya; KIRCHNER, Lauren. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 de maio de 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acessado em 16.04.2021.

parte da prática jurídica, tal como demonstrado no caso do Uber, surge "a necessidade de se pensar em alternativas viáveis, exequíveis e que sejam capazes de conciliar os direitos envolvidos, tanto por parte das empresas, como parte dos terceiros que são afetados por suas decisões" (Frazão, 2021b). Isto envolve ponderar o direito ao segredo empresarial com outros direitos fundamentais e sociais, ou até mesmo as repercussões jurídicas em não abrir qualquer tipo de informações sobre seus algoritmos.[191]

Além da opacidade que se refere ao segredo empresarial do algoritmo, Burrell indica um segundo tipo, que se refere ao conhecimento técnico e especializado necessário para escrever, ler e desenvolver o código dos algoritmos de ML (Burrell, 2016, p. 4), assim como os modelos matemáticos aplicados ao mesmo. Tal conhecimento engloba o conjunto de regras que constitui o algoritmo, assim como a sua documentação, apontando para a importância das boas práticas para permitir que a tradução dos objetivos e interesses dos criadores possam ser compreendidos não apenas por máquinas, mas por pessoas. A autora destaca a importância da documentação da construção do algoritmo e do modelo matemático implementado, visto que esta é a maneira pela qual o desenvolvedor registra, indica e descreve as regras e os seus componentes para explicar como o código funciona. Contudo, apenas a documentação pode não ser suficiente quando se está diante de algoritmos que não são interpretáveis ou inteligíveis.

Uma alternativa, que possui um viés ético, para superar esta opacidade pode ser encontrada nos cartões modelo (*model cards - for model reporting)* (Mitchell et al., 2019), uma proposta de padronização de documentação de modelos de aprendizado de máquina apresentada por pesquisadores do Google. O *model card* é um formulário que reúne diversas informações sobre o modelo de ML, tais como: resultados do seu comportamento (*performance*), os casos para os quais o mesmo foi desenvolvido, o contexto que será aplicado e seus potenciais riscos, métricas para avaliar vieses e discriminações, resultados dos seus testes, quais foram os

[191] Sobre este ponto, Ana Frazão faz um paralelo interessante envolvendo o ônus probatório, com a decisão do STF que determinou que a paternidade do réu fosse relativamente presumida quando o mesmo se negasse a realizar o teste de paternidade. O tema é pacífico na jurisprudência, sendo previsto até mesmo na Súmula n. 301 do STJ ("Em ação investigatória, a recusa do suposto pai a submeter-se ao exame de DNA induz presunção juris tantum de paternidade"). A autora levanta questionamentos sobre a possibilidade de haver repercussões ao ônus probatório de empresas que se recusem a abrir informações sobre o seu algoritmo sob o argumento de proteger seu segredo empresarial, tal como presumir relativa da relação empregatícia, no caso do Uber (Frazão, 2021b).

dados que foram utilizados, entre outras informações. Esta proposta traz diversos benefícios para os desafios relacionados à explicação de modelos de ML, razão pela qual serão apresentados maiores detalhes sobre a mesma no subitem a seguir.

Por fim, o terceiro tipo de opacidade elencado por Burrell é a "complexidade inevitável" atribuída ao funcionamento e aos resultados gerados pelos algoritmos de ML (Burrell, 2016, p. 5), em que o aprendizado da máquina parece ser "não intuitivo, aleatório e desorganizado"[192] (Burrell, 2016, p. 7), e impossível de explicar o porquê das classificações atribuídas[193] (Burrell, 2016, p. 9). Esta complexidade aumenta proporcionalmente à diversidade de dados utilizados, e aos diferentes resultados que se busca alcançar. Dependendo do contexto que o algoritmo será aplicado, e o potencial impacto que ele terá, será necessário restringir ou limitar o uso de algoritmos de ML que não permitem explicações (Burrell, 2016, p. 9), especialmente quando a ausência de explicações possa causar riscos elevados para direitos fundamentais e liberdades.

Quando Jack Balkin e Ryan Calo se referem ao comportamento emergente, ou seja, à imprevisibilidade e inescrutabilidade da IA, os autores estão preocupados com a complexidade destes artefatos, e a dificuldade de constranger, limitar e prever os resultados de sua ação. Esta característica está fortemente associada a este terceiro tipo de opacidade, inerente à "complexidade inevitável" de certas aplicações de aprendizado de máquina (como é o caso do aprendizado profundo). Neste sentido, a proposta dos *model cards* apresentada acima pode servir, também, como uma ferramenta para auxiliar a desvendar os motivos que ocasionaram uma determinada classificação, pela verificação dos tipos de dados utilizados, os objetivos para o qual o algoritmo foi construído, o contexto no qual ele será aplicado, ou a exibição de resultados a partir da aplicação de métodos capazes de interpretar ou explicar estes resultados, por exemplo. Outras maneiras para tratar desta opacidade serão aprofundadas no subitem seguinte sobre prestação de contas e responsabilização.

Ademais, pode-se falar, ainda, em um quarto tipo de sigilo capaz de gerar opacidade, que se refere ao conhecimento ou não do sujeito de que seus dados pessoais estão sendo coletados, e que irão compor um conjunto de dados que será utilizado como insumo para o aprendizado de um algoritmo de tomada de decisão,

[192] Tradução livre de: "unintuitive, random and disorganized".
[193] Como era o caso apresentado pela autora, de classificação de spam.

ou, utilizado para a formação e classificação de um perfil destinado à tomada de uma decisão automatizada. Esta obrigatoriedade de informação e transparência do uso de dados pessoais, e a finalidade da sua coleta, é um dos princípios e deveres basilares de normas de proteção de dados pessoais, em especial, quando há o uso de decisões automatizadas, tal como previsto no próprio GDPR, e podendo ser feita uma interpretação semelhante para a LGPD, a partir dos art. 6º, incisos I e VI,[194] e art. 9º,[195] da Lei.

Percebe-se, portanto, que a opacidade possui diferentes camadas, com distintas naturezas, e por isso diferentes propostas para desvendar as razões de decidir dos algoritmos de aprendizado de máquina, e justificar a sua apresentação, serão necessárias.

Conforme já se manifestou o Conselho de Direitos Humanos da Organização das Nações Unidas ao tratar do uso de IA no contexto da moderação de conteúdo, em um documento da Relatoria Especial na promoção e proteção do direito à liberdade de opinião e de expressão, a transparência deve ocorrer ao longo do desenvolvimento do sistema de IA. Segundo o Conselho, dada a complexidade e as diferentes camadas, que funcionam como embreagens distribuídas e conectadas, compreender o resultado destes sistemas de tomada de decisão depende de uma visão holística, e não segmentada de determinado processo dentro desta cadeia (ONU, 2018, p. 17).

2.1.1
Categorizando a transparência em etapas, pessoas e instituições

Neste sentido, é importante identificar as diferentes etapas de desenvolvimento destes algoritmos, e o público que busca ter acesso às informações

[194] Art. 6º, incisos I e VI, da LGPD: "As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios: I - finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades; VI - transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial".

[195] Art. 9º, da LGPD: "O titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados, que deverão ser disponibilizadas de forma clara, adequada e ostensiva acerca de, entre outras características previstas em regulamentação para o atendimento do princípio do livre acesso: I - finalidade específica do tratamento; II - forma e duração do tratamento, observados os segredos comercial e industrial; III - identificação do controlador; IV - informações de contato do controlador; V - informações acerca do uso compartilhado de dados pelo controlador e a finalidade; VI - responsabilidades dos agentes que realizarão o tratamento; e VII - direitos do titular, com menção explícita aos direitos contidos no art. 18 desta Lei".

relacionadas ao processo de tomada de decisão do algoritmo. De acordo com Zarsky[196], é importante assegurar que a transparência ocorra nas etapas de: coleta e junção de dados, análise dos resultados e nas estratégias e políticas na escolha de modelos preditivos[197]. Isto implica, portanto, no reconhecimento de que há a necessidade de promover transparência ao longo de toda cadeia de valor do desenvolvimento de um sistema de IA (Comissão Europeia, 2019a).

Na etapa de coleta e junção dos dados, a transparência se referiria à apresentação dos tipos de dados que são utilizados para realizar a análise preditiva, bem como quais foram as decisões sobre como realizar a junção destes dados, quando estes tiverem origem de diferentes conjuntos de dados. Nesta etapa, pode ser exigido o acesso aos dados utilizados (Zarsky, 2013, pp. 1.523-1.524)[198]. Embora crítico e cético quanto a esta posição[199], Zarsky aponta que o acesso a esta informação interessaria ao público em geral, tendo em vista que os titulares dos dados pessoais coletados têm o direito de saber que os mesmos foram coletados e utilizados, e a finalidade para a qual foram usados. A necessidade de transparência estaria intrinsicamente relacionada a uma maneira de assegurar a proteção da privacidade, aos dados pessoais e à autodeterminação informativa, sob uma perspectiva de que os titulares de dados devem "conhecer, controlar, endereçar, interromper o fluxo das informações a ele relacionadas" (Rodotà, 2008, p. 92).

[196] Zarsky defende a transparência, mas também justifica a manutenção da opacidade dos algoritmos em determinadas situações. Tendo em vista que Zarsky tem como objeto a transparência de entes públicos, os argumentos que o autor apresenta para justificar a manutenção da opacidade seriam de ordem de segurança pública e interesse público. Por exemplo, o autor avalia que a transparência em relação aos critérios de formação de perfis de fraudadores ou potenciais criminosos podem motivar a adoção de uma postura que tente enganar estes sistemas, ameaçando o próprio objetivo do governo ao implementar estes sistemas.

[197] Embora o autor esteja se referindo especificamente à tarefa de mineração de dados, este mesmo argumento pode ser aplicado ao uso de algoritmos de ML em geral.

[198] No Brasil, por exemplo, este pedido de acesso aos dados já é algo que encontra previsão legal em diferentes normas, como no art. 5º, II da Lei de Cadastro Positivo (Lei 12.414/2011), no art. 43, do Código de Defesa do Consumidor (Lei 8.087/1990 – CDC), e os arts. 6º, IV e 19, da LGPD, o habeas data, regulado pela Lei n. 9.507/1997, com previsão constitucional expressa (art. 5º, LXXII, da CF).

[199] Zarsky, no entanto, é cético quanto à construção teórica de que os titulares de dados seriam capazes de controlar o fluxo dos seus dados pessoais, pois alega que a premissa desta afirmação é que as pessoas valorizam a sua autonomia, e isto não ocorre no atual cenário tecnológico, pois o mercado não permite, de fato, que este controle seja exercido. Sustenta, ainda, que mesmo que fosse possível ter informações sobre o momento em que há a coleta e os posteriores usos dos dados pessoais, análises e usos subsequentes não seriam passíveis de serem controlados (Zarsky, 2013, pp. 1543-1545). No entanto, algumas normas de proteção de dados pessoais, como o GDPR, e a LGPD, buscam justamente regular este uso indiscriminado e ilimitado dos dados pessoais. A LGPD prevê princípios que cumprem com esta finalidade, tal como os princípios da finalidade, adequação e necessidade.

Ademais, na etapa da análise dos dados, a transparência estaria relacionada à abertura de informações sobre perspectivas técnicas e humanas. Na primeira, a transparência exigiria a disponibilização de informações sobre os programas de *software* utilizados na construção destes algoritmos, bem como se os mesmos foram desenvolvidos internamente ou se apoiados em sistemas externos. Sob uma perspectiva humana, a abertura exigiria a apresentação de dados sobre a confiança e a relevância do resultado estatístico da predição apresentada, como, por exemplo, na identificação de falsos positivos e negativos e como estes foram enfrentados, quais foram os resultados que se basearam em relações de correlação, e não de causalidade (Zarsky, 2013, p. 1.526) [200] . Esta transparência interessaria às instituições, no caso analisado pelo autor, ao Poder Público, para avaliar a qualidade e confiabilidade dos resultados.

Por fim, a transparência na etapa que diz respeito à finalidade para a qual os dados serão utilizados interessa àqueles impactados diretamente pelas decisões (Zarksy, 2013, p. 1.526). A transparência exigida nesta etapa não se refere à mera disponibilização de informações sobre o sistema em abstrato, mas informações sobre os resultados que surgem a partir do seu uso, ou seja, um resultado concreto e *ex post*.[201] Isto porque apenas desta forma será possível identificar os elementos que ocasionaram um determinado resultado, não sendo suficiente uma análise *ex ante* sobre a arquitetura e funcionalidade do sistema.

Em relação à transparência exigida pelo público em geral e pelos sujeitos impactados pelas decisões, esta se justifica por ser uma forma de garantir a proteção à privacidade, aos dados pessoais, à liberdade e à autonomia dos sujeitos, tanto os impactados diretamente, quanto os que potencialmente podem ser impactados, visto que toda a coletividade está sujeita ao tratamento de dados pessoais, conforme indicado no item que trata sobre a coletivização dos dados pessoais. Por isso, o acesso a estas informações deveria ser um direito estendido à coletividade. A ausência de transparência pode limitar o exercício da liberdade, e ocasionar a estigmatização social dos sujeitos, posto que impede a possibilidade de correção de

[200] O autor defende que a dignidade e a autonomia do sujeito apenas serão alcançadas quando os processos algorítmicos alcançarem conclusões causais, e não correlacionais.
[201] Esta discussão se assemelha ao debate iniciado por Wachter et al (2017) sobre o momento em que uma explicação sobre uma decisão automatizada deveria ser apresentada, se antes de uma decisão ter sido de fato tomada (*ex ante*) ou após a mesma ter sido apresentada (*ex post*).

eventuais inacurácias, categorizações errôneas ou inexatidão dos dados que ocasionaram tais erros (Rodotà, 2008, p. 92).

Aqueles impactados pelas decisões têm o direito de saber por que foram afetados, sob pena de restrição à sua liberdade e autonomia, e devem "receber uma explicação sobre o critério da decisão e da lógica por trás dessas ações"[202]. Isto implica em alguma forma de transparência, e o conhecimento sobre esta informação empodera o indivíduo (Zarsky, 2013, p. 1.545). Até mesmo as informações sobre as análises dos dados devem interessar ao público em geral como uma forma de garantir o letramento técnico e gerar uma confiança do público em relação aos seus resultados. Assim, uma maneira de viabilizar a transparência neste caso seria garantir uma espécie de devido processo legal[203], que permitiria que a pessoa impactada pela decisão tenha conhecimento, primeiro, de que ela está sujeita a um processo de análise algorítmica e, segundo, seja capaz de avaliar a discricionariedade deste processo, com a possibilidade de se opor sobre eventuais dados inexatos ou incorretos (Zarsky, 2013, p. 1.547-1548).

No mais, o respeito à autonomia e à dignidade apenas é possível quando a transparência ocorre de maneira que as informações disponibilizadas sejam passíveis de serem interpretáveis e explicáveis para a pessoa (Zarsky, 2013, p. 1.548)[204]. Neste sentido, o aspecto mais relevante para a transparência é a capacidade de interpretação dos processos algorítmicos (Zarsky, 2013, p. 1.566), mesmo que esta interpretação dependa do auxílio de outros processos.

Nesta perspectiva, o documento produzido pela Relatoria Especial de liberdade de expressão do Conselho de Direitos Humanos da ONU joga luz principalmente para as explicações que devem ser oferecidas para o público em geral, defendendo que a "transparência não precisa ser complexa para ser efetiva"[205] (ONU, 2018, p. 17). As explicações devem buscar contribuir para educar os sujeitos

[202] Tradução livre de: "They should receive an explanation as to the decision criteria and to the logic behind these actions".
[203] Zarsky pondera se o devido processo legal previsto na Quarta Emenda da Constituição poderia ser aplicado a processos algorítmicos. Inicialmente, o autor discorda de que isto seria possível, sustentando que a Quarta Emenda visa proteger danos causados à vida, liberdade e propriedade, e não à autonomia. Contudo, aponta que a previsão constitucional de devido processo legal poderia servir como um guia para legislações futuras. Para balancear esta opinião, são apresentados trabalhos de outros autores que defendem a possibilidade de se aplicar a referida disposição constitucional em casos em que o Governo estivesse se valendo de tais processos algorítmicos.
[204] Isto implica que "To assure dignity and lack of targeting, the individual should receive assurances as to the precision, effectiveness, and lack of discrimination in the process" (Zarsky, 2013, p. 1.548)
[205] Tradução livre de: "Transparency need not be complex to be effective"

sobre sistemas que se valem de IA, e não necessariamente com detalhes técnicos – que não podem ser descartados, mas exigidos em fóruns específicos, como, em perícias, por exemplo –, mas *insights* não técnicos como: a existência do sistema, seu propósito e seu impacto (ONU, 2018, p. 17). De acordo com o referido Relatório, a transparência radical de sistemas de IA, no contexto de moderação de conteúdo, que é o objeto do documento, envolveria apresentar informações sobre: a quantidade de conteúdo que é removido, a frequência com que este conteúdo é removido e contestado, e quando esta contestação é aceita, exemplos dos motivos que levaram um conteúdo ser priorizado sobre o outro, e limitações dos sistemas de IA (falhas, erros e limitações sobre o seu uso).

Os diferentes tipos de opacidade estruturam os desafios que os processos algorítmicos de aprendizado de máquina apresentam, quais são os direitos afetados pelo sigilo e pela opacidade, e dão indicativos de quais e como as informações sobre tais sistemas deverão ser apresentadas. O conteúdo, assim como a forma com que as informações envolvendo os sistemas e os resultados de sistemas de IA são apresentadas, são o cerne da questão em demandas envolvendo transparência.

2.1.2
A análise judicial do uso de um algoritmo de tomada de decisão na União Europa

A fim de ilustrar um exemplo sobre a importância do tema da transparência, em fevereiro de 2020, foi proferida a primeira decisão judicial no âmbito europeu que analisou a legitimidade e legalidade do uso de um algoritmo pelo Poder Público. Esta decisão é interessante, pois ela joga luz sobre as nuances das informações que podem ser exigidas sob o princípio da transparência, a fim de proteger os dados pessoais e a privacidade de cidadãos. A decisão[206], proferida pelo Tribunal Distrital de Haia, na Holanda, analisava a lei *SUWI Act*, que implementava o Sistema de Indicação de Risco (*System for Risk Indication* -- "SyRI"), um instrumento utilizado pelo Governo holandês para identificar diferentes tipos de fraudes relacionadas a

[206] HOLANDA. Corte Distrital de Haia (Rb. Den Haag). Processo n. C/09/550982/HA ZA 18/388. Julgado e publicado em 02.05.2020. Disponível em: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865&showbutton=true&keyword=AVG. Acessado em 12.02.2022.

benefícios de seguridade social (como de segurança e renda), trabalhistas e de impostos.

O SyRI foi criado pelo Ministério de Assuntos Sociais na Holanda, em 2014[207], para identificar pessoas que eram consideradas como tendo um alto risco de estarem fraudando os benefícios sociais oferecidos pelo governo. O instrumento, regulado em lei própria, previa o uso de dezessete categorias de dados coletados e armazenados por diferentes órgãos do poder público[208], tal como, dados de impostos, de registros de propriedade, registro de veículos, dados educacionais, entre outros, para aplicar um modelo algorítmico para formação de perfis e atribuição de pontuação a cada cidadão em relação ao seu potencial de estar cometendo algum tipo de fraude. Esta investigação conduzida pelo sistema ocorria independentemente de haver uma conduta suspeita do sujeito.[209]

O que motivou o ajuizamento da ação foi uma grande insatisfação manifestada por organizações locais e organizações da sociedade civil, que demonstraram que o monitoramento e a investigação realizado pelo SyRI ocorria em quatro cidades holandesas, exclusivamente em regiões carentes, que possuíam uma população majoritariamente de imigrantes[210]. Formou-se uma coalizão de entidades da sociedade civil[211] ("NJCM et al.") que, em conjunto com outros dois

[207] Como indicado na manifestação do Relator Especial das Organizações das Nações Unidas (ONU), Professor Philip Alston, na qualidade de amicus curiae, o SyRI tem origem na cooperação já existente entre diferentes órgãos do governo holandês para combater e investigar estas fraudes. Dentro desta, e outras, cooperações, iniciou-se uma intensa troca e análise de dados entre as autoridades, sob o programa chamado de 'Landelijike Stuurgroep Interventiens' (LSI). Os métodos utilizados no LSI pavimentaram o caminho para a criação do SyRI. Ver em: NAÇÕES UNIDAS. Landmark ruling by Dutch court stops government attempts to spy on the poor – UN expert. Escritório do Alto Comissariado das Nações Unidas para os Direitos Humanos, 05 de fev. de 2020. Disponível em: shorturl.at/arxFN. Acessado em 13.02.2021.

[208] Em relação a pessoas físicas, eram coletados o seguintes dados: trabalhistas; sanções ou medidas administrativas; imposto de renda; de registro de bens móveis e imóveis; exclusão de benefícios de assistência social ou outros; residenciais; de identificação (nome, endereço, data de nascimento, gênero, e características administrativas); integração cívica, compliance regulatório; educacionais; pensão; reintegração; dívidas; benefícios sociais pensões ou subsídios; permissões e isenções; e plano de saúde.

[209] Ver em: NAÇÕES UNIDAS. Landmark ruling by Dutch court stops government attempts to spy on the poor – UN expert. Escritório do Alto Comissariado das Nações Unidas para os Direitos Humanos, 05 de fev. de 2020. Disponível em: shorturl.at/arxFN. Acessado em 13.02.2021.

[210] Ver em: SIMONITE, Tom. Europe Limits Government by Algorithm. The US, Not So Much. Wired, 02 de julho de 2020. Disponível em: https://www.wired.com/story/europe-limits-government-algorithm-us-not-much/. Acessado em 13.02.2021.

[211] A coalizão é representada pela Seção Holandesa da Comissão Internacional de Juristas (Nederlands Juristen Comité Voor De Mensenrechten - NJCM), o Conselho Nacional de Participação do Cliente (Landelijke Cliëntenraad), e a Confederação Sindical da Holanda (Federatie Nederlandse Vakbeweging - FNV).

autores[212], ajuizaram em 2018 a ação em referência em face do Estado Holandês, alegando que o uso do SyRI violava as normas regionais e internacionais de direitos humanos[213].

Conforme relatado pelo Tribunal Distrital de Haia, o tratamento dos dados pessoais pelo SyRI ocorria em duas etapas. A primeira cuidava do processamento de dados (como o nome da pessoa física ou jurídica, número de identificação social, equivalente ao CPF e CNPJ no Brasil, e o respectivo endereço), eram pseudonomizados[214]. A segunda etapa consistia na geração de relatórios de risco pelo algoritmo a partir destes dados, havendo a re-identificação das pessoas em caso de pontuação elevada. Este relatório servia como subsídio e indicativo de que uma pessoa deveria ser investigada por potencial fraude ou desrespeito à legislação sobre o tema, razão pela qual tais relatórios eram enviados e avaliados pela unidade competente (no caso concreto, a unidade de Inspeção dos Assuntos Sociais e de Emprego). Ou seja, ocorria um processamento automatizado, sem interferência humana, das diferentes categorias de dados inseridos no sistema destinado à formação de perfis, atribuição de pontuação, geração de relatórios e, em seguida, um humano avaliava o relatório, e decidia se uma investigação deveria seguir adiante ou não.

O tribunal, ao analisar a demanda, concluiu que o SyRI violava o art. 8º, parágrafo 2º, da Convenção Europeia dos Direitos do Homem (CEDH). Em sua fundamentação, o tribunal se dedicou a analisar se a legislação do SyRI atendia os requisitos apresentados pelo art. 8, parágrafo 2º, da CEDH, indicando que a sua interpretação seria baseada nos princípios da Carta de Direitos Fundamentais da UE (arts. 7º e 8º), bem como no GDPR (em especial, a decisão levou em consideração os princípios transparência, da finalidade (*purpose limitation*) e minimização dos dados (*data minimisation*)).

[212] Os autores foram considerados como partes ilegítimas para figurarem na demanda.

[213] Segundo NJCM et al., o uso do SyRI constituiria uma violação aos direitos humanos, especialmente ao art. 8º da Convenção Europeia dos Direitos do Homem (CEDH), arts. 7º e 8º da Carta de Direitos Fundamentais da União Europeia (CDFUE), o art. 17 do Pacto Internacional dos Direitos Civis e Políticos, que tratam da proteção à vida privada e dos dados pessoais, e aos arts. 5, 6, 13, 14 e 22 GDPR.

[214] De acordo com Diego Machado e Danilo Doneda, a técnica de pseudonimização "opera de maneira que as informações não podem ser conectadas a um titular de dados específico sem que se recorra a informações suplementares, desde que estas sejam mantidas separadamente, empregadas medidas organizativas e de segurança." (Machado; Doneda, 2018, p. 112).

O tribunal indicou que qualquer restrição à vida privada deveria atender aos critérios indicados no parágrafo 2º do art. 8º da CDFUE: estar prevista em lei, ser necessária, adequada e proporcional. No caso analisado, verificou-se que não foram adotados requisitos necessários que justificassem uma interferência no direito à vida privada dos cidadãos, visto que os Estados têm uma responsabilidade especial quando se valem de novas tecnologias para monitorar seus cidadãos, pela demasiada interferência que estas causam na vida privada deles.[215] Um ponto interessante da decisão foi a reiterada manifestação de que a ausência de transparência em relação aos parâmetros do sistema inviabilizava (i) a análise sobre a necessidade e proporcionalidade da interferência na vida privada de uma pessoa, e (ii) se o seu uso estaria justificado para atender aos objetivos que a lei visava alcançar.

Assim, considerou-se que não houve transparência por parte do Estado, até mesmo porque a legislação do SyRI não exigia que fossem fornecidas quaisquer informações sobre o funcionamento de tal sistema, do algoritmo utilizado ou dos métodos de análise de risco. Consequentemente, não seria possível realizar a validação da conformidade, adequação, necessidade e proporcionalidade do modelo de risco adotado e de seus indicadores. Embora o tribunal tenha reconhecido que o modelo havia sido validado pelos órgãos internos do governo, os critérios utilizados para esta validação e os indicadores utilizados para atribuir a pontuação do risco não estavam disponíveis para verificação. Assim, o tribunal apontou que isto inviabilizaria a defesa do cidadão em relação a um relatório produzido sobre ele, bem como eventuais correções e retificações de seus dados.

O tribunal ressaltou, ainda, que a transparência também era fundamental para verificar se houve ou não discriminação. Inclusive, este era um dos principais argumentos do NJCM, de que o SyRI apenas teria sido utilizado em áreas estigmatizadas, consideradas como "problemáticas", o que reforçaria esta visão negativa do uso do sistema com potencial de gerar discriminação. Como não seria possível verificar os indicadores e o modelo de risco do algoritmo, tampouco seria possível avaliar o risco de vieses estarem sendo mitigados ou neutralizados. Assim,

[215] Como indicado pelo Tribunal Distrital de Haia, este é o entendimento do TEDH, que no caso S. E Marper vs. Reino Unido, foi considerado que: "The Court considers that any State claiming a pioneer role in the development of new technologies bears special responsibility for striking the right balance in this regard".

concluiu-se que a ferramenta não foi considerada transparente e verificável, além de não apresentar salvaguardas suficientes para justificar uma interferência na vida privada dos sujeitos monitorados.

Este caso é interessante para a discussão da transparência, pois nele é debatida a qualidade da transparência, ou seja, sobre o tipo de conteúdo que o tribunal destacou que seria adequado que fosse disponibilizado, e como a ausência destas informações afetava a privacidade, a proteção de dados, a autonomia, a liberdade e a igualdade dos sujeitos. Sob uma perspectiva subjetiva de quem recebe a informação, foi destacado que a transparência e a adequação do sistema em relação aos direitos humanos e outras legislações apenas pode ser verificada se houver a devida prestação de contas (*accountability*) a terceiros.

Desta forma, soluções técnicas e jurídicas serão necessárias para viabilizar esta demanda por transparência, sendo esta necessária para que haja a verificação e, consequentemente, a prestação de contas de como estes agentes inteligentes foram implementados e funcionam, conforme se verá a seguir.

2.2
O princípio da prestação de contas e responsabilização (*accountability):*

> "A prestação de contas (...) está ligada à noção de responsabilidade, justeza e processo devido no uso de algoritmos"[216]

Diversos sentidos têm sido atribuídos ao termo *accountability*, com o intuito de alcançar diferentes finalidades: responsabilização, prestação de contas, fiscalização ou sanção, *answerability* (Augusto; Rizzardi, 2014). A origem do termo remete ao contexto de direito público, e a exigência de prestação de contas dos Entes públicos e de fiscalização de seus atos, seja pelo próprio Estado e/ou por seus cidadãos (Augusto; Rizzardi, 2014). Contudo, atualmente, tal conceito vem sendo exigido e aplicado a atos e práticas exercidos por entes privados. Especialmente na temática envolvendo regulação da inteligência artificial, o termo *accountability* se coloca como uma demanda recorrente, embora seu conceito também nesta seara possua certa polissemia.

[216] Doneda; Almeida, 2018, p. 146.

Por exemplo, a própria transparência tem sido compreendida como uma forma de promover esta prestação de contas e responsabilização (Zarsky, 2013, p. 1.532; Ananny; Crawford, 2018), e muitas vezes estes dois princípios são tratados como sinônimos, embora autores prefiram marcar esta diferença (Zarsky, 2013; Kaminski, 2019b). Diakopoulos, por exemplo, indica que a "transparência pode ser um mecanismo que facilita a *accountability*, um que devemos exigir do governo e encorajar a indústria"[217] (Diakopoulos, 2016). A primeira, muitas vezes, é apontada como uma mera forma de estruturar o fluxo de informações para diferentes públicos, enquanto a *accountability* pode assumir diferentes formas, para atender a objetivos de substância e procedimento, e envolve fiscalização e monitoramento, análise de experts, além de conhecimento e envolvimento do público (Kaminski, 2019b, p. 1.566-1.567). A transparência é necessária, mas não suficiente, uma vez que são necessários outros mecanismos de avaliar a prestação de contas e responsabilização de um algoritmo de tomada de decisão.[218]

Especialmente no que se refere à explicação de decisões algorítmicas, a exigência de transparência pode ser diferenciada de uma exigência por explicação. Enquanto a primeira exige a abertura de certas informações que correm em um sistema de IA, a última busca explicar como certos fatores influenciaram no resultado de um algoritmo (Doshi-Velez; Kortz, 2017, p. 6). Ao final, ambas são maneiras de viabilizar uma prestação de contas e responsabilização daqueles que desenvolveram e utilizam o algoritmo.

A prestação de contas e responsabilização no contexto dos algoritmos de tomada de decisão de ML, está relacionada (i) à responsabilidade assumida pelos agentes econômicos dos resultados gerados, tanto em relação a questões éticas (Diakopoulos, 2016), quanto jurídicas (Citron; Pasquale, 2014; Pasquale, 2015), e (ii) a apresentação de evidências verificáveis que demonstrem os cuidados adotados para evitar resultados danosos, permitindo formas de (quando possível) compreender como, em que medida e porquê um certo comportamento ocorreu, além de quem seria responsável por ele (Desai; Kroll, 2017, p. 10).

[217] Tradução livre de: "Transparency can be a mechanism that facilitates accountability, one that we should demand from government and exhort from industry".
[218] Neste sentido, Ananny e Crawford (2018, p. 984) indicam que: "If we recognize that transparency alone cannot create accountable systems and engaging with the reasons behind this limitation, we may be able to use the limits of transparency as conceptual tools for understanding how algorithmic assemblages might be held accountable."

A completa e total abertura do código fonte dos algoritmos para atender a esta finalidade é fortemente desestimulada (Diakopoulos; 2016; Desai; Kroll, 2017; Zarsky, 2013; Ananny; Crawford, 2018; Kroll et al., 2017). Por isso, formas de apresentar informações consideradas chaves, tais como resultados agregados, testes implementados, e análises de potencial impacto do uso de algoritmos, seriam muito mais eficazes na comunicação ao público em geral (Diakopoulos; 2016), bem como para os reguladores.

Como indicado pela decisão do Tribunal de Haia, a análise da legitimidade e licitude do uso do algoritmo de tomada de decisão por parte do Estado dependia de uma verificação externa, de terceiros, e não apenas aqueles responsáveis pela sua aplicação. Isto aponta para uma necessidade de transparência e prestação de contas destinadas a apresentar evidências sobre as medidas adotadas para, naquele caso, mitigar o impacto à proteção de dados pessoais e à privacidade das pessoas sujeitas a serem monitoradas pelo algoritmo.

Ademais, a prestação de contas é relevante não apenas para aqueles impactados pela decisão, mas é também uma forma de regular e estabelecer sistemas de IA que sejam legítimos, justos e que funcionem adequadamente (Kaminski, 2019b). Embora em um primeiro momento a discussão sobre o direito à explicação no âmbito do GDPR tenha caminhado para discutir qual seria o tipo de explicação que poderia ser exigido, atualmente, há uma discussão mais ampla sobre *accountability* algorítmica. O direito à explicação tem se colocado como uma das maneiras de implementar esta prestação de contas, (Doshi-Velez; Mason, 2017, p. 2),[219] contribuindo para questões que transcendem a discussão sobre proteção de dados, tal como pontuação de crédito, na moderação de conteúdo das plataformas de aplicação, e no debate sobre o desenvolvimento de uma IA ética. Percebe-se, portanto, que o direito à explicação ganha importância como uma maneira de assegurar formas mais amplas de prestação de contas e responsabilização, em relação aos aspectos da "explicabilidade" e "interpretabilidade" da IA, em especial, de algoritmos de tomada de decisão de ML (Frajhof, 2021, p. 470).

[219] Como indicado por Doshi-Velez e Mason: "While there are many tools to increasing accountability in AI systems, we shall focus on one in this report: explanation. (We briefly discuss alternatives in Section 7.) By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute" (Doshi-Velez; Mason, 2017, p. 2).

Neste sentido, Margot Kaminski (2019b), em um artigo fundamental, realizou um mapeamento da literatura sobre as dificuldades que os algoritmos de tomada de decisão de ML apresentam para o Direito, e as soluções jurídicas que devem ser implementadas para proteger a dignidade, a autonomia, a liberdade das pessoas impactadas ou que potencialmente sejam impactadas pelas decisões algorítmicas, e maneiras de verificar a legitimidade e legalidade de tais sistemas. A autora organiza as principais motivações e propostas que têm surgido pleiteando por uma regulação dos algoritmos de tomada de decisão.

A autora identifica três demandas que justificam a regulação de algoritmos de tomada de decisão. São elas: proteger a dignidade da pessoa afetada pela decisão algorítmica (dignidade)[220], acessar as razões e os fundamentos desta decisão para verificar sua legitimidade e legalidade (justificativa), e evitar resultados errados, enviesados ou discriminatórios do sistema como um todo (instrumental). As demandas que envolvem a dignidade e justificativa estão fundamentadas na proteção de direitos individuais das pessoas afetadas pelas decisões, e a instrumental em uma motivação de interesse público.

Estas diferentes demandas, e fundamentos, formam o que Kaminski chama de governança binária dos algoritmos de tomada de decisão, constituída de dois regimes diferentes: (1) um sistema que protege os direitos individuais, e pleiteia por medidas que se assemelhem a um devido processo legal, nomeado como um devido processo algorítmico como forma de proteção (dignidade e justificativa), e (2) uma regulação sistêmica, chamada de governança colaborativa[221], entre o poder público e agentes privados, exigindo transparência sobre o processo de desenvolvimento

[220] A autora traz diferentes perspectivas e abordagens da compreensão do que seria o conceito de dignidade da pessoa humana, o relacionando com a proteção assegurado ao sujeito de não ser tratado como um objeto, como uma forma de tratamento igualitário, afastando tratamentos discriminatórios e abusivos, em favor da liberdade e autonomia (Kaminski, 2019, pp. 1.541-1.545).

[221] Nas palavras da autora, o conceito de governança colaborativa pode ser compreendido da seguinte maneira: "deploys private- public partnerships towards public governance goals. Collaborative governance should not be confused with self-regulation, though it may include or even rely in substantial part on private governance. In its ideal form, collaborative governance is not hands-off or deregulatory. It exists on a spectrum between traditional command-and-control regulation and private ordering, and may employ significant aspects of each (….). Collaborative governance represents a hybrid or, in the case of "responsive regulation," an escalating approach. It may include but is not limited to formal coregulation through the adoption of codes of conduct or certification mechanisms Collaborative governance is, at best, a highly tailored, site-calibrated regulatory system that aims to pull inputs from, obtain buy-in from, and affect the internal institutional structures and decision-making heuristics of the private sector, while maintaining the legitimacy, efficacy, and public-interest orientation of public sector governance (Kaminski, 2019b, pp. 1.559 – 1.560 removidas as referências originais do texto).

dos algoritmos de tomada de decisão (instrumental), e que também se preocupa com a legitimidade dos resultados de tais sistemas (justificativa) (Kaminski, 2019b, p. 1529-1531). Enquanto o primeiro regime estaria preocupado em formas de assegurar uma prestação de contas de uma decisão concreta (*ex post*), o segundo regime tem como foco principal medidas de prestação de contas que buscam avaliar o processo de criação e desenvolvimento de algoritmos (*ex ante*), além das pessoas envolvidas neste processo (Kaminski, 2019b, p. 1.557).

Este conceito de governança binária, composto por estes dois regimes, é uma maneira de viabilizar o direito à explicação, como apontado pela própria autora, que acredita que o GDPR oferece previsões que servem de exemplo para este modelo. O que é interessante nesta proposta, para fins deste capítulo, é a construção da autora da possibilidade de que sejam utilizados instrumentos organizacionais e corporativos para registro e documentação do processo de tomada de decisão na elaboração destes algoritmos de tomada de decisão. Considerando a opacidade inerente a alguns resultados de algoritmos de ML, exigir outros tipos de evidência, em momentos que precedem o desenvolvimento e implementação dos mesmos, torna-se necessário como uma forma de permitir a inteligibilidade, compreensão e avaliação da legitimidade e legalidade de tais sistemas e dos seus resultados. As motivações, objetivos, finalidades, intenções, e decisões tomadas ao longo do momento de desenvolvimento, externadas e documentadas, são provas importantes para compreender estes algoritmos de ML.

Assim, partindo da proposta de governança binária de Kaminski, propõe-se que a prestação de contas de algoritmos de tomada de decisão de ML, voltada para o direito à explicação, deva ocorrer em dois momentos, chamando-os de *ex ante* e de *ex post* (ou *post hoc*). O primeiro momento (*ex ante*) se refere à apresentação de informações estáticas relacionadas ao algoritmo de ML e o seu processo de desenvolvimento, tais como informações sobre o modelo em si, quais foram os dados utilizados, a documentação em abstrato, e outras informações consideradas relevantes sobre o processo de criação e desenvolvimento do algoritmo[222]. A

[222] Neste sentido: "That is, although we cannot guarantee that we can always analyze programs for correctness, we can guarantee that properly designed programs are correct. This supports arguments for capturing important values in software at the point of de- sign and construction" (Desai; Kroll, 2017, p. 35).

segunda (*ex post*) se relaciona ao comportamento e dinamismo do funcionamento de fato do algoritmo, e se refere aos resultados do seu comportamento em concreto.

No que se refere a este momento *ex ante* uma série de mecanismos e instrumentos se colocam à disposição, tais como: o Relatório de Impacto (de Proteção de Dados Pessoais ou específicos para sistemas de IA), Códigos de Boas Práticas e Códigos de Conduta. Por sua vez, uma análise de uma decisão *post hoc* (ou *ex post*) do algoritmo, dependerá da produção de: documentação, auditoria, e métodos de interpretação e explicação dos comportamentos e resultados de algoritmos (Frajhof, 2021, p. 470).

A perspectiva *ex ante* estaria relacionada com a ideia de governança colaborativa e instrumental, que analisa o processo de elaboração do sistema, avaliações de impacto e de risco do seu uso e medidas adotadas para mitigá-los, enquanto a visão *ex post* se refere especificamente a maneiras de viabilizar explicações sobre uma decisão algorítmica. Esta proposta difere do que fora proposto por Kaminski[223], posto que a autora visa que a governança binária seja uma alternativa de um desenho regulatório para algoritmos de tomada de decisão. Por sua vez, este trabalho parte do regime binário proposto por ela para organizar, justificar, e detalhar formas de viabilizar interpretações e explicações de resultados algorítmicos, além de salvaguardas e documentos que devem ser elaborados. O debate mais aprofundado e específico sobre o conteúdo e forma de exercer o direito à explicação sob a perspectiva de direitos individuais (dignidade e justificativa), será abordado no próximo capítulo.

No mais, as dificuldades e limitações em "abrir a caixa-preta" diante das diferentes opacidades envolvendo os algoritmos de ML implicam em ter de assumir frentes de naturezas diferentes. Isto envolve levar em consideração qual é o conteúdo das informações que devem (ou podem) ser apresentadas, a forma como estas informações são disponibilizadas, e quais interesses, e de quem, a transparência visa proteger. Assim, para realizar a prestação de contas, é necessário investigar os algoritmos de tomada de decisão de ML, e isto deve ocorrer por meio

[223] A autora, inclusive, propõe outros instrumentos para assegurar tais medidas, como implementação de conselho especializado dentro de companhias (*expert oversight board*), ou táticas de *whistleblowers,* funcionários que revelam informações das atividades de empresas públicas ou privadas que até então eram mantidas privadas. Contudo, este trabalho optou por não incluí-las, para se dedicar em mais detalhes a outras propostas, principalmente aquelas que já possuem previsão normativa na LGPD.

de evidências verificáveis, de diversas maneiras. A seguir, serão analisados cada um destes mecanismos, as salvaguardas e ferramentas de prestação de contas e responsabilização do direito à explicação.

2.2.1
Prestação de contas e responsabilização *ex ante*:

2.2.1.a Relatório de Impacto

Algumas legislações têm exigido a elaboração de relatórios de impacto que antecedem o início de certas atividades, pelo risco de a mesma causar danos a determinados bens jurídicos protegidos pelo ordenamento. São exemplos o Estudo Prévio de Impacto Ambiental e seu respectivo relatório (EIA/RIMA)[224], exigido quando uma atividade seja potencialmente causadora de significativa degradação ambiental (Moreira; Oliveira, 2019), ou o Relatório de Impacto à Proteção de Dados Pessoais (RIPDP) previsto na LGPD, solicitado quando o tratamento de dados pessoais gerar riscos às liberdades civis e aos direitos fundamentais dos titulares de dados. Até mesmo no contexto de direitos humanos já houve manifestação da ONU recomendando [225] que, além do próprio poder público, a iniciativa privada elabore um relatório de impacto sempre que determinada política ou atividade carreada por tais entidades possam impactar direitos humanos.

No contexto da inteligência artificial, tem sido defendida a importância de se elaborar um relatório de impacto quando houver o desenvolvimento e aplicação de sistemas de IA[226]. O AI Now Institute, por exemplo, desenvolveu o *Algorithm Impact Assessment* (AIA) para orientar o poder público na aquisição ou no desenvolvimento de algoritmos de tomada de decisão (Reisman et al., 2018). Selbst

[224] Nos termos do art. 225 § 1º, IV, da CF, em que a Resolução CONAMA 001/1986, recepcionada pela Constituição Federal de 1988, estabelece as normas para elaboração do Estudo Prévio de Impacto Ambiental e o respectivo Relatório de Impacto Ambiental (EIA/RIMA) (Moreira; Oliveira, 2019).

[225] Ver em: NAÇÕES UNIDAS. *Guiding Principles on Business and Human Rights:* Implementing the United Nations "Protect, Respect and Remedy' Framework, 20-24, 2011. Disponível em: http://www.ohchr.org/Documents/Publications/ GuidingPrinciplesBusinessHR_EN.pdf. Acessado em 01.05.2021.

[226] Em seu modelo de governança colaborativa dos algoritmos de tomada de decisão, Kaminski aponta que relatórios de impacto podem ser elaborados para que entidades, e aqui a autora se refere especialmente a entes privados: "to both make substantive commitments about their systems and create impact assessments before their deployment. We could require those impact assessments to clearly describe the system, detail decisions around its design, address risk-mitigation measures, establish potential benchmarks, and detail considered but rejected alternatives (…)" (Kaminski, 2019, p. 1.574)

(2017) propõe a elaboração do que denomina como *algorithmic impact statements*, quando algoritmos preditivos forem considerados pela polícia para serem aplicados, que demanda a avaliação do impacto da tecnologia e as alternativas existentes ao seu uso. A organização de direitos humanos, Article 19, propõe a realização de um relatório de impacto de direitos humanos quando estes estiverem ameaçados pelo desenvolvimento e uso de sistemas de IA (Article 19, 2020).

Dois comentários gerais sobre estas propostas merecem ser destacados. O primeiro é que, tendo em vista que sistemas de IA que aplicam ML evoluem com o seu comportamento conforme aprendem com os dados que analisam, necessariamente estas avaliações devem ocorrer de maneira continua. O segundo é uma exigência de que a sua elaboração seja interativa, e que documento possa estar sujeito para consulta pública, para que diferentes *stakeholders,* como reguladores, pesquisadores, sociedade civil, representantes de classe, etc, possam contribuir e apresentar sugestões ao documento. Esta publicidade e participação social atendem ao que Kaminski chama da *accountability* de segunda ordem, destinada à sociedade e ao interesse público, em que é assegurada a sua participação e contribuição por meio de procedimentos que legitimem as definições e políticas adotadas pelos entes privados (Kaminski, 2019b, p. 1.563).

A elaboração destes relatórios de impacto quando do desenvolvimento e implementação de algoritmos de ML de tomada de decisão a determinados contextos tem sido fortemente indicada como um mecanismo de *accountability* e transparência da tecnologia. Este debate tem interessado reguladores e pesquisadores sobre o tema, que têm questionado quando que tais sistemas de decisão automatizadas devem ser aplicados, e em que contexto isto pode e deve ocorrer (Reisman et al., 2018, p. 3). O resultado da análise destes algoritmos auxilia na formação de conhecimento sobre o impacto que os mesmos causam na sociedade, criando elementos para uma pergunta bastante delicada, mas fundamental: enquanto sociedade, como desejamos que estes algoritmos funcionem? (Sandvig et al., 2014, p. 18).

O AI Now Institute, ao propor seu *framework* do AIA[227], está atento a este objetivo. Em seu documento sobre o tema, este instituto é taxativo ao indicar que

[227] Como informado pelo relatório, o governo canadense adotou o *Algorithm Impact Assessment* (AIA) desenvolvido pelo AI NOW, no seu processo de compra de sistemas de IA. Ver em:

visa contribuir para um debate mais bem informado sobre o uso de algoritmos e sistemas de IA, especialmente pelo Poder Público. Embora o documento reconheça as limitações desta avaliação para evitar potenciais danos e impactos, o documento aponta que o seu valor está em servir como um mecanismo para informar o público e engajar a sociedade civil, reguladores e pesquisadores a fiscalizarem, criticarem e questionarem o desenvolvimento e uso de algoritmos de tomada de decisão. Mesmo que o documento se volte para o Poder Público, as relações entre público e privado no desenvolvimento de sistemas de IA é imbricada, em que não é irrealista imaginar situações onde o Estado poderia ser o financiador, idealizador e usuário destas iniciativas, e posteriormente o seu uso poderia ser difundido e utilizado em um contexto diferente pela iniciativa privada, ou vice-versa. Assim, os elementos chaves na análise do AIA, que serão melhor identificados abaixo, também merecem especial atenção da iniciativa privada.

Na ausência de uma legislação específica sobre o tema e, considerando que a maior parte de sistemas de IA e seus respectivos algoritmos de tomada de decisão têm como insumo uma quantidade enorme de dados pessoais, a elaboração do RIPDP deve ser fortemente recomendada na implementação destes algoritmos no contexto de tratamento de dados pessoais, especialmente quando houver "decisões tomadas unicamente com base em tratamento automatizado de dados pessoais" (art. 20, da LGPD). Mesmo que tais sistemas não tenham como insumo dados pessoais, o RIPDP também pode servir como exemplo para a elaboração de um relatório de impacto próprio para algoritmos de ML. Até mesmo porque relatórios de impacto no contexto de dados pessoais não devem se limitar a questões estritamente relacionadas à proteção de dados, mas devem considerar riscos à liberdade e aos direitos da pessoa natural de forma mais ampla (Rovatsos, 2020, p. 28), não se restringindo a um mero documento para atestar uma conformidade legal.[228]

A LGPD prevê que, quando um determinado tratamento de dados pessoais implicar em um risco para os direitos e as liberdades dos titulares de dados pessoais,

https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/ en/algorithmic-impact-assessment.html. Acessado em 15.05.2021.

[228] Neste sentido, vale a lição de Maria Cecília O. Gomes (2019, p. 182): "A proposta da UE com o relatório de impacto, que foi posteriormente tropicalizada no Brasil por meio da LGPD, não foi gerar um calhamaço de documentos para comprovar que está tudo certo dentro das organizações, mas sim e, essencialmente, figurar como uma ferramenta dinâmica, capaz de auxiliar na governança de dados e na mitigação dos riscos associados às operações de tratamento de dados. Guardá-lo em caixinhas é submetê-lo ao esquecimento, esvaziando o seu propósito".

será necessário elaborar o RIPDP, também conhecido como *Data Protection Impact Assessment* (DPIA), em razão da sua previsão no art. 35 do GDPR. Este relatório consiste na elaboração de uma documentação por parte do controlador que contém: (i) a descrição dos processos de tratamento de dados pessoais que podem gerar riscos às liberdades civis e aos direitos fundamentais, e (ii) as medidas, salvaguardas e mecanismos de mitigação de risco (art. 5º, XVII, da LGPD) (Frajhof, 2021, p. 480). Seu objetivo tem origem na ideia de "viabilizar a visualização de processos e procedimentos internos, bem como o tratamento de dados existentes, para que seja possível através dele realizar a prevenção de riscos e a mitigação desses" (Gomes, 2019, p. 177). Os riscos se referem a ameaças tanto às liberdades civis dos titulares de dados (liberdade de expressão, de reunião e de associação, liberdade de consciência e de religião, direito à vida privada e à intimidade), quanto aos direitos fundamentais, previstos no art. 5º, da CF. O importante, como bem destaca Sérgio Negri, é que os mecanismos de gerenciamento de riscos envolvendo novas tecnologias tenham como foco os "detentores dos direitos que deverão ser respeitados", e não a sociedade empresária e os riscos de negócio.[229]

Sobre este tema, a LGPD não apresentou qualquer requisito ou diretriz do que deve constar no referido relatório, indicando apenas que será a ANPD a responsável por definir quando este será mandatório (art. 38, da LGPD). Assim, caberá à ANPD a discricionariedade em exigir a sua elaboração[230], não ficando "claro qual é o grau de comprometimento dos controladores com a referida obrigação" (Frazão, 2018c). Nesta indeterminação legislativa e, diante da semelhança entre a legislação brasileira e a norma europeia, e a ampla experiência da UE em elaborar tal instrumento, cabe avaliar os termos desta obrigação à luz do GDPR e das diretrizes do GTA29 específicas sobre o DPIA. Inclusive, o GDPR destaca expressamente a necessidade da elaboração deste documento na incidência do art. 22, do GDPR, que cuida do direito à explicação (Frajhof, 2021, p. 480).

[229] Ver em: NEGRI, Sérgio Marcos Carvalho de Ávila. Personalidade, responsabilidade e classificação dos riscos na Inteligência Artificial e na robótica. *Migalhas,* 01 de julho de 2021. Disponível em: https://www.migalhas.com.br/coluna/migalhas-de-responsabilidade-civil/347862/personalidade-responsabilidade-e-classificacao-dos-riscos-na-ia. Acessado em 02.07.2021.

[230] O art. 10, da LGPD, que traz parâmetros para a aplicação da base legal de tratamento de dados pessoais do legítimo interesse do controlador, indica em seu § 3º que a ANPD poderá solicitar ao controlador o RIPDP, "quando o tratamento tiver como fundamento seu interesse legítimo, observados os segredos comercial e industrial"

Portanto, a análise a seguir pode apresentar diretrizes de como este relatório pode ser aplicado quando houver o uso de dados pessoais no tratamento automatizado, em que há a aplicação de algoritmos de tomada de decisão de ML. Além disso, pode pautar os parâmetros a serem aplicados especificamente para o desenvolvimento de sistemas de IA que se valem de ML. O que deverá ser considerado, no caso concreto, é a necessidade de se desenvolver um relatório de impacto específico e próprio quando houver o desenvolvimento de IA, ou se aquele indicado pela legislação de proteção de dados pessoais será suficiente para avaliar os riscos inerentes às liberdades civis, direitos fundamentais e direitos de titulares de dados pessoais no contexto da IA.

2.2.1.a.i
A ótica do GTA29 sobre o Relatório de Impacto de Proteção de Dados Pessoais no âmbito do GDPR

Como ressaltado pelo antigo GTA29, o DPIA tem por objetivo "descrever o processamento, acessar sua necessidade e proporcionalidade e auxiliar no gerenciamento de riscos aos direitos e liberdades"[231] dos titulares de dados que estão tendo seus dados tratados, e determinar quais são as medidas que devem ser tomadas para lidar com tais riscos (GTA29, 2017, p. 4). O parecer faz uma abordagem sobre a necessidade ou não de elaborar este documento baseado no risco de um determinado cenário, das consequências, severidade e probabilidade deste ocorrer. Esta análise não se limita ao risco envolvendo a proteção de dados pessoais e a privacidade dos indivíduos, mas também outros direitos fundamentais, tais como o direito à liberdade de expressão e liberdade religiosa, por exemplo.

De acordo com o GTA29, o DPIA elaborado em um único projeto pode ser utilizado para outros contextos similares em termos de natureza, escopo, contexto, propósito e risco, tal como quando tecnologias similares são utilizadas para coletar determinados tipos de dados para o mesmo propósito (GTA29, 2017, p. 7). Seria o caso, por exemplo, de uma empresa que administra uma estação de trem, que poderia realizar um único DPIA para tratar das filmagens realizadas em todas as suas estações, sem a necessidade de elaborar este relatório para cada estação.

[231] Tradução livre de: "Describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedom"

O DPIA também poderia ser útil para avaliar o impacto de um produto de tecnologia[232], principalmente quando houver dois controladores diferentes que vão implementar processamentos diversos: o responsável pelo seu desenvolvimento e aquele que de fato irá implementá-lo. O indicado é que este último realize um DPIA, e que ambos troquem informações sobre o mesmo, com os devidos cuidados com os segredos empresariais envolvidos. Esta orientação é especialmente relevante no contexto de algoritmos de tomada de decisão, tendo em vista que os responsáveis pelo seu desenvolvimento nem sempre serão os agentes econômicos que o implementarão.

Apesar de não ser mandatória a sua realização em toda e qualquer atividade de processamento de dados pessoais, o DPIA é um instrumento que demonstra a aderência dos controladores com o GDPR, sendo, portanto, uma boa prática de *compliance* a ser adotada. O art. 35[233], do GDPR, traz uma lista não exaustiva, mas exemplificativa, de hipóteses que considera apresentar um nível elevado de risco, justificando a elaboração do relatório. Portanto, outras atividades de tratamento de dados ainda assim podem exigir a elaboração de um DPIA por apresentarem uma alta probabilidade de risco aos direitos e liberdades dos titulares de dados.

O GTA29 lista nove hipóteses que poderiam exigir a preparação do DPIA, sendo que a maior parte das situações listadas são tipicamente atividades que precedem o uso de algoritmos de tomada de decisão de ML, diante da eficiência e benefício dos seus resultados. É o caso, por exemplo, do uso de dados pessoais para fazer uma avaliação, pontuação ou predição, inclusive com formação de perfil, relacionados a aspectos de uma pessoa e o seu "trabalho, situação econômica, saúde, preferências pessoais, comportamento ou confiabilidade, localização ou movimentos", tal como ocorre na formação de perfis comportamentais para fins de *marketing* (GTA29, 2017, p. 9). Outros exemplos envolvem o uso de dados pessoais

[232] Os Considerando 89 e 91 do GDPR indicam que quando um tratamento de dados pessoais envolver o uso de novas tecnologias será necessário realizar a avaliação do impacto sobre a proteção de dados pessoais, para avaliar a probabilidade ou gravidade de risco que podem advir do referido tratamento.

[233] (a) Avaliação sistemática e completa dos aspectos pessoais relacionados com pessoas singulares, baseada no tratamento automatizado, incluindo a definição de perfis, sendo com base nela adotadas decisões que produzem efeitos jurídicos relativamente à pessoa singular ou que a afetem significativamente de forma similar; b) Operações de tratamento em grande escala de categorias especiais de dados a que se refere o artigo 9o n. 1, ou de dados pessoais relacionados com condenações penais e infrações a que se refere o artigo 10o; ou c) Controlo sistemático de zonas acessíveis ao público em grande escala.

sensíveis,[234] dados considerados personalíssimos em sua natureza[235], ou de pessoas vulneráveis (como crianças), diante do impacto que o tratamento de dados pessoais pode causar ao exercício de outros direitos fundamentais ou no cotidiano destas pessoas (GTA29, 2017, p. 9-10). Recomenda-se, ainda, a elaboração do DPIA quando houver o processamento de dados em larga escala, o que é bastante rotineiro em projetos de *big data*, ou na combinação de dois conjuntos de dados diferentes, utilizados por diferentes controladores ou com finalidades diversas, que superem a expectativa do titular de dados. Ademais, o uso de novas tecnologias, como, biometria ou reconhecimento facial, também justificam a elaboração do DPIA.

A recomendação da própria norma (arts. 35(1), 35(10), Considerandos 90 e 93) é que o DPIA seja conduzido antes do início do processamento de dados, pois o mesmo "é visto como uma ferramenta para ajudar na tomada de decisão referente ao processamento"[236] (GTA29, 2017, p. 14). A mesma lição pode ser transposta para o desenvolvimento de algoritmos e modelos de ML. Havendo evidências de que o uso do algoritmo cria uma ameaça concreta aos direitos e às liberdades das pessoas, seja desproporcional ao fim almejado com o seu uso, e que as salvaguardas adotadas não sejam suficientes para mitigá-la, o projeto deverá ser repensado, remodelado ou até mesmo abandonado. Esta mesma discussão tem ocorrido no uso de tecnologias de reconhecimento facial em espaços públicos[237], diante dos riscos inerentes à democracia e aos direitos humanos, como privacidade, liberdade e igualdade, que ocorrem com o constante monitoramento e registro da população, além dos diversos relatos de discriminações e erros[238] que elas têm causado.

---

[234] Art. 9, do GDPR: "É proibido o tratamento de dados pessoais que revelem a origem racial ou étnica, as opiniões políticas, as convicções religiosas ou filosóficas, ou a filiação sindical, bem como o tratamento de dados genéticos, dados biométricos para identificar uma pessoa de forma inequívoca, dados relativos à saúde ou dados relativos à vida sexual ou orientação sexual de uma pessoa."

[235] Como aqueles envolvendo condenações criminosas, indicadas no art. 10, do GDPR: "O tratamento de dados pessoais relacionados com condenações penais e infrações ou com medidas de segurança conexas com base no artigo 6o, n. 1, só é efetuado sob o controle de uma autoridade pública ou se o tratamento for autorizado por disposições do direito da União ou de um Estado-Membro que prevejam garantias adequadas para os direitos e liberdades dos titulares dos dados. Os registos completos das condenações penais só são conservados sob o controle das autoridades públicas".

[236] Tradução livre de: "as a tool for helping decision-making concerning the processing".

[237] Sobre o tema, ver: TEFFÉ, Chiara Spadaccini de. FERNANDES, Elora Raad. Reconhecimento Facial: laissez-faire, regular ou banir?. *Migalhas*, 16 de jul. de 2020. Disponível em: https://www.migalhas.com.br/coluna/migalhas-de-vulnerabilidade/330766/reconhecimento-facial--laissez-faire--regular-ou-banir. Acessado em 15.05.2021.

[238] Ver: SILVA, Tarcizio. Reconhecimento facial na Bahia: mais erros policiais contra negros e pobres. *Blog do Tarcizio*. São Paulo, 21 de nov. de 2019. Disponível em: <https://tarciziosilva.com.br/blog/reconhecimento-facial-na-bahia-mais-erros-policiais-contra-

Em uma alternativa à perspectiva Europeia, as Autoridades de Proteção de Dados Pessoais da Argentina e do Uruguai[239] elaboraram um documento conjunto chamado de "*Guía de Evaluación de Impacto en la Protección de Datos"* (AAIP; URCDP, 2020), que traz uma matriz para avaliação da probabilidade de risco, indicando quando o referido relatório deverá ser organizado. Para estimar o nível do impacto (podendo ser baixo, médio, alto ou muito alto), que pode ser de ordem moral ou material, o documento traz uma série de exemplos concretos que ajudam nesta classificação.

Por exemplo, o recebimento de e-mails *spam* seria classificado como um exemplo de baixo impacto material aos direitos dos titulares de dados; a negativa de acesso a serviços administrativos ou comerciais seria considerada como de impacto médio; dificuldade financeira a médio ou longo prazo por conta do tratamento de dados, e a perda de oportunidades únicas e não recorrentes seriam consideradas como de danos materiais com alto impacto; a sensação de violação a direitos fundamentais, como igualdade, ocasionando discriminação, e liberdade de expressão, seriam exemplos de impactos morais altos; risco financeiro, dúvidas, incapacidade de trabalhar ou perda de acesso a bens estruturais, como água e energia, seriam considerados danos materiais críticos, enquanto condenações penais, perda de vínculos familiares e de amizade seriam exemplos de impactos morais críticos (AAIP; URCDP, 2020, p. 21-23). Note-se que a preocupação neste tratamento de dados é que o mesmo afete aspectos pessoais dos titulares de dados, ou a fruição de direitos fundamentais.

Voltando ao contexto europeu, o GTA29 indica que, para elaborar o relatório de impacto, o GDPR sugere que a perspectiva dos titulares de dados ou seus representantes devam ser coletadas (art. 35(9), do GDPR), quando adequado. Caso isto não seja feito, deve haver uma justificativa, de que isto seria impraticável ou desproporcional (GTA29, 2017, p. 15). Embora seja uma recomendação, o AIA elaborado pelo IA Now Institute indica esta consulta pública como sendo necessária, bem como o regime de governança colaborativa de Kaminski (2019b).

Ainda, devem ser documentadas e definidas as responsabilidades e os papéis desempenhados para a elaboração do DPIA, indicando as unidades

negros-e-pobres/>. Acessado em 15.05.2021.

[239] Ambos os países são signatários da Convenção para a Proteção das Pessoas relativamente ao Tratamento Automatizado de Dados de Caráter Pessoal (Convenção 108) do Conselho da Europa.

envolvidas no tratamento de dados e seus responsáveis, delimitar a responsabilidade dos agentes de tratamento, e contar com o auxílio do operador na elaboração do relatório (quando houver). Ainda, é sugerido: procurar recomendações especializadas de agentes externos e independentes de outras profissões (como advogados, sociólogos ou especialistas em computação) para auxiliarem na sua elaboração, que encarregados possam sugerir quando um DPIA deva ser elaborado em uma determinada operação, indicando a metodologia, auxiliar na avaliação, entre outros.

De acordo com o art. 35(7) e os Considerandos 84 e 90, todos do GDPR, a metodologia para aplicação do DPIA dever ser flexível, de maneira que cada controlador pode estabelecer a sua estrutura e forma, a fim de atender às práticas de um determinado setor (GTA29, 2017, p. 17). Os requisitos mínimos são: (i) a descrição sistemática das operações de tratamento e a finalidade, (ii) uma avaliação da necessidade e proporcionalidade do tratamento em relação à sua finalidade; (iii) uma avaliação dos riscos para os direitos e liberdades dos titulares de dados, (iv) as medidas adotadas frente aos riscos mapeados, e (v) a demonstração de conformidade com o GDPR. Estas etapas devem ser realizadas de maneira interativa e devem ser revisitadas constantemente até serem completadas (GTA29, 2017, p. 16). Esta avaliação de risco deve levar em consideração a natureza, escopo, contexto e finalidade do tratamento, origem destes riscos, probabilidade deles ocorrerem, com a indicação de como foram mitigados e endereçados pelas salvaguardas adotadas (GTA29, 2017, p. 17).

Embora o GDPR não estabeleça a obrigatoriedade de publicação do DPIA[240], é considerada uma boa prática a divulgação de um sumário ou uma conclusão, com o objetivo de estabelecer confiança, realizar a prestação de contas e assegurar a transparência da atividade de tratamento conduzida pelo controlador. Ainda, o GTA29 indica que uma autoridade de controle deverá ser consultada quando houver alto risco aos titulares de dados ou quando mesmo após a implementação de controles e salvaguardas ainda houver "riscos residuais"

[240] Art. 36(5), do GDPR: "As legislações de proteção de dados dos Estados-Membros podem exigir esta consulta obrigatória, sendo necessária uma autorização prévia, quando o tratamento envolver interesse público, tal como quando relacionado ao tratamento por motivos de proteção social e de saúde pública".

consideráveis, como, não ser capaz de evitar um grande acesso aos dados quando este é compartilhado de maneira distribuída (GTA29, 2017, p. 18-19).

Comparando as exigências do DPIA com o AIA, proposto pelo AI Now Institute, destacamos que este último possui cinco elementos chaves: (i) avaliar os impactos dos sistemas automatizados existentes ou que vão ser desenvolvidos, que deverão se pautar, entre outros, em princípios como justiça e igualdade, a fim de evitar discriminações (ii) desenvolver processos de revisão externos para avaliar impactos ao longo do uso do sistema, (iii) tornar o público ciente do que é compreendido como um sistema de tomada de decisão automatizado e das avaliações feitas sobre os mesmos, (iv) realizar consultas públicas para recebimento de críticas e mostrar disponibilidade para responder dúvidas sobre o sistema, (v) estruturar um procedimento para assegurar um devido processo legal em relação à decisão automatizada, envolvendo mecanismos para que pessoas afetadas possam contestar estes resultados (Reisman, 2018, p. 4).

Percebe-se que o AIA propõe um escrutínio e participação maior do público, e a adoção expressa de meios que permitam que uma pessoa possa contestar uma decisão tomada por um algoritmo que a afete. O GTA29 recomenda, mas não indica ser mandatória, esta comunicação com o público e uma espécie de prestação de contas do resultado do DPIA. A transparência e a prestação de contas e responsabilização, como demonstrado ao longo deste capítulo, são pressupostos para a legitimidade do uso de algoritmos de ML, além de serem condições necessárias para o exercício, efetividade e avaliação do respeito a direitos fundamentais. Por isso, a avaliação de impacto nestes contextos deve exigir a participação de terceiros interessados, com a abertura de informações relevantes, sem que sejam mantidas em sigilo informações importantes sob a justificativa do segredo empresarial (Reisman, 2018, p. 14).

2.2.1.b
Boas práticas e códigos de conduta

Na "caixa de ferramentas" (*toolkit*) proposta por Kaminski (2019b), códigos de conduta e estabelecimento de boas práticas constituem instrumentos que compõem a governança colaborativa, assim como certificações, e são vistas como meios para apoiar a fixação de parâmetros de governança para o desenvolvimento

de algoritmos. Para que este instrumento tenha força coercitiva, é importante que a sua elaboração possa estar sujeita ao escrutínio do público uma vez publicados, tal qual recomendado na elaboração do relatório de impacto, além de permitir formas de assegurar que os mesmos estão sendo cumpridos pela organização.

Em que pese a importância deste tema, este subcapítulo não se atentará a uma análise mais aprofundada sobre o mesmo, sob pena de fugir do escopo deste trabalho[241]. Serão apontadas, apenas, algumas orientações e previsões legislativas que podem auxiliar no estabelecimento destas diretrizes no contexto de algoritmos de tomada de decisão de ML.

Assim como o GDPR, a LGPD também prevê tais mecanismos. O artigo 50 e seguintes tratam dos Código de Boas Práticas e Governança[242], e do Código de Condutas [243] , que são considerados como instrumentos importantes para o *compliance* e para demonstrar a prestação de contas dos agentes de tratamento em relação à legislação. Ambos buscam oferecer maneiras de operacionalizar e concretizar comandos abertos propostos pela lei (Frajhof, 2021, p. 482). Estes instrumentos estimulam as organizações "a trabalhar[em] na linha de corregulação, em uma associação entre a legislação e um conjunto de instrumentos capaz de auxiliá-las na aplicação e na eventual comprovação da correta aplicação da Lei" (Palmeira, 2020, p. 336). Assim, estes documentos corporativos podem servir como instrumentos adequados para estabelecer, também, procedimentos e padrões técnicos que devem ser observados quando do desenvolvimento de algoritmos de

[241] Sobre o tema, ver: PALMEIRA, Mariana de Moraes. A segurança e as boas práticas no tratamento de dados pessoais. *In*: Caitlin Mulholland. (Org.). *A LGPD e o novo marco normativo no Brasil*. 1a ed. Porto Alegre: Arquipélago Editorial, 2020, pp. 319-342. Um exemplo brasileiro, é a elaboração pela ANPD do recente Guia de Boas Práticas da LGPD no âmbito do serviço público federal. Disponível em: <https://www.gov.br/governodigital/pt-br/seguranca-e-protecao-de-dados/guias/guia_lgpd.pdf/>. Acessado em 12.01.2022. Um exemplo no contexto da UE foi produzido pelo CEPD, em seu Guidelines 1/2019 on Codes of Conduct and Monitoring Bodies under Regulation 2016/679. Disponível em: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-12019-codes-conduct-and-monitoring-bodies-0_en. Acessado em 12.01.2022.

[242] Previsto no art. 50, *caput,* da LGPD: "Os controladores e operadores, no âmbito de suas competências, pelo tratamento de dados pessoais, individualmente ou por meio de associações, poderão formular regras de boas práticas e de governança que estabeleçam as condições de organização, o regime de funcionamento, os procedimentos, incluindo reclamações e petições de titulares, as normas de segurança, os padrões técnicos, as obrigações específicas para os diversos envolvidos no tratamento, as ações educativas, os mecanismos internos de supervisão e de mitigação de riscos e outros aspectos relacionados ao tratamento de dados pessoais."

[243] Art. 50, § 2º, II, da LGPD: "demonstrar a efetividade de seu programa de governança em privacidade quando apropriado e, em especial, a pedido da autoridade nacional ou de outra entidade responsável por promover o cumprimento de boas práticas ou códigos de conduta, os quais, de forma independente, promovam o cumprimento desta Lei".

ML no contexto de tratamento de dados pessoais. Sua adoção pode "funcionar como um gerenciamento de qualidade de sistemas de IA para usuários, consumidores, organizações, instituições de pesquisa e governos" (Comissão Europeia, 2019a, p. 22)[244] (Frajhof, 2021, p. 484).

O GTA29 em seu *Guideline* específico sobre o tratamento de decisões automatizadas aponta que controladores possam se valer, entre outras medidas, de códigos de conduta para prever processos para auditar aplicações de aprendizado de máquina e certificações para operações de tratamento de dados (GTA29, 106, p. 30). É importante que estes códigos de conduta sejam estruturados para estabelecer maneiras de implementar esta auditoria, bem como para indicar as medidas que devem ser tomadas para evitar erros e discriminações pelo uso destes algoritmos ou violação de direitos (Kaminski, 2019b, p. 1.600).

O Código de Conduta visa especificar detalhes do procedimento de tratamento de dados pessoais, estabelecendo instruções gerais e os valores a serem seguidos pelos funcionários e pela alta administração em relação às decisões relativas ao tratamento de dados (Frazão et al., 2020, p. 693-694) (Frajhof, 2021, p. 483). Tais documentos são tipicamente diretrizes éticas indicando como a atividade de uma corporação deve ser conduzida. No contexto do desenvolvimento de sistemas de IA, e especialmente para o direito à explicação, a indicação de princípios como o próprio FAT, ou de outros indicados nos mais variados guias de recomendação de princípios éticos já elaborados,[245] seria uma demonstração de prestação de contas importante. Além do FAT, outros princípios específicos da IA, como o princípio de acurácia (*accuracy*) e da inteligibilidade (*intelligibility*), implicam em medidas que:

> (i) impeçam a aplicação de sistemas de IA que violem o princípio da igualdade de tratamento; (ii) permitam reconhecer que os insumos utilizados pela IA e os resultados que advém de seu tratamento sejam precisos; e (iii) proporcionem à pessoa humana o conhecimento dos processos de decisão tomados pela IA" (Mulholland; Frajhof, 2021, p. 73).

[244] Tradução livre de: "function as a quality management system for AI users, consumers, organisations, research institutions and governments".
[245] Ver em: Mulholland; Frajhof, 2021.

Estes princípios também são basilares para o fim que o direito à explicação visa alcançar e contém um elemento importante, que é estabelecer confiança em relação ao funcionamento e resultado destes sistemas.

A LGPD não trouxe maiores detalhes sobre a elaboração do código de conduta, apenas indicando que a ANPD ou outra entidade poderá requerer a elaboração de boas práticas ou códigos de conduta, "os quais, de forma independente, promovam o cumprimento da Lei" (art. 50, § 2º, II, da LGPD). Comparativamente, o GDPR trouxe um artigo próprio para tratar dos códigos de conduta (art. 40, da LGPD), detalhando o conteúdo que o mesmo poderá trazer, propondo até mesmo que tais códigos estejam sujeitos à aprovação de uma autoridade de controle (art. 40(5), da LGPD).

Por sua vez, as regras de boas práticas, instrumentalizadas em códigos, também se destinam a estabelecer questões operacionais do processamento de dados, definindo padrões técnicos e os mecanismos que devem ser observados para estruturar um sistema (Frajhof, 2021, p. 482). A LGPD prevê que os agentes de tratamento, em conjunto ou não, poderão formular regras de boas práticas e de governança que estabeleçam "condições de organização, o regime de funcionamento, os procedimentos, incluindo reclamações e petições de titulares, as normas de segurança, os padrões técnicos, as obrigações específicas para os diversos envolvidos no tratamento, as ações educativas, os mecanismos internos de supervisão e de mitigação de riscos e outros aspectos relacionados ao tratamento de dados pessoais" (art. 50, da LGPD). Devem ser levadas em consideração a natureza, o escopo, a finalidade, a probabilidade e a gravidade dos riscos, bem como os benefícios decorrentes do tratamento de dados (art. 50 § 1º, da LGPD).

Em relação à aplicação dos princípios da segurança e prevenção (art. 6, incisos, VII e VIII, da LGPD), o controlador deverá se atentar à estrutura, escala e volume das suas operações, a sensibilidade dos dados tratados e a probabilidade e gravidade dos danos que poderão decorrer da sua atividade de tratamento de dados. Para tanto, poderá ser implementado um programa de governança (art. 50 § 2º, inciso I, da LGPD), que deve atender a alguns requisitos mínimos, entre eles: estabelecimento de políticas e salvaguardas para avaliação sistemática de impactos e riscos à privacidade (alínea d); construção de uma relação de confiança com o titular, com atuação transparente e mecanismos de participação (alínea e);

implementação e aplicação de mecanismos de supervisão internos e externos (alínea f); e, atualização e avaliação periódica do referido programa (alínea h).

As regras de boas práticas e de governança deverão ser publicadas e atualizadas periodicamente, podendo ser reconhecidas e divulgadas pela ANPD (art. 50 § 3º, da LGPD).

A elaboração daquelas podem ser uma oportunidade para o estabelecimento de processos que devem ser adotados quando houver o uso de dados pessoais por algoritmos de tomada de decisão de ML, além da definição de padrões técnicos adotados nestes contextos. Especificamente em relação ao direito à explicação, as seguintes determinações podem estar previstas neste documento: orientações para que se analise se o uso de uma decisão automatizada permite que seja utilizado um algoritmo considerado caixa-preta ou não, diante do potencial risco aos titulares de dados; necessidade de se estabelecer formas de transparência e explicação dos seus resultados; regras para definir se haverá uma intervenção humana em um processo decisório, ou se a interferência humana apenas ocorrerá em eventual pedido de revisão da decisão; como este direito poderá ser exercido e o procedimento para tanto; fixar o procedimento para limpeza de dados quando há o uso de algoritmos de ML para evitar discriminações, padrões de documentação, entre outras medidas que busquem respostas para as perguntas indicadas acima. Quanto a este último ponto, é importante que as organizações tenham uma equipe diversa e inclusiva, com poder decisório e influência, contribuindo com pontos de vista diferentes e novas perspectivas, evitando a reprodução de estigmas e discriminações nos algoritmos de tomada de decisão[246] (Frajhof, 2021, p. 484).

Portanto, na ausência de uma legislação ou diretrizes específicas sobre o tema aplicado a algoritmos de tomada de decisão, estes instrumentos previstos em leis gerais de proteção de dados pessoais podem ser utilizados como base para a propositura de diretrizes específicas para o contexto da inteligência artificial, principalmente quando estes algoritmos se valerem de dados pessoais como insumos. Apesar de a LGPD trazer princípios, direito e regras importantes para o tratamento de dados pessoais, passível de incidir em aplicações de IA, a propositura de uma legislação específica para regular a inteligência artificial seria importante

[246] Sobre a importância desta abordagem, ver: UNESCO. *I'd blush if I could*: closing gender divides in digital skills through education. 2019. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000367416>. Acessado em 12.09.2019.

para estabelecer questões específicas deste contexto, como a definição de responsabilidade, e a previsão de princípios e direitos próprios.

2.2.2
Prestação de contas e responsabilização *ex post*

2.2.2.a
Técnicas de interpretação e explicação de ML

Atualmente, há um grande esforço de pesquisadores da área de Ciência da Computação para estudar métodos de interpretação de algoritmos de aprendizado de máquina, em uma área denominada como *Explainable Artificial Intelligence* (XAI). A XAI visa "compreender e interpretar o comportamento de sistemas de IA" (Linadartos, 2020, p. 2) por meio de diferentes métodos, que buscam atender diferentes finalidades (Frajhof, 2021, p. 488). As propostas de explicação nesta área têm como objetivo gerar explicações para um usuário especializado, e não para usuários leigos (Maranhão et al., 2021, p. 146), embora alguns trabalhos recentes estejam propondo abordagens neste sentido (Chari et al., 2020).

Métodos diferentes têm sido aventados como maneiras de interpretar e explicar o resultado de um algoritmo de aprendizado de máquina. Em um mapeamento sistemático[247] sobre os trabalhos relacionados na área de Ciência da Computação sobre este tema, Linardatos et al. (2020) identificaram quatro principais tipos de métodos que vêm sendo estudados e testados por pesquisadores. Estes visam oferecer indícios sobre o resultado de modelos e algoritmos de ML, relacionado à sua lógica interna, e como tais indícios devem ser apresentados para que sejam compreensíveis por um humano (Linadartos et al., 2020, p. 2-3). Estes quatro métodos, ou a taxonomia para interpretação e explicação de modelos de aprendizado de máquina, são: métodos interpretativos para explicar modelos *black-box*, métodos para criar modelos *white-box*, métodos para promover justiça e igualdade (*fairness*) evitando discriminações, e métodos para analisar a sensibilidade da predição de modelos.

[247] Este tipo de estudo é comum na área de Ciência da Computação, e consiste em uma metodologia estruturada, em que pesquisadores buscam em diferentes bases de dados de publicações acadêmicas os artigos já publicados sobre determinado tema. O resultado destes mapeamentos é a apresentação de um amplo panorama sobre os tipos de pesquisas que têm sido conduzidas sobre um assunto, sendo possível identificar os principais achados e as principais oportunidades de pesquisa. Assim, a utilização deste estudo é um suporte para compreender este panorama.

Em breve síntese, o primeiro busca aplicar métodos que forneçam explicações e interpretem o resultado dos modelos de ML, com alguns estudos voltados para verificar a relação entre os dados de entrada (*input*) e o respectivo resultado (*output*); o segundo visa propor modelos de ML que considerem desde a sua concepção uma preocupação em fornecer explicações e interpretações sobre os seus resultados; o terceiro busca criar métricas para definir o que é *fairness*, analisando os dados e verificando a presença de vieses e se o resultado pode ser considerado justo; e o último visa aplicar métodos para avaliar a confiança na predição, e como alterações nos dados de entrada alteraram os dados de saída (Frajhof, 2021, p. 488).

Os métodos interpretativos para explicar modelos caixa-preta tentam apresentar formas de interpretar modelos complexos, e, por isso, são aplicados a resultados concretos do modelo, conhecidos como métodos *post hoc*. Alguns destes métodos buscam analisar os dados de entrada e os resultados gerados por um sistema, mas não permitem visualizar a operação interna do algoritmo (Desai; Kroll, 2017, p. 36). A maior parte dos trabalhos que tratam sobre o tema tem dado especial atenção a modelos aplicados ao reconhecimento de imagem, com o uso de redes neurais e a aplicação de um conceito chamado de *saliency,* em que a partir de mapas topográficos é possível observar como que o modelo representa a imagem ao reproduzi-la em *pixels* (Linadartos et al., 2020, p. 5-17)[248].

Neste sentido, Maranhão, Cozman e Almada dividem em três categorias as técnicas que são exploradas em IA para gerar explicações. A primeira seriam técnicas baseadas em sensibilidade, em que se busca responder a seguinte pergunta: "a decisão deste modelo é mais sensível a que itens?". Como resposta, uma lista é gerada, e seus itens são considerados como "explicações". O exemplo indicado no parágrafo acima é uma aplicação desta técnica (Maranhão et al., 2021, p. 145).

As duas outras categorias para gerar explicações são estratégias decomposicional ou agnóstica. A estratégia decomposicional escolhe um tipo de modelo (tal como redes profundas ou florestas aleatórias) e o decompõe em diferentes partes, e um módulo interpretador é criado para ser responsável por gerar

[248] Alguns dos métodos aplicados são: *gradient, integrated gradient, DeepLIFT, Guided BackPropagaton; Deconvolution; Class Activation Maps (CAMs), Grad-CAM, Grad-CAM++, Layer-wise Relevance Propagation (LRP), moothGrad, RISE algorithm, Concept Activation Vectors (CAVs), Deep Taylor*.

explicações para as diferentes frações do modelo a qual está associado. Por exemplo, a explicação decomposicional de uma rede profunda é um módulo que trabalha de maneira paralela ao modelo que visa ser explicado, gerando as explicações. Por sua vez, a estratégia agnóstica não analisa o processo interior do modelo, acessando apenas os dados de entrada e os dados de saída (decisões). O modelo interpretador é mais simples que o modelo que visa ser explicado, e é desenvolvido para que seja interpretável, gerando as explicações para os usuários (Maranhão et al., 2021, p. 146).

Um dos métodos mais populares[249] / [250], que tem sido replicado pela literatura, é o *Local Interpretable Model-Agnostic Explanations* (LIME) (Ribeiro et al., 2016), que fornece uma explicação local sobre decisões específicas de modelos de ML. Este identifica o que está sendo considerado relevante pelo algoritmo para que o mesmo realize a sua classificação. Embora alguns estudos tenham indicado a baixa qualidade do LIME,[251] motivando o desenvolvimento do DLIME para refinar sua capacidade de interpretação,[252] a sua técnica se apresenta como um marco importante para explicar as predições e resultados de decisões tomadas por algoritmos de ML em diferentes contextos.

Este modelo busca resolver duas questões relacionadas à confiança de algoritmos que se valem de ML. Uma diz respeito à confiança na predição, ou seja, se uma pessoa confia no resultado de um modelo, a motivando a tomar uma ação baseada nesta predição. A outra se refere à confiança no modelo, de que este deverá se comportar de maneira adequada quando implementado "*in the wild*". Estas duas preocupações de confiança se relacionam com questões de prestações de contas e

[249] Esta método foi indicado como bibliografia indicada no projeto chamado de "The AI Blindspot", desenvolvido no Berkman Klein Center e o MIT Lab, em 2019, por Ania Calderon, Dan Taber, Hong Qu, and Jeff Wen. O objetivo do projeto é desvendar vieses inconscientes em sistemas de IA e desigualdades estruturais. Disponível em: https://aiblindspot.media.mit.edu/index.html. Acessado em 19.04.2021.

[250] Linadartos et al., também indicam que há um outro método bastante citado na literatura para auxiliar na interpretação das explicações, chamado de SHAP (SHapley Additive exPlanation), que, assim como LIME, atribui a cada funcionalidade um valor de importância para uma determinada predição. Sobre o SHAP, ver em: LUNDBERG, S.M.; LEE, S.I. A unified approach to interpreting model predictions. In: Proceedings of the Advances. In: *Neural Information Processing Systems*, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.

[251] Conforme identificado por Lindartos et al., com a indicação do seguinte trabalho: Garreau, D.; von Luxburg, U. Explaining the Explainer: A First Theoretical Analysis of LIME. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, Palermo, Sicília, Itália, 26–28, agosto de 2020, Volume 108, pp. 1287–1296 .

[252] Conforme identificado por Lindartos et al., com a indicação do seguinte trabalho: Zafar, M.R.; Khan, N.M. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv 2019, arXiv:1906.10263.

responsabilização, pois os autores estão interessados em investigar o adequado, correto e ético funcionamento do modelo, investigando diferentes maneiras de apresentar explicações sobre resultados de modelos de ML.

Além do LIME, outros métodos, como o Métodos de Explicações Contrastivas (*Contrastive Explanations Methods* -- CEM) ou Explicações Contrafactuais (*Contrafactual Explanations)*, também se propõem a oferecer explicações sobre resultados de modelos de ML. Este último se propõe a analisar o dado de entrada do modelo e verificar quais dados devem estar presentes para que um determinado resultado ocorra, apontando quais são as mudanças que podem ser realizadas para que o resultado possa ser alterado e seja diferente do que o esperado. As explicações contrafactuais, inclusive, têm ganhado destaque pelo seu potencial de fornecer elementos que fortalecem o exercício da autonomia de sujeitos, a transparência e a prestação de contas e responsabilização de modelos de ML (Mittelstadt et al., 2019; Wachter et al., 2018). Diante da sua importância, este tipo de explicação será mais bem analisado no próximo capítulo.

No mais, como apontam Linadartos et al. (2020), a literatura indica o desenvolvimento de explicações para diferentes tipos de modelos e formatos de dados, com foco especial para imagem e texto. Não é possível indicar a existência de um método melhor, pois cada um se destina a oferecer um tipo de explicação (ora uma explicação local, ora uma explicação global[253]) para modelos diferentes.

Ademais, a aplicação de métodos caixa-preta em algoritmos de aprendizado de máquina encontra um desafio relevante: o aprendizado implica na entrada constante de novos dados, que incrementam e evoluem seu aprendizado, ocasionando a sua modificação de forma imprevisível, não-programada e inescrutável (Desai; Kroll, 2017, pp. 41-42). A prestação de contas neste contexto deve envolver o conhecimento do problema que o algoritmo visa resolver, além do

[253] Uma explicação local é "focada em uma particular decisão, procurando por exemplo esclarecer a razão de uma particular imagem ser classificada como paisagem, ou uma particular pessoa ser considerada apreciadora de feijoada". Por sua vez, uma explicação global "tenta capturar o comportamento de todo o modelo explicado, por exemplo, afirmando que algumas observações são levadas em conta e outras não, ou indicando como as decisões se relacionam com padrões de entrada" (Maranhão et al., 2021, p. 146). Esta diferenciação se assemelha, portanto, à diferenciação feita por Wachter et al. (2017) sobre explicações *ex ante* e *ex post*, embora os autores não usem esta expressão (Sobre a escolha em não utilizar esta nomenclatura técnica, ver: Edwards; Veale, 2017)

que tentar compreender como que os dados de entrada influenciam os seus resultados (Desai; Kroll, 2017, pp. 53-54) [254].

Outro método interpretativo seriam os chamados modelos caixa-branca, que se referem, na realidade, a soluções para o desenvolvimento de modelos de aprendizado de máquina que sejam passíveis de terem o seu funcionamento compreendido por humanos, sendo a transparência inerente a ele. Contudo, diante de suas limitações[255] quanto à sua eficiência e a qualidade dos resultados de seu comportamento quando comparado aos modelos mais complexos e considerados caixa-preta, aqueles nem sempre são escolhidos para serem desenvolvidos (Linadartos et al., 2020, p. 17).

Como apontam Linadartos et al. (2020, p. 18-26), tem sido objeto de muita pesquisa o desenvolvimento de métodos de interpretação para evitar discriminações e assegurar justiça e igualdade em modelos de ML. Tais métodos buscam maneiras de interpretar o resultado para avaliar seu impacto social e ético, propondo métricas e avaliação de resultados discriminatórios, abusivos ou ilícitos, motivados por questões de gênero, etnia ou raça. Os diferentes métodos propostos buscam remover vieses dos dados de treinamento e das predições do modelo, além de buscarem treinar algoritmos que não sejam discriminatórios ou injustos, sendo certo que todos eles têm suas vantagens e limitações. Apesar das diferentes abordagens para tentar mitigar e evitar discriminações, ainda é uma área de pesquisa relativamente nova, e ainda não se sabe se tais soluções vão ter resultados na prática (Rovatsos et al., 2020, p. 23). O subcapítulo 3.3. deste capítulo abordará em maior profundidade este tópico.

2.2.2.b Auditoria

[254] Os autores identificam outras formas de *accountability:* (i) proteções em relação a whistleblowers, que são pessoas que tiveram acesso a informações sobre um determinado sistema e vem a público apresentar informações até então sigilosas sobre o algoritmo, tal como Christopher Wylie e Brittany Kaiser, no caso do Cambridge Analytica, (ii) criação de ações motivadas por interesse público.

[255] Isto porque tais métodos podem não ser adequados para testar problemas complexos, realizar predição de resultados, identificar como certos dados influenciam um determinado resultado, qual conjunto de *input-output* vai apresentar as informações necessárias para avaliar o algoritmo, e se um código conterá erros (Desai; Kroll, 2017, pp. 37-38).

A auditoria tem sido uma solução que tem ganhado certa aceitação para investigar resultados e predições algorítmicas, pois seria uma forma de assegurar uma transparência qualificada (*qualified transparency*) da caixa-preta, como indicado no início deste capítulo. Apesar de ter sido classificada aqui como um mecanismo que deve ocorrer após a implementação de um algoritmo de tomada de decisão, há quem defenda que a auditoria possa ocorrer antes e ao longo do desenvolvimento dos algoritmos, funcionando como uma espécie de conselho especializado (*expert board*) e de supervisão (Kaminski, 2019b, p. 1.540), o que também é viável e poderia ser uma boa prática a ser adotada.

A auditoria consistiria na reunião de um grupo de pessoas, tidas como "auditores confiáveis", internos ou externos à organização, para que possam avaliar a legitimidade e conformidade do algoritmo (Pasquale, 2015, p. 142). Foi neste sentido que o art. 20, § 2º, da LGPD, pareceu caminhar, ao prever que caso o controlador não ofereça informações relacionadas à decisão automatizada, invocando a proteção ao segredo comercial e industrial, a ANPD poderá realizar uma auditoria "para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais". Uma crítica a esta redação, e uma dúvida sobre como que a ANPD irá interpretar este artigo, é se a alegação de qualquer outro motivo para fornecimento de informações, como, inviabilidade ou complexidade técnica, poderá eximir a autoridade de conduzir uma auditoria, como destacado no capítulo 1 (Mulholland; Frajhof, 2019, p. 272).

A auditoria, no entanto, por ser restrita a um grupo de experts, não implica na disponibilização de explicações de uma maneira mais ampla, tampouco implica na abertura para um público mais geral sobre os seus resultados. A auditoria é um instrumento adequado para inspecionar potenciais discriminações e erros, se a decisão foi legítima e aderente aos comandos normativos (por exemplo, não se valendo de conteúdos protegidos em suas análises)[256], mas ela não necessariamente resulta no fornecimento de uma explicação sobre o resultado de uma decisão. Essa falta de abertura pode prejudicar a confiança do público em geral, incluído aqui a sociedade, pesquisadores, sociedade civil, legisladores, etc, sobre o comportamento

[256] Como determina o art. 7º, inciso I, da Lei de Cadastro Positivo (Lei n. 12.414/2011), que veda o uso de dados relacionados "à origem social e étnica, à saúde, à informação genética, ao sexo e às convicções políticas, religiosas e filosóficas" na análise de risco de crédito.

e resultados gerados pelos algoritmos (Zarsky, 2016, p. 130) [257], criando uma crença – que por vezes pode não refletir o resultado real – de que o mesmo age de maneira arbitrária.

Um exemplo de como a disponibilização destes resultados pode ser feita disso se refere à publicação do relatório produzido pelo Departamento Estadual de Nova Iorque de Serviços Financeiros (*New York State Department of Financial Services*) que investigou a acusação de discriminação de gênero do cartão de crédito emitido pela empresa Apple, em parceria com a instituição financeira Goldman Sachs[258]. O relatório, que concluiu que não houve discriminação intencional em relação às mulheres, ou um impacto em relação a este grupo, narra a maneira como a investigação foi conduzida, sem publicar dados sensíveis das empresas ou de seus clientes.[259]

Assim, formas de inspecionar algoritmos não se limitam apenas a esta proposta de sua abertura, do modelo utilizado, dos seus dados, da documentação e informações relevantes sobre o mesmo, para um grupo especializado. Além deste, Sandvig et al. (2014) propõem cinco tipos de auditorias para inspecionar plataformas digitais. Os autores desenvolvem seus métodos inspirados em pesquisas de Estudos de Auditorias criadas na década de 70 nos EUA, para investigar práticas discriminatórias no contexto imobiliário, que ocorriam por meio de entrevistas e grupos de controle. As auditorias são realizadas em plataformas de Internet, em especial, aquelas intermediárias que possuem um grande repositório de dados, como Google, YouTube, Facebook, para verificar se estas "estão conduzindo discriminação danosa por classe, gênero, e investigar as consequências

[257] Neste sentido: "(…) lack of transparency leads affected persons to speculate that algorithmic decision making is arbitrary. Opacity further hinders their ability to question the outcome and understand the process. Here too, transparency measures—such as disclosing which factors were used or the rate of statistical error in predicting the outcome— could be adopted to mitigate these concerns" (Zarsky, 2016, p. 130)

[258] Um empresário acusou haver uma discriminação de gênero em relação às mulheres, após ele receber um valor de crédito 20 vezes superior ao da sua mulher. Ver em: Goldman Sachs é investigado por suposta discriminação de gênero do Apple Card. O globo, 10 de novembro de 2019. Disponível em: <https://oglobo.globo.com/economia/tecnologia/goldman-sachs-investigado-por-suposta-discriminacao-de-genero-do-apple-card-24073289>. Acessado em 01.01.2021

[259] FARELL, Greg; NASIRIPOUR, Shahien. Goldman Cleared of Bias in New York Review of Apple Card. *Bloomberg*, 23 de março de 2021. Disponível em: https://www.bloomberg.com/news/articles/2021-03-23/goldman-didn-t-discriminate-with-apple-card-n-y-regulator-says. Acessado em 01.12.2021.

da operação de seus algoritmos em outras preocupações normativas[260]" (Sandvig et al., 2014, p. 6).[261]

São cinco tipos de auditorias propostas pelos autores: auditoria de código (*code audit - algorithm transparency*), auditoria de usuário não-invasiva (*noninvasive user audit*), auditoria por raspagem (*scraping audit*), auditorias por fantoche (*sock puppet audit*, auditoria) colaborativa e coletiva (*crowdsourced audit/collaborative audit*). A primeira delas se assemelha à ideia de Pasquale, e depende da abertura do algoritmo para investigação por um grupo de pessoas. No entanto, os autores reconhecem que as chances de empresas abrirem seus algoritmos é mínima. Além disso, reconhecem a dificuldade deste ser compreendido pela mera leitura do código, e da insuficiência em avaliar os seus resultados e seu funcionamento sem o acesso aos dados que foram utilizados para treinamento, considerando a influência que estes têm nos resultados gerados.

A auditoria não invasiva se baseia na utilização de dados de interação dos usuários com estas plataformas (i.e. históricos de busca e palavras-chave utilizadas), que seriam compartilhados com pesquisadores, que poderiam fazer inferências sobre como o algoritmo se comporta. Este tipo de auditoria não depende da plataforma, o que facilitaria a condução da auditoria. Entretanto, há algumas desvantagens: não há uma "manipulação" ou grupo de controle dos dados; é difícil estabelecer uma causalidade entre os resultados, pois o comportamento é altamente dependente de cada usuário, podendo não consistir no estabelecimento de um padrão, além do risco de que os relatos sejam enviesados por serem pessoais (Sandvig ar al., 2014, p. 11). Um exemplo de aplicação de auditoria não invasiva foi quando houve o relato de que o algoritmo do YouTube acabava recomendando vídeos de crianças dançando ou fazendo yoga seguidamente, animando comentários de pedófilos que se aproveitam desta recomendação[262] (Silveira; Silva, 2020, p. 5).

Este exemplo demonstra a importância de realização de auditorias, e a comunicação dos seus resultados, como uma forma de melhor informar o debate

[260] Aqui os autores estão se referindo a leis que proíbem a raspagem de dados de plataformas digitais, assim como guias éticos para condução de pesquisa acadêmicas.
[261] Tradução livre de: "whether they are conducting harmful discrimination by class, race, gender, and to investigate the operation of their algorithms consequences on other normative concerns"
[262] ORPHANIDES, K.G. On YouTube, a network of paedophiles is hiding in plain sight. *Wired*, 28 de fev. de 2019. Disponível em: <https://www.wired.co.uk/article/youtube-pedophile-videos-advertising> Acessado em 15.05.2021.

público sobre os problemas e dificuldades do universo dos algoritmos, levando os responsáveis a tomarem medidas para enfrentar este comportamento. Ainda, a transparência neste caso acaba incentivando que outras entidades realizem inspeções e testes para verificar seus próprios algoritmos.

A auditoria de raspagem de dados se vale de uma técnica em que um programa de computador extrai dados de um *site* na internet, ou de uma API[263], e os armazena localmente (ou em serviço de nuvem) para serem utilizados para uma certa finalidade. Isto, contudo, pode ser uma atividade proibida pelos termos de uso da plataforma[264], e pode até mesmo ser considerado ilícito em alguns países[265]. Um exemplo de raspagem de dados foi realizado pela Amazon quando a empresa desenvolveu um algoritmo para recrutar profissionais. Para treiná-lo, foi feita uma raspagem dos dados dos currículos de seus funcionários dos últimos 10 anos, em que se descobriu que o mesmo tinha um viés de gênero, tendo preferência em selecionar candidatos homens em relação às mulheres.[266]

Aindaa, auditoria de fantoche (*sock puppet*) consistiria na criação de contas para os pesquisadores nas plataformas para simular interações, criando um tráfego no sistema, e manipulando o mesmo. Contudo, isto poderia levantar os mesmos questionamentos legais da raspagem de dados e as limitações impostas pelos termos de uso. Os autores até sugerem a contratação de pessoas que já possuem conta na plataforma para realizar os testes, apesar de reconhecer que ainda assim possam ocorrer os mesmos problemas listados na auditoria não-invasiva. Exemplos de aplicação desta auditoria seriam as denúncias de usuários do Airbnb, que criaram contas falsas com certas características demográficas para comprovar a ocorrência de discriminação racial e de outras etnias cometidas pelos proprietários dos imóveis[267] (Silveira; Silva, 2020, p. 6).

---

[263] De forma simplista, *Application Programming Interface* (API), é um algoritmo que permite que dois computadores ou dois sistemas diferentes de comuniquem e troquem informações.

[264] O YouTube e o Facebook, por exemplo, proíbem em seus termos de uso meios automatizados para acessar dados da sua plataforma. O YouTube nomeia as técnicas utilizadas para tanto, entre elas o scrapping, e prevê duas exceções para quando esta poderá ocorrer: aplicada aos motores de busca públicos, nos termos do arquivo robots.txt do YouTube, e com a autorização prévia e por escrito da plataforma.

[265] Nos EUA, a violação seria ao *US Computer Fraud and Abuse Act* (CFAA) que criminaliza acesso não autorizado a qualquer computador, e que motivou a condenação de pessoas a pena de prisão pela atividade de raspagem de dados (Sandvig, 2014, p. 12)

[266] DASTIN, Jeffrey. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 10 de out. de 2018..Disponível em: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Acessado em 15.05.2021.

[267] GUYNN, Jessica. Airbnb's 'belong anywhere' undercut by bias complaints. *USA Today*, 06 de

Por fim, a última auditoria proposta por Sandvig et al. seria o que eles chamam de colaborativa e coletiva, em que pesquisadores poderiam contratar usuários para realizar testes. No entanto, alguns problemas poderiam ser enfrentados: nem todas as plataformas viabilizam oportunidades para realização de testes, tal como ocorre com a Amazon[268], e não seriam gerados dados suficientes de maneira rápida suficiente. O lado positivo seria que este tipo de auditoria não seria classificado como uma atividade de raspagem de dados, tampouco seriam criadas contas falsas, embora mesmo assim questões éticas possam ser levantadas, visto que seriam criados dados falsos que poderiam congestionar e interferir no sistema. Apesar destes questionamentos, esta é a auditoria considerada como a mais promissora das cinco pelos autores.

Estas alternativas para auditar algoritmos podem ser saídas para a típica defesa das empresas em restringir o acesso aos algoritmos e dados utilizados. Para viabilizar algumas dessas soluções será necessário estabelecer propostas normativas, ou iniciativas de fomento, para proceder com a remuneração tanto dos pesquisadores que realizam a auditoria, quanto os seus voluntários. Quanto a esse último ponto, merece ser ressaltado que no Brasil não é permitida a remuneração de voluntários de pesquisa[269], diferente do que ocorre em outros países, como nos EUA.

Outro entrave para estas propostas de auditorias é que elas nem sempre serão passíveis de serem aplicadas a todos os contextos em que algoritmos são utilizados, pois apenas se aplicam a plataformas digitais que prestam serviços ou produtos a seus usuários. Isto significa que estas auditorias não podem ser aplicadas a algoritmos proprietários e sigilosos, como é o caso daquele utilizado pelo Governo Holandês no contexto de fraudes de seguridade social. Isto é um argumento de reforço para a exigência da transparência e prestação de contas e

jun. de 2016. Disponível em: <https://www.usatoday.com/story/tech/news/2016/06/06/airbnb-openair-diversity-racism- airbnb-connect/85490536/.> Acesso em 15.05.2021.

[268] Eles indicam que a Amazon disponibiliza o *Amazon Mechanical Turk*, que permite enormes quantidades de usuários para usar este serviço, mas que isto é uma iniciativa que não se repete em outras plataformas.

[269] Nos termos da Resolução do Conselho Nacional de Saúde (CNS) n. 466, de 12 de dezembro de 2012, que tinha como objetivo apresentar revisões necessárias das áreas tecnocientíficas e ética, conforme indicado pela Resolução do CNP n. 196, de 10 de outubro de 1996. Sobre o assunto, ver: ALBUQUERQUE, Aline; BARBOZA, Heloisa Helena. Remuneração dos participantes de pesquisas clínicas: considerações à luz da Constituição. *Revista Bioética* [online]. 2016, v. 24, n. 1, pp. 29-36. Disponível em: <https://doi.org/10.1590/1983-80422016241103>. Acessado em 28.06.2021.

responsabilização de algoritmos de tomada de decisão, e a necessidade de que as pessoas sejam informadas sobre quando estão sujeitas ao mesmo[270], para que possam ser exigidas garantias em relação ao responsável pelo uso, tal como as salvaguardas envolvidas no direito à explicação.

Desta forma, algumas soluções podem ser desenhadas a partir desse cenário. Mesmo que o segredo empresarial se imponha e uma entidade se negue a abrir seu código ou qualquer outra informação sobre o seu algoritmo, será possível que esta seja submetida a uma auditoria no formato previsto por Pasquale (2015), de um grupo seleto e especializado, por exemplo, tal como determinado pela LGPD. Uma entidade pode, no entanto, resolver compartilhar alguns elementos do seu algoritmo (como os dados utilizados para treinamento, metodologia preditiva e especificações do objetivo do algoritmo) para que pesquisadores independentes realizem pesquisas, e trabalhem em conjunto para que eventuais discriminações ou erros cometidos pelo algoritmo sejam corrigidos (Obermeyer, 2019). Ou seja, alguns dados selecionados são compartilhados para fins estritamente de investigação e análise do algoritmo, e outros podem ser disponibilizados para avaliação de terceiros. Por fim, os métodos apresentados por Sandvig et al. (2014) podem ser aplicados no contexto de plataformas digitais, ou se houver outra maneira de acessar os dados, como por meio de APIs. É curioso, no entanto, que empresas privadas estejam dispostas a organizar competições abertas para otimizar seus algoritmos, tal como realizado pelo Netflix, em 2009,[271] ou contratar empresas especializadas em "hackear" sistemas, como a Hackerone[272], Bugcrowd[273] ou Intigriti[274], para descobrirem falhas de segurança, mas têm pouca disposição em abrir estas informações para outros interesses que não sejam os seus próprios.

[270] Recentemente o governo do Reino Unido publicou uma diretriz sobre transparência algorítmica que determina que entes públicos devem explicar quando um algoritmo está sendo utilizado, porque ele foi utilizado, se o mesmo atingiu seu objetivo, além de ser necessário revelar a arquitetura do mesmo. Ver em: MILMO, Dan. Working of algorithms used in government decision-making to be revealed. *The Guardian*, 25 de nov., de 2021. Disponível em: <https://www.theguardian.com/technology/2021/nov/29/working-of-algorithms-used-in-government-decision-making-to-be-revealed>. Acessado em 12.02.2022.

[271] Tal como feito pelo Netflix, em 2009, que levantou diversos questionamento sobre a efetividade da técnica de anonimização. Isto porque na competição publicada na internet para melhorar seu algoritmo, foi possível fazer a reidentificação dos dados, que estavam supostamente anonimizados, com a junção de outra base de dados disponíveis gratuitamente na internet. Sobre o tema, ver: BIONI, 2029, p. 73.

[272] Ver em: https://www.hackerone.com

[273] Ver em: https://www.bugcrowd.com

[274] Ver em: https://www.intigriti.com

Seja em qualquer um dos casos, é importante que haja a comunicação do seu resultado e conclusões, como uma forma de melhor informar o debate público sobre os problemas e dificuldades do universo de IA, bem como viabilizar uma espécie de revisão por pares (*peer review*), como pesquisadores ou representantes da sociedade civil, para questionarem o uso destes sistemas (Reisman et al., 2018, p. 19). Esta publicidade, como já reiterado diversas vezes ao longo deste trabalho, contribui para a construção sobre a confiança nos resultados e uso destes algoritmos de tomada de decisão, principalmente para evidenciar as modificações implementadas para corrigir os problemas detectados[275].

2.2.2.c
Documentação

A documentação do desenvolvimento de um sistema é fundamental para que outras pessoas, que não apenas os seus desenvolvedores, compreendam o seu funcionamento. Seu objetivo, portanto, é redigir em linguagem humana as escolhas das regras, dos métodos, das funcionalidades e das variáveis do(s) algoritmo(s) que compõe(m) o sistema. A mera apresentação do código de um determinado algoritmo, o que muitas vezes não irá ocorrer diante da proteção assegurada pelo segredo empresarial ou propriedade intelectual, não significa que terceiros, inclusive profissionais com expertise técnica, conseguirão interpretar e compreender seu funcionamento (Frajhof, 2021, p. 485). Assim, a documentação se coloca como uma salvaguarda importante tanto sob a perspectiva da prestação de contas, como da transparência. No entanto, optou-se por colocá-la neste subitem com o intuito de organizar os tipos de documentos que podem ser apresentados para fins de prestação de contas.

É comum que não haja padronização de como uma documentação deve ser feita. O objetivo da propositura de um padrão para documentação é estabelecer as informações básicas sobre o algoritmo, para que o processo de tomada de decisão, desenvolvimento e implementação estejam disponíveis para outros desenvolvedores, pesquisadores ou terceiros interessados em auditar, verificar e analisar o algoritmo. Esta padronização poderia auxiliar na identificação de vieses

[275] No caso do art. 20, § 2º, da LGPD, a própria ANPD poderá aplicar sanções e exigir esta *accountability* como recomendação de sua auditoria.

de quando há a coleta de dados e a observação de considerações éticas (Rovatsos, 2020, p. 31).

Empresas como Microsoft Research e Google já propuseram uma documentação padrão no que se refere à coleta de dados para treinar modelos algorítmicos, chamado de *datasheets for datasets,*[276] a fim de mapear e identificar potenciais vieses. Um complemento a este estudo, elaborado por pesquisadores do Google, são os cartões modelo (para comunicar) (*model cards* (*for model reporting*)) (Mitchell et al., 2019), para reportar medidas de comportamento e usos pretendidos com modelos de ML. Seu objetivo é estabelecer um padrão ético a ser observados pelos responsáveis pelo desenvolvimento de modelos de ML, além de visar que pessoas não técnicas possam compreender os seus resultados (Mitchell et al., 2019, p. 221). Os *model cards* buscam trazer informações sobre o comportamento do modelo, os casos para o qual o mesmo foi desenvolvido, o contexto que será analisado, potenciais riscos de aplicação, métricas para avaliar vieses, discriminações, entre outras (Mitchell et al., 2019, p. 220). A proposta é permitir maior transparência sobre como os modelos de aprendizado de máquina funcionam, como que os usuários compreendem o que o modelo pode ou não fazer, e os tipos de erros, discriminações e injustiças que podem ocorrer (Mitchell et al., 2019, p. 221).

A proposta é interessante, pois tanto o conteúdo, como a forma que as informações são apresentadas são relevantes. De acordo com Mitchell et al., o *model card* deve seguir um padrão de apresentação, com a reunião das informações em um documento que deve ter entre uma e duas páginas, e não dispensam a documentação do conjunto de dados. A elaboração desta documentação prescinde a implementação, testagem e uso do modelo, ou seja, ela é elaborada após a sua aplicação e apresentação de resultados. A tabela abaixo indica as informações que devem ser resumidas e reunidas sobre o modelo de ML para o *model card* (Mitchell et al., 2019, p. 222):

[276] Ver em: Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*; Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Model Card

- **Model Details**. Basic information about the model
  - Person or organization developing model
  - Model Date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, [or others listed in Section 4.3]
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to respect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Tabela 1 - Figura extraída de Mitchell et al (2019) sumarizando as seções e sugestões do conteúdo do *model card*

O exemplo acima é apenas uma proposta sobre os detalhes do modelo que são considerados relevantes, e não pretende ser completa ou exaustiva. A elaboração deste documento estimula o questionamento dos impactos éticos, além

de auxiliar tanto os responsáveis pelo desenvolvimento do modelo, como demais agentes externos, na avaliação da adequação dos algoritmos de ML a marcos normativos. Os próprios autores reconhecem que cartão modelo é apenas uma das muitas ferramentas para transparência, que pode incluir auditorias, aplicações de metodologias técnicas e até mesmo comentários de usuários (Mitchell et al., 2019, p. 228).

O próprio exemplo trazido no *model card* pelos autores chama atenção para a importância de abrir essas informações para realizar alguns questionamentos. De acordo com a documentação apresentada, a intenção de uso do modelo de ML seria em "aplicações divertidas, como para a criação de desenhos felizes baseados em imagens reais; aplicações de acessibilidade para fornecer detalhes para pessoas cegas; ou aplicações que reconhecem automaticamente sorrisos em fotos"[277] (Mitchell et al., 2019, p. 227). Chama atenção as diferentes naturezas de finalidades do desenvolvimento: aplicações divertidas para crianças e apoio para o desenvolvimento de aplicações para pessoas cegas. Ambas parecem estar em espectros distantes de interesses.

Além disso, embora o próprio *model card* deste exemplo faça uma ressalva ao modelo, de que o mesmo não "captura raça ou cor de pele, que tem sido reportado como uma fonte desproporcional de erros"[278] (Mitchell et al., 2019, p. 227), o uso de um banco de dados de celebridades para realizar o treinamento deste modelo levanta uma série de dúvidas éticas quanto ao seu uso. O conjunto de pessoas que aparecem nas imagens em que o modelo será treinado são altamente preocupadas com a sua imagem física, extremamente vaidosas, e possuem o que pode ser considerado como o sorriso "perfeito" (possuem todos os dentes, com uma cor artificialmente branca, que tampouco é natural).[279] Certamente o padrão das celebridades não corresponde a outros tipos de sorrisos, tal quais aqueles que não possuem todos os dentes, ou com dentes de diferentes tons, que fazem uso de aparelho, ou sorriso de pessoas idosas.[280] Portanto, até mesmo o próprio modelo

[277] Tradução livre de: "(...) fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos".
[278] Tradução livre de: Does not capture race or skin type, which has been reported as a source of disproportionate errors"
[279] Conforme pode ser verificado a partir de alguns exemplos disponibilizados do dataset: ver em: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.
[280] Vale um agradecimento especial à professora Clarisse de Souza, que ao apresentar o referido artigo teceu estas críticas valiosas sobre o modelo apresentado pelos autores, em sua disciplina

apresentado pelos autores está sujeito a questionamentos éticos, o que revela os benefícios deste tipo de documentação para realizar estas análises.

Assim, as informações que constam no *model card* de fato podem beneficiar este nivelamento de conhecimento entre pessoas não técnicas e técnicas, e permitem que sejam realizados diversos questionamentos sobre a adequação da sua finalidade, o contexto que o mesmo será aplicado, suas limitações e a avaliação sobre suas taxas de acerto e de erro. A maneira como estas informações são apresentadas facilitam a compreensão do modelo de ML, misturando visualizações gráficas, texto em uma linguagem acessível, e com informações sucintas, mas relevantes.

A documentação apresentada desta forma é interessante por dois motivos principais. O primeiro é que devem ser fornecidas informações pertinentes para avaliar o modelo de aprendizado de máquina, sem necessariamente exigir a abertura de informações que estariam protegidas pelo segredo empresarial e/ou propriedade intelectual. O segundo é que exige também informações e métricas sobre os resultados dos testes do modelo, oferecendo insumos para apoiar uma investigação sobre os motivos dos resultados. Ainda, como o *model card* pretende avaliar os resultados e o comportamento do modelo de ML, diante da apresentação de resultados insatisfatórios, tanto em relação à sua qualidade, quanto em termos éticos, a decisão do seu uso pode ser postergada até que resultados melhores se façam presentes, podendo até mesmo ocasionar a rejeição do seu uso. Embora esta documentação tenha sido desenvolvida para o contexto de ML, ela ainda poderia ser utilizada por algoritmos mais simples, que não se valem de ML, tal qual uma árvore de decisão (Frajhof, 2021, p. 486).

Desta forma, especificamente sobre a "explicabilidade", tais instrumentos devem exigir que sejam documentados os seguintes aspectos do uso de algoritmos de tomada de decisão: (i) apresentação de justificativas que motivaram o uso de sistemas de IA para aquele determinado contexto; (ii) indicação de quais tipos de métodos e modelos de interpretação e explicação do resultado do modelo de ML foram considerados e implementados; (iii) registro da origem dos dados utilizados para o treinamento e teste do modelo, (iv) indicação de como foi realizado o pré-processamento dos dados, e como potenciais discriminações foram consideradas; (v) apontamento de quais foram os critérios para a escolha do modelo em relação à

Linguística Computacional Interativa, oferecida pelo Departamento de Informática da PUC-Rio, em 27.06.2019.

capacidade de interpretação e explicação dos seus resultados, e as ferramentas utilizadas para tanto; (vi) apresentação dos resultados da aprendizagem e testagem do modelo; e (vii) indicação de como e qual tipo de explicação está sendo fornecida para o titular de dados[281] (Frajhof, 2021, p. 483)[282].

Por fim, antes de ir em frente para o próximo subcapítulo, cabe fazer um breve fechamento do presente item. A apresentação destas ferramentas e documentos visa auxiliar na prestação de contas sobre o desenvolvimento de algoritmos de tomada de decisão e os seus resultados, com o intuito de tentar revelar informações úteis que permitam que uma série de atores possam compreender e contestar decisões algorítmicas. Tendo em vista que a própria técnica vai inviabilizar a abertura e inteligibilidade da caixa-preta, cabe ao Direito apresentar meios para exigir, demandar e regular como e quais informações devem ser apresentadas.

2.3
O Princípio da Justiça e igualdade *(Fairness):*

> "desigualdade e vieses não são achados em único lugar, como um bug que pode ser localizado e consertado. Essas questões são sistêmicas"[283]

Os vieses e discriminações ocorridas pelo uso da tecnologia têm alcançado o debate público, principalmente por meio de publicação de notícias em meios jornalísticos, que jogam luz para a ocorrência destas discriminações (Silveira; Silva, 2020, p. 4; Silva, 2019). É até mesmo frustrante a ausência de dados empíricos para apoiar estas reclamações (Rovatsos et al., 2020, p. 8). Isto ocorre, principalmente, pela manutenção do sigilo do uso e dos códigos algorítmicos de plataformas e sistemas de IA. Mesmo diante da ausência de uma pesquisa mais estruturada e

[281] INFORMATION COMMISSIONER'S OFFICE (ICO); THE ALAN TURING INSTITUTE. *Explaining decisions made with AI*. Draft Guidance for Consultation. Part 3: What explaining AI means for your organization. Disponível em: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-3-what-explaining-ai-means-for-your-organisation/documentation/. Acessado em 12.01.2022.

[282] Na primeira publicação sobre o tema (Frajhof, 2021), havia indicando que as respostas a estas perguntas deveriam ser apresentadas nos Códigos de Boas Práticas e Governança e nos Códigos de Conduta. No entanto, atualmente acredito que a documentação parece ser o local mais adequado para apresentá-las, sendo que os Códigos de Boas Práticas podem exigir que este tipo de documentação seja elaborado.

[283] Tradução livre de: "(...)inequity and bias are not to be found in a single place, like a bug that can be located and fixed. These issues are systemic" (West et al., 2019, p. 9).

desenvolvida sobre o tema, a exposição destas discriminações causadas por algoritmos a populações historicamente marginalizadas tem instado que empresas privadas se manifestem sobre estas ocorrências, mostrando a relevância da pressão pública e da imprensa em chamar a atenção (Silveira; Silva, 2020, p. 8; Zarsky, 2016).

Não há, ainda, uma definição estanque ou precisa do que significa um resultado justo (*fair)* e equitativos (Rovatsos et al., 2020, p. 8; Wachter et al., 2020). Um resultado que não seja considerado como justo é comumente atribuído como uma "discriminação algorítmica" (*algorithm bias*), um termo que também não possui um conceito estreito e bem definido, que pode ser compreendido como um tratamento injusto dado a um grupo (i.e. minoria étnica, gênero ou tipo de trabalhador) causado pelo uso de um algoritmo de tomada de decisão (Rovatsos et al., 2020, p. 11).

Bianca Kremer N. Corrêa aponta que o termo "discriminação algorítmica" pode ser compreendido como a incorporação de visões de mundo à tecnologia, que acabam reproduzindo estereótipos, vieses e preconceitos existentes socialmente, atribuindo um peso desproporcional em detrimento ou em favorecimento de uma pessoa (Corrêa, 2021, p. 136). No caso de aprendizado de máquina, estas discriminações podem ocorrer por uma série de motivos, sendo os casos mais bem mapeados aqueles que envolvem discriminações ou vieses oriundos de banco de dados utilizados para treinamento, tal como as classificações absurdas e preconceituosas feita pelo serviço do Google Fotos, em 2015, que rotulava pessoas negras como "gorilas"[284], ou no oferecimento de entregas rápidas pela empresa Amazon, pelo seu serviço Prime, que ocorrem no mesmo dia em CEPs considerados brancos, excluindo distritos predominantemente negros.[285]

Estes são apenas dois exemplos, entre tantos outros já mapeados[286], que evidenciam o mito sobre a objetividade, neutralidade, racionalidade e

[284] KASPERKEVIC, Jana. Google says sorry for racist auto-tag in photo app. *The Guardian*, 1º de julho de 2015. Disponível em: <https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app> Acessado em 11.05.2021.

[285] INGOLD, David; SOPER, Spencer. Amazon Doesn't Consider the Race of Its Customers. Should It?. Disponível em: https://www.bloomberg.com/graphics/2016-amazon-same-day/ Acesso em 11.05.2021.

[286] Por exemplo: ARAÚJO, Camila Souza; JUNIOR, Wagner Meira; ALMEIDA, Virgilio. I*dentifying Stereotypes in the Online Perception of Physical Attractiveness.* 8 de agosto de 2016. Disponível em <https://arxiv.org/pdf/1608.02499v1.pdf> Acessado em 04.09.2017; SLAM-CALISKAN, Aylin; BRYSON, Joanna J.; NARAYANAN, Arvind. S*emantics derived automatically from language corpora necessarily contain human biases*. Science, Vol. 356, Issue

imparcialidade dos algoritmos (Kitchin, 2014). Estas notícias e pesquisas têm demonstrado que os vieses existentes na cultura humana são inevitavelmente replicados nos algoritmos, pois estes acabam reproduzindo, em larga escala, preconceitos e estereótipos (O'Neil, 2016) que repercutem negativamente na mediação entre o humano e a máquina. Assim como nós humanos estamos sujeitos a heurísticas e a vieses em nossas tomadas de decisão,[287] os algoritmos também estão (Mulholland; Frajhof, 2019, p. 267).

Tarcízio Silva se vale do conceito "microagressões raciais"[288] aplicado ao contexto das plataformas digitais para demonstrar como estas ocorrem de maneiras sutis, podem ser conscientes ou inconscientes, são sistêmicas e cotidianas. A partir de denúncias documentadas e demonstradas, Silva mapeia doze casos de racismo algorítmico que ocorreram em diversas aplicações que se valem de técnicas de IA, tal como oferecimento de anúncios em plataformas digitais, buscadores de imagens, processamento de linguagem natural, visão computacional e robôs (*bots*) conversacionais (Silva, 2019). São indicados alguns tipos de microagressões racistas [289], como suposição de criminalidade, negação de realidades raciais/democracia racial, suposição de inferioridade intelectual, patologização de valores culturais, exotização e exclusão ou isolamento, e como todas elas ocorreram no ambiente digital. Por exemplo, o caso da imagem do Gorila descrita acima é classificado como uma microagressão de negação de cidadania; o caso do *software* que não reconhece rostos negros (Buolamwini; Gebut, 2018) é classificado como uma negação de cidadania, exclusão e isolamento. Tendo em vista que o direito à explicação também se coloca como um direito para revelar resultados discriminatórios, manter-se vigilante sobre o potencial que estes sistemas têm de causar discriminações é fundamental.

6334, 14 de abril de 2017, p. 183-186. Este artigo também está disponível em <https://www.princeton.edu/~aylinc/papers/caliskan-islam_semantics.pdf>. Acessado em 06.09.2017; How I'm fighting bias in algorithms. Disponível em <https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-26257> Acessado em 04.08.2017.

[287] Ver: SCOTT, Plous. *The Psychology of Judgement and Decision Making*. McGraw-Hill Education, 1993.

[288] Este conceito é proposto pelo psiquiatra Chester Pierce para descrever mecanismos ofensivos de grupos opressores que se assemelhavam a práticas psiquiátrica. Como indica Tarcízio Silva, as situações do cotidiano que evidenciam microagressões racistas é acrescida de outros relatos documentados, experiementos e organizados em tipologias (Silva, 2019).

[289] O autor também se vale de outra classificação a partir de Levchak, que classifica estas ocorriencias em macro e microagressões, se estas foram causadas de maneira intencional ou não intenrcional, e se são consideradas racismo encoberto ou explícito.

Para evitar que resultados não sejam considerados justos e equitativos, os princípios da igualdade e da não discriminação devem ser observados ao longo do desenvolvimento de um algoritmo de tomada de decisão. Por exemplo, se um recrutador desenvolve um algoritmo para analisar e selecionar currículos para uma entrevista, fixando como parâmetro que o/a candidata/o tenha inglês fluente (compreensão, conversação, leitura e escrita) isto certamente vai excluir uma parcela de pessoas que não tiveram a mesma oportunidade de estudar em escolas que não tinham em sua grade aulas de inglês, ou de frequentar cursos de línguas (Rovatsos et al., 2020, p. 11), o que é uma realidade sócia para a população brasileira mais vulnerável. Neste caso, embora a análise dos currículos dependesse de um critério "igualitário", isto acaba favorecendo pessoas com um determinado tipo de educação, que é mais privilegiada, deixando de fora uma parcela de potenciais candidatos/as.

Até mesmo quando dados sensíveis, como aqueles relacionados à cor, etnia, raça, gênero, por exemplo, são afastados do conjunto de dados, há um impacto no aprendizado do algoritmo e em seus resultados. Considerando que mulheres possuem uma taxa de reincidência de crime menor que homens, caso gênero seja um dado excluído, é provável que um algoritmo que realiza esta predição atribua taxas de reincidências desproporcionais (Rovastos et al., 2020, p. 12). Da mesma forma, quando o Google propôs corrigir o erro da marcação das fotos de gorila removendo estas fotos da sua base de dados[290], isto não resolve o problema estrutural do racismo neste caso, pois invisibilizar não seria uma solução.

Da mesma maneira, mesmo que os dados não sejam tratados para evitar estes vieses e discriminações, pode haver uma forte correlação entre dados com atributos sensíveis como, por exemplo, raça e residência[291], nome e raça[292], profissão e gênero[293]. Um estudo publicado em 2019 na renomada revista Science, identificou que o uso de um algoritmo de ML utilizado por entidades de saúde para predizer o risco de cuidado necessário de um determinado paciente, e direcionar

[290] VINCENT, James. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. *The Verge*, 12 de janeiro de 2018. Disponível em: < https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>

[291] Como visto no caso da entrega da Amazon.

[292] FAWCETT, Amanda. Understanding racial bias in machine learning algorithms. Educative, 08 de junho de 2020. Disponível em: <https://www.educative.io/blog/racial-bias-machine-learning-algorithms#wrap-up> Acessado em 11.05.2021.

[293] SLAM-CALISKAN et al, 2017.

mais recursos para realizar esta assistência médica e garantir uma assistência de qualidade, continha um viés que discriminava pessoas negras (Obermeyer et al., 2019).

Os pesquisadores tiveram uma oportunidade única de ter acesso aos dados utilizados pelo algoritmo, seus resultados e as predições feitas pelo mesmo, permitindo um acesso mais amplo que viabilizava a compreensão de como o mesmo funcionava. Ao ter acesso a estas informações, e a função objetiva do algoritmo, os pesquisadores conseguiram identificar a disparidade racial e quando a mesma surgia. Importante notar que o algoritmo analisava um conjunto de dados envolvendo informações sobre o uso do seguro pela pessoa ao longo dos anos, dados demográficos (ano, sexo), tipo de seguro, diagnósticos, código de procedimentos, medicações, detalhes de custos, e excluía especificamente o dado de raça. Para proceder com a pesquisa, os pesquisadores tiveram acesso à lista de pessoas submetidas a este programa, identificando pacientes negros e brancos a partir da autodeclaração de raça existente nos registros do hospital.

Aos pacientes eram atribuídos uma pontuação de risco de acordo com estes dados coletados ao longo dos anos. Pacientes que possuíam um risco acima de 97% eram automaticamente cadastrados no programa, e os que possuíam um risco acima de 55% eram indicados a um médico de assistência primária, que analisaria se estes pacientes deveriam ou não participar do programa. Ao calcular o nível de saúde das pessoas com base em raça, identificou-se que pessoas negras tinham 26,3% mais doenças que pessoas brancas contudo, a despeito disso, pessoas brancas tinham maior indicação para serem incluídas no programa.

Isto ocorria por conta dos dados que o algoritmo utilizou para treinamento, fortemente baseado nos custos gerados para o plano de saúde. Os custos de pessoas negras com saúde eram inferiores aos custos de pessoas brancas por ano (em média $1.801 por ano). Os tipos de custos de pessoas negras eram diferentes (mais gastos de emergência e diálise do que intervenção cirúrgica). Ou seja, mesmo que pessoas negras tivessem mais comorbidades, elas geravam menos gastos para o plano de saúde, e acabavam sendo preteridas pelo programa.

Os autores resolveram verificar porque isto ocorria. O primeiro motivo identificado seriam as barreiras de acesso a cuidados médicos motivados por condições sócio-econômicas, como a distância até a assistência médica, necessidade de transporte, trabalho ou cuidado com crianças que impedem a

locomoção, e até mesmo a falta de conhecimento sobre a necessidade de procurar ajuda. A segunda estaria motivada pela discriminação baseada na relação médico/paciente, que afetava o tratamento ofertado aos mesmos.

Os autores do estudo reconhecem que o uso desta métrica até seria razoável, já que o próprio objetivo do programa é reduzir custos ao monitorar pacientes. Contudo, para tentar reverter estes resultados, os pesquisadores treinaram novamente o algoritmo, alterando o seu objetivo para que o mesmo fizesse uma predição com base em doenças crônicas, ao invés de custos com saúde. Isto resultou na alteração da pontuação de pacientes negros, que subiriam de 14,1% para 26,7% dentro da margem de risco, classificando-os como candidatos a serem considerados no programa. Isto motivou que a empresa proprietária do algoritmo o alterasse. Esta mudança fez com que o viés em relação a pessoas negras diminuísse em até 84%. A lição apontada pelos autores para evitar estes resultados discriminatórios é a importância em ter pessoas que detenham conhecimento sobre o domínio de aplicação, saibam identificar e extrair quais são os dados relevantes, e tenham a capacidade de experimentar.

Isto também evidencia que, ainda que existam técnicas para mitigar ou evitar discriminações, com o desenvolvimento de algoritmos considerados justos e equitativos, como anti-classificação[294], classificação ou resultado por erro de paridade[295] e calibração[296] (Rovatsos, 2020, p. 12), que foram até mesmo aplicadas pelos pesquisadores deste estudo, fica evidente que os resultados produzidos pelo algoritmo refletem problemas sociais existentes "fora do seu constructo".[297]

A dificuldade em evitar a reprodução destes vieses nos algoritmos não se limita apenas ao desafio de definir estatisticamente, quantitativamente, metricamente ou computacionalmente o que é considerado uma discriminação algorítmica, e como tecnicamente evitá-la, mas se estende a compreender quem são as pessoas por trás da sua elaboração. As decisões sobre o desenvolvimento do

[294] O algoritmo será considerado como justo/equitativo se o mesmo não utilizar dados sensíveis como proxies ou características que permitem fazer a inferência destas características.
[295] Deve ser garantido tratamento igualitário e equitativo a grupos protegidos, seja na proporção de resultados positivos ou a erros cometidos.
[296] Esta técnica foi aplicada por Obermeyer et al., (2019), em que o score atribuído a um algoritmo deve ser o mesmo score que na vida real uma pessoa seria atribuída. O algoritmo deve ser calibrado para dar resultados semelhantes para grupos protegidos.
[297] É neste sentido que Bianca Kremer N. Corrêa chama atenção que para qualquer tipo de regulação algorítmica, seja esta exercida por meio da ética, tecno-regulação ou governança algorítmica, será necessário reivindicar uma luta contra todas as formas de opressão, que se refere de maneira mais ampla a uma efetiva democracia racial verdadeira (Corrêa, 2021, p. 212).

algoritmo vão refletir necessariamente os valores e prioridades da equipe responsável pela criação do código. O caso do algoritmo que não reconhecia rostos negros retrata este problema, e a resposta apresentada por Joy Buolamwini e Timnit Gebru (2018) é um exemplo de como estas questões devem ser resolvidas: tendo pesquisadores e desenvolvedores cientes do potencial discriminatório destas tecnologias. Inclusive, Gebru[298] é uma das autoras que propuseram os *model cards*, indicando sua preocupação em desenvolver, elaborar e propor novas alternativas para a implementação de algoritmos de ML mais éticos.

Assim como ocorre com todas as tecnologias, a inteligência artificial inevitavelmente replicará os valores sociais de seus criadores[299]. Não é apenas o que está embutido no código que apresenta problemas, pois os "algoritmos não são apenas o que os programadores criam, ou os efeitos que eles criam a partir de determinado *input*, mas eles também são o que os usuários fazem dele diariamente." [300] (Boyd; Crawford, 2012, p. 13). Um relatório sobre o tema produzido em abril de 2019 pelo AI Now Institute, reúne números interessantes para revelar o cenário das discriminações de gênero e de raça que ocorrem no mercado de tecnologia (West et al., 2019). O documento sinaliza a existência de uma crise de diversidade nesse setor, tanto no contexto acadêmico quanto no mercado, o que tem tido um reflexo direto nos sistemas que estão sendo desenvolvidos, implementados e vastamente utilizados. Nesse mesmo sentido caminha o relatório elaborado pela UNESCO (2019), que evidencia os impactos que o desenvolvimento de assistentes virtuais de gênero feminino está causando na percepção social do que são e do que devem ser as mulheres.

Por exemplo, constatou-se que: nas principais conferências acadêmicas na área de aprendizado de máquina no ano de 2018, apenas 18% das publicações tinham como autoras mulheres (Kiser; Mantha, 2019), e em universidades

[298] Recentemente a pesquisadora esteve envolvida em uma polêmica envolvendo a sua saída do Google motivada pela acusação de Gebru de que a empresa estava "silenciando vozes marginalizadas", diante de divergências em relação à produção de um artigo acadêmico em que Gebru figurava como uma autora. Ver em: BBC. *Timnit Gebru: Google staff rally behind fired AI researcher*. BBC, 20 de dezembro de 2020. Disponível em: https://www.bbc.com/news/technology-55187611

[299] CRAWFORD, Kate. Artificial Intelligence's White Guy Problem. *The New York Times*, 25 de jun. 2016. Disponível em: "https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html". Acessado em 13.09.2019.

[300] Tradução livre de: "algorithms are not just what programmers create, or the effects they create based on certain input, they are also what users make of them on a daily basis*"*.

estadunidenses e europeias[301], 80% dos professores que lecionavam cursos sobre inteligência artificial em 2018 eram homens (Shoham et al., 2018, p. 25). Apesar da amostra limitada (em relação à seleção das universidades, bem como dos cursos selecionados), a pesquisa realizada por Shoham et al., aponta a baixa presença de mulheres inscritas em cursos introdutórios de inteligência artificial e de aprendizado de máquina entre os anos de 2010 e 2017, indicando que mais de 70% dos inscritos são do gênero masculino (Shoham et al., 2018, p. 21). No Brasil, de acordo com dados do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), de 2010 a 2017, mulheres representavam 17% e 27% dos bacharéis dos cursos de Computação e Engenharias, respectivamente[302].

A presença masculina também se sobressai no mercado de trabalho, onde se constata que nas maiores empresas globais de tecnologia apenas 15% dos pesquisadores do Facebook e 10% do Google são mulheres, sendo que a presença de pessoas negras é ainda menor: 2,5% dos funcionários do Google, e 4% no Facebook e na Microsoft [303] . Além disso, percebe-se que em fóruns de desenvolvedores de software os homens têm 15% mais de probabilidade do que as mulheres de serem desenvolvedores sênior, o dobro de chance de assumirem papéis de gerência nas respectivas empresas que trabalham, e quase quatro vezes mais chances de se tornarem executivos.

Ainda que a pressão midiática e de pesquisadores da área tenha chamado a atenção para estes problemas, motivando algumas respostas – não que sejam satisfatórias, ou ataquem o problema de fundo – esta ressalta a importância da transparência. Esta exigência tem sido o valor invocado como essencial para investigar a ocorrência de discriminações e reproduções de vieses, visto que com ela é possível prestar contas, responsabilizar e avaliar a ocorrência destas discriminações ocasionadas pelo algoritmo. Neste sentido, o direito à explicação e à revisão algorítmica se coloca como uma via para enfrentar discriminações codificadas em algoritmos (Silveira; Silva, 2020, p. 4).

[301] UC Berkeley, Stanford, UIUC, CMU, UC London, Oxford e ETH Zurich.
[302] Ver em: COTRIM, Cícero; PIOVESANA, Matheus; ALBUQUERQUE, Naiara. Mulheres ainda não têm lugar na tecnologia — e essa diferença vai demorar para acabar. *Jornal Estadão*, 12 de jul 2019. Disponível em: <https://arte.estadao.com.br/focas/estadaoqr/materia/mulheres-ainda-nao-tem-lugar-na-tecnologia-e-essa-diferenca-ainda-vai-demorar-para-acabar> Acessado em 13.09.2019.
[303] SIMONITE, T.. *AI is the future - but where are the women?* Revista *WIRED*, 17 de agosto de 2018. Disponível em: <https://www.wired.com/story/arti cial-intelligence- researchers-gender-imbalance/> Acessado em 12.09.2019.

Como uma maneira de viabilizar esta explicação, especialmente em casos em que há o uso de ML, algumas medidas, que não são mutuamente excludentes, podem ser adotadas para detectar e mitigar vieses (i) abordagens estatísticas e caixas de ferramenta de *software*[304]; (ii) *framework* discursivo (workshops e fóruns de discussão), ferramentas para avaliações e material de aprendizado; (iii) padrões de documentação (tal qual proposto pelo *model cards*), (iv) auditoria, e (v) desenvolvimento de parâmetros técnicos e certificação (Rovatsos et al., 2020, p. 24). Assim como proposto como ferramentas de prestação de contas e responsabilização, as soluções apresentadas aqui também visam unir medidas técnicas e não técnicas para investigar e identificar estas discriminações.

[304] Tal como, o Accenture's 'Fairness Tool', IBM's 'AI Fairness 360 Open Source Toolkit', indicado por Rovatsos et al. (2020)

3.
Afinal, o que é um direito à explicação de decisões algorítmicas?

Nos primeiros capítulos deste trabalho foi apresentado o conceito de inteligência artificial, com especial ênfase à área de aprendizado de máquina, e como o funcionamento e os resultados gerados pelos seus usos estão afetando o Direito, em especial, os direitos fundamentais. Isto ocorre principalmente pela presença cada vez maior de algoritmos em espaços de tomada de decisão que antes eram ocupados por humanos[305]. Embora ainda existam resquícios de um certo mito sobre a objetividade, neutralidade, racionalidade e imparcialidade dos resultados gerados por estes algoritmos, pesquisas acadêmicas [306] têm paulatinamente descontruído esta crença.

Esse desvendar acaba revelando três principais problemas relacionados ao uso de algoritmos de tomada de decisão representados pela sigla FAT[307]: (i) as dificuldades em realizar uma prestação de contas e atribuir uma responsabilização a eventuais danos causados por tais artefatos; (ii) a ausência de um nível de transparência adequado do processo de desenvolvimento, aplicação e resultados de algoritmos; e, (iii) os desafios envolvendo resultados discriminatórios, enviesados e incorretos. A complexidade inerente a estes sistemas tem gerado certo ceticismo quanto aos entraves que devem ser superados para atender ao FAT, especialmente pela opacidade técnica do funcionamento e resultados gerados pelos mesmos, bem como pela opacidade jurídica garantida para proteger o sigilo de tais códigos. Ciente destes desafios, o capítulo 2 propõe diferentes mecanismos para contornar esta opacidade e complexidade, propondo uma abordagem multidisciplinar, a fim de permitir que haja certa tradução e compreensão por humanos do processo e dos

[305] Balkin indica que haverá uma substituição de humanos por agentes inteligentes, diante dos benefícios e baixos custos que isto irá significar. Contudo, esta substituição se dará de maneira parcial, e apenas para exercer certas capacidades humanas. Nas palavras do autor: "A key feature of robotic substitution is that it is *partial*. Robots and AI entities take on particular aspects and capacities of persons (….)" (Balkin, 2015, p. 49).

[306] Conforme reconhecido por Casey et al. (2019, p. 147): "In recent years, however, society's deferential attitude toward algorithmic objectivity has begun to wane—thanks, in no small part, to a flurry of influential publications examining bias within complex computational systems. Particularly in the last five years, numerous studies across multiple industry sectors and social domains have revealed the potential for algorithmic systems to produce disparate real world impacts on vulnerable groups" (removidas as referências do original)

[307] *Como se sabe, também* são três princípios comumente presentes nas cartas éticas propostas para o desenvolvimento da inteligência artificial.

resultados de sistemas de IA de maneira ampla, e dos algoritmos de tomada de decisão de ML, em particular.

Como visto no capítulo 1, o debate sobre um direito à explicação de decisões automatizadas pode ter sua origem remetida à promulgação do GDPR, motivada por um debate travado na doutrina, apesar da existência de previsões semelhantes na antiga Diretiva 95/46/CE, em seus arts. 12 e 15 [308]. Contudo, como colocado por Kaminski (2019b), o debate sobre o direito à explicação deve ser compreendido dentro de uma discussão mais ampla de *accountability* de sistemas de IA[309]. Por isso, formas de implementar e assegurar este direito exigem esta abordagem, ampla e multifacetada, de uma documentação, justificação e monitoramento do desenvolvimento, implementação e aplicação antes (*ex ante*) e depois (*ex post*) de tais sistemas.

Esta abordagem, contudo, não exime de responsabilidade e não diminui a necessidade de definir, conceituar e justificar o que é um direito à explicação de decisões algorítmicas, em especial, aquelas que se valem de modelos de aprendizado de máquina. É preciso considerar o que seria uma explicação adequada. É o que este capítulo visa enfrentar. Neste sentido, serão apresentados os diferentes tipos de explicações de decisões algorítmicas que têm sido mapeados pela literatura, a fim de identificar qual é o tipo de explicação mais adequada e qual pode ser exigida à luz do ordenamento jurídico brasileiro. Esta análise deve observar que o ordenamento já ensaiou a regulação sobre o tema em legislações setoriais e, recentemente, positivou tal direito na LGPD. Tendo em vista que o debate sobre o direito à explicação ganhou notoriedade em razão da sua previsão na LGPD, se buscará aprofundar e oferecer o que deve ser compreendido como um direito à explicação nesta legislação.

3.1
Os tipos de explicações mapeados pela doutrina

[308] Diversos artigos sobre o tema tratam da antiga previsão na referida Diretiva, como: Wachter et al., 2017; Malgieri; Comendé, 2017; Edwards; Veale 2017.
[309] A explicação de sistemas de IA também é considerado como uma das ferramentas, ou maneiras, de garantir a sua *accountability* (Doshi-Velez; Kortz, 2017, p. 2).

3.1.1
Diferenciações no tipo e conteúdo de explicações, ainda sob à luz do GDPR

A promulgação do GDPR e a sua previsão de um direito à explicação, que deve ser compreendido e derivado a partir de uma leitura sistemática da norma, motivou a discussão sobre o que exatamente deveria – ou poderia – ser fornecido ao titular de dados quando tal direito fosse exercido. Parte da discussão gira em torno do aspecto substancial e de mérito do direito à explicação, ou seja, quais informações devem ser apresentadas quando tal direito é exercido? Devem ser oferecidas informações que permitam que um humano possa compreender como um resultado foi alcançado. Isto significa saber (i) os fatores que motivaram aquela decisão (como, os dados que influenciaram ou as regras do modelo) e (ii) como estas informações serão apresentadas. Doshi-Velez e Kortz, apesar de limitarem o conteúdo da explicação à relação de como os dados de entrada determinaram ou influenciaram os dados de saída, apresentam três perguntas relevantes que devem ser respondidas quando uma explicação é exigida (Doshi-Velez; Kortz, 2017, pp. 3-5).

A primeira pergunta seria: quais são os principais fatores que causaram uma decisão? Segundo os autores, esta colocação é quase um senso comum do que se espera de uma explicação. A resposta envolve uma lista de fatores que devem ou não serem levados em consideração, que seriam determinantes para alcançar um ou mais resultados, e necessariamente estariam relacionados com a correlação entre eventos e resultados. Por exemplo, seria importante compreender como certos dados foram considerados em uma decisão de um algoritmo envolvendo a concessão de crédito, oferecimento de oportunidades de emprego, ou para predizer a ocorrência de abuso infantil em um núcleo familiar[310]. Seria importante compreender se os dados de entrada não constituem uma categoria protegida que não poderia ser utilizada na análise, e a adequação deles para a finalidade que se visa alcançar.

A segunda pergunta é formulada da seguinte maneira: trocar um determinado fator teria alterado a decisão? Para os autores, esta pergunta busca

[310] MCINTYRE, Niamh; PEGG, David. Councils use 377,000 people's data in efforts to predict child abuse. *The Guardian*. Disponível em: https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict- child-abuse. Acessado em 12.01.2022.

responder se alguma coisa, e qual coisa, foi determinante para que aquele resultado fosse alcançado. O efeito deste evento/fator/dado, e se isto impacta ou não no seu resultado, parece estar relacionado com as explicações contrafactuais, que serão melhor apresentadas abaixo, que buscam evidenciar como o mundo poderia ser diferente para que a decisão desejada ocorresse. Deve ser indicado como determinados fatores influenciaram neste resultado, e como ele poderia ser diferente se outros eventos fossem considerados (Wachter et al., 2018).

A terceira e última pergunta consiste em saber: por que dois casos semelhantes alcançaram resultados diferentes (e por que dois casos diferentes alcançaram resultados semelhantes)? De acordo com Doshi-Velez e Kortz (2017, p. 5), o que se busca saber com essa pergunta é se um certo evento foi determinante em uma outra decisão, e avaliar a consistência e integridade do processo de tomada de decisão e, consequentemente, a sua confiabilidade.

Embora tecnicamente seja possível responder algumas dessas perguntas, a maneira como uma explicação pode ser fornecida tem variado bastante na literatura, que tem buscado identificar, de acordo com as previsões do GDPR sobre o tema, qual seria o tipo de explicação que deve fornecida à luz da norma europeia (Wachter et al., 2017; Selbst; Powles; 2017; Malgieri; Comandé, 2017; Edwards; Veale, 2017; Wachter et al., 2018).

A discussão sobre o tipo de explicação que deve ser fornecida iniciou-se com Wachter et al. (2017), conforme apresentado no capítulo 1, em que se argumentou pela existência de dois tipos de explicação: uma sobre funcionalidade do sistema, que deveria ser apresentada antes que de fato uma decisão automatizada ocorresse (*ex ante*), e outra sobre as razões (*reasoning)* de uma decisão em particular (*ex post)*. Esta diferenciação se assemelha ao conceito usado por cientistas de computação para se referirem a explicações globais e locais. A primeira do modelo e suas funcionalidades, enquanto explicações locais expõem informações sobre uma parte específica do modelo, e visam apresentar maneiras de como um modelo de aprendizado de máquina pode ser explicado e interpretado.

Lilian Edwards e Michael Veale propõem uma divisão semelhante, denominando a explicação sobre funcionalidade como explicação centrada no modelo (*model-centric explanation* – MCE), e a de uma decisão em particular como explicação centrada no sujeito (*subject-centric explanation* –SCE). Inclusive, os autores manifestaram expressamente rejeitar o uso do termo "explicação global e

local" para evitar confusões conceituais (Edwards; Veale, 2017, 55). Edwards e Veale avançam na classificação proposta por Wachter et al. (2017) ao especificarem o que exatamente poderia ser exigido em cada uma delas a fim de que atendam aos requisitos do GDPR, de que as explicações devem ser consideradas significativas (*meaningful*).

Para tanto, os autores indicam que as MCE devem incluir (i) informações sobre a maneira como o modelo foi desenvolvido, a intenção no processo de modelagem, o tipo de modelo, e os parâmetros utilizados; (ii) metadados do treinamento, como resumo das estatísticas, descrição da qualidade dos dados utilizados para treinar o modelo, onde e como os dados foram obtidos, e quais seriam os dados de saída ou as classificações que foram realizadas; (iii) métricas de comportamento, o que envolve apresentar resultados sobre a habilidade de predizer dados nunca antes analisados, as falhas e acertos; (iv) explicações que possam ser compreendidas por um humano sobre como certos dados de entrada influenciaram ou são transformados em dados de saída, e (v) informação sobre o processo, tais como se o modelo foi testado e treinado para situações "inesperadas e não desejadas", e como ele deveria se comportar nestas hipóteses (Edwards; Veale, 2017, pp. 55-56).

As SCE seriam explicações que se refeririam a uma pequena parte do sistema, variam de acordo com o tipo de pergunta que se visa responder, e suas respostas são limitadas a informações relacionadas aos dados de entrada. Ao contrário de Wachter et al. (2017), Edwards e Veale apontam que este tipo de explicação poderia ser fornecido tanto antes, enquanto uma decisão ainda não foi de fato tomada, quanto depois, pois seria apenas necessário apresentar os dados que seriam utilizados para responder uma determinada pergunta.

Seria possível, ainda, classificar a SCEs em quatro tipos (Edwards; Veale, 2017, p. 58): (i) baseada na sensibilidade (*sensitivity-based)*: que visa identificar quais dados de entrada deveriam ter sido ou devem ser alterados para modificar a decisão; (ii) baseada no caso (*case-based),* que visa saber quais dados utilizados para treinar o modelo são similares ao do indivíduo pleiteando a explicação; (iii) baseados na demografia (*demographic-based*), visa conhecer quais são os sujeitos que tiveram tratamento semelhante ao do indivíduo pleiteando a explicação; (iv) baseado no comportamento (*performance-based)*, que busca conhecer informações sobre a confiança estatística de uma decisão, entre outros dados, como quantos

casos semelhantes ao do indivíduo pleiteando a explicação foram identificados como similares, mas que deveriam ter um resultado diferente. Percebe-se aqui uma semelhança com as perguntas formuladas por Doshi-Velez e Kortz.

Na perspectiva dos autores, as explicações SCEs são menos úteis para questões relacionadas à regularidade do procedimento, mas possuem uma utilidade para os usuários afetados por elas. Contudo, os autores reconhecem que nem sempre fornecer estas explicações será possível por questões técnicas. Modelos muito complexos, como quando há muitos dados de entrada, podem inviabilizar explicações SCEs, e a pretensão causal entre dados de entrada e como eles influenciam seus resultados. Ainda, explicações locais só representam partes específicas do sistema, de forma que resultados chamados de *outliers*[311] podem não estar representados, inviabilizando explicações sobre eles. Isso também é verdade para explicações sobre a lógica básica do modelo, em que pessoas que têm um perfil muito distante da média podem não receber uma explicação correta e adequada do sistema.

Por tais motivos é que Edwards e Veale defendem que, diante da dificuldade em fornecer explicações sobre modelos de ML, devem ser utilizados outros direitos individuais previstos no GDPR, como o direito a apagar dados e o direito à portabilidade de dados, para garantir um maior controle do sujeito sobre os seus dados pessoais no contexto de ML. Além destes direitos individuais, os autores defendem a implementação da privacidade por *design (privacy by design*)[312]*,* que vai gerar resultados que vão proteger e beneficiar a sociedade como um todo, ao invés de reforçar direitos e ferramentas que vão ajudar indivíduos. Além disso, os autores também defendem a implementação de um devido processo algorítmico (Crawford; Schultz, 2014), sugerindo a formação de um painel de árbitros qualificados para julgar estas demandas, além da elaboração de relatórios de

[311] No contexto de aprendizado de máquina, *outliers* são dados que se diferenciam de maneira significativa de outros dados analisados, ou seja, que estariam distantes da média ou da mediana dos dados analisados.

[312] Este conceito foi desenvolvido na década de 90, por Ann Cavoukian, à época comissária de informação e privacidade de Ontário, no Canadá, e vem ganhando notoriedade e relevância normativa, ao ser previsto em legislações, especialmente em normas de proteção de dados pessoais. Conforme organizado por Mariana Palmeira, este conceito é composto por sete princípios: "(i) proatividade e prevenção; (ii) privacidade como configuração padrão; (ii) privacidade incorporada ao design; (iv) funcionalidade completa – soma positiva, não soma zero; (v) segurança ponta a ponta; (vi) visibilidade e transparência; e (vii) respeito à privacidade do usuário" (Palmeira, 2019, p. 327).

impacto de proteção de dados (conhecidos como DPIAs) e certificações, em uma posição similar ao que Kaminski (2019b) propõe.

Apesar do esforço dos autores em propor e diferenciar tipos de explicações consideradas significativas, nos termos do GDPR, eles adotam uma posição cética sobre a efetividade de um direito à explicação, no qual acreditam que há um perigo em se criar "um paradigma da 'transparência insignificante'"[313], tal como ocorreu com o consentimento[314] (Edwards; Veale, 2017, p. 23). Apesar de não concordar com tal ceticismo em relação à explicação, diante do papel mais amplo que o direito à explicação exerce na prestação de contas de sistemas de IA e seus algoritmos de tomada de decisão, as propostas apresentadas pelos autores sobre as informações que podem ser disponibilizadas são insumos que agregam valor e conhecimento sobre o processo de tomada de decisão automatizada. As MCEs apoiam uma explicação voltada ao modelo de maneira mais ampla, uma visão sistêmica, enquanto as SCEs focam em uma explicação específica, voltada para o usuário. Ambas as informações são relevantes para fins de explicação. O que será determinante, no entanto, é identificar para quem e para qual finalidade se deseja uma explicação, a fim de que a informação apresentada seja inteligível e interpretável para quem esteja recebendo.

Ademais, outros autores, como Malgieri e Comande (2017), também possuem uma visão própria sobre como o direito à explicação deve ser compreendido, propondo uma alternativa para superar a distinção feita por Wachter et al. (2017)[315]. Para tanto, é proposto um "teste de legibilidade" (*legibility test)* que deve ser aplicado pelos controladores para avaliar se a exigência do GDPR, sobre o que seria uma "informação significativa" (*meaningful information*) sobre a lógica envolvida em processos de tomada de decisão automatizados, foi observada. Este teste, que deve ser aplicado pelo controlador, avaliaria as informações apresentadas tanto sob um ponto de vista de arquitetura (lógica interna do algoritmo, antes que

[313] Tradução livre de: "a 'meaningless transparency' paradigm (…)"
[314] OLIVEIRA, Caio César de; FILHO, Paulo César Tavares Filho. A LGPD e o início do fim da cultura do consentimento. *Revista Jota*, 28 de jul., de 2021. Disponível em: https://www.jota.info/opiniao-e-analise/artigos/lgpd-e-o-inicio-do-fim-da-cultura-do-consentimento-28062021. Acessado em 12.01.2022.
[315] Merece ser destacado que os autores concordam com a distinção feita por Wachter et al. (2017) sobre explicações *ex ante* e *ex post*, com base nos artigos 13 e 14 (*ex ante*) e 15 (*ex post*) do GDPR, mas discordam de que o art. 15 estaria se referindo a uma explicação sobre a funcionalidade, pois entendem que seria uma explicação referente à lógica envolvida em uma decisão, diante do tempo verbal utilizado no artigo (orientado para o futuro, e não para o passado) (Malgieri; Comande, 2017, p. 22).

uma decisão seja tomada), quanto de implementação (significado da decisão e das consequências por ela, quando esta já tenha sido tomada) do algoritmo. O objetivo é que a aplicação deste teste possibilite que um sujeito tenha a autonomia e, de maneira concreta, compreenda os dados e as análises realizadas de fato pelo algoritmo.

Este teste de legibilidade possui duas partes. A primeira trata de um questionário sobre a arquitetura do sistema, e busca obter respostas sobre a criação do algoritmo, suas funcionalidades e os resultados esperados. A segunda apresenta um questionário sobre a implementação, com perguntas sobre a finalidade para a qual o algoritmo foi desenvolvido, o nível de tarefas que exigem presença humana, a natureza comercial da decisão, dados estatísticos, a possibilidade de revisão do resultado, e o contexto no qual o algoritmo é aplicado. Dois objetivos pretendem ser alcançados com o preenchimento destes questionários. O primeiro é que eles sirvam como um relatório de impacto sobre o algoritmo, além de serem uma forma de o controlador demonstrar a sua adequação e aderência (*compliance)* ao GDPR. O segundo é que seria um insumo importante a ser apresentado caso seja conduzida uma auditoria sobre os possíveis vieses do algoritmo. Na visão dos autores, os arts. 13 (2)(f), 14(2)(g), 15(1)(h) estabelecem um dever do controlador de realizar auditorias, e defendem que o teste de legibilidade poderia ser aplicado nesses casos. Ainda, argumentam que tal teste também evitaria discussões envolvendo segredo empresarial, visto que poderiam ser apresentadas respostas sem a necessidade de "abrir a caixa-preta".[316]

Apesar da boa intenção dos autores, o questionário é extenso, e não se compromete a realizar qualquer cálculo sobre os riscos mapeados de acordo com as respostas apresentadas. Ou seja, não há ao final uma análise mais qualitativa das consequências de cada resposta, e como elas contribuem para avaliar se um algoritmo é ou não mais compreensível. Em que pese isto, o questionário e suas

[316] Os autores fazem uma interpretação dos Considerandos do GDPR para indicar que, em caso de conflito entre direitos empresariais/comerciais, e a privacidade dos indivíduos, este último deve prevalecer. Reconhecem que, embora o segredo empresarial inclua a proteção ao código do algoritmo, a predição realizada, e as práticas do negócio, há certas informações que não podem ser incluídas nesta categoria, como a abertura de informações sobre o resultado de uma auditoria do algoritmo ou decisões específicas, pois estas últimas não podem ser consideradas como aptas a causar um efeito adverso aos direitos do controlador. Ademais, defendem que informações sobre funcionalidades (*ex ante*) não poderiam ser limitadas, pois o direito dos titulares de dados deve prevalecer em relação aos dos controladores, enquanto algumas explicações *ex post* poderiam ser restritas, mas sem que isto implique em uma total negativa em fornecer informações (Malgieri; Comande, 2017, p. 34).

respostas não deixam de ser mais uma maneira de gerar uma documentação sobre o desenvolvimento e a aplicação de um algoritmo de tomada de decisão. O que é relevante na proposta dos autores em relação ao presente trabalho, é o reconhecimento de que é necessário produzir documentos que sejam capazes de acessar informações sobre o processo de desenvolvimento do algoritmo, e dos resultados gerados, para fins de análise de impacto e de realização de auditorias.

Nesta mesma ideia de oferecer uma explicação sem abrir a caixa-preta, Sandra Wachter, Brent Mittelstadt e Chris Russell (2018), apesar de continuarem mantendo a controversa posição de que não seria possível derivar um direito à explicação no GDPR[317], ainda assim defendem que seria possível exigir certas informações do controlador sobre os resultados gerados do algoritmo de decisão. Na visão dos autores, existiria um valor social e ético em oferecer explicações, que consiste em garantir que seja estabelecida uma confiança sobre o algoritmo de tomada de decisão, pela diminuição da assimetria de informação, gerando uma aceitação social do mesmo (Wachter et al., 2018, p. 4). O objetivo precípuo de um direito à explicação, ou do direito de acesso à informação, é que ele deve ser capaz de ajudar o titular de dados a tomar uma ação em relação a uma dada situação, e não meramente que ele compreenda uma decisão algorítmica. Neste contexto, os autores propõem que sejam asseguradas explicações contrafactuais de decisões individuais.

A partir da filosofia analítica, os autores buscam criar afirmações e regras lógicas para indicar como uma explicação contrafactual deveria ser formulada. Para os autores, o conceito de contrafactual é uma "afirmação sobre como o mundo poderia ser diferente para que uma decisão desejada ocorresse" (Wachter et al., 2018, p. 6)[318]. Devem ser oferecidas informações que sejam perto o suficiente, mas

[317] Os autores repetem os argumentos desenvolvidos no artigo de Watcher et al (2017), e fazem uma síntese dos mesmos: "First, a legally binding right to explanation does not exist in the GDPR. Second, even if legally binding, the right would only apply in limited cases (when a negative decision was solely automated and had legal or other similar significant effects). Third, explaining the functionality of complex algorithmic decision-making systems and their rationale in specific cases is a technically challenging problem. Explanations may likewise offer little meaningful information to data subjects, raising questions about their value. Finally, data controllers have an interest in not sharing details of their algorithms to avoid disclosing trade secrets, violating the rights and freedoms of others (e.g. privacy), and allowing data subjects to game or manipulate the decision- making system" (as referências originais do texto foram removidas) (Wachter et al, 2018, p. 1-2)

[318] Tradução livre de: "statement of how the world would have to be different for a desirable outcome to occur".

não o 'mais próximo possível' (Wachter et al., 2018, p. 6) [319]. A explicação contrafactual visa explicar como fatores externos influenciaram naquele resultado, e não expor o funcionamento lógico e/ou o estado interno do algoritmo (como estes autores vinham defendendo). O exemplo apresentado por Wachter et al. (2018) é o de uma pessoa que pleiteia um empréstimo no banco, e este se vale de um algoritmo de tomada de decisão para avaliar o pedido. O pedido é negado, e a pessoa solicita uma explicação do porquê seu pedido foi negado. A explicação contrafactual indicaria que o banco apenas fornece empréstimos para pessoas que possuem uma renda anual de 30 mil reais, e a renda do solicitante era de 24 mil reais.

De acordo com os autores, as explicações contrafactuais teriam uma série de vantagens em relação a explicações sobre a lógica interna do sistema. Primeiro, acreditam que a forma como a informação é apresentada permite que o titular de dados compreenda de maneira mais fácil e clara as razões de uma decisão, criando incentivos de como eles podem no futuro alterar seu comportamento para obter uma decisão desejada e mais favorável para ele ou ela. Segundo, este tipo de explicação atenderia a três objetivos que explicações de decisões automatizadas devem atender. Apesar de indicarem que as previsões do GDPR não oferecem um suporte para atender estes objetivos, defendem que a proposta de explicações contrafactuais contempla e supera os deveres previstos na norma. Assim, os objetivos são: (i) informar e ajudar o titular de dados a compreender por que uma decisão em particular foi alcançada; (ii) permitir a contestação da decisão; e (iii) compreender o que poderia ser alterado para alcançar uma decisão desejável no futuro Wachter et al. (2018).

Ao longo do artigo os autores vão indicando como que a explicação contrafactual atende aos artigos 13, 14 e 15, do GDPR, que tratam das informações que devem ser fornecidas quando há uma decisão automatizada (os autores defendem que o art. 22, do GDPR, não seria capaz de apoiar um direito à explicação). Eles ressaltam como devem ocorrer as notificações quando há uma decisão automatizada (de maneira ampla e não individualizada para cada pessoa, tal como ocorre com os avisos sobre *cookies* [320]). Ainda, as explicações

[319] Os autores reconhecem que nem toda variável vai ser considerada relevante. A relevância de uma variável para fins de explicação contrafactual depende de diferentes fatores externos à decisão ou da capacidade de a variável em questão ser alterada.

[320] Os *cookies* são pequenos arquivos de dados que são depositados em um site quando um usuário o acessa, com o objetivo de que este usuário possa ser reconhecido tanto naquele site, quanto em

contrafactuais seriam úteis para o exercício do art. 22(3), do GDPR, ao prever o direito de o titular de dados contestar uma decisão, para tentar revertê-la e alterá-la, além de permitir que possa verificar se o que foi considerado é de fato legítimo. No entanto, os autores reconhecem que a melhor maneira de avaliar se um algoritmo produz resultados discriminatórios ou enviesados é por meio de análises estatísticas, e não por meio de explicações contrafactuais, pois esta última se refere a uma decisão em particular dada em um certo contexto, e as estatísticas apresentam um comportamento sistêmico e geral do algoritmo.

As limitações das explicações contrafactuais não passam despercebidas pelos autores, que apontam que em algoritmos muito complexos, com diversas variáveis, ou em ambientes com várias mudanças ao longo do tempo, elas podem ser consideradas como tendo limitações relevantes (Wachter et al., 2018, p. 42). Além disso, quando for necessário apresentar explicações sobre as funcionalidades ou o funcionamento interno do algoritmo, além de evidências estatísticas para avaliar eventuais vieses ou resultados discriminatórios, o tipo de explicação proposto pode não ser suficiente

Complementando o referido artigo, Mittelstadt, Russell e Wachter (2019) sugerem que pesquisadores da área técnica de *Explainable AI* (xAI) devem explorar os métodos interativos de interpretação *post-hoc*, em especial, as explicações estudadas pelas ciências humanas (como a psicologia, filosofia e ciências cognitivas) para identificar a maneira mais adequada de oferecer explicações que permitam a contestação, compreensão e discussão dos resultados algorítmicos. A defesa de uma explicação contrafactual e contrastante é reiterada, e colocada como uma alternativa para métodos aplicados pela ciência da computação[321], em especial, modelos de aproximação.[322]

outros sites/aplicações na internet (Lessig, 2006, p. 48). Nas palavras de Lessig: "(...) strictly speaking, cookies are nothing more than a tracing technology. They make it simple to trace a machine across web pages. That tracing doesn't necessarily reveal any information about the user (...). But sometimes something important is revealed about the user by association with data stored elsewhere" (Lessig, 2006, p. 49).

[321] De acordo com os autores, duas formas de explicação têm sido exploradas pela literatura de Explainable AI (xAI): transparência e interpretação *post-hoc*. O primeiro se dirige a compreender como o modelo funciona internamente, enquanto o segundo se preocupa em como e por que o modelo se comporta de uma certa maneira.

[322] Tais como: modelos lineares em espaços contínuos (*Linear Models in Continuous Spaces*), sensibilidade gradiente vs. binarização (*gradient sensitivity verses binarization*), e modelos lineares em espaços de alta dimensão (*Linear models in high-dimensional spaces*).

Segundo os autores, uma das características da explicação humana é que elas são contrafactuais ou contrastantes. Ou seja, as pessoas não pedem explicação do porquê dado evento ocorreu (P), mas por quê (P) ocorreu no lugar de (Q). Segundo Mittelstadt, Russell e Wachter, isto ocorre pois, em termos psíquicos, humanos preferem explicações contrastantes do que explicações em cadeia, e preferem compreender um comportamento normal a partir de uma explicação do que é um evento anormal. Continuam os autores indicando que outras duas características sobre a explicação humana seriam a sua seletividade e sociabilidade (Mittelstadt et al., 2019, p. 5).

As explicações humanas são seletivas: explicações completas ou científicas são raramente apresentadas na prática. As pessoas não esperam uma explicação completa para um evento, em que todas as causas que motivaram a sua ocorrência são apresentadas, elas selecionam uma ou outra causa de um número de causas finitas para serem explicações. Isto é importante quando se considera quem está recebendo explicações, pois estas devem ser relevantes a quem pergunta. Como apontam os autores, em xAI, por exemplo, a seleção da informação considerada importante normalmente se baseia nas principais funcionalidades do modelo, ou em evidências enfatizadas em interfaces explicativas e interativas, e normalmente são baseadas no seu peso ou influência em certa predição ou resultado. No entanto, a relevância destas funcionalidades pode variar, e ser interpretada e compreendida diferentemente, de acordo com os interesses subjetivos de quem recebe a explicação e as suas expectativas, do que o seu peso estatístico real de fato, o que pode ser um problema (Mittelstadt et al., 2019, p. 5).

Segundo Mittelstadt et al. (2019, p. 6), as explicações humanas são sociais, pois envolvem a interação entre um ou mais explicadores e de quem recebe a explicação. Ou seja, a explicação depende de uma interação, em que há uma transferência de conhecimento que é apresentada de acordo com a crença e capacidade de compreensão de quem recebe a informação. Para os autores, no contexto de ML, explicações devem ser tratadas como um processo interativo, incialmente envolvendo a presença humana e atores automatizados.

Mittelstadt et al. (2019), no entanto, reconhecem os limites destas exigências. Caso a explicação contrastiva não se refira exatamente ao resultado que o sujeito deseja, isto pode não ser relevante para os objetivos que ele pretende com aquela explicação (como contestar e pedir a sua revisão). A maneira como a

explicação é apresentada e recebida é inerentemente subjetiva, impactando a sua compreensão, como, por exemplo, fazendo com que sejam apresentadas apenas explicações sobre as funcionalidades mais aceitáveis ou aquelas irrelevantes (Mittelstadt et al., 2019, p. 7).

Contudo, para Mittelstadt et al. (2019), estes riscos poderiam ser mitigados com a apresentação de argumentos críticos sobre os eventos que servem como justificativas para as explicações. Isto porque explicações conversacionais são essencialmente formas de argumentação, e são utilizadas para justificar eventos, a verdade e a relevância das suas causas. A teoria das explicações de todo o dia, por exemplo, envolve não apenas a transferência de informação ou argumentos causais, mas basicamente funciona como um suporte argumentativo para estas causas (Mittelstadt et al., 2019, p. 7). Mittelstadt et al. ressaltam que a apresentação de justificativas não é um tema que tem sido objeto de estudo de xAI, mas seria relevante para a prestação de contas e responsabilização algorítmica, e isto implica em saber quais eventos e registros devem ser mantidos e identificados para serem apresentados.

Em suma, uma explicação para modelos de ML deve ser contrastiva, seletiva e social, permitindo a troca de informação de maneira interativa, como uma conversa entre pessoas, mas também com a apresentação de argumentos críticos sobre as justificativas dos eventos selecionados para uma explicação.

Este aspecto da explicação pela interação é um aspecto fundamental e, no caso de explicações de decisões algorítmicas, isso necessariamente implica na presença de uma pessoa humana. Por exemplo, o uso de assistentes conversacionais demonstra como a interação, comunicação e compreensão entre a máquina e o humano pode ser limitado, principalmente diante do conhecimento pré-programado destes sistemas, que possuem um rol limitado de possíveis respostas para determinadas palavras utilizadas em uma pergunta. A própria limitação dos modelos de ML em compreender o contexto e a semântica das palavras e frases já se impõe como uma impossibilidade técnica. Se compreender uma pergunta e apresentar uma explicação adequada já é um desafio para os humanos (quantas vezes professores já responderam coisas totalmente diferentes do que foi perguntado por não compreender uma pergunta?), para algumas aplicações de IA isto certamente não será possível em um número grande de vezes e situações.

Ainda, percebe-se que as explicações contrastantes e contrafactuais comumente utilizadas por humanos está fortemente relacionada à possibilidade de se atribuir uma relação causal a um dado evento ou fator (Hoffman et al., 2017a; 2017b; Klein, 2018). Isto é, as pessoas se engajam em um raciocínio causal para explicar determinados acontecimentos, e o objetivo disso é obter uma explicação com algum nível de satisfação (Hoffman et al., 2017b). Ainda assim, mesmo nestas situações, é um verdadeiro desafio identificar uma só causa que seja responsável por gerar certo evento. O mundo é complexo e nem sempre é possível selecionar um único motivo para responsabilizar o acontecimento de um cenário, sendo certo que múltiplas variáveis se relacionam e se influenciam. Por exemplo, explicar por que o aquecimento global está ocorrendo implica em selecionar uma série de fatores, de diferentes naturezas, e organizá-los para contar uma história, que deve ser coerente, científica e racional. Isto vale também para a disciplina da responsabilidade civil[323], em que a causalidade é amplamente discutida para o reconhecimento da responsabilização.

Quando uma explicação é oferecida, ela visa articular todas essas causas para formar um argumento que tenta justificar, e não apenas expor,[324] a ocorrência de algo. Estudos da psicologia cognitiva demonstram que as explicações podem ser apresentadas em formato de: eventos contrafactuais, abstrações, condicionais, lista, e história contada em uma cadeia de reações (Hoffman et al., 2017b, p. 82). Em geral, as pessoas tendem a não preferir explicações complexas, com múltiplas variáveis causais e diferentes interações, e sim explicações simples, com uma única causa[325], podendo adicionar novas causas e variáveis para tornar a explicação mais complexa e completa, mas com o mínimo necessário de informação (Hoffman et al., 2017b, p. 82-83). Isto porque explicações curtas e simples são "mais fáceis de comunicar, de lembrar e de usar como base de projeção para o futuro".[326] (Hoffman

[323] Neste sentido, ver: MULHOLLAND, Caitlin. *Responsabilidade civil por presunção de causalidade*. 1a edição. Rio de Janeiro: GZ Editora, 2009.

[324] Maranhão et al. apontam que as técnicas de explicação em IA visam apenas explicar ao usuário como uma decisão foi tomada, e não justificá-la, ou seja, "o ponto não é convencer o usuário de que a máquina está correta, mas sim mostrar como a máquina chegou a uma decisão de forma que o usuário possa entender o processo e corrigi-lo se for o caso" (Maranhão et al., 2021, pp. 144-145)

[325] Abstrata ou reversível, em que a primeira leva em consideração diversas causas, às vezes de maneira contrafactual, e sintetiza em uma única explicação, enquanto a segunda se refere à explicação contrafactual, em que se um determinada causa deixar de ocorrer, o evento será afetado (Hoffman et., 2017b, pp. 78 e 82).

[326] Tradução livre de: "the simpler explanation is easier to communicate, to remember, and to use as a basis of projection to the future".

et al., 2017b, p. 83). Apesar disso, quando for possível comparar a situação que está sendo explicada com outras situações, a necessidade de explicações mais complexas cresce (Hoffman et al., 2017b, p. 84). Estes achados, de acordo com pesquisadores do tema, possuem importância para explicações de sistemas inteligentes, para que no lugar de explicações abstratas e complexas sobre aplicações de IA sejam apresentadas explicações simples para que os usuários possam compreender e interagir com a tecnologia[327].

Apesar da importância da apresentação de elementos causais para explicações (*causal reasoning* - argumentação causal), já se sabe que isto pode não ser possível em resultados algorítmicos de ML. É neste sentido que Selbst e Barocas (2018, p. 1.091)[328] afirmam que resultados de algoritmos de tomada de decisão de aprendizado de máquina seriam contra-intuitivos, desafiando a compreensão sobre a relevância de certos critérios da decisão, pois, apesar de ser possível identificar as relações estatísticas da tomada de decisão[329], o porquê destas relações existirem ainda é turvo e opaco para os humanos. Para os autores, a capacidade de justificar a explicação de uma (ou muitas) decisão(ões) algorítmica(s) é relevante diante dos diferentes objetivos que ela visa alcançar, e os valores que ela deve proteger. Isto porque, é importante avaliar normativamente se a maneira como aquele está operando e os seus resultados são legítimos e legais.

3.1.2
Uma taxonomia de explicações exigidas no contexto de algoritmos que se valem do aprendizado de máquina

[327] Ver: Hoffman et al., 2018 e Klein, 2018.

[328] Para os autores, dentre as características dos algoritmos de ML que eles chamam atenção, e colocam como um desafio, é que os mesmos são inescrutáveis. Ou seja, eles desafiam a compreensão humana em relação ao seu funcionamento, que é complexo (com numerosas variáveis interdependentes) e sofisticado. Explicações voltadas para a inescrutabilidade estariam relacionadas a uma explicação com o modelo como um todo, e não sobre uma decisão em particular (Selbst; Barocas, 2018, p. 1.091).

[329] Brennan-Marquez (2017) explora a teoria de "plausible cause" no direito penal norte americano, indicando que a existência de correlação estatística e a sua acurácia é importante, mas não o essencial para realizar o nexo causal entre um fato e um evento danoso. O autor indica a importância de se considerar o contexto e as exceções na análise de casos concretos, e como a apresentação das razões de decidir é fundamental para proteger princípios constitucionais, especialmente para verificar a legitimidade de ações do poder público. O autor chama a atenção para o fato de que nem todas explicações se qualificam como justificativas, e ressalta a importância de que estas sejam apresentadas como uma maneira de avaliar estas ações (p. 1288).

Percebe-se que houve uma mudança de chave sobre o tipo de explicação que vem sendo exigida dos algoritmos de tomada de decisão de ML. Nos últimos anos, há uma demanda por um tipo de explicação que é mais centrada no usuário, e não tanto em explicações mecânicas sobre o funcionamento do sistema. Há preocupações, entre outras, relacionadas à confiança e à capacidade de compreensão e interpretação sobre uma decisão algorítmica (Chari et al., 2020, p. 1)[330]. Neste sentido, Chari et al. (2020) reuniram diferentes tipos de explicações que têm sido mapeados pela doutrina no contexto de sistemas especialistas, e apontam que, embora haja diferentes ideias sobre o que seria uma explicação, todas buscam apresentar algum tipo de conhecimento. As explicações podem tratar sobre o modelo ou o seu contexto de aplicação; podem ser formuladas sob uma perspectiva mais científica ou de maneira mais corriqueira, sendo estas últimas conhecidas como explicações do dia a dia. O interessante no trabalho de Chari et al. (2020) é a variedade de perguntas que visam ser respondidas com cada uma das explicações. O desafio regulatório será compreender *quais* dessas deverão e poderão ser demandadas, e consideradas como um dever a ser atendido pelo agente econômico responsável pela aplicação do sistema de IA.

Em atenção à perspectiva da governança binária de Kaminski, e os interesses que ela visa atender tanto sob uma perspectiva subjetiva e individual e de justificativa de uma decisão algorítmica, quanto sistêmica e coletiva, a taxonomia de Chari et al. (2020), apesar de apresentar tipos de explicações centradas no indivíduo, também tem a capacidade de atender ao aspecto instrumental em evitar/prevenir erros do sistema. Os resultados e as informações de uma decisão também são capazes de documentar e revelar aspectos do sistema como um todo. Como sugere Zarsky (2013), conhecer *quem* é o seu usuário ou o público para o qual será apresentada a explicação é essencial para fornecer um ou mais tipos de explicações.

A taxonomia construída pelos autores consiste em nove tipos de explicações. É apresentada a definição de cada uma, e uma pergunta que visa ser respondida por

[330] Esta perspectiva também é compartilhada pelos autores do Direito, como Edwards e Veale (2017) e Wachter et al. (2018). Os autores reconhecem que, enquanto a literatura de IA e aprendizado de máquina busca oferecer explicações voltadas para *bugs* e funcionalidades do sistema, o Direito busca explicações que sejam capazes de avaliar a legitimidade de uma decisão, contestá-la, direcionar responsabilidades e verificar a sua prestação de contas e responsabilização de maneira geral (Wachter et al., 2018, p. 7).

cada tipo de explicação, ampliando os tipos de perguntas e respostas pensadas por Doshi-Velez e Kortz e Edwards e Veale. Ao final, os autores sugerem a aplicação de algumas técnicas [331] para promover explicações que sejam personalizadas, baseadas em contexto e confiáveis para os usuários de sistema de IA. Assim, os tipos de explicações da taxonomia são (Chari et al, 2020, p. 5):

- **Baseada em casos** (*case-based*), o sistema deve ser capaz de verificar outros casos semelhantes que possuem o mesmo resultado.
- **Contextual** (*contextual*), devem ser fornecidas informações mais amplas, que não envolvam os *outputs* e *inputs* do sistema, tais como informações relacionadas ao usuário, situações e informações contextuais que podem ter afetado a maneira que aquele resultado foi alcançado.
- **Contrastante** (*contrastive*), visa apresentar uma explicação sobre um resultado que pretende ser alcançado por meio de uma pergunta que faça uma comparação entre o fato (o evento que ocorreu), o resultado (*output*) e o evento que ainda não ocorreu (que é o resultado de interesse).
- **Contrafactual** (*counterfactual*), visa obter explicações sobre como um resultado no futuro poderia ser diferente caso outros dados de entrada (*inputs),* diferentes daqueles utilizados, fossem levados em consideração na análise. Aqui se pressupõe uma relação causal entre eventos e certos resultados;
- **Dia a dia** (*everyday*), oferece explicações associadas ao conhecimento comum do usuário, envolvidas em experiências do seu dia a dia, para explicar porque um determinado evento ocorreu.
- **Científica** (*scientific*), são explicações que dependem da aplicação de métodos científicos para explicar algo no mundo natural.
- **Baseada em simulação** (*simulation-based*), busca oferecer explicações baseadas na aplicação de métodos que simulam e implementam o funcionamento de um sistema ou processo pelo uso de dados de entrada semelhantes. Esta explicação pode englobar aspectos das explicações estatísticas e baseadas em rastreio de evidências (*trace-based*).

[331] São elas: métodos causais, métodos neuro-simbólicos de IA, representação semântica do espaço de explicação e o uso de tecnologias de registro distribuído (*Distributed Ledger Technologies* – DLT).

- **Estatística** (*statistical),* visa apresentar evidências matemáticas, numéricas, baseadas em dados, que sejam capazes de indicar a probabilidade de que determinado evento irá ou não ocorrer. Isto inclui explicação sobre os fatores que influenciam este resultado.
- **Baseada em rastreio de evidências** (*trace-based*), envolve explicações que apresentam uma sequência de passos que devem ser seguidos para que um sistema alcance um resultado específico (respondem à pergunta de por quê e como uma aplicação alcançou um resultado).

Esta taxonomia auxilia no objetivo de descobrir quais foram os fatores que motivaram uma decisão, e especificar quais são estes fatores e o conteúdo que deve ser apresentado. Esses tipos de explicações também fornecem as respostas colocadas por Doshi-Velez, uma vez que identifica diferentes causas que podem ser relevantes e devem ser apresentados para o usuário obter uma informação. Na realidade, é possível afirmar que a taxonomia é capaz de responder a novas perguntas, e se preocupa principalmente com o conteúdo do que deve ser apresentado, e a finalidade que visa atender.

Sob uma perspectiva individual de justificativa de uma tomada de decisão algorítmica, para um sujeito sem qualquer tipo de conhecimento técnico sobre o tema, as seguintes explicações deveriam ser fornecidas: baseada em casos, contextuais, contrastantes, contrafactual, e de dia a dia. Estas explicações permitem que o sujeito tenha elementos mínimos para compreender como que certos aspectos pessoais foram considerados pelo algoritmo, e lhe informam sobre como ele pode alterar uma situação indesejada alterando certos eventos, dados ou situações. Estas explicações podem ser geradas a partir de uma decisão concreta que tenha impactado a vida de uma pessoa em uma situação real, como podem resultar de testes criados pelos desenvolvedores do sistema como uma maneira de avaliar os seus resultados.

Em relação às outras explicações – científica, baseada em simulação, estatística e baseada em rastreio de evidências –, seriam conteúdos que atenderiam melhor a um público mais técnico, com expertise sobre o tema, como uma maneira de prestar contas tanto sobre aspectos ligados à arquitetura e às funcionalidades de um sistema, quanto à qualidade dos resultados do sistema. Conhecer a maneira como ele é modelado auxilia na identificação dos elementos que devem ser

investigados para que se possa obter explicações sobre uma decisão. Em um exemplo hipotético, se um sistema possui um componente de IA que fornece os parâmetros para avaliar o perfil de um menor órfão para definir em qual lar adotivo ele deverá residir, será importante compreender ao longo do sistema em que momento estes parâmetros serão utilizados (i.e. no momento de definição do perfil do menor e dos pais adotivos, ou na decisão em conceder a guarda para um dos candidatos a pais adotivos), e as saídas/resultados de cada parâmetro.

Fica evidente haver um amplo esforço da literatura, tanto jurídica, quanto técnica, em propor diferentes definições e delimitações do que seria o conteúdo adequado para explicar uma decisão algorítmica. Tendo em vista que um debate mais estruturado e direcionado sobre o direito à explicação tem origem na publicação do GDPR, o esforço da literatura jurídica naturalmente se dirige à interpretação do que a norma jurídica exige, seja sob uma perspectiva substancial, seja em relação às salvaguardas que devem ser garantidas ao titular de dados.

Neste sentido, o Information Commissioner's Office (ICO), autoridade de proteção de dados pessoais do Reino Unido, em conjunto com o Alan Turing Institute, um instituto governamental inglês que realiza pesquisas orientadas a desafios sociais, científicos e econômicos causados pelas inteligência artificial e ciência de dados, elaboraram um guia consultivo sobre explicação de sistemas de IA e decisões automatizadas. O trabalho é dividido em três partes: a primeira estabelece conceitos básicos sobre explicação de sistemas de IA; a segunda aborda maneiras de implementar explicações na prática; e a terceira estrutura medidas organizacionais que podem ser incorporadas para prover explicações.

Seguindo a linha do trabalho de Chari et al., a primeira parte do Relatório do ICO apresenta uma taxonomia com seis diferentes tipos de explicações:

- **Explicação das razões** (*rationale)*: as razões que levaram a uma decisão, apresentada de forma acessível e não técnica, o que pode incluir a influência que os dados de entrada tiveram nos resultados gerados pela IA. Seus objetivos são permitir que a decisão seja contestada, e viabilizar uma ação da pessoa impactada.
- **Explicação de responsabilidade**: informações sobre quem está envolvido no desenvolvimento, gerenciamento e implementação de um sistema de IA, e a entidade (pessoa, equipe ou departamento) responsável por realizar a

revisão humana de uma decisão. O seu propósito é informativo, e viabilizar que a decisão seja contestada.

- **Explicação dos dados**: informações sobre como e quais dados foram considerados em uma determinada decisão, além de como e quais dados foram utilizados para treinar e testar o modelo de IA.
- **Explicação de justiça/igualdade**: indica quais foram as medidas tomadas ao longo do desenvolvimento e produção de um sistema de IA para evitar discriminações. Seus objetivos são gerar confiança e permitir que a decisão possa ser contestada.
- **Explicação de segurança e desempenho**: indicar quais foram as medidas tomadas ao longo do desenvolvimento e produção de um sistema de IA para maximizar, garantir e demonstrar a precisão, confiabilidade, segurança e robustez de seus resultados. O seu propósito é informativo, visa permitir que a decisão seja contestada, e assegurar a adoção de medidas para gerar resultados seguros.
- **Explicação do impacto**: aponta qual é o impacto que o uso de um sistema de IA e suas decisões têm ou pode ter sobre uma pessoa em particular, e sobre a sociedade, em geral. Seus objetivos são garantir a autodeterminação do sujeito, e assegurar a adoção de medidas para gerar resultados seguros.

O conteúdo para estas explicações sugeridas pelo Relatório do ICO está aderente à proposta apresentada no capítulo segundo deste trabalho, que também visa organizar maneiras de fornecer uma explicação por meio de diferentes medidas e documentos. A necessidade de ter diferentes explicações, para atender aos seus respectivos propósitos, é justificada no FAT, bem como em outros princípios formulados pelo ICO: consideração do contexto operacional e de impactos. Desta forma, cada explicação está associada a um destes princípios, orientando e justificando qual o tipo de explicação que deve ser providenciada.

Conforme apresentado pelo ICO, a fim de atender ao princípio da transparência, devem ser apresentadas a explicação das razões e de dados, em um formato que seja compreensível e adequado, com um conteúdo que seja confiável e significativo. O princípio da prestação de contas e responsabilidade exige uma explicação de responsabilidade, explicitando quem são os responsáveis nas

diferentes fases de desenvolvimento do sistema IA, admitindo a identificação, o rastreio e a auditabilidade dessa responsabilidade. O princípio da consideração do contexto operacional exige que seja avaliada uma série de variáveis, como: o contexto em que a decisão automatizada ocorre, e as circunstâncias individuais de quem é impactado pela decisão, com o intuito de identificar o que deve ser fornecido em termos de explicação. Por fim, o princípio da consideração dos impactos envolve a reflexão sobre as repercussões causadas pelas decisões automatizadas (i.e. à integridade física, emocional e mental do titular de dados, à autodeterminação informativa e à autonomia da pessoa, entre outros), e exige explicações de justiça/igualdade, segurança e desempenho, além de impacto.

Conforme indicado na segunda parte do Relatório do ICO, as explicações exigem a criação de documentação, envolvendo a elaboração da explicação em si, como também a justificativa da escolha ou da priorização do tipo de explicação em relação às outras. É recomendado que as decisões algorítmicas em contextos que têm um maior impacto a direitos fundamentais exigem um esforço elevado de documentação[332]. Ainda, é ressaltada a importância de que sejam formalizados procedimentos que viabilizem o acesso a estas explicações, sugerindo a adoção do que vem sendo chamado de devido processo aplicado a usos de algoritmos (Citron; Pasquale, 2014; Crawford; Shultz, 2014; Kaminski, 2019b). Como se verá mais a frente, isto implicaria em garantir alguns direitos, como: notificação de quando uma pessoa estiver sujeita a uma decisão automatizada; acesso a informações que permitam que ela possa compreender como aquele resultado foi alcançado, quando possível; garantir a oportunidade de se manifestar, pedir uma revisão ou contestar a decisão; e, que um grupo de pessoas, ou uma pessoa humana, de preferência neutra, possa avaliar estes últimos pedidos (Kaminski, 2019b, p. 1.549)

Percebe-se que o presente trabalho e a proposta do Relatório do ICO parecem estar caminhando na mesma direção, pois ambos reconhecem a necessidade de uma abordagem multifacetada[333] para oferecer explicações. Isto envolve a elaboração de diferentes documentos e, consequentemente, a produção

[332] Além da documentação exigida em lei, é recomendado que sejam elaboradas documentação complementa para explicações de IAs, como Códigos de Boas Práticas, e implementação de medidas organizacionais adequadas para reduzir eventuais danos.

[333] Margot Kaminski usa a expressão "aggregate accountability" para se referir às diversas ferramentas/mecanismos que devem ser aplicados, diante das diferentes preocupações e agentes envolvidos no monitoramento e fiscalização dos algoritmos de tomada de decisão (Kaminski, 2019b, p. 1569).

de diversas informações, com o objetivo de atender ao FAT. Nota-se uma diferença nas taxonomias do ICO e aquela desenvolvida por Chari et al., em que na primeira as documentações que devem ser produzidas foram classificadas como tipos de explicações, enquanto no segundo há um foco maior no que ora tem se chamado de parte substancial da explicação. Diferentemente da classificação proposta pelo ICO, este trabalho organizou maneiras de operacionalizar explicações em tipos de documentações e medidas que podem ser adotadas antes da implementação de uma decisão algorítmica (*ex ante*), e depois que esta já ocorreu (*ex post*). Ainda, o presente trabalho aprofundou-se no debate da explicação das razões (*rationale*), algo que também é explorado por Chari et al. (2020). Apesar das diferentes maneiras de organizar este conteúdo, há uma convergência sobre a direção que o estudo sobre explicações de sistemas de IA deve seguir.

No mais, pode-se dizer que o ponto comum entre a literatura técnica e jurídica é mais a preocupação de que a explicação seja direcionada ao usuário/sujeito impactado, ou que potencialmente poderá ser impactado, pela decisão algorítmica, e menos os elementos técnicos e internos associados aos algoritmos de tomada de decisão. Expandindo esta perspectiva, a doutrina jurídica vem demonstrando a necessidade de que as informações referentes aos sistemas de IA não se limitem apenas aos usuários. A exigência de que sejam elaborados relatórios de impacto, sejam aqueles regulados pelas legislações de proteção de dados pessoais, ou quando há o desenvolvimento de sistemas de IA, que são documentos que possuem a finalidade de demonstrar o *compliance* regulatório para as instituições competentes, já demonstram uma inclinação em expandir a qualificação deste usuário. Ou seja, quando pensamos nas explicações e abertura de informações envolvendo decisões algorítmicas, há um espectro mais amplo sobre quem é este usuário e os seus interesses, e isto será determinante para definir que tipo de explicação deve ser apresentada.

Apesar disso, há elementos básicos de uma explicação que devem ser assegurados, independentemente da regra jurídica aplicável ao caso, até mesmo sob um ponto de vista ético[334]. Para tanto, é importante identificar por que uma

[334] "Essa opacidade dos sistemas de inteligência artificial, em especial daqueles baseados em aprendizado de máquina, faz com que o fornecimento de explicações adequadas do funcionamento de um sistema seja necessário, de um ponto de vista ético, mesmo quando não existe obrigação jurídica neste sentido" (Maranhão et al., 2021, p. 140)

explicação é necessária, e quais são os valores e interesses que ela visa garantir. Com estas respostas, será possível avaliar os elementos básicos para assegurar uma explicação de decisões algorítmicas adequada.

3.2
Porque é preciso explicar e os valores que se visa proteger

A necessidade de explicação de decisões algorítmicas possui distintos objetivos, e pode ser organizada em três principais razões (Maranhão et al., 2021, p. 147). De acordo com Maranhão et al., a primeira seriam os deveres éticos e jurídicos, que reconhecem a necessidade de um dever de transparência e a consequente interpretabilidade e inteligibilidade dos resultados algorítmicos, diante do impacto que estes resultados podem ter em interesses e direitos das pessoas. A segunda é uma razão técnica, de que explicações auxiliam na manutenção e no diagnóstico de decisões concretas, e permitem o acompanhamento da qualidade de seus resultados. A terceira seria o seu valor epistêmico, em que, em certas ocasiões, é possível que as explicações sejam capazes de indicar como os dados influenciaram um resultado produzido[335] (Maranhão et al., 2021, pp. 147-148).

Estes três objetivos não são silos, e dependem um do outro para que eles possam ser atendidos. O dever ético e jurídico da explicação apenas poderá ser atendido se uma explicação técnica for apresentada, exigindo que os resultados sejam passíveis de serem interpretados e explicados por uma pessoa. Por sua vez, a explicação epistêmica é importante para avaliar a legitimidade dos resultados gerados sob uma ótica ético-jurídica, sendo certo que a explicação técnica deve se modificar e se ajustar para atender a tais demandas. Portanto, esta interação reflete bem o quadro traçado por Lessig de como as diferentes ferramentas e instrumentos regulatórios se relacionam e influenciam um ao outro, além de demonstrar a interação destas áreas no contexto da governança algorítmica.

Neste sentido, em relação ao primeiro objetivo apontado acima, a seguir serão apresentados quais são os valores, direitos e interesses que se visa proteger quando uma explicação de decisões algorítmicas é exigida.

[335] Discussões travadas por boyd e Crawford (2012) e Mittelstadt et al. (2016) apontam sobre o tipo de conhecimento que é gerado, e a desconfiança dos resultados gerados no contexto do *big data* e no uso de modelos de aprendizado de máquina.

3.2.1
Liberdade e Direitos Fundamentais individuais e coletivos

Sob a razão ético-jurídica, a explicação visa atender alguns objetivos. A doutrina tem apontado que um destes é a importância em proteger a liberdade, autonomia, a dignidade da pessoa humana [336], o livre desenvolvimento da personalidade e a igualdade (Kaminski, 2019b; Selbst; Barocas, 2018; Doneda et al., 2018). O argumento se volta para a relação entre o desenvolvimento de perfis e a tomada de decisão algorítmica, e a opacidade deste processo, diante da ausência de qualquer notificação pelo agente econômico às pessoas impactadas, ou que podem ser impactadas, de que seus dados estão sendo utilizados para tanto, ou que elas já foram categorizadas.

Como se viu, estes perfis são comumente desenvolvidos com o apoio de técnicas de IA, além de serem um elemento essencial para a tomada de decisão algorítmica. A ausência de transparência também se refere à falta de informação sobre quais são os critérios relevantes para que uma pessoa seja categorizada e classificada dentro de um ou outro perfil. Especialmente quando há o uso de dados pessoais para a constituição e formação destes perfis, a opacidade deste processo como um todo acaba desrespeitando a autodeterminação informativa do titular de dados, ante a impossibilidade de que ele possa exercer o controle do fluxo de seus dados pessoais.

A ausência de notificação da aplicação desta técnica e os resultados gerados elimina a possibilidade de que a pessoa possa de certa forma participar deste processo e interferir nele. Os dados pessoais atualmente se constituem na principal maneira de representação das pessoas perante entidades públicas e privadas, podendo afetar oportunidades (Doneda et al., 2018, p. 4). Por isso, a representação equivocada da pessoa, que desconhece e/ou não participa deste processo, afeta sua personalidade e integridade moral. A possível perda de oportunidade pode acabar criando restrições que limitam sua autonomia, liberdade, e escolhas tanto econômicas, quanto existenciais (Doneda et al., 2018, p. 5). Isto implica, também, em ter o direito de ser tratado de forma igualitária e equânime, no sentido de que não seja gerada uma discriminação.

[336] Deve ser lembrado que são corolários do princípio da dignidade da pessoa humana, entre outros, os princípios jurídicos da igualdade e liberdade (Moraes, 2006, p. 17).

Há uma ideia de que a explicação é um "bem por si só" (*inherent good*) (Selbst; Barocas, 2016, p. 1.120), sendo importante que a pessoa possa participar deste processo decisório, avaliando quais são as categorizações feitas sobre ela e como ela vem sendo representada, tomar conhecimento sobre o uso de decisões automatizadas, e como que aspectos pessoais foram considerados pelo algoritmo de tomada de decisão. O acesso à informação fornece elementos importantes para que a pessoa possa tomar uma decisão informada, como, por exemplo, pedir uma revisão de uma decisão algorítmica, pois ela verificou que não foram levados em consideração o seu atual histórico de crédito. Portanto, o acesso à informação é um pressuposto para o exercício da liberdade e da autonomia, e estes últimos são condições chaves para o exercício de outros direitos. Aqui fica evidente uma outra característica do direito à explicação: seu valor funcional/instrumental (Selbst; Barocas, 2016; Selbst; Powles, 2017).

Vale ressaltar que a apresentação destas informações, como atualmente é feita, por exemplo, com as políticas de privacidade, pode gerar o mesmo problema que atualmente existe com o consentimento para tratar dados pessoais, em que informações essenciais são apresentadas em documentos longos, complexos, e com uma linguagem hermética, em que o dinamismo do modelo de negócio da internet não incentiva que as pessoas os leiam e compreendam. Por isso, os diferentes tipos de documentação que são propostos no capítulo 2 cumprem um papel importante, especialmente sob a dimensão objetiva dos direitos fundamentais, em que as instituições e órgãos competentes também deverão ser responsáveis por fiscalizar e supervisionar a atividade de controladores e desenvolvedores, na defesa destes direitos fundamentais.

Como colocado por Kaminski (2019b), exigir esta transparência com base em direitos individuais – como fundado na dignidade da pessoa, o direito à liberdade e à autonomia e no desenvolvimento da personalidade – é uma forma de dar conta das preocupações sobre justificativa e legitimidade destas decisões. Contudo, o exercício destes direitos também viabilizam, justificam e instrumentalizam outras demandas por explicação.

3.2.2
O valor funcional e instrumental

O exercício do direito à explicação em seu aspecto substancial tem o potencial de fornecer às pessoas impactadas, bem como àquelas que podem ser impactadas por decisões algorítmicas, informações para que elas possam tomar uma ação e proteger seus direitos e interesses, caso estes estejam ameaçados ou tenham sido violados. Ou seja, acessar a justificativa, as motivações, as "razões de decidir" deste algoritmo, é uma premissa e uma condição para que a pessoa possa exercer bens jurídicos tutelados pelo ordenamento, especialmente, direitos fundamentais.

O caso de moderação de conteúdo nas plataformas digitais, como visto no capítulo 1, julgado pelo Comitê Supervisor do Facebook, deixa evidente a relação intrínseca entre a transparência e, mais especificamente, a explicação dos motivos que justificaram a remoção e restrição de conteúdo, e direitos fundamentais. Compreender as motivações do algoritmo daquela rede social foi considerado essencial para verificar a sua aderência ou não às regras da plataforma, além de servir como uma maneira de avaliar o seu respeito a padrões internacionais de direitos humanos, especialmente sobre uma possível restrição indevida à liberdade de expressão. Isto é, saber como que determinada diretriz estava sendo interpretada e aplicada de maneira automatizada para fins de remoção de conteúdo é relevante para avaliar se há ou não o respeito a direitos humanos.

A relação de IA com a liberdade de expressão também foi pauta perante a Assembleia Geral da Organização das Nações Unidas (ONU), no já referido Relatório Especial sobre a promoção e proteção do direito à liberdade de opinião e expressão. De acordo com o Relatório, é inerente às discussões sobre regulação da IA as preocupações relacionadas aos direitos humanos, que devem ser sempre levadas em consideração, quando do desenvolvimento de regulações éticas ou de códigos de conduta ou setoriais. Uma abordagem de IA que considere os direitos humanos, de acordo com a ONU, deve ter como parâmetro dois princípios fundamentais: proteger e respeitar a agência individual e a autonomia (uma condição chave para exercer outros direitos, como a liberdade de expressão e opinião) e abertura significativa da tecnologia para que seja possível explicar a IA ao público (ONU, 2018, p. 16-17).

A ideia de que a explicação é um meio para o exercício de outros direitos encontra-se positivada no próprio GDPR, que explicitamente relaciona a necessidade de um direito à explicação com a aplicação de salvaguardas, e o seu impacto em outros direitos. O próprio GTA29 (GTA29, 2016, p. 10) reconhece isto,

ao apontar a importância de se ter acesso a informações quando há o uso de decisões automatizadas em contextos de análise de benefícios sociais que sejam garantidos por lei (como benefícios de moradia ou para menores), e que tenham sido recusados, ou quando tais meios automatizados são utilizados para impedir uma pessoa de cruzar a fronteira. Isto demonstra como uma explicação possui não apenas um valor em si mesmo, mas como se coloca como uma condição para o exercício de outros direitos.

Quanto a este ponto, da explicação enquanto um meio que viabiliza o exercício de outros direitos fundamentais, a perspectiva substantiva, de conteúdo, do que exatamente será apresentado, começa a ganhar um destaque especial. Pois, se o acesso a esta informação e o seu conhecimento é uma premissa para o exercício de outros direitos, a sua qualidade, o que e como será fornecido, ganham especial relevância. É neste sentido que as explicações contrafactuais e contrastivas têm sido apresentadas como uma solução para atender a este aspecto funcional e instrumental do direito à explicação, pois um dos seus objetivos é permitir que o sujeito tenha elementos que o instrumentalizem a tomar uma ação para mudar sua atual situação para um resultado melhor, e contestar esta decisão (Selbst; Barocas, 2016; Wachter et al., 2018; Mittelstadt et al., 2019).

3.2.3
A justificativa para avaliar a legitimidade e legalidade da decisão algorítmica

A explicação também é importante como uma maneira de atribuir legitimidade (Kaminski, 2019b) e gerar confiança nos resultados gerados por algoritmos de tomada de decisão (Kaminski, 2019b; Wachter et al., 2018; Edwards; Veale, 2017; Casey et al., 2019; Doshi-Velez; Kortz, 2017; Ribeiro et al., 2016). Isto é, explicações permitem que sejam avaliadas a legalidade e legitimidade do processo de tomada de decisão (Kaminski, 2019b, p. 1.546), uma vez que se pressupõe que sejam apresentados meios que permitem que uma pessoa tenha a capacidade de interpretar um modelo e seus resultados (Selbst; Barocas, 2016, p. 1.122). Este tipo de argumento se relaciona a princípios fundantes do Estado de Direito, devido processo legal ou um sistema amplo de *accountability* (Kaminski, 2019b, p. 1.545).

É o caso, por exemplo, quando a própria legislação protege determinados tipos de conteúdo, vedando o uso de certas informações para a realização de análises, tal como ocorre na Lei de Cadastro Positivo (Lei n. 12.414/2011 – LCP), que proíbe o uso de dados sobre saúde, origem social ou étnica, entre outros, para composição da nota ou pontuação do crédito da pessoa cadastrada em um banco de dados[337]. Este é um exemplo em que o legislador reconheceu que o uso de informações sensíveis para a atribuição de pontuação de crédito não seria justificável, portanto, não seria legítimo, proibindo o seu uso e o tornando ilícito.

A importância das explicações como motivações dos resultados algorítmicos pode ser comparada aos objetivos que a motivação de decisões judiciais visa alcançar. Não se pressupõe, tampouco se defende, que todas as decisões algorítmicas estejam sujeitas às mesmas regras, princípios, limites e escrutínios de uma decisão judicial, tendo em vista a formalidade exigida de um processo judicial, regulado formal e materialmente por um arcabouço próprio e específico. A exceção seria o caso em que magistrados estejam se valendo de algoritmos de tomada de decisão para apoiar a prestação jurisdicional ou para substituí-la, como já vem ocorrendo em outros países[338]. Tampouco está sendo sugerido que as explicações exigidas dos algoritmos sejam as mesmas que exigimos de magistrados. O argumento que se busca desenvolver é apenas observar os valores e objetivos que visa serem alcançados e atendidos por tal princípio, um pilar do Estado Democrático de Direito, e como estes se assemelham ao objetivo de justificar uma decisão algorítmica[339].

A motivação das decisões judiciais visa trazer a racionalização para o pronunciamento judicial, visando combater "arbitrariedades, autoritarismos e subjetivismos no momento de decidir" (Oliveira, 2020, p. 124), a fim de que seja garantida uma decisão que esteja justificada, seja legítima, legal e gere confiança da sociedade na prestação jurisdicional. Conforme aponta Humberto S. Oliveira, a

[337] Art. 7º-A, da LCP: "Nos elementos e critérios considerados para composição da nota ou pontuação de crédito de pessoa cadastrada em banco de dados de que trata esta Lei, não podem ser utilizadas informações (…) I - que não estiverem vinculadas à análise de risco de crédito e aquelas relacionadas à origem social e étnica, à saúde, à informação genética, ao sexo e às convicções políticas, religiosas e filosóficas; (…)"

[338] Como o uso do COMPAS nos Estados Unidos da América, explicado na nota de rodapé 20 e no capítulo primeiro desta tese

[339] Esta comparação também ganha especial relevância uma vez que o direito à motivação das decisões judiciais possui previsão constitucional expressa (art. 93, IX, da CF/88), sendo considerado como um direito fundamental. Além disso, há previsão infraconstitucional, no art. 11 do Novo Código de Processo Civil (NCPC).

motivação judicial visa alcançar algumas funções. Uma delas seria convencer as partes sobre a validade e qualidade da decisão, além de fornecer as justificativas para as partes recorrerem[340]. Outra função tem como objetivo avaliar: o processo como caráter instrumental (em que se busca avaliar se as garantias processuais foram atendidas e concretizadas), a determinação objetiva do julgado (se a decisão do juiz é justa e boa, e está coerente com o ordenamento jurídico vigente) e a racionalização do julgado à luz da jurisprudência (verificar a conformidade da decisão com a jurisprudência, visando à sua uniformização)[341] (Oliveira, 2020, pp. 131-136). Ou seja, a motivação de uma decisão judicial é uma espécie de prestação de contas e responsabilização do processo decisório do magistrado, tanto às partes processuais, como à sociedade. Isto também se coaduna com a proposta de governança binária, que tem demandas e objetivos tanto de caráter individual, como coletivo.

Outro pressuposto para uma motivação judicial adequada[342] é a participação das partes, por meio do contraditório, para que o magistrado possa avaliar e se manifestar sobre as evidências e argumentos apresentados pelas partes. Tal contraditório deve ser democrático e participativo, com garantias que permitam o envolvimento dos cidadãos no processo, possuindo um significado dialógico. Ou seja, é uma: "norma fundamental que assegura o direito das partes de manifestar, colaborar e influenciar a decisão judicial, além de permitir o controle dos pronunciamentos do magistrado" (Oliveira, 2020, p. 98-99).

Mais uma vez, estes direitos constitucionais e fundamentais, do que se busca concretizar com o princípio do contraditório e com o direito à motivação judicial, encontram-se bastante alinhados com as demandas de decisões algorítmicas, especialmente na literatura norte-americana, em que tem sido defendida a importância de se garantir um devido processo algorítmico (Kaminski, 2019b; Citron; Pasquale, 2014; Crawford; Shultz, 2014). Isto é, a fim de evitar uma

[340] Esta função estaria dentro da concepção endoprocessual, que visa analisar o caráter persuasivo da motivação, destinado às partes processuais (Oliveira, 2020, p. 131).
[341] Estes três objetivos estão aderentes à concepção extraprocessual, que tem por objetivo permitir que a sociedade possa realizar um controle externo das decisões judiciais (Oliveira, 2020, p. 134).
[342] "Uma adequada e completa motivação judicial pressupõe não apenas a apresentação de argumentos que corroborem ou demonstrem a correção das escolhas feitas pelo magistrado, mas também é necessário apontar justificativas que indiquem os motivos pelos quais outros fatos ou outros direitos discutidos foram relegados. A fundamentação das decisões inerente ao Estado Democrático de Direito deve abordar tanto as questões que apoiam a decisão tomada, quanto os motivos pelos quais determinados fatos e direitos relacionados à discussão foram rejeitados ou ignorados" (Oliveira, 2020, p. 145).

"'tirania' dos julgamentos automatizados" (Frazão, 2018a), devem ser asseguradas garantias mínimas para proteger o sujeito impactado por uma decisão automatizada. Isto implica: (i) notificar que a pessoa está sujeita à uma decisão automatizada, (ii) que ela seja informada sobre a decisão em si, (iii) possibilidade de que esta pessoa se manifeste, conteste e solicite a revisão da decisão; (iv) conheça a justificativa da mesma, e (v) acesse os dados utilizados para que a decisão tenha sido alcançada, conheça como ela está sendo representada neste processo, caso haja uso de perfis, e como seus dados influenciaram neste processo, e (vi) que uma pessoa humana possa analisar e conhecer a sua manifestação. Garantir a ciência e a participação da pessoa que foi impactada pela decisão algorítmica é uma maneira de assegurar a proteção de seus direitos individuais, e um pressuposto para o aspecto instrumental do sistema, pois permite uma avaliação sistêmica de eventuais erros e vieses que podem ser corrigidos, tendo também um impacto coletivo.[343]/[344]

Em suma, o que um direito à explicação visa proteger é que a pessoa esteja implicada neste processo. Ou seja, a pessoa humana deve ser o "ponto de chegada e de partida na análise do que consiste em ser uma explicação substantiva e significativa" (Mulholland; Frajhof, 2019, p. 287). Isto significa implementar e tornar eficaz todas as garantias citadas acima.

3.3
Previsões de explicação em normas setoriais no ordenamento jurídico brasileiro

Como já mencionado, pode-se remeter a origem da discussão sobre o direito à explicação à promulgação do GDPR e, consequentemente no Brasil, à LGPD. Apesar disso, no ordenamento jurídico brasileiro, é possível apontar para regulações setoriais que estabelecem o direito de uma pessoa obter informações quando há a aplicação de processos de decisões automatizadas (Leite, 2018), que derivam do princípio da transparência e do direito de acesso à informação. Estas previsões se relacionam especialmente com o contexto dos sistemas de pontuação de crédito.[345] O tema já possui jurisprudência consolidada do Superior Tribunal de

[343] De acordo com Kaminski, algoritmos exigem não apenas justificativas individuais, mas coletivas "(Broad policies for many people coupled with individual decisions with deep effects on individuals)" (Kaminski, 2019b, p. 1549)
[344] Conforme demonstrado no Capítulo 1.
[345] Sobre o tema, ver: ITS-Rio, 2017 e IDEC, 2017.

Justiça (STJ), no Recurso Especial n. 1.419.697/RS, que foi escolhido como caso paradigma, representativo desta controvérsia, e afetado pelo incidente de recursos repetitivos (art. 543, do antigo Código de Processo Civil – CPC). Desta forma, a interpretação sobre o assunto pode apontar para possíveis interpretações da previsão do direito à explicação da LGPD.

Assim, este subcapítulo irá apresentar a existência de previsões setoriais que estabelecem direitos e deveres no contexto de decisões automatizadas, e, ao final, retomar a análise do art. 20 da LGPD, para que sejam apresentadas propostas de diretrizes interpretativas do referido artigo. Ademais, serão apresentados os projetos de lei que ora tramitam no Congresso Nacional, e que trazem previsões relacionadas ao direito à explicação de algoritmos de tomada de decisão.

3.3.1
Pontuação de crédito: explicação no Código de Defesa do Consumidor (CDC) e na Lei de Cadastro Positivo (LCP)

O contexto de pontuação de crédito tem sido um exemplo corriqueiro para aplicações que envolvem algoritmos de tomada de decisão e usos de aplicações de inteligência artificial, em especial, aprendizado de máquina. Esta análise, tipicamente feita por birôs de crédito, como as empresas Boa Vista SPS e Serasa Experian, busca verificar a probabilidade de inadimplência de uma pessoa para categorizá-la dentro de um determinado perfil de risco. Tal perfil irá fundamentar a sua classificação dentro de uma certa pontuação de crédito, para analisar uma oferta de empréstimo, fixar taxa de juros, e eventualmente concluir um contrato. Para realizar toda sorte de previsões, são utilizados os dados pessoais de indivíduos (potenciais clientes, consumidores e titulares de dados) que podem estar relacionados ou não com sua capacidade creditícia, tal como dados relacionados à sua saúde. Estes birôs atualmente funcionam como verdadeiros *data brokers*[346], possuindo bancos de dados de informações provenientes de fontes públicas e privadas de dados, responsáveis não apenas por realizar avaliações creditícias, mas para atividades de marketing, prospecções de mercado e até mesmo transacionar a informação em si (ITS-Rio, 2017, pp. 4-5).

[346] De acordo com o ITS-Rio, *data brokers* são: "entidades que procuram extrair para os seus clientes conteúdo e utilidade da gama de informações às quais têm acesso, o que inclui, muitas vezes, transacionar a própria informação" (ITS-Rio, 2017, p. 5).

Para o exercício de suas atividades, estes birôs aplicam as técnicas de perfilamento e mineração de dados, já apresentadas no capítulo 1. Esta prática atrai a incidência das normas do Código de Defesa do Consumidor (Lei n. 8.078/1990 – CDC) e a LCP, criando um microssistema[347] regulatório da pontuação de crédito. Neste contexto, alguns direitos básicos devem ser garantidos ao consumidor, como: compreender as bases de dados pessoais utilizadas para desenvolver a pontuação de crédito, retificar as informações utilizadas que eventualmente estiverem erradas ou imprecisas, respeito ao prazo limite de até quando estas informações podem ser utilizadas, pedir a revisão da decisão automatizada, e impedir que sejam utilizados, na análise creditícia, dados com conteúdos protegidos (IDEC, 2017, pp. 7-8).

O CDC tem como princípios básicos a transparência[348] e a boa-fé[349]. A transparência nas relações de consumo significa o fornecimento de "informação clara e correta sobre o produto a ser vendido, sobre o contrato a ser firmado, significa lealdade e respeito nas relações entre fornecedor e consumidor, mesmo na fase pré-contratual (...)" (Marques, 2008, p. 286). Ou seja, é o direito de o consumidor ter acesso de maneira clara e simplificada às informações sobre o produto ou serviço oferecido pelo fornecedor, durante todas as etapas negociais.

A boa-fé, "– padrão de conduta ético entre o fornecedor e o consumidor – também inspira o *direito básico à informação* como um dos pilares do Código de Defesa do Consumidor" (IDEC, 2017, p. 26). Ambos os princípios, transparência e boa-fé, mostram-se indissociáveis, podendo-se afirmar "genericamente que a boa-fé é o princípio máximo orientador do CDC", e a transparência "um reflexo da boa-fé exigida aos agentes contratuais" (Marques, 2008, pp. 342-343), de forma que a facilidade no acesso à informação é uma presunção de quem conduz relações negociais de boa-fé.

O acesso à informação adequada e clara em relação aos diferentes produtos e serviços, além da apresentação dos riscos que os mesmos apresentem é um pressuposto, portanto, de toda relação de consumo. Tal direito é reiterado no art. 43,

[347] Gustavo Tepedino trata do tema ao escrever sobre a força normativa e a incidência de princípios constitucionais no direito privado, e a proposta interpretativa do direito civil constitucional. Nesta linha, o autor ressalta o seguinte: "Se o conceito de ordenamento pudesse se reduzir ao conjunto de normas de um mesmo nível hierárquico, poder-se-ia admiti-lo como universo técnico homogêneo e fechado em si mesmo. Sendo, ao contrário, o ordenamento jurídico composto por uma pluralidade de fontes normativas, apresenta-se necessariamente como sistema heterogêneo e aberto (...)". (Tepedino, 2009, p. 9)

[348] Art. 4º, *caput*, e inciso III, do CDC.

[349] Art. 6º, inciso III, do CDC

do CDC[350], que trata dos bancos de dados e cadastros de consumidores, sejam estes arquivos analógicos ou digitais, arrolando e especificando uma série de direitos básicos que surgem para o consumidor neste contexto (IDEC, 2017, pp. 26-27).

O primeiro deles se refere ao direito de acesso aos cadastros e dados, em relação "às informações existentes em cadastros, fichas, registros e dados pessoais e de consumo arquivadas sobre ele, bem como sobre as suas respectivas fontes". O segundo é a exigência da clareza sobre as informações, exigindo-se que "os cadastros e dados de consumidores devem ser objetivos, claros, verdadeiros e em linguagem de fácil compreensão". O terceiro é a vedação de que informações negativas (tal como restrições de crédito) possam ser mantidas por período superior a cinco anos[351] (art. 43, § 1º, do CDC), restringindo, portanto, a utilização destas informações por tempo indeterminado. O quarto é o direito que o consumidor possui de ser notificado, por escrito, quando houver "a abertura de cadastro, ficha, registro e dados pessoais e de consumidor" quando tenha expressamente solicitado tal abertura (art. 43, § 2º, do CDC). O quinto se refere à garantia de retificar informações incorretas ou inexatas, o que deverá ser atendido no prazo de cinco dias úteis, com o dever de que isto seja comunicado aos outros destinatários das informações incorretas (art. 43, § 3º, do CDC).

No contexto dos birôs de crédito, deve ser aplicada também a LCP, complementando e ampliando as garantias e direitos dos consumidores, além dos deveres das empresas. É possível estruturar o seguinte conjunto de deveres, já reconhecido pela jurisprudência, conforme será detalhado abaixo, que deve ser atendido pelos agentes econômicos quando da aplicação da metodologia de pontuação de crédito:

> a) dever de veracidade (as informações devem ser verdadeiras, ou seja, "exatas, completas e sujeitas à comprovação");
> b) dever de clareza (as informações devem permitir "o imediato entendimento do cadastrado independentemente de remissão a anexos, fórmulas, siglas, símbolos, termos técnicos ou nomenclatura específica");

[350] Como aponta Danilo Doneda, o art. 43 do CDC estabelece uma "série de direitos e garantias para o consumidor em relação às suas informações pessoais presentes em "bancos de dados e cadastros", implementando uma sistemática baseada nos *Fair Information Principles* à matéria de concessão de crédito e possibilitando que parte da doutrina verifique que neste texto legal o marco normativo dos princípios de proteção de dados pessoais no direito brasileiro" (Doneda, 2011, p. 103).

[351] O nome do devedor não pode ser mantido nos cadastros restritivos de crédito por um período superior a cinco anos (Súmula 323 do STJ), e o consumidor deverá ser notificado antes de realizar a sua inscrição (anotação) (Súmula 359 do STJ). (IDEC, 2017, p. 27).

> c) dever de objetividade (as informações devem descrever "fatos" e não informações "que envolvam juízo de valor");
> d) vedação de informações excessivas (os cadastros não podem envolver anotações "que não estiverem vinculadas à análise de risco de crédito ao consumidor");
> e) vedação de informações sensíveis (não se pode utilizar informações "pertinentes à origem social e étnica, à saúde, à informação genética, à orientação sexual e às convicções políticas, religiosas e filosóficas") (IDEC, 2017, p. 27).

Merece ser adicionado a este conjunto o que ora se considera como previsões verdadeiramente percursoras do direito à explicação. A LCP, em seu art. 5º, incisos IV e VI, respectivamente, estabelece que o consulente possa "conhecer os principais elementos e critérios considerados para a análise de risco, resguardado o segredo empresarial", além de ter garantida a "revisão de decisão realizada exclusivamente por meios automatizados". Ambos os direitos possuem redação similar ao art. 20, da LGPD, e têm sido considerados dentre os direitos que formam "a espinha dorsal do direito à explicação de decisões automatizadas em relações de consumo" (Leite, 2018, p. 8).

Como apontado pelo IDEC, a revisão é importante até mesmo para identificar eventuais dados errôneos, desatualizados ou que não poderiam ter sido coletados e armazenados, sendo possível pleitear o compartilhamento dos principais elementos e critérios que foram considerados na análise de risco. Por fim, esta revisão, e o acesso a tais informações, deve ser feita por uma pessoa humana, e não por sistemas de IA (IDEC, 2017, p. 31).

Este tema já foi analisado pelo STJ, por meio do Recurso Especial n. 1.419.697/RS. O julgamento, sob a relatoria do Ministro Paulo de Tarso Sanseverino, reconheceu a legalidade da prática de *credit scoring* (pontuação de crédito), que fora compreendida como uma metodologia que se vale de "fórmulas matemáticas para avaliação do risco de crédito, a partir de modelos estatísticos, considerando diversas variáveis de decisão" (STJ, 2014, p. 11), e utilizada para atribuir uma nota ao consumidor sobre o seu potencial de inadimplência. Embora não tenha sido utilizado expressamente o termo "algoritmo de tomada de decisão", ou "decisão automatizada", a utilização do termo *modelos estatísticos* indica que foram usados métodos de tomada de decisão criados a partir generalizações feitas por um conjunto de dados, em que estes métodos aprendem automaticamente as variáveis importantes de decisão para a tarefa em questão. Nestes casos, o próprio método em si é um algoritmo. Por isto, pode-se afirmar que a metodologia da

pontuação de crédito nada mais é do que a aplicação de um algoritmo preditivo que visa prever o grau de probabilidade de adimplência ou inadimplência de um sujeito a partir de uma análise estatística. Como entrada, são inseridos dados pessoais, com atribuição de pesos a cada uma destas informações de acordo com a sua relevância para o fim almejado, e a saída é a pontuação do consumidor, que será considerado mais ou menos confiável para adimplir, e alocado dentro de um dos perfis de consumo.

Ao analisar o caso concreto, o Relator reconheceu que a boa-fé objetiva é considerada como um dos princípios fundamentais que regem as relações de consumo, servindo como cláusula geral para avaliar a abusividade dos contratos consumeristas. Além do reconhecimento de que o dever de informação decorreria da função integrativa da boa-fé objetiva, foi ressaltada pelo STJ a importância de que a linguagem utilizada pelos fornecedores seja de fácil compreensão para que o consumidor seja informado de maneira clara e objetiva sobre todas as etapas negociais, o que envolve, também, a necessária abertura de quais dados da pessoa encontram-se armazenados nos arquivos de consumo.

A Corte destacou que tanto a LCP, como o CDC têm como finalidade assegurar a proteção da privacidade do consumidor quando há o uso de suas informações pessoais em bancos de dados (art. 43, do CDC), ou para a constituição do cadastro positivo (art. 3o, § 3º, II, da LCP). Para o STJ, no caso de pontuação de crédito, a privacidade do consumidor estaria sendo violada quando informações privadas sobre ele estiverem sendo utilizadas para uma finalidade diversa da análise da probabilidade de inadimplência (informações excessivas), ou quando informações sensíveis forem utilizadas para tal fim[352]. De acordo com o Ministro Relator, estas previsões da LCP buscam proteger a privacidade dos consumidores e, junto com o dever de clareza, objetividade e veracidade previstos no CDC, estabelecem um dever de transparência entre as partes contratantes. A não observância destes deveres – privacidade e transparência – na aplicação da metodologia de pontuação de crédito acarretaria na sua ilegalidade e abusividade (art. 187, do CC 2002), fazendo surgir a responsabilidade civil pelos danos

[352] Percebe-se, aqui, que "mesmo antes de haver uma previsão legal incorporando os princípios da finalidade e da necessidade (art. 6º, II e III, LGPD), o STJ já entendia ser abusiva a utilização de informações excessivas" (VIOLA; MATOS, 2018, p. 60).

materiais e morais causados, em que o banco de dados, a fonte e o consulente deverão responder objetiva e solidariamente (art. 16, da LCP).

Especificamente sobre o acesso a informações, o STJ compreendeu que os agentes envolvidos na prática de pontuação de crédito, ao utilizarem algoritmos de tomada de decisão para a classificação de consumidores como adimplentes ou inadimplentes, quando requisitados, devem prover aos sujeitos informações que sejam "claras, precisas e pormenorizadas acerca dos dados considerados e as respectivas fontes para atribuição da nota (histórico de crédito)" (STJ, 2014, p. 37). Foi ressaltado, no entanto, que a metodologia em si estaria protegida pelo segredo da atividade empresarial, além de ter sido dispensada a necessidade de obtenção do consentimento do consumidor para a aplicação da metodologia. O argumento para tanto é que, no caso concreto, o recorrido não havia sido inscrito no banco de dados de cadastro positivo, tendo ocorrido apenas a aplicação do método estatístico, compreensão que foi reproduzida na súmula do STJ n. 550.[353]

Percebe-se que, neste caso analisado pelo STJ, há uma clara opção da Corte em delimitar o dever de transparência dos gestores destes bancos de dados ao fornecimento das seguintes informações: dados dos consumidores que foram considerados pela fórmula matemática (endereço, idade, gênero, etc) e o histórico de crédito que serviu como parâmetro pela avaliação estatística para dizer se o sujeito é um bom ou mau pagador. Não foi mencionada, no caso, a necessidade de que fosse dada qualquer explicação sobre a forma de funcionamento do modelo utilizado, até porque entendeu-se que o mesmo estaria protegido pelo segredo comercial.

Uma crítica a esta decisão, no contexto da literatura sobre o direito à explicação, é que o acesso a estas informações não permite que o consumidor tenha uma real compreensão dos motivos que levaram o algoritmo a atribuir uma ou outra nota. Isto porque há apenas o dever de que sejam informados os dados pessoais considerados e o histórico do crédito, mas não a relevância de cada uma destas variáveis, os pesos que cada dado possui na formulação da nota final, por exemplo. Fornecer estas informações é uma maneira de garantir que tais critérios não estão

[353] Súmula 550, do STJ: "a utilização de escore de crédito, método estatístico de avaliação de risco que não constitui banco de dados, dispensa o consentimento do consumidor, que terá o direito de solicitar esclarecimentos sobre as informações pessoais valoradas e as fontes dos dados considerados no respectivo cálculo".

ocasionando resultados abusivos (IDEC, 2017, p. 23). O mero conhecimento sobre os dados utilizados não permite ao consumidor entender porque ele foi considerado como um bom ou mau pagador, pois ele não irá saber como estes dados foram valorados, o que foi relevante e ou não. Portanto, o tipo de informação indicado pelo STJ não permitiria que o consumidor tivesse plena compreensão das razões de decidir do algoritmo. Resta acompanhar se esta compreensão, com a promulgação e a vigência da LGPD, além do atual desenvolvimento social e tecnológico, será mantida ou atualizada quando casos semelhantes alcançarem o STJ.

Em suma, no contexto de pontuação de crédito, e aplicação do CDC e da LCP, podemos reconhecer a construção deu um rol de direitos importantes que sustentam uma "procedimentalização" para o exercício de um direito à explicação: (1) direito de acesso aos dados utilizados, (2) retificação de dados caso estejam inexatos ou incorretos, (3) acesso facilitado a informações, sendo que estas devem ser objetivas, claras, verdadeiras e mantidas em linguagem de fácil compreensão, (4) direito a ser notificado quando uma informação é inserida em um banco de dados, podendo estender esta interpretação para que se compreenda que tal dever também se aplica quando há um tomada de decisão automatizada, (5) acesso aos principais elementos e critérios considerados para a análise de risco, e (6) direito de pedir uma revisão por um humano de decisões automatizadas.

3.3.2
Os Projetos de Lei que regulam questões relacionadas ao direito à explicação

3.3.2.a
O PL n. 2.630/2020 – o "PL das Fake News"

O Projeto de Lei (PL) n. 2.630/2020,[354] proposto pelo Senador Alessandro Vieira (Cidadania/SE), e apresentado em conjunto com os Deputados Federais Tábata Amaral (PDT/SP) e Felipe Rigoni (PSB/SP), institui a "Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet". Este PL, que ficou conhecido como o "PL das Fake News", tem como principal objetivo evitar a disseminação de desinformação por meio da proibição de contas inautênticas,

[354] Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735. Acessado em 01.12.2021.

prevendo a possibilidade de que provedores de aplicação, em especial redes sociais e aplicativos de mensageria privada, sejam responsabilizados e sancionados pelo conteúdo disponibilizado em seu ambiente digital.[355] Contudo, com o passar do tempo, foram sendo apresentados diversos substitutivos, e adicionadas previsões que foram alargando o escopo do PL, levantando preocupações sobre a constitucionalidade e legalidade destas novas adições, em especial àquelas que se referem a agentes públicos.

Até o momento, a última versão que se tem do PL[356], que, após ter sido encaminhado para a Câmara dos Deputados, teve um substitutivo apresentado pelo Deputado Federal Orlando Silva, prevê a inclusão da Seção IV, denominada como "Dos Procedimentos do Devido Processo". Esta seção regula a moderação de conteúdo realizada pelas plataformas, que são hipóteses em que os provedores de aplicação e aplicativos de mensageria privada excluem, indisponibilizam, reduzem o alcance ou sinalizam conteúdos gerados por terceiros ou de suas contas, com base em seus termos de uso ou previsão legal. Pode-se dizer que as previsões desta seção estabelecem o que temos chamado neste trabalho de devido processo algorítmico, apesar de exclusivas para o contexto de moderação de conteúdos. Nestas situações, são estabelecidos direitos mínimos aos seus usuários, como: a notificação, prestação de informações, e a revisão.

Em relação à notificação (art. 15, inciso I, do PL), o PL indica que o usuário deverá ser notificado sobre: a natureza da medida aplicada e o seu âmbito territorial de aplicação (alínea "a"); a fundamentação da moderação, devendo ser identificada a cláusula das regras da aplicação ou a base legal que justificou a moderação, além do conteúdo ou a conta que motivou a decisão (alínea "b"); indicar o procedimento e o prazo para o exercício do direito à revisão; e se a decisão for tomada exclusivamente por meios automatizados, devem ser fornecidas informações claras e adequadas em relação aos critérios e procedimentos utilizados na decisão, nos termos do art. 20 § 1º, da LGPD (alínea "d"). A expressa referência ao artigo da LGPD que trata do direito à explicação contribui para a formação de um microssistema para o exercício deste direito, havendo uma complementação

[355] SILVA, Priscilla. A desinformação do PL das Fake News. *Revista Jota,* 01 de junho de 2020. Disponível em: https://www.jota.info/coberturas-especiais/liberdade-de-expressao/a-desinformacao-do-pl-das-fake-news-01062020. Acessado em 01.12.2021.
[356] Ver em: https://www.migalhas.com.br/arquivos/2022/4/A8C8933E3C5BA7_relatorio-fake-news.pdf. Acessado em 12.01.2022.

importante entre ambas as leis, no qual este PL estabelece as normas formais e de procedimento, e a LGPD fundamenta apresenta os direitos que podem ser exercidos, quando envolver o uso de dados pessoais.

O inciso II, do art. 15, do PL, prevê um dever de transparência, indicando que devem ser mantidas públicas e facilmente acessíveis, por um prazo mínimo de seis meses, as informações que foram prestadas, as denúncias de conteúdo apresentadas, as contas em operação e o envio de pedido de revisão de decisões. A importância de manter pública essas informações também foi reconhecida pelo Comitê Supervisor do Facebook como uma forma de prestar contas sobre o funcionamento da plataforma, além de dar publicidade à interpretação que tem sido feita das suas regras internas, ou outras previsões normativas.

As análises dos pedidos de revisão deverão apresentar respostas fundamentadas e objetivas, com a necessária reversão imediata da decisão original, caso constatado um equívoco (inciso III)[357]. Contudo, assim como a LGPD, o PL também se omite quanto à necessidade de uma revisão feita por pessoa natural (o que era uma garantia trazida pelo texto original do PL em seu art. 13, §1º, inciso VII). O legislador brasileiro em todas as oportunidades que teve em (re)estabelecer esta previsão no contexto de sistemas de inteligência artificial se absteve em assim fazê-lo.

Apesar de ainda estar em tramitação, e de regular outros temas relevantes para o contexto da Internet, este PL demonstra como a discussão de um direito à explicação de decisões automatizadas acaba extrapolando os limites de uma regulação sobre proteção de dados pessoais. Caso aprovado este substitutivo, ficará configurado mais um passo importante na garantia de direitos e deveres no contexto de decisões algorítmicas.

[357] O art. 15, inciso III, ainda estabelece os seguintes direitos: "§ 1º O código de conduta previsto no inciso III do art. 33 da presente Lei deverá dispor sobre os prazos razoáveis para cumprimento do inciso II do caput deste artigo. § 2º Os provedores devem observar as mesmas garantias do caput com relação às contas de que trata o art. 22 desta Lei. § 3º Caso constatado equívoco na aplicação de regras previstas no caput após avaliação de pedido de revisão, havendo dano individual, coletivo ou difusos a direitos fundamentais, os provedores de redes sociais ou mensageria instantânea devem, no âmbito e nos limites técnicos de seus serviços, informar os usuários sobre seu erro, na mesma proporção de alcance do conteúdo considerado inadequado, podendo esta obrigação ser requerida a autoridade judicial. § 4º Em caso de provimento do pedido de revisão, as medidas aplicadas devem ser imediatamente revogadas, devendo ser dada publicidade ao conteúdo restaurado".

3.3.2.b.
O PL n. 21/2020 – O Marco Legal da Inteligência Artificial

Tramita no Congresso Nacional o PL n. 21/2020, proposto pelo Deputado Federal Eduardo Bismark (PDT/CE), que estabelece os fundamentos e princípios para o desenvolvimento e aplicação da IA no Brasil. Diferentemente do que ocorreu com o MCI e a LGPD, normas que regulam o ambiente de inovação, empreendedorismo e desenvolvimento tecnológico, o PL n. 21/2020 não foi precedido de um amplo debate ao longo da sua tramitação, com a participação do setor público, da iniciativa privada, de pesquisadores e da sociedade civil. Por este motivo, pesquisadores têm vindo a público manifestando preocupações sobre a celeridade imposta à sua tramitação, o que atropelou a organização de um debate informado sobre o tema.[358]

As críticas ao PL se dirigem à completa ausência de concretude da proposta, que prevê apenas princípios abstratos, vagos e indeterminados, sem apresentar qualquer diretriz de como os mesmos deverão ser interpretados e aplicados pelos responsáveis pelo desenvolvimento de IA. Tampouco a norma estabelece sanções para quando houver o seu descumprimento. Como apontam especialistas[359], o PL tem se apresentado como uma carta de intenções muitas vezes afastando e restringindo direitos e garantias previstas no ordenamento jurídico, como aponta Laura Mendes, como é o caso do art. 6º, VI do PL, em relação à responsabilidade por risco do art. 927, do CC de 2002[360], ou o princípio da transparência que restringe o direito à explicação previsto na LGPD, o direito de acesso à informação do consumidor (art. 6º, III, do CDC), entre outros[361].

[358] MENDES, Laura Schertel. Projeto de Lei da Inteligência Artificial: armadilhas à vista. *Jornal o Globo,* 26 de nov., de 2021. Disponível em: https://blogs.oglobo.globo.com/fumus-boni-iuris/post/laura-schertel-mendes-pl-da-inteligencia-artificial-armadilhas-vista.html. Acessado em 12.01.2022. Entrevista de Bruno Bioni para o Jornal Folha de São Paulo, em 07 de dez., de 2021. Disponível em: https://link.estadao.com.br/noticias/cultura-digital,mais-importante-lei-de-tecnologia-no-brasil-nao-esta-sendo-debatida-diz-especialista,70003918886. Acessado em 12.01.2022.

[359] Conforme destacado pelos pesquisadores na matéria e entrevista indicadas acima.

[360] Civilistas apresentaram uma carta aberta ao Senado Federal, criticando a previsão do PL que define a responsabilidade civil como subjetiva como padrão em caso de danos decorrentes do uso de IA.

[361] MENDES, Laura Schertel. Projeto de Lei da Inteligência Artificial: armadilhas à vista. *Jornal o Globo,* 26 de nov., de 2021. Disponível em: https://blogs.oglobo.globo.com/fumus-boni-iuris/post/laura-schertel-mendes-pl-da-inteligencia-artificial-armadilhas-vista.html. Acessado em 12.01.2022.

Especialmente quanto a este último ponto, a versão do PL aprovada pelo Senado Federal em 2021 chama atenção pela ausência de qualquer expressão que faça menção à explicação ou revisão de sistemas automatizados. O princípio da centralidade do ser humano (art. 5º, II, do PL) indica a necessidade de respeitar a dignidade humana, a privacidade, a proteção de dados e os direitos fundamentais, quando o sistema tratar de questões relacionadas ao ser humano. A primeira crítica é a ausência de tecnicidade desta previsão. Indaga-se qual teria sido o intuito de limitar as situações em que estes direitos devem ser observados e, consequentemente, quando eles não devam ser respeitados. Será que o uso de sistemas de IA para o agronegócio seria considerado como "relacionado ao ser humano"? Ou o uso de sistemas preditivos para definir políticas públicas de orçamento ensejaria a necessidade de proteger tais direitos? Considerando tratar-se de direitos fundamentais, a restrição aos mesmos deve possuir autorização constitucional imediata (prevista expressamente na Constituição[362]) ou mediata (por reserva legal[363]) (Mendes, 2011, p. 227)[364], ou, sendo o caso de conflitos entre direitos fundamentais e outros valores constitucionais[365], deve ser realizada uma ponderação para resolução do caso concreto, justificando a sua limitação.

A segunda crítica à redação deste princípio é que não há qualquer menção à necessidade de uma supervisão humana de sistemas de IA, um princípio comumente presente nas cartas éticas de IA. Para atender a este objetivo, alguns documentos éticos (Comissão Europeia, 2019a), têm exigido a supervisão e agência humana (*human agency and oversight*) de sistemas de IA, como uma maneira de assegurar o respeito à autonomia humana e à sua autodeterminação. Como amplamente debatido, a omissão normativa da obrigatoriedade da presença humana na supervisão de algoritmos de tomada de decisão acaba minando a

[362] É o caso, por exemplo, do sigilo das comunicações telefônicas, que podem ser suspensas apenas mediante ordem judicial, "nas hipóteses e na forma que a lei estabelecer para fins de investigação criminal ou instrução processual penal" (art. 5º, inciso XII, da CF) (Mendes, 2011, p. 224)

[363] É o caso, por exemplo, da garantia do "livre o exercício de qualquer trabalho, ofício ou profissão, atendidas as qualificações profissionais que a lei estabelecer" (art. 5º inciso XIII, da CF) (Mendes, 2011, p. 225).

[364] "Os direitos individuais enquanto direitos de hierarquia constitucional somente podem ser limitados por expressa disposição constitucional (*restrição imediata*) ou mediante lei ordinária promulgada com fundamento imediato na própria Constituição (*restrição mediata)*" (Mendes, 2011, p. 227).

[365] Sobre o tema: "As situações de embates entre princípios podem assumir tanto a forma de colisão de direitos fundamentais, como a de conflito entre um direito fundamental e um outro valor consagrado da Constituição" (Mendes, 2011, p. 268).

autodeterminação informativa do sujeito, bem como sua autonomia e liberdade, criando um embaraço para o acesso à informação.

No mais, o princípio da transparência (art. 5º, V, do PL) aponta para a necessidade de que as pessoas sejam informadas de maneira clara, acessível e precisa sobre a utilização de sistemas de IA, salvo quando houver previsão legal em sentido contrário, observando-se o segredo comercial e industrial. Esta notificação deve: (a) ocorrer quando a pessoa estiver se comunicando com e utilizando diretamente sistemas de IA, como assistentes conversacionais (*chatbots*); (b) informar a identidade da pessoa natural ou pessoa jurídica responsável pela operação do sistema de IA; e (c) apresentar os critérios gerais que orientam o funcionamento do sistema de IA quando houver potencial de risco relevante para os direitos fundamentais, garantida a proteção ao segredo comercial e industrial. Embora possa ser elogiada a previsão que torna mandatória a notificação do uso de sistemas de IA, e a identificação do responsável pela operação do sistema, o que está em linha com o que tem sido defendido sob uma perspectiva procedimental do direito à explicação, o legislador foi silente quanto à implementação de outras salvaguardas, como o direito à revisão, de manifestar ou contestar os resultados de tais sistemas.

Ademais, não há qualquer menção no PL sobre a exigência de que estes sistemas sejam interpretáveis ou explicáveis. Quando tal exigência é prevista no art. 5º, V, alínea c, do PL, apenas se estabelece que sejam apresentados "critérios gerais que orientam o funcionamento do sistema de IA", que ainda assim podem não ser disponibilizados em razão do segredo comercial e industrial. Isto parece ser um convite para a elaboração de documentos rasos e básicos, sem quaisquer detalhes, que não trazem informações substanciais sobre o próprio sistema, tampouco sobre os impactos que podem ser causados pela IA. Além disso, tal previsão parecer sugerir que as informações a serem apresentadas estariam relacionadas meramente à funcionalidade do sistema. Teria sido positivo se tivesse previsto, expressamente, a necessidade de que fosse elaborado um relatório de impacto, tal como ocorre no contexto de proteção de dados pessoais.

O PL, que poderia resolver as lacunas deixadas pela LGPD, em especial, em relação ao veto da revisão humana de decisões automatizadas, à possibilidade de melhor delimitar o que seria uma decisão totalmente automatizada, e esclarecer o tipo de explicação e informação que deveriam ser disponibilizadas em relação aos

resultados e funcionamento de sistemas de IA, acaba restringindo e limitando o microssistema regulatório do direito à explicação de decisões algorítmicas, constituído pelas leis de proteção de dados pessoais, consumeristas e de cadastro positivo. Resta acompanhar o debate que se seguirá do substitutivo no Senado Federal, verificar se as críticas que vêm sendo feitas ao PL e à sua condução serão sanadas ao longo processo legislativo.

3.4
A importância em dar publicidade para as explicações

No primeiro capítulo o conceito de inteligência artificial foi apresentado, ficando nítido como o aspecto humano é uma característica essencial, tanto sob uma perspectiva filosófica, de a IA ter como objetivo emular uma ação ou pensamento humano, quanto sob o aspecto técnico, de a pesquisa ter caminhado para tentar se assemelhar ao funcionamento do cérebro humano. Esta relação também fica evidente na atual abrangência do uso da IA e seus algoritmos, que vêm paulatinamente se instalando em processos de decisões antes tomadas por humanos (diagnóstico médico, contratação de pessoas, condenação judicial, etc).

Esta onipresença tem motivado a exigência de transparência, prestação de contas e responsabilização e, mais especificamente, de uma explicação. Como visto, estudos na área de psicologia têm apontado para uma forte necessidade de que explicações devam apresentar uma relação causal entre a ocorrência de um determinado fato e os motivos e eventos que o justificam (Hoffman et al., 2017a; 2017b; Klein, 2018). Pode-se dizer que existe esta mesma expectativa para a demanda por explicação de um comportamento ou um pensamento, como, por exemplo, quando exigimos uma explicação do porquê uma transeunte não doou dinheiro para uma pessoa em situação de rua, ou porque uma pessoa não levou seu guarda-chuva ao sair de casa quando havia fortes indícios de que iria chover.

Explicações e justificativas[366] são apresentadas por meio do uso da razão, em que a apresentação de bons argumentos que justifiquem um comportamento ou

[366] Há autores que diferenciam a justificativa da explicação em contextos de racionalização, como Jesse Summers (2017). Para o autor, a justificativa é composta pelas considerações que militaram a favor da nossa ação, enquanto a explicação apresentaria os fatores causais que explicam porque nos comportamos de tal maneira. O autor admite que a verdade não seria o único critério para decidir o que seria uma melhor explicação. As justificativas apresentadas como motivos para agir nem sempre são as explicações corretas ou completas. Elas podem ser parciais e incompletas, e o fato de uma

pensamento é importante como uma maneira de gerar confiança em nossas ações e discursos. Neste sentido, o uso da razão tem sido objeto de estudo da área de filosofia do direito e da psicologia comportamental.

Mercier e Sperber, em seu livro "*The enigma of reason",* revisitam o estudo sobre a razão, e o papel que ela exerce nas relações sociais humanas. Segundo os autores, a razão se desenvolveu como um atributo humano para exercer duas funções diferentes: ela é utilizada para que possamos justificar a nós mesmos e para produzir argumentos para convencer, e avaliar criticamente a justificativa apresentada pelos outros (Mercier; Sperber, p. 18). A razão também possui um uso explicativo, que está a serviço da justificativa, e tem a função de relacionar não apenas a razão a um pensamento ou comportamento de maneira teórica, mas ao seu agente (Mercier; Sperber, p. 313). Esta posição busca se diferenciar da tradicional defesa que é feita de que a razão teria evoluído e é utilizada para que o indivíduo se desenvolva e alcance maior conhecimento e tome as melhores decisões individualmente. Os autores defendem o papel social, de cooperação e comunicação, dialógico e interativo que a razão possui.

De acordo com os autores, a razão é sempre apresentada de maneira *ad hoc* para justificar e explicar um determinado comportamento ou pensamento (Mercier; Sperber, 2017, p. 184-185). Os autores acreditam que a razão não seria responsável por guiar nossas crenças e decisões primariamente, mas sim para justificar nossas ações e pensamentos, e avaliar criticamente as justificativas apresentadas pelos outros (Mercier; Sperber, 2017, p. 192-198). Estas justificativas, que seriam meras racionalizações[367] após o fato[368], nem sempre representam os verdadeiros motivos que ocasionaram determinada crença ou tomada de decisão, e a explicação do processo que motivou o julgamento sob o ponto de vista causal. Os autores defendem que, mesmo nos casos em que decisões sejam tomadas de maneira

justificativa explicar a ação não implica em ser uma razão suficiente para esta ação (Summers, 2017, p. 26). No entanto, esta diferenciação nem sempre fica muito clara, razão pela qual não será feita uma distinção entre elas.

[367] Diferentes autores atribuem diferentes conceitos sobre o que é a racionalização, ora atribuindo efeitos positivos a este processo (Summers, 2016 e Cushman, 2020), ora negativos (Schwitzgebel; Ellis, 2016).

[368] Como indica Haidt, ao mencionar o estudo de Nisbett; Wilson, 1977: "when asked to explain their behaviors, people engage in an effortful search that may feel like a kind of introspection. However, what people are searching for is not a memory of the actual cognitive processes that caused their behaviors, because these processes are not accessible to consciousness. Rather, people are searching for plausible theories about why they might have done what they did" (Haidt, 2001, p. 822)

consciente, as reais motivações seriam inconscientes e tampouco estariam abertas para inspeção (Mercier; Sperber, 2017, p. 195).

O que é interessante no argumento dos autores é a importância na interação e avaliação de terceiros em relação a uma explicação, que pretensamente deve ser racional, apresentada para justificar uma determinada ação ou pensamento, como uma maneira de refinar e melhorar comportamentos e ideias. Este objetivo do uso da razão é especialmente relevante quando se discute o direito à explicação, e é uma lição que vem sendo abordada por outros autores especialistas sobre o tema (Mittelstadt et al., 2019).

Este posicionamento parece interessante para aplicar às explicações exigidas na IA. É comum que essas duas áreas sejam relacionadas (Cushman, 2020; Mercier; Sperber, 2017; Klein, 2018) sob uma perspectiva dos processos cognitivos voltados para compreender a maneira que a inteligência artificial aprende e evolui seu aprendizado e, portanto, desenvolve sua autonomia. A própria definição do que seja a IA perpassa por essa compreensão, embora não se ignore que tal parâmetro acaba superestimando o valor da razão, que é falha e enviesada, como demonstram diversos autores[369].

No mais, se reconhecemos que os humanos realizam racionalizações[370] e que a inteligência artificial, ao fim e ao cabo, tenta reproduzir o humano, assim como reconhecemos ser impossível inspecionar e compreender o funcionamento da mente humana e acessar os reais motivos que justificam um pensamento ou comportamento, nos satisfazendo com as razões externadas que tentam lhes explicar, será que não estaríamos inclinados a aceitar que, em alguns casos, mesmo que não seja possível acessar a maneira como os algoritmos de tomada de decisão funcionam, deveríamos explorar as razões que são apresentadas para explicar uma determinada decisão?

Sob esta perspectiva, independentemente de ser possível ou não obter uma explicação dos verdadeiros motivos que desencadearam uma decisão algorítmica, a proposta apresentada no capítulo dois demonstra a importância de produzir um conjunto de documentos e evidências como uma maneira de expor "razões" que

[369] Mercier; Sperber; Haidt, 2001, para citar apenas alguns.
[370] Existem diferentes conceitos de racionalizações, que varia de autor para autor. Aqui me refiro genericamente ao fato de que os humanos apresentam razões para justificar e explicar suas ações como motivações após terem praticado este ato, e que estas razões não necessariamente correspondem às reais motivações do agente, pois estas não seriam acessíveis para inspeção.

sejam capazes de explicar e justificar os resultados de um algoritmo de tomada de decisão de ML de maneira mais específica, ou um sistema de IA de forma geral. Se para explicar o comportamento humano, nossas escolhas e decisões, nós apresentamos nossas razões, *ex post facto,* fundadas nas mais diversas fontes, moralmente carregadas, enviesadas e falhas, é certo que as explicações em relação aos resultados algorítmicos, assim também serão.

Desta forma, para evitar seus efeitos negativos, independentemente do conteúdo que será apresentado, é preciso que estas razões estejam disponíveis para avaliação de terceiros, como uma forma de assegurar transparência e a prestação de contas e responsabilização de algoritmos de tomada de decisão. Isto funciona como uma maneira de permitir que aqueles que fornecem explicações refinem, ajustem, melhorem e desenvolvam seus argumentos e suas razões. A razão como uma maneira de justificar e explicar uma decisão algorítmica, também atende à finalidade de gerar aceitação social sobre aquele resultado, o legitimando, e, portanto, gerando uma confiança em relação a ele.

Conforme aponta Kaminski (2019b), o acesso a informações relacionadas a uma decisão em si é importante sob aqueles três aspectos: proteção da dignidade da pessoa humana[371], por questões de justificativas e instrumental. O importante é que a transparência assegure a proteção das liberdades, interesses e direitos que podem ser impactados pelas decisões algorítmicas. Mais uma vez, repita-se, de forma quase redundante e repetitiva, que, apenas será possível saber se um direito foi violado ou ameaçado se houver transparência e conhecimento sobre processo de tomada de decisão e a justificativa da decisão em si.

3.5
O direito à explicação: garantindo o mérito e o seu procedimento

[371] Como indica Maria Celina Bodin de Moraes, o princípio da dignidade da pessoa humana "visa garantir o respeito e a proteção da dignidade humana não apenas no sentido de assegurar um tratamento humano e não degradante, e tampouco conduz ao mero oferecimento de garantias à integridade física do ser humano (...)". A dignidade da pessoa humana vai exigir que as pessoas sejam tratadas em sua humanidade, e não como objeto. Continua a autora indicando que o substrato material de tal princípio pode ser entidade e desdobrada em quatro postulados: "i) o sujeito moral (ético) reconhece a existência dos outros como sujeitos iguais a ele; ii) merecedores do mesmo respeito à integridade psicofísica de que é titular; iii) é dotado de vontade livre, de autodeterminação; iv) é parte do grupo social, em relação ao qual tem a garantia de não vir a ser marginalizado" (Moraes, 2009, p. 16-17).

Como apresentado no capítulo 2, a eficácia de um direito à explicação não depende apenas do oferecimento de informações sobre como que determinado algoritmo alcançou uma decisão. É preciso viabilizar meios que permitam o exercício deste direito, além da necessidade de produção e disponibilização de um conjunto de documentos referentes ao sistema e ao algoritmo de tomada de decisão como um todo, produzidos antes (*ex ante*) e depois da sua implementação (*ex post*). Isto porque diante da inerente opacidade do algoritmo, a produção deste conjunto de elementos será fundamental para obter evidências sobre o seu funcionamento, com a possibilidade de que sejam avaliados. A previsão do art. 20 da LGPD permite que seja aplicado este entendimento sobre o direito à explicação.

Inicialmente, é preciso definir o nível de intervenção e envolvimento humano necessário para que uma decisão tomada por um algoritmo seja enquadrada dentro do escopo do art. 20, da LGPD. É dizer, quando uma decisão algorítmica será considerada como "tomada unicamente com base em tratamento automatizado de dados pessoais". O GDPR, em seu art. 22, na tradução portuguesa da norma, se vale do termo "exclusivamente", e o art. 20, da LGPD, usa o termo "unicamente". Ambos possuem definições semelhantes, e implicam em uma decisão exclusiva, excludente e única. Na União Europeia, o GTA29 já se posicionou quanto à interpretação do mencionado artigo (GTA29, 2017, p. 9-10), indicando as qualificadoras para que se caracterize o envolvimento humano substantivo, a fim de descartar a incidência do art. 22, do GDPR.

Estender esta interpretação para o art. 20, da LGPD, pode acabar restringindo a incidência da proteção trazida pela norma. É comum que algoritmos de tomada de decisão sejam utilizados para apoiar uma tomada de decisão (Edwards; Veale, 2017, p. 45), em que esta não é a resposta final e definitiva da predição ou do resultado, mas utilizada como uma evidência para decidir. Como aponta Edwards e Veale, algumas autoridades nacionais de proteção de dados na UE têm apontado preocupações sobre a possibilidade de afastar a incidência do art. 22, do GDPR, em casos em que há usos semiautomatizados, considerando que estes também trazem riscos aos titulares de dados. Neste sentido, ora se defende que, apesar da interpretação atribuída pelo GTA29, uma decisão tomada unicamente com base em tratamento automatizado na LGPD deva ser interpretada de forma ampla, a fim de que sejam consideradas as decisões algorítmicas que tenham ocorrido sem qualquer intervenção humana substantiva, quanto aquelas que foram

utilizadas para apoiar uma decisão humana, verificando-se se a participação humana foi significativa e adequada naquele contexto (Wimmer; Doneda, 2021, p. 393). Neste contexto, vale ressaltar o que tem se chamado de um viés de automação (*automation bias*), em que pesquisas apontam uma tendência humana em "assumir a validade de decisões feitas por algoritmos, mesmo quando apresentadas com informações que diretamente contradizem a decisão de aparente validade" (Casey et al., 2019, p. 146). [372] Soma-se a isto a complexidade em compeender estes resultados, o que pode gerar um ônus argumentativo excessivo para a pessoa, que pode não ser superado, para não adotar a solução apresentada pelo sistema automatizado (Wimmer; Doneda, 2021, p. 394).

Um paralelo sobre o assunto pode ser feito a partir do relatório produzido pela organização sem fins lucrativo *Human Rights Watch* e a Clínica Internacional de Direitos Humanos de Harvard chamado "*Losing Humanity: The Case against Killer Robots*"[373] (Burri, 2016). O relatório, elaborado em novembro de 2012, foi responsável por iniciar um movimento da sociedade civil, que ganhou forte apelo internacional, pela proibição de "robôs matadores" (*killer robots*) utilizados em ambiente de guerra[374]. Nesse documento, apesar de datado e específico para máquinas destinadas para uso militar e de guerra, são apresentadas três definições sobre o nível da autonomia dos robôs. São elas: (1) armas com humanos-*dentro*-do-loop (*humans-in-the-loop*), em que os robôs selecionam seus alvos e executam sua força com um humano no comando; (2) armas com humanos-*no*-loop (*humans-on-the-loop*), onde robôs selecionam seus alvos e executam sua força sob a supervisão de um humano, que pode não respeitar a ação da máquina; (3) armas com humanos-*fora*-do-loop (*humans-out-of-the-loop*), em que robôs selecionam seus alvos e executam sua força sem qualquer *input* ou interação com um humano[375]. De acordo com o relatório, robôs dentro dos grupos (2) e (3) seriam

[372] Tradução livre de: "assume the validity of decisions made by algorithms, even when presented with information that directly contradicts the decision's apparent validity" (as referêcias indicadas no original foram removidas).

[373] Human Rights Watch e Harvard International Human Rights Clinic. *Losing Humanity: The Case against Killer Robots.* Nova Iorque, 2012. Disponível em: <https://www.hrw.org/report/2012/11/19/losing-humanity/case -against-killer-robots>.

[374] Este relatório também é citado por Citron e Pasquale (2014), para fundamentar sua crítica sobre os potenciais danos causados por robôs e o risco de que estes estejam violando tratados internacionais.

[375] Tradução livre de: "Human-*in*-the-Loop Weapons: Robots that can select targets and deliver force only with a human command; Human-*on*-the-Loop Weapons: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots' actions; and

classificados como de plena autonomia, por conta da limitada supervisão que um humano tem no seu comportamento. Neste cenário, pode-se compreender que o uso de decisões automatizadas utilizadas para apoiar uma tomada de decisão podem ser classificadas dentro do grupo (2).

Ademais, um outro argumento favorável a esta posição é a associação de decisões unicamente automatizadas à afetação de interesses e aspectos da personalidade do titular de dados, bem como à formação de perfis, e a interpretação atribuída ao art. 12 § 2º, da LGPD. Isto é, os dados utilizados para a formação de perfis comportamentais, mesmo que não identifiquem uma pessoa, deverão ser considerados como dados pessoais, incidindo a LGPD, diante da necessidade de proteger a autodeterminação informativa e o livre desenvolvimento da personalidade da pessoa (art. 2º, II e VII).[376] Ou seja, a fim de assegurar tais direitos, que são fundamentos da LGPD, deve se compreender que as decisões tomadas unicamente com base em tratamento automatizado devam englobar decisões algorítmicas utilizadas para apoiar a tomada de decisão humana, tal como é feito com o uso de perfis.

Quando nos referimos à questão de mérito de uma explicação de uma decisão algorítmica, isto está relacionado ao conteúdo de uma decisão que tenha impactado uma pessoa e, portanto, poderia ser obtida, quando tecnicamente possível, por meio de evidências *ex post*. Caso seja tecnicamente viável compreender como que determinado resultado foi alcançado, será preciso verificar quem é o sujeito ou a entidade que está solicitando a explicação, e avaliar o tipo de conteúdo que deverá ser fornecido (i.e. explicações contrafactuais, estatísticas, baseada em casos, etc.), conforme indicado acima. Independentemente, caso não seja tecnicamente possível oferecer este conteúdo, a documentação gerada *ex ante* irá exercer um papel fundamental.

Especialmente no que se refere ao direito à explicação previsto na LGPD, pode-se dizer que o mesmo tem como fundamento três principais pontos: "o princípio da transparência, o direito de acesso à informação e como um pressuposto

Human-*out-of*-the-Loop Weapons: Robots that are capable of selecting targets and delivering force without any human input or interaction". Disponível em: Human Rights Watch e Harvard International Human Rights Clinic. *Losing Humanity: The Case against Killer Robots.* Nova Iorque, 2012. Disponível em: https://www.hrw.org/report/2012/11/19/losing-humanity/case -against-killer-robots, p. 2.

[376] Este argumento foi melhor desenvolvido no capítulo 1, subitem 2.6. O Direito à Explicação previsto na LGPD.

para o exercício de outros direito e, particularmente, do direito a requerer revisão de decisões automatizadas" (Souza et al., 2021, p. 478). Inclsuive, o acesso à informação é um pressuposto para o exercício do direito à revisão, pois esta "somente se legitima quando é capaz de explicitar os critérios e vetores que a inspiraram (...)*"* (Wimmer; Doneda, 2021, p. 386). Assim, combinando tais fundamentos com os direitos e interesses que a explicação de decisões algorítmicas visa assegurar – proteção de direitos fundamentais, seu valor funcional/instrumental e de justificativa –, é preciso contemplar um direito à explicação que forneça informações sobre as "razões de decidir" da decisão algorítmica.

Em relação ao conteúdo da explicação em si, o art. 20, § 1º, da LGPD, prevê que "o controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada". Souza et al. (2021, p. 480) apontam que o termo "critérios e procedimentos" pode aludir a uma interpretação de que o controlador deveria apresentar informações sobre como o algoritmo opera e sua funcionalidade. Contudo, os autores ressaltam que não há a obrigatoriedade de que sejam divulgados elementos técnicos aos titulares de dados, mas o suficiente para que estes possam exercer outros direitos, tal como o próprio direito de revisão previsto no *caput* do art. 20, e o direito de oposição (Souza et al., 2021, p. 479).[377] No mais, não há nada na LGPD que indique em que momento uma explicação poderá ser exigida, se antes ou depois de uma decisão automatizada, afastando-se da discussão existente no âmbito do GDPR (Souza et al., 2021, p. 480).

Apesar das contundentes e acertadas críticas à proteção ao segredo comercial e industrial previsto pela norma, com o intuito de evitar que o agente de tratamento faça uso deste direito, seria oportuno que fosse constituído um fórum técnico qualificado na ANPD capaz de receber estes tipos de documentos e proceder com a devida análise. Como sugerido, a publicação destes resultados seria fundamental para que a sociedade civil, o terceiro setor, pesquisadores e outros interessados pudessem acessar este conteúdo e analisá-lo criticamente[378], sendo uma maneira de legitimar o sistema de IA como um todo e seus resultados.

[377] Art. 18, da LGPD: "O titular dos dados pessoais tem direito a obter do controlador, em relação aos dados do titular por ele tratados, a qualquer momento e mediante requisição: (…) § 2º O titular pode opor-se a tratamento realizado com fundamento em uma das hipóteses de dispensa de consentimento, em caso de descumprimento ao disposto nesta Lei".

[378] Tal como ocorreu na investigação do cartão de crédito da Apple, apresentada no capítulo anterior.

No mais, este mesmo fórum técnico seria responsável por realizar a auditoria prevista do art. 20 § 2º. A LGPD, mais uma vez, acaba limitando o direito dos titulares de dados, ao prever que uma auditoria poderá – portanto, de maneira discricionária – ser realizada pela ANPD para verificar aspectos discriminatórios no tratamento de dados, e apenas quando as informações do § 1º, do art. 20 não forem compartilhadas quando justificadas pelo segredo comercial e industrial. Como já visto no capítulo 2, estas limitações são verdadeiros empecilhos que fragilizam o direito à explicação e revisão do titular de dados, além dos princípios da finalidade, visto que a ausência de auditoria não permite que sejam avaliados os propósitos legítimos e específicos do tratamento, da transparência, não discriminação e responsabilização e prestação de contas (art. 6º, I, VI, IX e X)[379]. Por tais motivos, seria adequado que, a ANPD, no exercício de sua competência de realizar auditorias ou determinar que estas ocorram[380], ampliasse a interpretação do art. 20, § 2º, para que a auditoria fosse mandatória, e não discricionária, e não se limitasse à hipótese em que o agente de tratamento se nega a fornecer as informações solicitadas com base no segredo comercial e industrial. No mais, as lacunas e as indeterminações deixadas pelo legislador na LGPD acabam dando margem para que estas sejam preenchidas e definidas pela ANPD e pelo Poder Judiciário.

Ainda, as evidências de prestação de contas produzidas quando do desenvolvimento e implementação do sistema de IA e seu respectivo algoritmo de tomada de decisão também podem ser exigidas e, neste caso, deve haver um evidente interesse público no fornecimento destas informações. Caso estas evidências não tenham sido disponibilizadas ao público, pode ser exigido que estes

[379] Art. 6º, da LGPD: "As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios:
I - finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades; (…) VI - transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial; (…) IX - não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos; X - responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas."
[380] Art. 55-J, da LGPD: "XVI - realizar auditorias, ou determinar sua realização, no âmbito da atividade de fiscalização de que trata o inciso IV e com a devida observância do disposto no inciso II do caput deste artigo, sobre o tratamento de dados pessoais efetuado pelos agentes de tratamento, incluído o poder publico".

documentos sejam compartilhados, sob a forma da transparência qualificada. Nesta hipótese, as informações seriam compartilhadas para uma instituição competente, como a ANPD, ou mediante um pedido judicial, para que pessoas especializadas e com competência realizem uma análise do algoritmo de tomada de decisão. O objetivo é que sejam oferecidas explicações sobre uma decisão em particular, respondendo a questões referentes ao funcionamento do sistema como um todo.

Neste sentido, a taxonomia desenvolvida por Zarksy em relação à transparência pode ser útil para avaliar o tipo de informação que deverá ser apresentada, definindo o tipo de explicação e as evidências que devem ser fornecidas, como o relatório de impacto, os *model cards*, e os resultados de eventuais auditorias e testes internos. Isto pode ser especialmente relevante no caso do direito à explicação previsto na LGPD, por exemplo, que limita o fornecimento de informações para preservar o segredo comercial e industrial.

O direito à explicação também possui um aspecto de procedimento, já previsto pelo GDPR, por exemplo, que estabelece a necessidade de implementar salvaguardas quando há a tomada de decisão exclusivamente por meios automatizados. Nesta situação, o titular de dados deverá ser informado quando estiver sujeito à uma decisão totalmente automatizada, manifestar seu ponto de vista, obter uma intervenção humana, obter uma explicação, ter acesso a informações úteis relativas à lógica subjacente, e conhecer a importância e as consequências do tratamento, além de poder contestar a decisão. Mesmo fora do contexto de tratamento de dados pessoais, o Comitê Supervisor do Facebook, por exemplo, recomendou a implementação de uma série de medidas neste sentido, pois compreendeu que para avaliar a proporcionalidade da moderação de conteúdo em relação ao direito à liberdade de expressão é necessário transparência para compreender a sua implementação.

Além dessas medidas, uma defesa prevalente na literatura, é a necessidade de que, quando uma pessoa é impactada pela decisão algorítmica, sejam disponibilizados os dados que foram considerados pelo algoritmo de tomada de decisão. Contudo, não basta que sejam meramente apresentados os dados que foram utilizados, mas como eles foram avaliados, hierarquizados em relevância, sopesados e considerados pelo algoritmo, visto que o mero acesso aos dados utilizados não permite a inteligibilidade e interpretação da pessoa sobre como cada uma das informações utilizadas contribuiu para aquele resultado. Este é o caso do

uso de dados pessoais para a formação de perfis no contexto de pontuação de crédito, em que acessar o histórico de crédito que foi considerado pelo modelo matemático não informa, tampouco permite a compreensão do consumidor, sobre porque ele foi classificado e segmentado como uma pessoa rica ou sofisticada, próspero morador urbano, aspirante social, etc (IDEC, 2017, p. 22). Como apontado pelo IDEC:

> É importante que isso fique claro: o cidadão brasileiro possui o direito de saber muito mais do que a "nota" ou a "pontuação" a qual ele é relacionado (ex: score de valor 640, em uma escala de 0 a 1000). Todos nós possuímos o direito de entender como os computadores chegaram a essa nota, quais foram os fatores que mais tiveram peso e quais foram as bases de dados utilizadas para se computar essa nota (IDEC, 2017, p. 24).

No que se refere ao aspecto procedimental, ou ao devido processo algorítmico, a LGPD traz previsões que também permitem que sejam exercidas estas garantias. Pode-se dizer que os princípios da finalidade e da transparência, além da boa-fé indicada no *caput* do art. 6º, da LGPD, exigem que o controlador notifique e informe ao titular de dados que o mesmo estará sujeito a decisões tomadas unicamente com base em tratamento automatizado, e o contexto em que isto ocorrerá. Ou seja, é possível interpretar tais artigos para defender que seria um direito do titular de dados de ser informado e notificado sobre decisões tomadas unicamente com base em tratamento automatizado.

O direito de acesso a dados (art. 18, II)[381] encontra fundamento no princípio de livre acesso (art. 6º, IV)[382], e garante ao titular de dados que ele possa consultar, de maneira facilitada e gratuita, a forma do tratamento de dados, e ter acesso à integralidade dos dados pessoais mantidos pelo controlador. O art. 19[383] detalha a

[381] Art. 18. O titular dos dados pessoais tem direito a obter do controlador, em relação aos dados do titular por ele tratados, a qualquer momento e mediante requisição: (…) II - acesso aos dados;
[382] Art. 6º, VI: IV - livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;
[383] Art. 19, da LGPD: "A confirmação de existência ou o acesso a dados pessoais serão providenciados, mediante requisição do titular: I - em formato simplificado, imediatamente; ou II - por meio de declaração clara e completa, que indique a origem dos dados, a inexistência de registro, os critérios utilizados e a finalidade do tratamento, observados os segredos comercial e industrial, fornecida no prazo de até 15 (quinze) dias, contado da data do requerimento do titular. § 1º Os dados pessoais serão armazenados em formato que favoreça o exercício do direito de acesso. § 2º As informações e os dados poderão ser fornecidos, a critério do titular: I - por meio eletrônico, seguro e idôneo para esse fim; ou II - sob forma impressa. § 3º Quando o tratamento tiver origem no consentimento do titular ou em contrato, o titular poderá solicitar cópia eletrônica integral de seus dados pessoais, observados os segredos comercial e industrial, nos termos de regulamentação da autoridade nacional, em formato que permita a sua utilização subsequente, inclusive em outras operações de tratamento. § 4º A autoridade nacional poderá dispor de forma diferenciada acerca dos prazos previstos nos incisos I e II do caput deste artigo para os setores específicos".

maneira como a confirmação ou o acesso a dados pessoais deverá ser providenciada quando houver solicitação pelo titular de dados. Este deverá receber os dados do controlador imediatamente em formato simplificado ou por meio de declaração clara e completa, indicando a origem dos dados, a inexistência de registro, os critérios utilizados e a finalidade do tratamento, observados os segredos comercial e industrial, em até 15 dias. Os dados devem ser armazenados em um formato que favoreça o exercício do direito de acesso, e podem ser recebidos em formato eletrônico ou impresso. Ou seja, este conjunto de artigos cria um "robusto regime de acesso à informação" (Souza et al., 2021, p. 479).

O acesso aos dados e a explicação em relação à decisão automatizada são elementos essenciais para que o titular de dados possa exercer a revisão da referida decisão, e tentar modificar o resultado apresentado. Repita-se, mais uma vez, que o veto à presença de um humano para receber e avaliar o pedido de revisão é mais um empecilho e óbice para o pleno exercício do direito à explicação. A sua ausência, contudo, apesar de não tornar mandatória a intervenção humana na hipótese de um pedido de revisão, não impede que seja uma prática adotada pelos agentes de tratamento. Como já manifestado, a supervisão humana é um princípio ético que vem sendo reiteradamente previsto nos diversos documentos sobre o tema, sendo uma fonte de *soft law* para motivar entes públicos e privados a implementarem esta prática.

Neste sentido, em uma tentativa de delimitar as situações em que a intervenção humana deve se fazer necessária, Miriam Wimmer e Danilo Doneda (2021) propõem dois critérios para avaliar a necessidade, a forma e o momento em que aquela deverá ocorrer no processo de tomada de decisão automatizada. O primeiro deles se refere à uma análise dos riscos e consequências, presentes e futuros, a indivíduos ou grupos impactados pela decisão, a direitos fundamentais, à discriminação e à possibilidade de reversão da decisão [384]. Tal análise fundamentaria uma escolha para garantir uma intervenção humana antes ou depois de uma decisão automatizada, ou se ao longo de todo o ciclo do sistema. O segundo se refere a situações em que a decisão automatizada deve realizar uma análise

[384] Quanto a este ponto, Wimmer e Doneda apontam que a "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" (Wimmer; Doneda, 2021, p. 396).

subjetiva do que seja considerado “certo” ou “errado”, ou em situações que dependem de uma análise contextual e de valores/percepções que as máquinas não sejam tecnicamente aptas a realizar (Wimmer; Doneda, 2021, p. 395).

Os autores apontam que estes critérios não seriam exaustivos, mas visam justamente oferecer uma direção mais concreta e específica das situações que devem ensejar a intervenção humana. O objetivo é tornar mais apurada a avaliação da legitimidade de decisões automatizadas, por meio de controles a parâmetros que permitem que uma pessoa possa compreendê-la e contestá-la (Wimmer; Doneda, 2021, p. 400). Assim, a proposta de prestação de contas apresentada neste trabalho pode auxiliar nesta avaliação apresentada pelos autores, que auxilia na melhor especificação de critérios sobre quando uma intervenção humana se faça necessária.

Este último ponto deixa evidente os problemas envolvendo o recebimento e a avaliação este pedido de revisão por meios automatizados. A impossibilidade de compreensão de contexto e da sintaxe da linguagem das máquinas põe em xeque a própria natureza dialética da explicação, que envolve naturalmente uma interação e comunicação entre os agentes envolvidos. Poderia até ser argumentada a possibilidade de que uma árvore de decisão pudesse ser elaborada para permitir uma explicação sobre como um algoritmo alcançou um resultado. Esta explicação ajudaria a compreender os caminhos, mas não a justificativa ou explicação dos resultados em si. Ainda, pode-se pensar que sistemas conversacionais (*chatbots*) possam ser uma primeira triagem para o recebimento de pedidos de revisão, e para casos mais complexos, tais solicitações devem ser direcionadas para pessoas humanas. É preciso considerar as dificuldades em oferecer uma explicação “universal”, ante as diferentes finalidades que o direito à explicação visa atender, os distintos públicos que a solicitam, e a natural diversidade cultural entre pessoas.

Apesar da diferença entre o GDPR e a LGPD na regulamentação de um direito à explicação, em que esta última não prevê todas as salvaguardas que devem ser implementadas pelo controlador quando uma decisão é tomada exclusivamente por meios automatizados, tal como, o direito de o titular de dados manifestar o seu ponto de vista e de contestar a decisão, ainda assim é possível extrair estas salvaguardas e garantias. Em semelhança, pode-se dizer que, em ambas, o pedido de revisão nada mais é do que uma maneira de o sujeito contestar a decisão do algoritmo, a fim de indicar porque a decisão deveria ser diferente daquela apresentada.

Neste sentido, Maranhão et al., (2021, p. 150) defendem que o objetivo da explicação é permitir a contestabilidade de uma decisão algorítmica, que busca: "fornecer o substrato para que o indivíduo – ou grupo – possa avaliar as consequências de uma decisão e, se for o caso, tenha as informações necessárias para contestar a decisão por vias judiciais ou extrajudiciais". Os autores continuam indicando que, para que este objetivo seja alcançado, a explicação deve envolver "a lógica que levou ao desfecho sob análise, de forma a permitir a identificação dos atores humanos que contribuíram para a produção do resultado computacional", não pressupõe uma explicação referente ao funcionamento do programa (tais como modelos utilizados, decomposicionais) mas "uma explicação capaz de legitimar o resultado" (Maranhão et al., 2021, p. 150).

Para os autores, a contestabilidade seria diferente das explicações contrastivas, ou que buscam contrastividade, em que a primeira estaria mais alinhada com as demandas do Direito[385]. Sob a perspectiva de Kaminski, contudo, esta diferenciação não seria relevante, pois um direito à explicação não visa atender apenas a um único objetivo focado no sujeito impactado como um todo, mas ele se insere dentro de uma discussão mais ampla de *accountability* de algoritmos.

Por fim, o direito à explicação pode ser exercido por um titular de dados, ou seu representante legal (art. 18, § 3º, da LGPD) frente a um controlador, sem a necessidade do ajuizamento de uma ação judicial. Tendo em vista que ora se compreende o direito à explicação sob uma perspectiva de procedimento e de mérito, é importante que o controlador tenha se estruturado internamente para receber pedidos por explicação, além de ser necessário notificar, previamente, o titular de dados quando seus dados serão coletados e utilizados para a tomada de decisão automatizada. A mera ausência destes deveres pode acarretar na compreensão de que o direito à explicação não está sendo adequada e eficazmente aplicado, contrariando a previsão legal. Em complemento ao direito à explicação, o titular de dados poderá exercer seu direito de acesso aos dados, e de obter a

[385] Isto porque, para os autores, a contrastividade "busca explicar como a diferença nos fatos leva a resultados distintos, enquanto a contestabilidade enfatiza as consequências jurídicas da operação do sistema inteligente". Ainda, ambas enfatizam os atores envolvidos de maneira diferente. A contrastividade estaria preocupada com a alteração no comportamento do sujeito, enquanto a contestabilidade deseja identificar os responsáveis sobre o sistema para responsabilizá-los por efeitos antijurídicos produzidos pelo sistema. Por sim, a contestabilidade seria considerada como uma prática instrumental, em que se busca conhecer o sistema para garantir a tutela de interesses juridicamente protegidos (Maranhão et al., 2021, p. 151). Apesar da diferenciação, os autores reconhecem a importância de outros tipos de explicação.

explicação de mérito sobre a decisão automatizada que o impactou, entre elas, baseada em casos, contextuais, contrastantes, contrafactual, e de dia a dia.

Caso o controlador se negue a fornecer estas informações, é possível que o titular de dados ajuíze uma ação judicial em face deste, ou dos agentes de tratamento. Neste contexto, será possível que, caso o magistrado avalie necessário, sejam compartilhados documentos que permitam que, em juízo, e sob sigilo judicial, caso pertinente, as razões de decidir do algoritmo possam ser compreendidas, bem como o sistema como um todo. Poderá ocorrer, caso o titular de dados pleiteie, e o magistrado conceda, a inversão do ônus da prova, desde que atenda aos requisitos estabelecidos pelo art. 42, § 2º, da LGPD[386]. Como já mencionado, também é possível que seja proposta uma ação coletiva para defender os interesses e direitos dos titulares de dados (art. 22, da LGPD). Em ambos os casos, o não atendimento ao direito à explicação pode gerar o dever de indenizar (art. 42, da LGPD). No âmbito da ação coletiva, é possível que um dos pedidos envolva a determinação de que sejam implementadas medidas organizacionais e as devidas salvaguardas que assegurem a transparência e prestação de contas das atividades envolvendo o algoritmo. Ademais, caso o controlador se negue a fornecer as explicações apontadas no art. 20, § 1º, da LGPD, pode ser demandada a realização de auditoria perante a ANPD, nos termos do 20, § 2º, da LGPD.

Ainda, ao invés da ação judicial, é possível que o titular de dados apresente uma petição para a ANPD (art. 18, § 1º, da LGPD). Neste caso, é possível que sejam fornecidos uma gama ainda maior de documentos, produzidos tanto *ex ante* (relatório de impacto e compartilhamento do código de boas práticas e de conduta), bem como *ex post* (documentação). O não atendimento ao direito à explicação, conforme ora compreendido, pode acarretar na aplicação de sanções administrativas, que irá variar de acordo com a gravidade da infração cometida (art. 52, da LGPD).

[386] Art. 42, da LGPD: "O controlador ou o operador que, em razão do exercício de atividade de tratamento de dados pessoais, causar a outrem dano patrimonial, moral, individual ou coletivo, em violação à legislação de proteção de dados pessoais, é obrigado a repará-lo (…):§ 2º O juiz, no processo civil, poderá inverter o ônus da prova a favor do titular dos dados quando, a seu juízo, for verossímil a alegação, houver hipossuficiência para fins de produção de prova ou quando a produção de prova pelo titular resultar-lhe excessivamente onerosa".

4.
Considerações Finais

A presença cada vez maior de sistemas de IA e seus respectivos algoritmos trazem transformações profundas sobre o tipo de conhecimento que pode ser gerado com os seus usos. Quando estas inovações são aplicadas a aspectos cotidianos, ou até mesmo para influenciar no exercício democrático de cidadãos, percebe-se a importância de se garantir direitos que protejam as pessoas da "tirania das decisões automatizadas" (Frazão, 2018a). Afirmar a necessidade de que deva ser garantido direitos às pessoas quando há a aplicação de algoritmos de tomada de decisão não pode significar a defesa de que isto desestimularia a inovação. Assim como as regulações de proteção de dados pessoais não visam obstar o tratamento e os usos de dados pessoais, mas justamente trazer um ambiente com regras claras para os atores envolvidos, diminuir a assimetria informacional entre o titular de dados e os agentes de tratamento, definindo os direitos e deveres desta atividade, garantir um direito à explicação, quando há o uso de IA, busca o mesmo objetivo.

Compreender como a IA e os algoritmos de tomada de decisão com aplicação de ML funcionam é fundamental para identificar os desafios e as maneiras de assegurar direitos, e as salvaguardas que devem ser implementadas para resguardá-los. Como visto, esta tese tem como foco a preocupação com o impacto desta tecnologia a direitos fundamentais, em especial, o direito à privacidade, proteção de dados, liberdade (autonomia) e igualdade, diante das decisões tomadas por algoritmos de ML. As características da IA mapeadas por Ryan Calo (2017) e Jack Balkin (2017) – emergência, valor social e materialidade – colocam novos desafios, especialmente a característica da emergência (foco deste trabalho), diante da imprevisibilidade e inescrutabilidade dos resultados e funcionamento destes sistemas. Portanto, indaga-se, como será possível compreender e avaliar os resultados de alguns tipos de algoritmos de tomada de decisão que se valem de ML, quando tampouco humanos são capazes de assim fazê-lo. Esta tese buscou oferecer meios para operacionalizar e viabilizar maneiras de explicar e compreender estes resultados, considerando os direitos que garantem uma proteção jurídica aos algoritmos, e seus desafios técnicos, com o objetivo de resguardar os mencionados direitos fundamentais.

Os principais desafios para os direitos fundamentais que são colocados por tais algoritmos de tomada de decisão podem ser resumidos sob os princípios de transparência, prestação de contas e responsabilização e justiça/igualdade (representados pelo acrônimo FAT). Isto é, os princípios que devem ser observados ao longo do desenvolvimento de sistemas e algoritmos de IA revelam os problemas que estes artefatos geram. E maneiras de assegurar a observância dos mesmos dependem de uma postura multifacetada e multidisciplinar. O fato de que alguns algoritmos de ML são considerados "caixas-pretas" torna necessário assegurar alguma maneira de prestar contas, que deve ocorrer antes do seu desenvolvimento, e depois, quando os mesmos são implementados. Esta prestação de contas deve ocorrer por meio da produção de documentos *ex ante* (relatório de impacto, boas práticas e códigos de conduta), e documentos e implementação de práticas *ex post* (documentação, implementação de técnica de interpretação e explicação de modelos de ML, e de investigação e avaliação da auditoria).

A prestação de contas e responsabilização *ex ante* e *ex post* têm o intuito de viabilizar o direito à explicação, isto é, permitir que haja uma tradução para a linguagem jurídica que permita uma compreensão de sistemas de IA e, mais especificamente, de algoritmos de ML, especialmente aqueles considerados "caixas-pretas". É necessário tornar possível avaliar os riscos deste artefato para as liberdades e direitos fundamentais das pessoas sujeitas ao mesmo, e adotar as medidas de segurança que visem mitigar ou eliminar tais riscos. Ainda, as boas práticas, assim como códigos de conduta, visam guiar de maneira padronizada, e com aderência ao arcabouço jurídico competente, o ambiente de desenvolvimento, de forma que atendam às balizas mínimas para eficientemente proteger direitos fundamentais e as liberdades das pessoas. A documentação do sistema e do algoritmo permitem compreender as funcionalidades, requisitos e os objetivos que buscam ser alcançados, justificando as decisões tomadas ao longo do desenvolvimento. A proposta do *model card* visa padronizar uma das maneiras de documentar e permitir dar publicidade sobre algumas informações do modelo de ML. O interessante é que o mesmo também traz preocupações éticas, atendendo a demandas relacionadas ao princípio de justiça e igualdade. Já a aplicação de técnicas que permitam a explicação e interpretação de resultados algorítmicos, quando possível, serão fundamentais para permitir o acesso à compreensão e explicação dos resultados algorítmicos. Por fim, a auditoria, sob a ideia de

transparência qualificada de Pasquale (2015) é fundamental para permitir algum escrutínio sobre tais sistemas.

Como visto, o direito à explicação também possui um aspecto importante que é definir exatamente o conteúdo que será fornecido em termos de explicação. Assim, dependendo de quem solicita uma explicação, o tipo de explicação exigida deverá ser diferente, de acordo com o objetivo que se visa alcançar, e a pergunta que se visa responder. A taxonomia proposta por Chari et al. (2020)[387], bem como as explicações apresentadas pelo relatório do ICO[388], visam organizar os diferentes tipos de conteúdo que podem ser fornecidos de acordo com as perguntas que se visa responder. Neste sentido, a posição de Zarsky (2013) se apresenta como um caminho a ser trilhado, quando defende que diferentes públicos vão exigir diferentes tipos de informações sobre algoritmos de tomada de decisão.

Isto também implica em reconhecer que existem níveis de transparência e prestação de contas, no sentido de que o mero exercício do direito à explicação não necessariamente irá ensejar abertura de toda e qualquer informação sobre o sistema de IA e algoritmos de tomada de decisão. Contudo, será considerado como uma boa prática e uma maneira de demonstrar uma adequação a princípios éticos, e previsões jurídicas, o compartilhamento de algum nível de informação como uma maneira de prestar contas, em atenção ao interesse público.

Percebe-se, portanto, que o direito à explicação possui uma dimensão coletiva, no sentido subjetivo de quem pode exercer este direito (i.e. titular de dados ou representantes de interesses coletivos e difuso), e objetivo, no sentido dos tipos de garantias que devem ser asseguradas. Quanto a este último ponto, ele se coaduna com a visão sistêmica da governança colaborativa de Kaminski (2019b), no sentido de que devem ser elaborados e fiscalizados pelas entidades competentes documentos que demonstrem a aderência dos responsáveis por sistemas de IA, e seus respectivos algoritmos de tomada de decisão, com a legislação infraconstitucional (como a LGPD, o CDC, ou o CPC). Isto é, existe um interesse de que o sistema, como um todo, esteja aderente à legislação, bem como a preceitos constitucionais, diante do seu potencial de afetar direitos fundamentais de uma

[387] Explicações (contextuais, baseadas em casos, contrastantes, contrafactual, de dia a dia, científica, baseada em simulações, estatísticas, e em rastreio de evidências.
[388] Explicações: das razões, da responsabilidade, dos dados, de justiça/igualdade, de segurança e desempenho e do impacto.

coletividade de pessoas, conforme visto no capítulo 1, sobre a perspectiva coletiva da proteção de dados pessoais e da privacidade.

Outro aspecto do direito à explicação se refere às salvaguardas e aos direitos que devem ser observados para assegurar a sua plena eficácia. Isto vai exigir que as pessoas sujeitas a decisões automatizadas tenham conhecimento e sejam notificadas quando isto estiver ocorrendo; possam solicitar o acesso aos dados utilizados para a tomada de decisão; possa pedir uma revisão da referida decisão, a fim de contestá-la; e tenha acesso, principalmente, às motivações para aquele resultado (por exemplo: os dados do sujeito que foram relevantes, as validações realizadas pelo sistema, as regras de negócio implementadas).

Quanto a este último ponto, a explicação é essencial como uma maneira de avaliar a legitimidade e legalidade de decisões algorítmicas, e assegurar a proteção e promoção dos direitos fundamentais envolvidos. Acessar estas informações é uma maneira de tentar garantir a autodeterminação informativa, com o intuito de atribuir ao titular de dados um maior controle sobre as suas informações pessoais e, por que não, dos usos que são feitos com os mesmos. Ademais, o valor instrumental/funcional de tal direito também é uma forma de verificar se outros direitos estão sendo violados ou ameaçados, como demonstrado no caso da moderação de conteúdo do Facebook.

Embora esta tese tenha como foco algoritmos de tomada de decisão e sistemas de IA que se valem de dados pessoais, percebe-se que o tema do direito à explicação não se limita a este contexto, conforme demonstram as leis setoriais que já regulam aspectos envolvidos em decisões automatizadas, bem como os projetos de lei apresentados no capítulo quatro. Em especial, o PL 21/2020, que propõe ser o marco regulatório de IA no Brasil, no que diz respeito ao direito à explicação, perdeu a oportunidade de especificar, detalhar e elaborar de maneira mais específica dito direito, e sequer faz menção ao art. 20, da LGPD. Infelizmente, tal PL é silente quanto um dos aspectos mais importantes de decisões automatizadas e sistemas de IA: a supervisão humana.

Como se viu, o direito à explicação e à revisão de decisões totalmente automatizadas prevista na LGPD também possui esta limitação, tendo em vista que a previsão de que a revisão fosse realizada por pessoa uma pessoa natural foi vetada, de forma que isto tem sido considerado como um verdadeiro impasse para o pleno exercício deste direito. Por isso, recomenda-se que nestes processos sejam

asseguradas, como regras de boas práticas e em atenção a princípios éticos voltados ao desenvolvimento de sistemas e aplicações de IA, a presença de pessoas naturais que tenham conhecimento sobre o domínio para avaliarem pedidos de revisão.

Por fim, a justificativa ou o mérito do direito à explicação exige transparência. Isto porque apenas com a abertura dessas informações será possível avaliar criticamente os argumentos e justificativas apresentadas, além da sua legitimidade e legalidade. Quando um titular de dados estiver solicitando uma explicação, estas informações poderão o ajudar a alterar a decisão (caso esta tenha lhe prejudicado) e mudar a situação que o mesmo se encontra. Por exemplo, ele poderá buscar mudar sua situação para obter um empréstimo perante uma instituição financeira, ou ser melhor alocado dentro de um perfil para uma melhor oferta de emprego. Sob a dimensão coletiva, quando solicitada, também será importante assegurar o escrutínio a agentes que assim o solicitem, em um foro específico e sigiloso, se necessário, com a elaboração de algum sumário que permita dar certa publicidade aos resultados da análise da explicação. Isto é importante para que pesquisadores, sociedade civil, associações, entes públicos, e outras entidades privadas ou públicas, conheçam, avaliem e questionem os documentos e análises realizadas.

5.
Referências Bibliográficas

ARGENTINA; URUGUAI. AGENCIA DE ACCESO A LA INFORMACIÓN PÚBLICA DE ARGENTINA (AAIP); UNIDAD REGULADORA Y DE CONTROL DE DATOS PERSONALES DE URUGUAY (URCDP). *Guía de Evaluación de Impacto en la Protección de Datos,* janeiro de 2020. Disponível em: https://www.argentina.gob.ar/sites/default/files/guia_final.pdf. Acessado em 01.05.2021.

ANANNY, Mike; CRAWFORD, Kate. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*. 2018, v. 20(3), pp. 973-989. Disponível em: https://journals.sagepub.com/doi/10.1177/1461444816676645. Acessado em 01.05.2021.

ANJOS, Lucas. *Decisões automatizadas e transparência algorítmica.* Belo Horizonte: Instituto de Referência em Internet e Sociedade (IRIS), em 6 de nov. 2019. Disponível em: https://irisbh.com.br/decisoes-automatizadas-e-transparencia-algoritmica/. Acessado em 03.02.2021.

ARAÚJO, Camila Souza; MEIRA, Wagner; ALMEIDA, Virgilio; SPIRO, Emma; AHN, Yong-Yeol. Identifying Stereotypes in the Online Perception of Physical Attractiveness. *Social Informatics*, v. 10046, Springer International Publishing, pp, 419-437. Disponível em: < https://arxiv.org/abs/1608.02499#:~:text=Our%20findings%20demonstrate%20the%20existence,in%20terms%20of%20physical%20attractiveness.> Acessado em 04.09.2019.

ARTICLE 19. *Governance with teeth:* How human rights can strengthen FAT and ethics initiatives on artificial intelligence. Londres, Reino Unido, abril de 2019. Disponível em: <https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf>. Acessado em 20.01.2020.

AGRAWAL, Ajay; GANS, Joshua; GOLDFARB, Avi. *Prediction Machines*: The Simple Economics of Artificial Intelligence. Boston: Harvard Business Review Press, 2017. Em formato digital Audible.

AUGUSTO, Luís Gustavo Henrique; RIZZARDI, Maira Martinelli. *Accountability segundo os Ministros dos Tribunais Superiores do Judiciário Brasileiro.* Apresentação de trabalho no XXIII Encontro Nacional do Conpedi - Direito e administração Pública. Florianópolis, 2014. Disponível em: https://www.researchgate.net/publication/280626416_Accountability_segundo_os_Ministros_dos_Tribunais_Superiores_do_Judiciario_Brasileiro_Accountability_according_to_the_Ministers_of_the_Superior_Courts_of_the_Brazilian_Judiciary> Acessado em 19.06.2021.

BALKIN, Jack M.. The Path of Robotics Law. *California Law Review,* vol. 6, Circuit 47, jun. de 2015. Disponível em: https://ssrn.com/abstract=2586570. Acessado em 21.04.2017.

BALKIN, Jack M., The Three Laws of Robotics in the Age of Big Data. *Ohio State Law Journal,* vol. 78, ago. de 2017, pp. 1217-1241. Disponivel em: <https://digitalcommons.law.yale.edu/fss_papers/5159> Acesso em 03.01.2020.

BOYD, Danah; CRAWFORD, Kate. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, & Society*, vol. 15, issue, 5, 2012, p. 662-679.

BRENNAN-MARQUEZ, Kiel. Plausible Cause: Explanatory Standards in the Age of Powerful Machines. *Vanderbilt Law Review,* vol. 70, 1249, 2017. Disponível em: <https://scholarship.law.vanderbilt.edu/vlr/vol70/iss4/2>. Acesso em 01.12.2021.

BURRELL, Jenna. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society,* 3 (1), jan. de 2016. Disponível em: <https://journals.sagepub.com/doi/pdf/10.1177/2053951715622512>. Acesso em 01.02.2021.

BURRI, Thomas. The Politics of Robot Autonomy. *European Journal of Risk Regulation*, vol. 7, 2, pp. 341-360, 2016. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2815082>. Acesso em 02.02.2020.

BLACK, Julia; MURRAY, Andrew D. Regulating AI and machine learning: setting the regulatory agenda. *European Journal of Law and Technology*, vol. 10, 3, 2019. Disponível em: http://eprints.lse.ac.uk/102953/. Acessado em 20.01.2020.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*. Proceeding of Machine Learning Research 81, pp, 1-15, 2018.

CALO, Ryan. Robotics and the Lessons of Cyberlaw. *California Law Review*, Vol. 103, n. 3, 2015, pp. 513-564. Disponível em: <https://ssrn.com/abstract=2402972>. Acessado em 21.04.2017.

CALO, Ryan. Artificial Intelligence policy: a primer and roadmap. 51 *U.C. Davis Law Review* 300, 2017, pp. 399-435. Disponível em: <https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Calo.pdf>. Acessado em 03.01.2020.

CARPENA, Heloísa. Os vazamentos de dados e a reparação dos danos à luz do Código de Defesa do Consumidor, no prelo.

CASEY, Bryan; FARHANGI, Ashkon; VOGL, Roland. Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal,* vol. 34, 2019. Disponível em <https://ssrn.com/abstract=3143325>. Acessado em 21.04.2021.

CASTILLO, Aída Ponde del. *A law on robotics and artificial intelligence in the EU*? Bruxelas: Foresight Brief (European Trade Union Institute -- ETUI), 2017. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3180004> Acessado em 21.04.2021.

CHARI, Shruthi; GRUEN, Daniel M.; SENEVIRATNE, Oshani; MCGUINNESS Deborah L. Directions for Explainable Knowledge-Enabled Systems. In: TIDDI, Ilaria; LECUE, Freddy; HITZLER, Pascal (Org.). *Knowledge Graphs for eXplainable AI -- Foundations, Applications and Challenges*. Studies on the Semantic Web, IOS Press, Amsterdam, 2020. No prelo. Disponível em: <https://arxiv.org/abs/2003.07523> Acessado em 20.10.2021.

CITRON, Danielle Keats; PASQUALE, Frank. The Scored Society: Due Process for Automated Predictions. *Washington Law Review,* Vol. 89, 2014, p. 1-, University of Maryland Legal Studies Research Paper No. 2014-8. Disponível em: <http://ssrn.com/abstract=2376209>. Acessado em 10.07.2020.

COMITÊ EUROPEU PARA A PROTEÇÃO DE DADOS (CEPD). *Guidelines 1/2018 on certification and identifying certification criteria in accordance with Articles 42 and 43 of the Regulation*. Versão 3.0, adotada em 04 de junho de 2019. Disponível em: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201801_v3.0_certificationcriteria_annex2_en.pdf. Acessado em 20.03.2021.

COPPIN, Ben. *Artificial Intelligence Illuminated.* 1ª Edição. Massachusetts: Jones & Bartlett Learning, 2004.

COMISSÃO EUROPEIA. *Ethics Guidelines for Trustworthy AI*. 2019a. Disponível em: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> Acessado em 14.04.2020.

_____. *A definition of AI: Main capabilities and scientific disciplines*. 2019b. Disponível em: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> Acessado em 14.04.2020.

CORRÊA, Bianca Kremer Nogueira; *Direito e Tecnologia em perspectiva amefricana*: autonomia, algoritmos e vieses raciais. Rio de Janeiro, 2021, 298 p. Tese de doutorado. Departamento de Direito, Pontifícia Universidade Católica do Rio de Janeiro.

CRAWFORD, Kate; SCHULTZ, Jason. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *55 B.C.L. Rev. 93*, 2014. Disponível em: <https://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4>. Acessado em 01.12.2021.

DESAI, Deven R.; KROLL, Joshua A.. Trust But Verify: A Guide to Algorithms and the Law. *Harvard Journal of Law & Technology*, vol. 31(2). Disponível em: https://ssrn.com/abstract=2959472. Acessado em 04.03.2021.

DIAKOPOULOS, Nicholas. Accountability in algorithmic decision making. *Communications of the ACM, vol.* 59(2), 2016, pp. 56–62. Disponível em: <https://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/fulltext>. Acessado em 03.02.2021.

DONEDA, Danilo. *Da privacidade à proteção de dados pessoais:* elementos da formação da lei geral de proteção de dados. 2ª Edição. São Paulo: Thomson Reuters Brasil, 2019.

_____, Danilo; MENDES, Laura Schertel; SOUZA, Carlos Affonso Pereira de; ANDRADE, Norberto Nuno Gomes de. Considerações sobre Inteligência Artificial, ética e autonomia pessoal. *Pensar – Revista de Ciências Jurídicas*, v. 23, p. 1-17, 2018.

_____, Danilo. A proteção dos dados pessoais como um direito fundamental. Espaço Jurídico, Joaçaba, v. 12, n. 2, p. 91-108, jul./dez. 2011, 2011.

_____, Danilo; ALMEIDA, Virgílio. O que é governança de algoritmos? In: BRUNO, Fernanda; CARDOSO, Bruno; KANASHIRO, Marta; GUILHON, Luciana; MELGAÇO, Lucas (Org). *Tecnolopolíticas da vigilância*: perspectivas da margem. São Paulo: Boitempo, 2018, pp. 141-148.

DOSHI-VELEZ, Finale; KORTZ, Mason. *Accountability of AI Under the Law*: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, 2017. Disponível em: <https://dash.harvard.edu/handle/1/34372584>. Acessado em 10.03.2021.

EDWARDS, Lilian; VEALE, Michael. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. 16 *Duke Law & Technology Review,* 18-84, 2017. Disponível em: <https://ssrn.com/abstract=2972855>. Acessado em 01.12.2021.

FERNANDES, Micaela B. B.; OLIVEIRA, Camila Helena M. B. de. O artigo 20 da LGPD e os desafios interpretativos ao direito à revisão das decisões dos agentes de tratamento pelos titulares de dados. *RDTec - Revista de Direito e as Novas Tecnologias*, v. 08, pp. 1-19, 2020.

FLORIDI, Luciano. Open data, data protection, and group privacy. *Philosophy & Technology*, v. 27, n. 1, p. 1-3, 2014.

FRAZÃO, Ana. O direito à explicação e à oposição diante de decisões totalmente automatizadas. *Revista JOTA*, 05 de dezembro de 2018a. Disponível em <shorturl.at/epAH7> Acessado em 02.03.2019.

_____, Ana. Controvérsias em torno do direito à explicação e à oposição diante de decisões totalmente automatizadas. *Revista JOTA*, 12 de dezembro de 2018b. Disponível em <shorturl.at/nCKU3> Acessado em 02.03.2019.

_____, Ana. Nova LGPD: ainda sobre a eficácia do direito à explicação e à oposição. *Revista JOTA*, 26 de dezembro de 2018c. Disponível em <shorturl.at/fquA6> Acessado em 02.03.2019.

_____, Ana; GOETTENAUER, Carlos. O jogo da imitação jurídica: o direito à revisão de decisões algorítimicas como um mecanismo para a necessária conciliação entre linguagem natural e infraestrutura matemática. In: TEPEDINO, Gustavo; SILVA, Rodrigo da Guia (Orgs.). *O Direito Civil na era da Inteligência Artificial*. São Paulo: Thomson Reuters Brasil, 2020, pp. 45-64.

_____, Ana; OLIVA, Milena Donato; ABILIO, Vivianne da Silveira. *Compliance* de dados pessoais. In: TEPEDINO, Gustavo; FRAZÃO, Ana; OLIVA, Milena Donato (Org.). *Lei Geral de Proteção de Dados Pessoais e as suas repercussões no direito brasileiro*. 2ª Edição. São Paulo: Thomson Reuters Brasil, 2020, pp. 669-704.

_____, Ana. Discriminação Algorítmica: compreendendo a "datificação" e a estruturação da sociedade da classificação. *Revista JOTA*, 23 de junho de 2021a. Disponível em <https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/discriminacao-algoritmica-2-23062021> Acessado em 24.06.2021.

_____, Ana. Transparência de algoritmos x segreso de empresa: as controvérsias a respeito das decisões judiciais trabalhistas que determinam a realização de perícia no algoritmo do Uber. *Revista JOTA*, 09 de junho de 2021b. Disponível em <https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/transparencia-de-algoritmos-x-segredo-de-empresa-09062021> Acessado em 24.06.2021.

FRAJHOF, Isabella Z. O papel dos mecanismos de compliance para a operacionalização do direito à explicação de decisões totalmente automatizadas. In: FRAZÃO, Frazão; CUEVA, Ricardo Villas Bôas. (Orgs.). *Compliance e Políticas de Proteção de Dados*. 1a edição. São Paulo: Thomson Reuters Brasil, 2021, pp. 467-494.

GASSER, Urs; SCHMITT, Carolyn. *The Role of Professional Norms in the Governance of Artificial Intelligence*. No prelo: DUBBER, Markus D.; PASQUALE, Frank; DAS, Sunit (Org.). *The Oxford Handbook of Ethics of AI*. Reino Unido: Oxford University Press, 2019. Disponível em: https://ssrn.com/abstract=3378267. Acessado em 26/04/2021.

GOODMAN, Bryce; FLAXMAN, Seth. *EU Regulations on Algorithmic Decision Making and "a Right to an Explanation*,". Apresentação de trabalho em Nova Iorque, no ICML Workshop on Human Interpretability in ML (WHI), 2016. Disponível em: <https://arxiv.org/abs/1606.08813>. Acessado em 21.04.2018.

GOMES, Maria Cecília Oliveira Gomes. Relatório de impacto à proteção de dados pessoais: uma breve análise da sua deinição e papel na LGPD. *Revista do Advogado*, n. 144, 2019, pp. 174-183.

GOMES, Rodrigo Dias de Pinho. *Big Data*: Desafios à tutela da pessoa humana na sociedade da informação. 2ª Edição. Rio de Janeiro: Lumen Juris, 2019.

GRUPO DE TRABALHO DO ARTIGO 29 (GTA29). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.* Disponível em <http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053> Acessado em 11.04.2020.

_____. *Guidelines for identifying a controller or processor's lead supervisory authority*. Disponível em <http://ec.europa.eu/newsroom/document.cfm?doc_id=44102>. Acessado em 11.04.2020.

_____. *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.* Adotado em 4 de abril de 2017. Disponível em <http://ec.europa.eu/newsroom/document.cfm?doc_id=44137> Acessado em 11.04.2020.

_____. *Guidelines on the right to data portability.* Adotado em 13 de dezembro de 2016b. Disponível em <https://ec.europa.eu/newsroom/article29/items/611233> Acessado em 11.04.2020.

HOFFMAN, Robert R.; KLEIN, Gary. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68-73, 2017a.

_____, Robert R.; MUELLER, Shane T.; KLEIN, Gary. Explaining Explanation Part 2: Empirical Foundations. *Intelligent Systems IEEE*, vol. 32, no. 4, pp. 78-86, 2017b.

_____, Robert R.; MILLER, Tim; MUELLER, Shane T.; KLEIN, Gary; CLANCEY William J.. Explaining Explanation Part 4: A Deep Dive on Deep Nets. *Intelligent Systems IEEE*, vol. 33, no. 3, pp. 87-95, 2018.

HILDEBRANDT, Mireille. *Law for Computer Scientists and Other Folk*. Oxford: Oxford University Press, 2020.

_____, Mireille. Defining Profiling: A New Type of Knowledge?. In: HILDEBRANDT, Mireille; GUTWIRTH, Serge. *Profiling the European Citizen:* Cross-Disciplinary Perspectives. Dordrecht: Springer Netherlands, 2008, pp. 17-45.

HOLANDA. Corte Distrital de Haia (Rb. Den Haag). Processo n. C/09/550982/HA ZA 18/388. Julgado e publicado em 02.05.2020. Disponível em: <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:8655&showbutton=true&keyword=AVG>. Acessado em 12.02.2022.

INFORMATION COMMISSIONER OFFICE (ICO); THE ALAN TURING INSTITUTE. *Explaining decisions made with AI*. Draft Guidance for Consultation. Part 1: The basics of explaining AI. Disponível em: <https://ico.org.uk/for-

organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-3-what-explaining-ai-means-for-your-organisation/documentation/>. Acessado em 12.01.2022.

_____. *Explaining decisions made with AI*. Draft Guidance for Consultation. Part 2: Explaining AI in practice. Disponível em: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-3-what-explaining-ai-means-for-your-organisation/documentation/>. Acessado em 12.01.2022.

_____. *Explaining decisions made with AI*. Draft Guidance for Consultation. Part 3: What explaining AI means for your organization. Disponível em: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-3-what-explaining-ai-means-for-your-organisation/documentation/>. Acessado em 12.01.2022.

INSTITUTO DE SOCIEDADE & TECNOLOGIA DO RIO DE JANEIRO (ITS-Rio). *Transparência e Governança nos algoritmos*: um estudo de caso sobre o setor de birôs de crédito. Rio de Janeiro: ITS-Rio, 2017. Dsiponível em: <https://itsrio.org/pt/publicacoes/transparencia-e-governanca-nos-algoritmos-um-estudo-de-caso/>. Acessado em 22.02.2021.

INSTITUTO BRASILEIRO DE DEFESA DO CONSUMIDOR (IDEC). ZANATTA, Rafael A. F. (Org.). *Por trás da pontuação de crédito*: conheça seus direitos. São Paulo: Idec, 2017.

KAMINSKI, Margot E.. The Right to Explanation, Explained. University of Colorado Law Legal Studies Research Paper No. 18-24. *Berkeley Technology Law Journal*, Vol. 34, No. 1, 2019a. Disponível em <https://btlj.org/data/articles2019/34_1/05_Kaminski_Web.pdf>. Acessado em 09.07.2021.

_____, Margot E.. Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. *Southern California Law Review*, Vol. 92, No. 6, 2019b 1529 U of Colorado Law Legal Studies Research Paper No. 19-9. Disponível em: <https://southerncalifornialawreview.com/wp-content/uploads/2019/12/92_6_Kaminski.pdf>. Acessado em 06.01.2021.

KISER, Grace; MANTHA, Yoan. *Global AI Talent Report 2019*, 2019. Disponível em: <https://jfgagne.ai/talent-2019/> Acessado em 12.09.2019.

KITCHIN, Rob. Thinking critically about and researching algorithms. *Information, Communication & Society*, v. 20, n. 1, 2017, pp. 14-29. Disponível em <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2515786> Acessado em 30.08.2017

KLEIN, Gary. Explaining Explanation Part 3: The Causal Landscape. *Intelligent Systems IEEE*, vol. 33, no. 2, pp. 83-88, 2018.

KOSINSKI, Michal; STILLWELL, David; GRAEPEL, Thore. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* (PNAS), 2013 110 (15), pp. 5802-5805, 2013.

KROLL, Joshua A.; HUEY, Joanna; BAROCAS; Solon; FELTEN; Edward W.; REIDENBERG, Joel R.; ROBINSON, David G.; YU, Harlan. Accountable Algorithms. 165 *U. Pa. L. Rev.* 633, 2017. Disponível em: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3. Acessado em 12.02.2022.

LEHR, David; OHM, Paul. Playing with the Data: What Legal Scholars Should Learn About Machine Learning. *University of California, Davis*, Vol. 51, 653, 2017. Disponível em: <https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf>. Acessado em 26.01.2020.

LESSIG, Lawrence. *Code (version 2.0)*. Nova York: Basic Books, 2006.

_____, Lawrence. The Law of the Horse: What cyberlaw might teach. *113 Harvard Law Review 501, 1999*. Disponível em <https://cyber.harvard.edu/works/lessig/finalhls.pdf> Acessado em 30.08.2017.

LINARDATOS, Pantelis; PAPASTEFANOPOULOS, Vasilis; KOTSIANTIS, Sotiris. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23, 18, 2021. Disponível em: <https://dx.doi.org/ 10.3390/e23010018>. Acessado em 14.04.2021.

MACHADO, Diego; DONEDA, Danilo. Proteção de dados pessoais e criptografia: tecnologias criptográficas entre anonimização e pseudonimização de dados. *Revista dos Tribunais*. vol. 998. Caderno Especial. São Paulo: Ed. RT, 2018, pp. 99-128.

MAGRANI, Eduardo. *A Internet das Coisas*. Rio de Janeiro: Editora FGV, 2018.

MALGIERI, Gianclaudio; COMANDÉ, Giovanni. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law*, Volume 7, Issue 4, pp. 243–265, 2017. Disponível em: <https://ssrn.com/abstract=3088976>. Acessado em 01.12.2021.

MANTELERO, Alessandro. Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. *Computer Law & Security Review*, v. 32, n. 2, p. 238-255, 2016.

MARANHÃO, Juliano; COZMAN, Fábio Gagliardi; ALMADA, Marco. Concepções de explicação e do direito à explicação de decisões automatizadas. In: VAINZOF, Rony; GUTIERREZ, Andriei Guerrero (Orgs.). *Inteligência Artificial*: Sociedade Economia e Estado. 1ª Edição. São Paulo: Thomson Reuters, 2021, pp. 137-154.

MARQUES, C. L.. *Contratos no Código de Defesa do Consumidor*: o novo regime das relações contratuais. São Paulo: Editora Revista dos Tribunais, 2008.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big Data:* A Revolution That Will Transform How We Live, Work, and Think. Londres: John Murray Publishers, 2013.

MAZUOLLI, Valerio de Oliveira. *Curso de direito internacional publico*. 13ª Edição. Rio de Janeiro: Forense, 2020.

MERCIER, Hugo; SPERBER, Dan. *The Enigma of Reason*. Cambridge: Harvard University. Edição iBook.

MENDES, Laura Schertel. *Privacidade, proteção de dados e defesa do consumidor*: linhas gerais de um novo direito fundamental. São Paulo: Saraiva, 2014.

MENDES, Gilmar Ferreira; COELHO, Inocêncio Mártires; BRANCO, Paulo Gustavo Gonet. *Curso de Direito Constitucional*. 6ª Edição. São Paulo: Saraiva, 2011.

MITTELSTADT, Brent; ALLO, Patrick; MARIAROSARIA, Taddeo; WACHTER, Sandra; FLORIDI, Luciano. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, Vol. 3(2), 2016. Disponível em: <https://journals.sagepub.com/doi/full/10.1177/2053951716679679>. Acessado em 03.01.2020.

_____, Brent; RUSSELL, Chris; WACHTER, Sandra. Explaining Explanations in AI. *Proceedings of FAT* '19: Conference on Fairness, Accountability, and Transparency* (FAT* '19), January 29-–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA. Disponível em: <https://ssrn.com/abstract=3278331>. Acesso em: 20.03.2021.

MITCHELL, Margareth; WU, Simone; ZALDIVAR, Andrew; BARNES, Parker; VASSEMAN, Lucy; HUTCHINSON, Ben; SPITZER, Elena; RAJI, Deborah Iniluwa; GEBRU, Timnit. Model Cards for Model Reporting. *FAT* '19*, Atlanta, GA, USA, 2019, pp. 220–229.

MONTEIRO, Renato Leite. *Existe um direito à explicação na Lei Geral de Proteção de Dados do Brasil*?. Instituto Igarapé. Artigo Estratégico 39, dezembro de 2018, p. 10. Disponível em: <shorturl.at/dtxU8> Acessado em 20.03.2021.

MORAES, Maria Celina Bodin de. Apresentação. In: RODOTÀ, Stefano. *A vida na sociedade de vigilância*, a privacidade hoje. Rio de Janeiro: Renovar, 2008.

_____, Maria Celina Bodin de. O Princípio da Dignidade Humana. In: MORAES, Maria Celina Bodin de (Org). *Princípios do direito civil contemporâneo*. Rio de Janeiro: Renovar, 2006, pp. 1-61.

MOREIRA, Danielle de Andrade; OLIVEIRA, Daniela Marques de Carvalho Oliveira. Sumário executivo da Nota técnica sobre alguns aspectos da 4a versão do Projeto de Lei acerca do licenciamento ambiental (PL 3.273/2004). 2019. (Relatório de Pesquisa).

MULHOLLAND, Caitlin. Dados pessoais sensíveis e a tutela de direitos fundamentais: uma análise à luz da lei geral de proteção de dados (Lei 13.709/18). *Revista de Direitos e Garantias Fundamentais*, v. 19, 2018.

_____, Caitlin; FRAJHOF, Isabella 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, Ana; MULHOLLAND, Caitlin (Orgs.). *Inteligência Artificial e Direito: Ética, Regulação e Responsabilidade*. 1ª Edição. São Paulo: Thomson Reuters, 2019, pp. 265-290.

_____, Caitlin; FRAJHOF, Isabella Z.. Entre as leis da robótica e a ética: regulação para o adequado desenvolvimento da Inteligência Artificial. In: BARBOSA, Mafalda Miranda; NETTO, Felipe Braga; SILVA, Michael César; JÚNIOR, José Luiz de Moura Faleiros. (Orgs.). *Direito Digital e Inteligência Artificial:* diálogos entre Brasil e Europa. 1ª Edição. Belo Horizonte: Editora Foco, 2021, v. 1, pp. 65-80.

OBERMEYER, Ziad; POWERS, Brian; VOGELI, Christine; MULLAINATHAN, Sendhil. Dissecting racial bias in an algorithm used to manage the health of populations. *American Association for the Advancement of Science*, v. 366, n. 6464, pp. 447--453, 2019. Disponível em: <https://science.sciencemag.org/content/366/6464/447>. Acessado em 11.05.2021.

OLIVEIRA, Humberto Santarosa de. *Motivação e Discricionariedade*: As razões de decidir e o contraditório como elementos legitimadores da atuação judicial. Rio de Janeiro: Lumen Juris, 2020.

O'NEIL, Cathy. *Weapons of Math Destruction*. Nova Iorque: Crown, 2016.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (A/73/348), de 29 de agosto de 2018. Disponível em: <https://daccess-ods.un.org/tmp/2168165.89236259.html> Acessado em: 11.05.2021.

PASQUALE, Frank. *The Black Box Society*. Cambridge: Harvard University Press, 2015.

PALMEIRA, Mariana de Moraes. A segurança e as boas práticas no tratamento de dados pessoais. In: MULHOLLAND, Caitlin. (Org.). *A LGPD e o novo marco normativo no Brasil*. 1ª Edição. Porto Alegre: Arquipélago Editorial, 2020, pp. 319-342.

RASO, Filippo A.; HILLIGOSS, Hannah; KRISHNAMURTH, Vivek; BAVITZ, Christopher; KIM, Levin. *Artificial Intelligence & Human Rights*: Opportunities & Risks. Berkman Klein Center for Internet & Society Research Publication Series, 2018. Disponível em: <https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights>. Acessado em 03.03.2020.

RESIMAN, Dillon; SCHULTZ, Jason; CRAWFORD, Kate; WHITTAKER, Meredith. Algorithmic Impact Assessments: A practical Framework for Public Agency Accountability. *AI Now*, abril de 2018. Disponível em: https://ainowinstitute.org/aiareport2018.pdf. Acessado em 01.05.2021.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

RODOTÀ, Stefano. *A vida na sociedade da vigilância*, a privacidade hoje. Rio de Janeiro: Renovar, 2008.

ROVATSOS, Michael; MITTELSTADT, Brent; KOENE, Ansgar. *Landscape Summary: Bias In Algorithmic Decision-Making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it*?. Governo do Reino Unido, 2019. Disponivel em: <http://Inhttps://www.gov.uk/government/publications/landscape-summaries-commissioned-by-the-centre-for-data-ethics-and-innovation>. Acessado em 12.01.2022.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence*: a modern approach. 3ª Edição. New Jersey: Pearson Education Inc., 2009.

SANDVIG, Christian; HAMILTON, Kevin; KARAHALIOS, Karrie; LANGBORT, Cedric. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In: *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, pré-conferência na "64th Annual Meeting of the International Communication Association", 22 de maio de 2014, Seattle, WA, EUA.

SARMENTO, Daniel. *Direitos Fundamentais e Relações Privadas*. Rio de Janeiro: Editora Lúmen Júris, 2008.

SAURWEIN, Florian; JUST, Natascha; LATZER, Michael. Governance of algorithms: options and limitations. *Social Science Research Network*, Vol. 17, Issu 6, 2015, pp. 35 – 49.

SCHERER, Matthew U., Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, Vol. 29, No. 2, Spring, 2016. Disponível em:<http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf>. Acessado em 13.09.2019.

SHOHAM, Yoav; PERRAULT, Raymond, BRYNJOLFSSON, Erik; CLARK, Jack; MANYIKA, James; NIEBLES, Juan Carlos; LYONS, Terah; ETCHEMENDY, John; GROSZ, Barbara; BAUER, Zoe. *The AI Index 2018 Annual Report*. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA, dezembro de 2018.

SELBST, Andrew D.; POWLES, Julia. *Meaningful Information and the Right to Explanation*. International Data Privacy Law, vol. 7(4), 233-242, 2017. Disponível em: <https://ssrn.com/abstract=3039125.> Acessado em 21.04.2018.

_____, Andrew D.; BAROCAS, Solon. Intuitive Appeal of Explainable Machines. 87 *Fordham L. Rev*. 1085, 2018. Disponível em: <https://ir.lawnet.fordham.edu/4r/vol87/iss3/11>. Acessado em 03.01.2020.

_____, Andrew. Disparate Impact in Big Data Policing. *52 Georgia Law Review 109,* 2017. Disponível em: <https://ssrn.com/abstract=2819182>. Acessado em 12.01.2022.

SILVA, Tarcízio. Racismo *Algorítmico em Plataformas Digitais*: microagressões e discriminações em código. In: VI Simpósio Internacional LAVITS, Salvador, Bahia, 2019.

SILVA, Regina Priscilla. Os direitos dos titulares de dados. In: MULHOLLAND, Caitlin (Org.). *A LGPD e o novo marco normativo*. 1ª Edição. Porto Alegre: Arquipélago, 2020.

SILVEIRA, Sergio Amadeu da; SILVA, Tarcizio Roberto da. Controvérsias sobre danos algorítmicos: discursos corporativos sobre discriminação codificada. *Revista Observatório* , v. 6, n. 4, p. a1pt, 1 jul. 2020.

SLAM-CALISKAN, Aylin; BRYSON, Joanna J.; NARAYANAN, Arvind. Semantics derived automatically from language corpora necessarily contain human biases. *Science, Vol. 356, Issue 6334*, 14 de abril de 2017, pp. 183-186. Disponível em <https://www.princeton.edu/~aylinc/papers/caliskan-islam_semantics.pdf>. Acessado em 06.09.2017.

SOUZA, Carlos Affonso; PERRONE, Cristian; MAGRANI, Eduardo. O direito à explicação entre a experiência europeia e a sua positivação na LGPD. In: DONEDA, Danilo; SARLET, Ingo Wolfgang; MENDES, Laura Schertel; JUNIOR, Otávio Luiz Rodrigues (Orgs.). *Tratado de proteção de dados pessoais*. Rio de Janeiro: Forense, 2021, pp. 454-484.

_____, Carlos Affonso de; LEMOS, Ronaldo. *Marco civil da internet: construção e aplicação*. Juiz de Fora: Editar Editora Associada Ltda, 2016.

STONE, Peter; BROOKS Rodney; BRYNJOLFSSON Erik; CALO, Ryan; ETZIONI, Oren; HAGER, Greg; HIRSCHBERG, Julia; KALYANAKRISHNAN Shivaram; KAMAR, Ece; KRAUS, Sarit; LEYTON-BROWN, Kevin; PARKES, David; PRESS, William; SAXENIAN, AnnaLee; SHAH, Julie; TAMBE, Milind; TELLER, Astro. *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence*: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, 2016. Disponível em: <http://ai100.stanford.edu/2016-report.> Acessado em 15.04.2020.

SURDEN, Harry. Machine Learning and Law. *89 Washington Law Review 87,* 2014, pp. 87-115. Disponível em: <https://scholar.law.colorado.edu/articles/81>.

Acessado em 03.01.2020.

_____, Harry. Artificial Intelligence and Law: An Overview. *Georgia State University Law Review, vol. 35, Iss. 4, Art. 8,* 2019, pp. 1306-1337. Disponível em: <https://readingroom.law.gsu.edu/gsulr/vol35/iss4/8>. Acessado em 03.01.2020.

TAYLOR, Linnet; FLORIDI, Luciano; VAN DER SLOOT, Bart. *Introduction*: A New Perspective on Privacy. In: TAYLOR, Linnet; FLORIDI, Luciano; VAN DER SLOOT, Bart (Orgs.). *Group privacy*: New challenges of data technologies. Springer, 2016, pp. 1-12.

TEPEDINO, Gustavo. *Temas de direito Civil – Tomo III*. Rio de Janeiro: Renovar, 2009.

VIOLA, Mario; MATOS, Leonardo Henriques. Proteção de Dados Pessoais no Setor de Seguros: boa-fé objetiva como fonte irradiadora de deveres e os reflexos da Lei n. 13.709/2018 na relação entre Segurado e Seguradora. In: *Revista Jurídica de Seguros* n. 9. Rio de Janeiro: CNSeg, novembro de 2018, pp. 36-61.

WACHTER, Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, Volume 7, Issue 2, Maio de 2017, pp. 76–99. Disponível em: <https://ssrn.com/abstract=2903469>. Acessado em 21.04.2018.

_____, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31 (2), 2018. Disponível em: <https://ssrn.com/abstract=3063289>. Acesso em: 20.03.2021.

_____, Sandra; MITTELSTADT, Brent. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, 2019(2). Disponível em: <https://ssrn.com/abstract=3248829>. Acessado em 02.02.2021.

_____, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, no prelo, 2020. Disponível em: <https://arxiv.org/pdf/2005.05906.pdf>. Acesso em: 20.03.2021.

WARREN, S. D.; BRANDEIS, L. D. The right to privacy. *Harvard Law Review*, 1890, pp. 193–220.

WEST, Mark; KRAUT, Rebecca; EI, Han Chew. I'd blush if I could: closing gender divides in digital skills through education. *UNESCO*. [S.l]: Equals, 2019. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000367416>. Acessado em 12.09.2019.

WEST, S. M., WHITTAKER, M; CRAWFORD, K. *Discriminating Systems*: Gender, Race and Power in AI. *AI Now Institute*, 2019. Disponível em: <https://ainowinstitute.org/ discriminatingsystems.html> Acessado em 07.11.2020.

WIMMER, Miriam ; DONEDA, Danilo C. M. . Falhas de IA e a intervenção humana em decisões automatizadas: parâmetros para a legitimação pela humanização. *Revista Direito Público* , v. 18, pp. 374-406, 2021.

ZANATTA, Rafael A. F.. A tutela coletiva na proteção de dados pessoais. In: *Revista do Advogado – Associação dos Advogados de São Paulo (AASP)*, n. 144, v. 39, nov. 2019, pp. 201-208.

ZANATTA, Rafael A. F.. Tutela coletiva e coletivização da proteção de dados pessoais. *In*: Felipe Palhares (Org.) *Temas Atuais de Proteção de Dados Pessoais*. São Paulo: Revista dos Tribunais, 2020, pp. 345-374.

ZARSKY, Tal. Transparent Predictions. *University of Illinois Law Review*, Vol. 2013, No. 4, 2013. Disponível em: <https://ssrn.com/abstract=2324240>. Acessado em 06.02.2021.

_____, Tal. 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):118-132, 2016. Disponível em: <https://journals.sagepub.com/doi/abs/10.1177/0162243915605575>. Acessado em 20.03.2021.

_____, Tal. Incompatible: The GDPR in the Age of Big Data. *Seton Hall Law Review,* Vol. 47, N. 4(2), 2017. Disponível em: <https://ssrn.com/abstract=3022646>. Acessado em 01.12.2021.