Sugestões ao marco regulatório da IA.

Cesar Beck <ocesarbeck@gmail.com>

qua 11/05/2022 09:48

Para: CJSUBIA <CJSUBIA@senado.leg.br>;

2 anexos

CV - Cesar Beck.pdf; Revisão de decisões automatizadas_ LGPD e mecanismos de Regulação em IA baseados em Ética.docx;

Prezados,

Bom dia,

O meu nome é Cesar Beck, sou Mestre em Direito e estudo Ciência de Dados. Maiores informações, como meu lattes, ORCID e LinkedIn estão em anexo, sob o documento CV.

Em tempo, seguem as minhas contribuições para o marco regulatório da IA no Brasil:
Revisão de decisões automatizadas: LGPD e mecanismos de Regulação em IA baseados em Ética.

Eu escrevi sob o eixo temático sob o eixo temático 4 (Accountability, governança e fiscalização), sub item 4.2. (Códigos éticos e melhores práticas).

Aguardo o recebimento do mesmo.

--
Respeitosamente,
Regards,

Cesar Beck
Empreendedor & Pesquisador
Entrepreneur & Researcher

[Twitter:**]** @ocesarbeck
[Instagram:**]** @ocesarbeck
[e-mail:**]** ocesarbeck@gmail.com

Cesar Augusto Moacyr Rutowitsch Beck

INFORMAÇÕES PESSOAIS

Estado civil: União Estável
Data de nascimento: 02/03/1985
Nacionalidade: Brasileiro.
Endereço: Rua Thomáz Gonzaga no. 799, apto. 1002, Annes, Passo Fundo/RS, CEP: 99020-170.
Natural do: Rio de Janeiro
Telefone: +55 21 99927 7707 **E-mail:** ocesarbeck@gmail.com
ID Currículo Lattes: 6488052912983557
ORCID ID: https://orcid.org/0000-0002-8625-6503
Perfil LinkedIn: https://www.linkedin.com/in/cesar-beck

FORMAÇÃO ACADÊMICA

- 2022 - Presente Pós-Graduação Lato Sensu MBA em Data Science & Analytics pela USP / ESALQ.
- 2020 - 2022 Pós-Graduação em Direitos Digitais, no Programa de Pós-Graduação Lato Sensu pela UERJ (CEPED) / ITS Rio.
- 2020 - 2021 Mestrado em Direitos Humanos (Conceito CAPES 4). Universidade Regional do Noroeste do Estado do Rio Grande do Sul, UNIJUI, Brasil.
- 2013 - 2014 Especialização em Pós-Graduação em Direito Constitucional. (Carga Horária: 360h). Universidade Candido Mendes, UCAM, Brasil. Título: Análise de Aspectos do Direito Constitucional Contemporâneo. Orientador: Rodrigo Padilha.
- 2006 - 2010 Graduação em Direito. Fundação Getúlio Vargas, FGV, Brasil. Título: Da recente aplicabilidade dos danos punitivos pelo STJ em casos de violação de direitos da personalidade; 2010. Orientador: Carlos Affonso Pereira de Souza.
- 2001 - 2004 Ensino Médio (2º grau). The British School of Rio de Janeiro, TBS, Brasil.

FORMAÇÃO COMPLEMENTAR

- 2021 - 2021 Extensão universitária em Dados e Segurança Pública. (Carga horária: 11h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2021 - 2021 Extensão universitária em Tokenização e o Mercado Financeiro. (Carga horária: 13h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2021 - 2021 Gaming Law. (Carga horária: 16h). Future Law, FL, Brasil.
- 2020 - 2020 Extensão universitária em Inteligência Artificial e o futuro do trabalho. (Carga horária: 17h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2020 - 2020 Extensão universitária em Lei Geral de Proteção de Dados: Aspectos Gerais e Desafios. (Carga horária: 22h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2020 - 2020 Extensão universitária em Relações Governamentais, Advocacy e novas tecnologias. (Carga horária: 17h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2020 - 2020 LGPD e GDPR: Transferência Internacional e Além. (Carga horária: 3h). Data Privacy Brasil, DATAPRIVACYBR, Brasil.

- 2020 - 2020 Venture Capital. (Carga horária: 8h). Future Law, FL, Brasil.
- 2020 - 2020 Design de Contratos. (Carga horária: 28h). Future Law, FL, Brasil.
- 2020 - 2020 Inteligência Artificial: aspectos práticos e teóricos de governança. (Carga horária: 24h). Data Privacy Brasil, DATAPRIVACYBR, Brasil.
- 2019 - 2019 Extensão universitária em Blockchain Tezos: Introdução à programação. (Carga horária: 11h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2019 - 2019 Extensão universitária em Distopias do Real e Imaginação. (Carga horária: 14h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2015 - 2015 Extensão universitária em Entendendo Bitcoin na Prática. (Carga horária: 45h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2014 - 2015 Extensão universitária em Curso de Atualização em Direito do Entretenimento. (Carga horária: 120h). Universidade do Estado do Rio de Janeiro, UERJ, Brasil.
- 2014 - 2014 CopyrightX 2014. (Carga horária: 60h). Instituto de Tecnologia e Sociedade, ITS Rio, Brasil.
- 2004 - 2005 Internacional Baccalaureate Organization Diploma. The British School of Rio de Janeiro, TBS, Brasil.

PUBLICAÇÕES

- BECK, Cesar A. M. R.; BOFF, Murilo M.; PIAIA, Thamy C. Os (ab) usos da tecnologia de reconhecimento facial na segurança pública e na prestação de serviços a partir da pandemia de COVID-19. In: Revista Pensamento Jurídico, v. 15, n. 02, maio-agosto, 2021, pp. 389-411, ISSN: 2238-944X.
- BECK, Cesar. O Mundo Pós-Covid-19: a Proteção de Dados e Inteligência Artificial à Luz das Violações dos Direitos Humanos. In: Teses e Dissertações, 2021, UNIJUÍ.
- BECK, Cesar; FORNASIER, Mateus de Oliveira. Cambridge Analytica: escândalo, legado e possíveis futuros para a democracia. Revista Direito em Debate, v. 29, p. 182-195, 2020.
- BECK, Cesar. Pela defesa de uma Inteligência Artificial humanística em plataformas e apps na prevenção de suicídio durante a pandemia do Covid-19 e o Novo Mundo *ex post* pandemia. In: Debates sobre Inteligência Artificial (d.I.A.), ocorrido no Salão do Conhecimento 2020 - Inteligência Artificial: A Nova Fronteira da Ciência Brasileira, 2020, Ijuí. Salão do Conhecimento 2020 - XXV Jornada de Pesquisa. Ijuí: Editora UNIJUÍ, 2020. v. 6. p. 1-4.
- BECK, Cesar. Pela defesa de uma Inteligência Artificial humanística em plataformas e apps na prevenção de suicídio durante a pandemia do Covid-19 e o Novo Mundo ex post pandemia. 2020. (Apresentação de Trabalho/Outra).

IDIOMAS

- Inglês: Compreende Bem, Fala Bem, Lê Bem, Escreve Bem.
- Espanhol: Compreende Bem, Fala Razoavelmente, Lê Bem, Escreve Razoavelmente.
- Italiano: Compreende Razoavelmente, Fala Pouco, Lê Razoavelmente, Escreve Pouco.
- Português: Compreende Bem, Fala Bem, Lê Bem, Escreve Bem.

Revisão de decisões automatizadas: LGPD e mecanismos de Regulação em IA baseados em Ética

Cesar Beck[1]

RESUMO

Esse artigo tem como objetivo geral analisar propostas para regulação sobre sistemas de decisões automatizadas, com base na regulação prevista na Lei Geral de Proteção de Dados para pedidos de revisão de decisões automatizadas. Os sistemas de inteligência artificial possuem camadas de opacidade, em especial aqueles que tomam decisões sem a interferência de seres humanos. Em alguns sistemas, não é possível oferecer precisão absoluta sobre os caminhos que o algoritmo faz para determinada decisão. A LGPD previu a possibilidade de que o indivíduo requeira direito à explicação sobre essas decisões totalmente automatizadas. Esse artigo discute como a previsão da LGPD pode ser uma prerrogativa para regulação em inteligência artificial, através de mecanismos de prestação de contas que inclua auditorias baseadas em ética, Oversight Board e autorregulação setorial, após uma avaliação dos riscos para definir o escopo da empresa e a natureza do tratamento de dados. Sugere-se que os comitês de supervisão (Oversight Board) são organizações independentes ideais para sistemas automatizados que representam alto risco de violação de direitos ou padrões inadequados de decisões. Para outros níveis de risco, sugere-se que a autorregulação setorial pode ser utilizada para uma combinação de responsabilidade, ética e custo-benefício dividido entre as empresas.

Palavras-chaves: Lei Geral de Proteção de Dados. Sistemas de tomada de decisões automatizadas. Oversight Board. Regulação em Inteligência Artificial.

ABSTRACT

This article aims to analyze proposals for regulation on automated decision-making systems, based on the regulation provided in the General Data Protection Law (LGPD) for requests on review of automated decisions. Artificial intelligence systems have layers of opacity, especially those that make decisions without the interference of humans. In some systems, it is not possible to provide absolute precision about the paths that the algorithm takes for a given decision. The LGPD has provided the possibility to individuals require a right to explanation about these fully automated decisions. This paper discusses how the LGPD can be a prerogative for regulation in artificial intelligence, through accountability mechanisms that include ethics-based audits, Oversight Board, and industry self-regulation, after a risk assessment to define the scope of the enterprise and the nature of the data processing. It is suggested that Oversight Boards are ideal independent organizations for automated systems that pose a high risk of rights violations or inappropriate decision patterns. For other levels of risk, it is suggested that industry self-regulation may be used for a combination of accountability, ethics, and cost-benefit split between companies.

Keywords: General Data Protection Law. Automated decision-making systems. Oversight Board. AI Regulation.

Introdução

Os avanços tecnológicos impactaram largamente a sociedade moderna, à medida que se alastram por diversos setores da vida social e essa interação gera problemáticas de ordem tecnológica, mas também ética. A transformação dos mecanismos tecnológicos é

[1] Mestre em Direito (área de concentração: Direitos Humanos; UNIJUÍ), Pós-Graduado em Direitos Digitais (UERJ/ITS RIO). ocesarbeck@gmail.com

tão rápida que o desafio humano têm sido acompanhar adequadamente essas mudanças e ampliar as formas de exercer controle sobre a tecnologia. Enquanto algumas tecnologias possuem maior facilidade de controle, outras são mais opacas até para os especialistas. Embora as problemáticas pareçam de ordem tecnológica, como explica Gutierrez (2020), elas são eminentemente humanas. Relacionam-se com os valores e princípios éticos que regem a sociedade e evocam direitos humanos fundamentais na manipulação da tecnologia.

A perspectiva da regulação tecnológica, amplamente discutida ao redor do mundo, tem como objetivo ajustar as condutas com a tecnologia com base nos princípios éticos que norteiam as práticas sociais, sobretudo a relação entre a Ética e o Direito. Outra forma de compreender esse fenômeno é partir da ideia de que a tecnologia propõe desafios sociais instigantes e exige respostas do Direito, de como a interpretação da tecnologia pode ser realizada para caber ou para alargar as noções já presentes no ordenamento jurídico (DONEDA, 2020). Novas tecnologias propõem alargamento das noções dentro do Direito, de forma a resguardar os direitos humanos, mas com proposição de respostas ao contexto tecnológico. Nessa perspectiva, a regulação das tecnologias para proteger Dados Pessoais e Direitos Humanos tornou-se uma responsabilidade urgente para diversos países, culminando em regulamentos pioneiros, como o GDPR, que forneceu as bases interpretativas para regramentos em outros países.

No Brasil, a LGPD, fortemente inspirada pelo regulamento europeu, foi responsável por estabelecer as bases para o tratamento de dados pessoais no país. Alguns pontos da lei, como as bases legais e a prevalência do consentimento como marco principal da LGPD foi longamente explorado por diversos autores. Ressalta-se que a LGPD tem como objetivo, e como tarefa, garantir transparência e responsabilidade na coleta de dados pessoais, assim como no tratamento desses dados e nas decisões que envolvem qualquer utilização de dados pessoais no Brasil.

Há algumas questões, no entanto, que seguem complexas e com necessidade de análises e produções acerca da problemática. Esse eixo envolve as inovações em inteligência artificial, que representa um grande avanço tecnológico e de capacidade técnica, mas gera diversas problemáticas quanto à auditação das operações que envolvem inteligência artificial. Enquanto há aplicações em IA que possibilitam maior transparência nas operações, outras, como as que envolvem *machine learning* e *deep learning*, possuem maior complexidade e compreende um ambiente onde nem sempre é possível explicar os

caminhos feitos pela máquina para gerar determinado resultado. A transparência, nesse caso, se refere à possibilidade de entender o caminho feito pela máquina para determinada operação. Nos algoritmos mais simples, como aqueles de análises de dados, a aplicação cruza dados já estruturados em busca de um padrão, como numa planilha de excel, por exemplo (GUTIERREZ, 2020). Há outros mais complexos, de aprendizagem de máquina.

A inteligência artificial é um campo vasto que compreende diversas tecnologias e técnicas para produção automática de dados ou soluções. Em linhas gerais, há dois tipos de construção algorítmica em *machine learning* para solucionar determinado problema ou executar uma ação, estes sistemas podem ser supervisionados ou não supervisionados. Os algoritmos de aprendizado supervisionado envolvem um método de análise que utiliza dados estruturados para "ensinar" a máquina o que fazer a partir de dados históricos. Nesse modelo, o algoritmo é treinado para realizar tarefas específicas, a interação humana é fundamental para definir como o sistema atuará, naquele contexto, para conseguir os dados necessários à aplicação. Por exemplo, um indivíduo pode treinar o sistema com diversas imagens a responder que o que está ali é um prédio e, através do código, ensinar a máquina a reconhecer imagens semelhantes a estas.

Nos sistemas de aprendizado não supervisionado, não é necessário controle inicial por parte dos seres humanos. Nesses sistemas, há uma criação em rede de múltiplos dados processados, em que o próprio sistema cria padrões e correlações automaticamente. Os sistemas de aprendizado não supervisionado são capazes de tomar decisões tendo como base a correlação dos dados, sem intervenção (GUTIERREZ, 2020). Nesse modelo, com base no exemplo acima, o ser humano não apresenta nenhum treinamento inicial, apenas carrega imagens de diferentes prédios, casas, construções em geral e até mesmo imagens de outras categorias, e permite que o algoritmo encontre as próprias correlações. Embora a forma como isso acontece seja mais complexa, essa explicação inicial é fundamental para entender o objeto desse artigo.

Diante da possibilidade, em inteligência artificial, de que o aprendizado de máquina seja realizado de forma automática, sem a intervenção humana, como compreender os padrões e correlações feitos por essa cadeia algorítmica para tomar determinada decisão? Em que medida os sistemas de decisões automatizadas operam segundo critérios éticos? É possível auditar o modelo de inteligência artificial que realiza o tratamento de dados pessoais? Como oferecer transparência nessas decisões automatizadas? Essas questões, de forma mais ou menos enfática, aparece na literatura

que busca analisar os eixos responsabilidade e transparência nas operações de inteligência artificial. A problemática central parece situar a possibilidade de que os padrões operados por esses sistemas possuam vieses ou critérios discriminatórios, assim como uma série de decisões que violam direitos humanos fundamentais.

A Lei Geral de Proteção de Dados (LGPD) brasileira resguarda o direito do titular de dados a solicitar revisões de decisões automatizadas, nos casos em que o tratamento automatizado afeta os seus interesses (BRASIL, 2018). Além disso, designou que o controlador deve fornecer as informações ao titular de dados e possibilitar explicações sobre como ocorrem os procedimentos para a decisão automatizada. A lei prevê que o encarregado é um canal de comunicação ampla, uma pessoa indicada pelo controlador para mediar a comunicação entre o controlador, o titular e a Autoridade Nacional de Proteção de Dados. O encarregado pelo tratamento dos dados, conhecido em regramentos internacionais como *Data Protection Officer* (DPO), executa funções de comunicação e deve, também, prestar esclarecimentos acerca do tratamento, quando solicitado. No âmbito desse trabalho, questiona-se os limites que o encarregado possui para mediar as explicações acerca de sistemas com decisões automatizadas.

Com base nessa breve explicação acerca da opacidade dos sistemas de inteligência artificial que realizam decisões automatizadas, o problema que estrutura esse artigo é: em que medida a regulação prevista na LGPD para decisões automatizadas é adequada à complexidade desse modelo de sistema? Em outras palavras, o objetivo geral desse trabalho é analisar os mecanismos adequados para oferecer transparência e regulação de decisões automatizadas para o tratamento de dados pessoais. Parte-se do princípio de que o controle e a fiscalização dos sistemas de IA que tratam dados pessoais é fundamental para proteger direitos humanos, para impedir ou mitigar a formação de vieses algorítmicos discriminatórios nesse modelo de decisão.

A hipótese que norteia essa pesquisa sugere que a regulação de IA carece de mecanismos mais complexos de estruturação das revisões de decisões automatizadas, em face da natureza do problema. A autorregulação presente na LGPD, em que uma figura interna é responsável pelo processo de prestar informações, parece insuficiente em casos mais complexos. Para aumentar a transparência e gerar maior *accountability*, a possibilidade de regulação autônoma, com base em conselhos de supervisão responsáveis por receber as reclamações e realizar auditorias em casos de controvérsia entre os titulares de dados e os agentes de tratamento ou encarregado.

Esse artigo se divide em três capítulos e dialoga com uma bibliografia especializada no campo das interfaces entre o Direito, a Ética, o tratamento de dados pessoais e a inteligência artificial. O primeiro capítulo trata do direito à explicação previsto no art. 18 da LGPD, que enseja a possibilidade de requerer informações acerca do tratamento realizado com os seus dados, esse capítulo busca situar a discussão sobre *accountability* e governança para tratar dados. O segundo capítulo versa sobre problemáticas que envolvem tratamento de dados em modelos de inteligência artificial, em especial o eixo transparência em *Machine Learning* e *Deep Learning*. O último capítulo trata do mecanismo de *Oversight Board* como uma forma de regulação e transparência no tratamento de dados.

1 Accountability e governança: o direito à explicação sob a LGPD

O direito que se convencionou intitular como "direito à explicação" no Brasil está previsto no artigo 20 da Lei Geral de Proteção de dados e permite ao titular dos dados solicitar a revisão de decisões automatizadas que afetam os seus interesses (BRASIL, 2018). Como uma série de normas gerais, não é possível precisar adequadamente como o titular poderá exercer o seu direito à explicação e quais as informações que estarão disponíveis nos casos em que o titular requerer tal direito.

No campo das decisões automatizadas, é preciso compreender o que a lei pretende defender com "direito à explicação", já que a tecnologia possui arquitetura inacessível para a maioria dos indivíduos. Dada a ampliação das decisões automatizadas nas mais diversas instituições e no cotidiano das pessoas, deve-se discutir os impactos reais que uma decisão automatizada pode gerar na vida de um indivíduo e problematizar as formas de exercício do direito à explicação. Não se trata, portanto, de alinhar os princípios da lei, dos quais decorre dois direitos fundamentais, da transparência e da informação, mas situar parâmetros concretos de aplicação da regulação no Brasil. Se a tecnologia é um sistema opaco por sua arquitetura, como prevenir e mitigar os riscos dessa tecnologia?

Embora a lei apresente previsão acerca do direito à explicação, há controvérsias acerca da assimetria informacional e da possibilidade de que o controlador não forneça as informações solicitadas, em face da alegação de "segredos comercial e industrial". Há, ainda, previsão de que a autoridade nacional possa realizar auditoria para verificar possíveis aspectos discriminatórios na automatização das decisões, mas questiona-se, no

âmbito desse trabalho, as desvantagens enfrentadas pelo titular de dados para acesso ao direito à explicação.

Partiremos do conceito de *accountability* para situar o direito à explicação e os seus desafios. Embora não haja tradução exata do termo para o português, esse conceito envolve a utilização de práticas responsáveis, éticas, que visam a transparência e a prestação de contas, quando necessário (GUTIERREZ, 2020). Trata-se de um conceito que abre a necessidade de que, no caso dos sistemas de inteligência artificial, até mesmo as empresas privadas sejam responsabilizadas pelo que estão fazendo, que apresentem passos adequados para que os titulares dos dados possam exercer o seu direito à informação. É um chamado a "prestar contas".

Para o exercício adequado do direito à explicação, há alguns desafios a serem enfrentados. O primeiro envolve barreiras jurídicas, já previstas pela LGPD, em que o segredo de negócios pode bloquear a possibilidade de que o titular de dados acesse informações mais complexas sobre o tratamento dos dados (FRAZÃO, 2021). Outro desafio é um problema da arquitetura do sistema. Como postulado por Lessig (2006), se a arquitetura do código é projetada com opacidade, o controle torna-se difícil, porque no campo dos sistemas tecnológicos, o código é a lei. No caso das decisões automatizadas, o nível de complexidade pode ser ininteligível até para os especialistas.

Há ainda outro problema de arquitetura a ser enfrentado: os sistemas realizam conexões lógico-racionais (VEGA, 2020). As correlações feitas pelo aprendizado de máquina são opacas e nesse sentido, há uma questão importante a ser discutida: os dados processados por um sistema algorítmico, a partir de determinados inputs - ou seja, como a máquina é alimentada com dados - pode ter como consequência outputs enviesados. A máquina não é capaz de gerar a interpretação dos dados, apenas de realizar correlações, cálculos probabilísticos e seguir um caminho tendo como base a arquitetura do seu código. O resultado gerado por essa máquina não é intencionalmente enviesado, mas se os dados possuem vieses, eles serão incorporados na decisão final.

Os modelos de inteligência artificial não supervisionados não podem oferecer explicação de forma razoável, especialmente porque eles não são capazes de interpretar

os vieses dos dados, apenas processá-los segundo critérios e padrões. Embora esse seja um problema abordado no capítulo dois, é necessário destacar que esse é um empecilho ao direito à explicação na LGPD. Para a regulação em inteligência artificial, há alguns passos possíveis. As questões que norteiam essa explicação são: o que deve ser regulado? Como deve ocorrer essa regulação? Quais os mecanismos de explicabilidade envolvidos nesse processo?

A primeira falha do dispositivo normativo que versa sobre as decisões automatizadas é a imprecisão conceitual e a ausência pragmática. Com base no princípio da precaução, Bioni e Luciano (2020) demonstram que é necessário reconhecer as assimetrias de poder e de informação existentes e escolher como a regulação pode enfrentar o conhecimento incompleto que envolve a utilização de decisões automatizadas. Se não é possível prever as decisões que serão tomadas ou até mesmo esclarecer o porquê de determinado viés, os mecanismos de regulação devem ser capazes de assumir a impossibilidade de conhecer completamente as decisões da máquina e implementar mecanismos de mitigação dos riscos.

A LGPD não versa adequadamente sobre como exercer o direito de informação sobre decisões automatizadas. Em uma regulação de modelos de inteligência artificial, o princípio da precaução pode ser utilizado como um mecanismo de avaliação do risco antes do design. Ou seja, é necessário definir regras para o direito à explicação com base no modelo de sistema empregado. A inteligência artificial representa um domínio vasto e cada um desses modelos depende de forma diferente da intervenção humana (GUTIERREZ, 2020).

Se o direito à explicação tem como base o modelo específico ou uma lógica de níveis de risco setorizados, a prestação de contas pode ser mais pragmática. Vale questionar se é necessário regular modelos supervisionados da mesma forma que se regula modelos não supervisionados. By design, o direito à explicação possui mais empecilhos em sistemas totalmente automatizados (FRAZÃO, 2021). É possível pensar que, no processo de explicação de determinada decisão, a responsabilização pode ser facilitada nos casos em que a interpretação depende da interferência humana.

Hartmann et al. (2019) demonstram que para permitir que os indivíduos tenham acesso aos critérios da decisão algorítmica, deve haver transparência e sistemas de IA projetados para respeitar o Estado de Direito, os direitos humanos e fundamentais. Se a

opacidade é um dado de um modelo de sistema, a intervenção humana deve ser um mecanismo de salvaguardar que a interpretação algorítmica não reproduza vieses discriminatórios. De qualquer forma, mesmo com a existência de decisões totalmente automatizadas, as organizações e empresas que operam esses sistemas devem ser responsáveis pelo design apropriado ao ordenamento jurídico, de forma a não comprometer ou violar direitos humanos.

O direito à explicação requer que as proteções adequadas sejam implementadas, em face da complexidade dos sistemas algorítmicos utilizados (FRAZÃO, 2021). No limite, é necessário assumir que o problema com as decisões automatizadas não é uma predição do futuro, mas um fenômeno instaurado que tem gerado diversas consequências para direitos fundamentais, sociais e individuais, em áreas que lidam com dados sensíveis. Se determinado indivíduo não recebe crédito de um banco porque o sistema algorítmico do banco prevê que *pessoas como ele* não pagam, há que se considerar de que forma os dados do *input* desse sistema podem ter gerado um *output* potencialmente discriminatório (CAMPISI, 2021). Ou seja, por vezes, os dados utilizados para alimentar determinado sistema têm discriminações sociais históricas em seu *framework* e geram decisões automatizadas que seguem os mesmos critérios discriminatórios utilizados quando o sistema não existia.

Nesse sentido, dividem-se as soluções técnicas em duas possibilidades, embora ainda seja necessário avaliar a viabilidade da segunda solução. A primeira envolve o conceito *by design*, em que os princípios que devem reger o sistema de IA são incorporados na arquitetura do sistema. As outras medidas são do eixo de *accountability*, que pode incorporar formas de medir e fiscalizar a utilização dos sistemas seguindo alguns mecanismos. É necessário destacar que essas são apenas as soluções técnicas, os princípios éticos devem ser discutidos em torno da regulação (SCHERER, 2016).

Há duas abordagens possíveis para inserção de critérios éticos nos sistemas de decisão em inteligência artificial. Uma abordagem *bottom-up* e uma abordagem *top-down*. Na primeira, os sistemas de IA são treinados para observar o comportamento humano e aprender a tomar decisões com base no padrão de comportamento apreendido. Há um problema direto dessa abordagem, que é a adoção de um comportamento padronizado, com base numa média de ações comuns. Decisões éticas não podem estar a serviço da assimilação de um comportamento corriqueiro, já que esse comportamento pode representar vieses discriminatórios. No geral, a abordagem *bottom-up* seria

desaconselhada, porque se o sistema de inteligência artificial aprende com uma média do comportamento humano e essa média representa corrupção, discriminação ou comportamentos antiéticos afins, a tecnologia seguirá decisões inadequadas (HARTMANN et al., 2019).

Uma alternativa são as abordagens *top-down*, que consiste em incorporar os princípios e regras diretamente na arquitetura do sistema de inteligência artificial. Nessa abordagem, os princípios e as vedações são traduzidos em restrições incorporadas no design. Há uma complexidade que envolve essa alternativa, já que ela, de certa forma, requer que o sistema seja programado para responder a casos concretos. O sistema deve ser programado para reagir a determinada situação ou a um conjunto de situações. Nessa abordagem, a finalidade para a qual o sistema será desenvolvido deve contar para que a programação do sistema seja alinhada para casos concretos (WORLD ECONOMIC FORUM, 2019).

Treinar o sistema para fazer correlações de princípios éticos em casos concretos para aplicar um padrão nas decisões automatizadas parece mais frutífero do que assumir que a média dos comportamentos utilizados possa solucionar o problema da violação de direitos. No que se refere ao direito à explicação, essa alternativa poderia possibilitar maior transparência. Deve-se refletir, no entanto, em que medida esse sistema poderia ser utilizado para operações mais amplas em determinado setor.

Não é possível regular o direito à explicação em decisões automatizadas sem definir o que é uma inteligência artificial automatizada e sem propor regulação para esses sistemas. E esse é o primeiro problema relevante de diversas problemáticas que envolvem a regulação (SCHERER, 2016). A problemática sobre o design desenvolvida acima compreende o que Scherer entende como desafios *ex-ante*, que se referem à concepção da inteligência artificial. Entre esses problemas, é possível mencionar a imprevisibilidade do sistema, que pode operar de formas que os programadores não previram, gerando um labirinto não solucionável. Como as decisões têm implicação real na vida dos indivíduos, essa lacuna pode deixar os dados vulneráveis (SCHERER, 2016). Outros problemas são *ex-post* e envolvem especialmente a implementação de inteligência artificial. Deve-se discutir essas questões do ponto de vista regulatório.

Regular a inteligência artificial com base no princípio da precaução tem duas dimensões: que o debate teve ser amplo e afetar a todos os atores potencialmente

envolvidos na implementação de determinada tecnologia e, de outro lado, medidas para reduzir as incertezas relacionadas ao uso da tecnologia, com vistas a estabelecer regras para o uso ou não de IA numa avaliação dos riscos reais (BIONI; LUCIANO, 2020). Com isso, o que se torna importante no processo não é apenas garantir que o titular tenha direitos sobre os seus dados, mas sobretudo a construção de um processo de gerenciamento dos riscos no tratamento de dados. O risco não pode ser descartado, então regular com o princípio da precaução e com mecanismos de *accountability* parece uma das formas de possibilitar maior proteção em aplicações de inteligência artificial.

O gerenciamento de risco deve ser utilizado como uma forma de avaliar e medir os níveis de risco aos quais os dados dos titulares podem estar submetidos. No limite, as tomadas de decisão nesse contexto tratam do risco tolerável na utilização de sistemas de IA. Como parte da avaliação do risco, deve-se estabelecer uma escala do potencial menor ao maior de violação de direitos ou de tratamentos inadequados de forma geral. É preciso estabelecer princípios éticos em operações de inteligência artificial e utilizar esses princípios para avaliar os riscos em relação ao tratamento de dados. Há uma lacuna do direito à explicação na LGPD, porque o dispositivo não versa sobre as formas de exercício do direito, isso representa, ao mesmo tempo, uma fraca noção de *accountability* na lei, mas uma possibilidade de regulação acessória que adote a discussão pública como um ponto de partida para regular.

O ideal para operacionalizar a ética em decisões automatizadas é ampliar a arena do debate sobre o tema e os atores envolvidos (BIONI; LUCIANO, 2020). Em face da previsão da LGPD de que os titulares solicitem "informações claras e adequadas" sobre decisões automatizadas, é preciso definir que tipos de informações podem ser solicitadas e se essas informações são o suficiente para o exercício dos princípios da transparência e do direito à informação. Ressalta-se que, por exemplo, se as informações prestadas têm como base a publicização do código-fonte do sistema para o titular, essa informação pode não se converter adequadamente em conhecimento e explicação sobre o tratamento dos dados. A precaução como uma abordagem pragmática implica que até mesmo a definição de quais informações explicam as decisões tomadas é um objeto de debate público. A explicação não termina em si mesmo, ela deve pressupor inteligibilidade como um requisito.

2 Transparência sobre o Tratamento de Dados em Modelos de IA por Machine Learning e Deep Learning

Enfrentar os problemas relacionados à transparência e governabilidade nos sistemas de inteligência artificial é a única forma de tornar essas aplicações confiáveis para serem utilizadas para tomar decisões. Há algumas diferenças, já relatadas no capítulo anterior, sobre os sistemas de Machine Learning e Deep Learning. Ressalvadas as diferenças destacadas, esses podem ser entendidos como sistemas automatizados de tomada de decisão e têm como característica em comum a autoaprendizagem e coleta de dados para realizar julgamentos com pouca ou nenhuma intervenção humana. Do ponto de vista técnico, há diferenças entre os sistemas, que variam de árvores de decisão a redes neutrais, mas possuem em comum os desafios éticos por suas características centrais.

A máquina pode aprender a tomar decisões com base num ensinamento direto, em que o indivíduo testa o sistema para produzir determinado resultado, ou pode ser programada para realizar correlações, encontrar padrões e julgar qualitativamente qual decisão deve ser tomada, sem intervenções. Esse potencial é, ao mesmo tempo, o que permite rápidas soluções para problemáticas complexas, mas que carrega no pano de fundo a possibilidade de que a máquina tome decisões que não seriam aceitáveis aos humanos. Mais do que isso, a máquina pode reproduzir um comportamento histórico considerado inadequado.

Supervisionar decisões adotadas por máquinas com os mesmos parâmetros com os quais se avalia decisões humanas não é o melhor caminho. A primeira pergunta necessária é se o sistema que toma a decisão de forma automatizada permite um direito à explicação de fato inteligível. Há autores que apontam que, talvez, a opacidade seja um dado irrevogável que implica na impossibilidade de que decisões inteiramente automatizadas sigam padrões éticos adequados (FRAZÃO, 2021). Para essa problemática, sugere-se a intervenção humana como forma de garantir uma camada de supervisão e revisão da decisão tomada pela máquina. De qualquer forma, a pergunta que se apresenta é: será tangível um mecanismo de explicação em um sistema automatizado? Dada a complexidade da questão, é possível fazer apontamentos e indicações pragmáticas sobre como reduzir os impactos da ação desses sistemas para tomar decisões.

O primeiro desafio para compreender como determinar as soluções técnicas viáveis para sistemas de inteligência artificial é entender como a máquina "pensa" (BURREL, 2016). A discussão sobre o direito à explicação frequentemente gira em torno da inteligibilidade ou da opacidade, mas é preciso entender de que forma a opacidade opera e impede o exercício desse direito. Para Burrel (2016), conhecer as formas de

opacidade é o passo inicial para definir como agir pragmaticamente. A autora distingue três formas de opacidade: opacidade como sigilo intencional corporativo, opacidade por analfabetismo técnico e opacidade pelas características do próprio algoritmo.

A primeira forma de opacidade ocorre amplamente nas corporações, que buscam a opacidade para manter vantagem competitiva em face dos seus concorrentes. Esses algoritmos são utilizados especialmente no campo de mecanismos de pesquisa, para filtrar, classificar e moldar a atenção do público de forma dominante. Esses sistemas funcionam como aprendizagem adversária e, de forma geral, são algoritmos que raramente estarão à disposição dos titulares. A opacidade nesse caso não é um erro do design, mas um elemento intencional. Segundo Burrel (2016), uma alternativa seria modelos de negócio que trabalham com software de código aberto.

Pasquale (2016) explica que a opacidade algorítmica pode ser uma forma consciente de evitar a regulamentação dos algoritmos. A opacidade no ramo competitivo é uma forma de dissimulação e tem implicações práticas na vida de milhares de indivíduos, mas as consequências dessas ações, sobretudo no que se refere aos Direitos Humanos, estão escondidas por trás de um "sigilo comercial". Notadamente, a cortina de fumaça gerada por esse campo da informação sigilosa é conveniente para instituições que tratam os dados dos titulares com fins lucrativos. Como defende o autor, as leis de proteção as informações no campo comercial são dominantes frente aos regramentos que tratam da privacidade dos indivíduos. A incongruência é clara, mas as práticas se tornam, de certa forma, quase invisíveis, pois não é possível medir diretamente os potenciais danosos dos algoritmos usados por esses setores.

A segunda forma de opacidade é resultado do analfabetismo técnico. Conhecimento é poder (PASQUALE, 2016). A opacidade como analfabetismo técnico tem como consequência direta a assimetria informacional. Não apenas escrever códigos é uma habilidade especializada, mas ler código e interpretar os seus resultados é uma capacidade técnica indisponível para a maioria da população. Nos cursos de engenharia de software, Burrel (2016) afirma que a escrita limpa e inteligível do código é incentivada, mas a linguagem de programação possui sintaxes diferentes das linguagens humanas. A comunicação via código deve ser exata, lógica e formal, com o risco de o código ser mal projetado caso não siga regras necessárias ao dispositivo computacional. Uma das regras "de boa convivência" no campo da programação envolve a inclusão de comentários ao

código, com a criação de caminhos para que o código seja formulado de forma mais simples e possa ser lido por outros programadores.

Apelos de diversos setores tem sugerido a necessidade de que o pensamento computacional seja desenvolvido em todos os níveis educacionais. O desafio deve ser aplicar esforços para tornar os indivíduos mais informados acerca do tratamento dos dados e da lógica computacional de maneira geral. Traduzir soluções técnicas em informações palatáveis ao conhecimento médio dos indivíduos deve ser um objetivo, para que a autodeterminação informativa possa ser realizada de forma crítica. Esse tipo de opacidade carece de um esforço de longo prazo (BURREL, 2016).

A última forma de opacidade tem relação com a maneira como os algoritmos operam na aplicação. Essa forma de opacidade pode ser entendida como a própria complexidade do sistema, multicomponentes, que até mesmo os programadores daquele algoritmo específico podem enfrentá-la. Embora seja possível desvendar, a muito custo, a lógica da decisão de um software complexo, isso implica em um esforço de horas ou dias. Nesse sentido, Burrel (2016) argumenta que há desafios e escalas de complexidade que são inerentes aos algoritmos de *Machine Learning*. O desafio não é apenas de design de códigos complexos, mas do algoritmo em ação, sobre como esse algoritmo opera nos dados e a complexidade inevitável de algumas decisões.

A lógica é que quanto mais dados um sistema tiver que tratar e correlacionar, mais complexidade resultará na decisão tomada automaticamente. Se um sistema lida com muitos dados com propriedades heterogêneas, há maior complexidade na correlação que o sistema fará, para buscar padrões e gerar um resultado. A opacidade é um resultado da implementação do código e dos dados que alimentam o sistema. A interação entre esses dois mecanismos é o que produz a complexidade final. Não há como simplificar demasiadamente o código, parte do esforço em explicabilidade em inteligência artificial é assumir que uma parcela de opacidade é inevitável (BURREL, 2016).

Situados os desafios da opacidade, há um debate central de natureza principiológica. Floridi e Cowls (2019) estruturaram cinco princípios abrangentes e fundamentais para uma inteligência artificial ética. A beneficência, não maleficência, autonomia, justiça e explicabilidade. Da explicabilidade, decorre o princípio da inteligibilidade e do dever de prestar contas. A literatura especializada se dedicou a

apresentar conjuntos de princípios relevantes que envolve a mesma natureza semântica, com ideias repetitivas.

Em face disso, Floridi e Cowls (2019) declaram que uma proliferação de princípios não pode resolver o problema da ética em IA, embora a principiologia seja fundamental para estruturar qual caminho seguir. É preciso admitir que os princípios são a base para as práticas, portanto, deve-se posicionar e seguir com a operacionalização dos princípios nas ações. A ideia, como defendem Raji et al. (2020), é que traduzir os princípios em prática tem sido desafiador, porque os princípios são vagos e não fornecem muitas ideias de como empregar mecanismos de responsabilização. Há que se considerar a importância dos princípios, mas o intuito deve ser traduzir os princípios em práticas para operacionalizar a mudança desejada.

Para Morley et al. (2021), os guias de ética baseados em princípios produziram um cenário de indicações abstratas e pouca aplicabilidade prática, em especial para os responsáveis por projetar algoritmos e gerir os sistemas de inteligência artificial. Isso tende a criar a ideia de que os algoritmos são bons ou ruins, em um julgamento moral, ao invés de entender que esses algoritmos são bem ou mal projetados por conta de ser desenhado de forma potencialmente discriminatória. Sistemas de inteligência artificial não devem ser objetivamente avaliados como bons ou ruins, éticos ou antiéticos, porque isso implica uma responsabilidade ao sistema de IA em si, que não pode atuar sem os agentes humanos, mas, nesse desenho, os sistemas parecem agentes independentes.

A questão, para os autores, é a invisibilização da agência social dos desenvolvedores, engenheiros e programadores, que são algumas das figuras centrais no processo de operacionalização da ética. O impacto que um sistema de IA possui é definido pelas escolhas no seu desenvolvimento e na alimentação do sistema com os dados. Os profissionais que realizam essas operações são encorajados a trabalhar de forma ética, mas possuem pouca clareza sobre o que, de fato, isso significa na prática (MORLEY ET AL, 2021).

Os profissionais do campo da inteligência artificial têm o dever de cuidado com aqueles que recebem as decisões tomadas com base nos algoritmos. Apesar disso, os profissionais de IA não podem assumir sozinhos a responsabilidade pela neutralidade das decisões sem saber os resultados sociais possíveis daquelas ações. Diretrizes gerais são fundamentais para as práticas profissionais desses atores, mas devem incorporar um

conjunto de ferramentas de aplicação. Os profissionais devem poder obter uma assistência com base no status normativo do tema, mas que seja aplicável as decisões do mundo real (FLORIDI; TADDEO, 2016; MORLEY ET AL., 2021).

Se os desenvolvedores veem o código como uma série de linhas lógicas com base numa sintaxe específica e apenas isso, não estarão engajados na mudança das lógicas discriminatórias. Explicar que o comportamento deve ser ético não é o suficiente para tornar a ética palatável no sistema desenvolvido. Garantir que esses profissionais estão envolvidos nas decisões de um projeto de operacionalização da IA, assim como outras partes interessadas, é fundamental para ampliar o escopo da discussão e utilizar as repercussões práticas como facilitadores do desenvolvimento de uma inteligência artificial ética (MORLEY ET AL., 2021).

Na pesquisa conduzida por Morley et al. (2021), alguns dados demonstram como os profissionais que trabalham diretamente com o desenvolvimento de sistemas de inteligência artificial tem conhecimento restrito dos princípios gerais que devem reger a concepção do sistema. Embora a grande maioria acredite que a ética é relevante, especialmente para garantir a confiança e a satisfação do consumidor e melhorar o impacto social, muitos deles reconhecem apenas a privacidade e a segurança como princípios de proteção de dados. A autodeterminação informativa, que tem sido um feixe importante de discussão no campo do Direito e das Ciências Sociais, é o princípio menos reconhecido.

No entanto, o impacto público do enviesamento dos algoritmos ressoa na prática dos profissionais, já que parte deles definem uma IA ética como uma "IA não tendenciosa", resultado da cobertura da mídia sobre casos de IA com vieses algorítmicos. Os atores envolvidos na pesquisa acreditam que um design pró-ético tem custos adicionais, mas tem outros pontos positivos. É interessante pensar como no campo do aprendizado de máquina a prática de mitigar os danos parece menos custosa do que a concepção de um design ético (MORLEY ET AL., 2021).

O ponto mais crítico acerca dessa pesquisa envolve a falta de clareza sobre as responsabilidades cabíveis a cada ator envolvido no processo. Os desenvolvedores entendem que são capazes de projetar IA de forma ética quando são responsáveis pela decisão. Porém, os responsáveis finais pelo sistema não entendem o alinhamento entre design e princípios. Essa problemática é particularmente relevante porque são os

responsáveis finais que avaliam e sancionam as ações dos desenvolvedores. Grande parte dos profissionais que desenvolvem e implantam IA não estão confortáveis com os canais de denúncia em caso de violação dos direitos humanos ou de clara opacidade das implicações éticas de determinado sistema. Falta, nesse sentido, mecanismos de apoio aos profissionais responsáveis, como um eixo importante da transparência em Machine Learning e Deep Learning (MORLEY ET AL. 2021).

Cabe ressaltar que a transparência, nesse caso, não envolve apenas a característica do modelo de IA, mas de todo o processo de responsabilização de um design ético. Não parece claro, para os profissionais, sobre quais são os responsáveis por responder às preocupações éticas, até mesmo por risco de retaliação. Nesse sentido, a ênfase dada por Pasquale (2016) na opacidade como um projeto de tornar a inteligência artificial uma caixa preta é algo relevante, já que mesmo para especialistas a responsabilização dos seus superiores não parece disponível.

Não há ferramentas suficientes disponíveis para os profissionais de IA desenvolver sistema com design pró-ético. O que está disponível para esses atores não são os recursos que eles consideram úteis. A parcela dos entrevistados que possui algum tipo de estrutura ética para desenhar o sistema é menor do que a proporção que considera que ter um recurso para projetar produtos de forma ética é útil. Há uma lacuna, portanto, no alcance do debate e da operacionalização ética na comunidade de IA. Esses dados, mesmo limitados, fornecem caminhos importantes: a comunidade de IA que discute ética parece estar completamente afastada das necessidades reais dos profissionais que desenham os sistemas (MORLEY ET AL., 2021).

Para que as estruturas de ética forneçam sistemas mais transparentes, deve-se desenvolver padrões exequíveis, não meras compreensões abstratas. Muitas pesquisas têm destacado que os mecanismos de responsabilização externa podem ser alternativas para traduzir princípios em práticas de design, mas esse mecanismo possui debate mais recente. (MORLEY ET AL., 2021; RAJI ET AL., 2020; MOKANDER ET AL., 2021).

3 Regulação em inteligência artificial e Oversight Board como mecanismo regulatório

As organizações que projetam e implantam algoritmos devem ser responsabilizadas por potentes estruturas de governança (RAJI et al., 2020). Ao longo do trabalho, destacaram-se vários desafios para a concepção de ADMS éticos e esse

ambiente possibilita maior reflexão acerca dos mecanismos de governança, fiscalização e controle dos sistemas para conformidade ética. As auditorias são ferramentas interessantes para avaliar processos dentro de uma empresa, com vistas a confirmar se os processos da empresa estão de acordo com a própria política empresarial e com o ordenamento jurídico. As auditorias podem ser divididas em dois grupos: para confiabilidade do sistema e para avaliar danos sociais. Enquanto uma prevê uma fiscalização *ex-ante*, para garantir que o sistema foi desenhado de forma ética e para avaliar possíveis riscos, o outro modelo avalia os danos causados e impactos gerados, *ex-post*. Ressalta-se que essas duas formas de avaliação são complementares.

Como visto anteriormente em Morley et al. (2021), é indispensável haver uma cadeia de responsabilização para conformidade ética, em que os responsáveis por cada etapa saibam perfeitamente a quem recorrer em caso de cenários atípicos. Delegar a responsabilidade de avaliar o impacto de um sistema a uma figura específica, como a figura do DPO na LGPD, é insustentável em ADMS, porque a cadeia de tratamento é demasiada complexa para avaliação. Além disso, sugere-se que uma equipe diversa tenha maior capacidade de avaliação de potenciais vieses discriminatórios, pois é possível que uma equipe pouco diversa sequer entenda a repercussão de determinado tratamento de dados na proteção dos direitos humanos.

O que denomino como "cadeia de responsabilização" tem como base o esquema feito por Mokander et al. (2021), que trata de uma rede complexa, em especial: organizações que implantam ADMS e que são responsáveis pelo comportamento dos sistemas; a gestão das organizações, responsáveis pela adesão aos valores éticos; auditores independentes, que devem avaliar a aderência da organização aos princípios e normas; e os reguladores, que monitoram a conformidade de diversas organizações na prática. Esse modelo pode ser entendido como uma inspiração para regulação em inteligência artificial, não necessariamente deve ser empregado, mas tem boas pistas de como seguir. A ideia de que a auditoria deve ser independente, por exemplo, é algo fundamental para barrar as falhas éticas dentro da organização. Essa autonomia tem duas consequências, ou impactos: dirimir o risco de enviesamento na análise dos auditores, já que eles não estão operacionalmente dependentes do comando das empresas, e facilitar a visualização da responsabilidade por diferentes falhas no sistema. Uma visão "de fora" pode fornecer maneiras de detectar consequências inadequadas dos ADMS.

Um desenvolvimento recente que pode ser um mecanismo promissor para regulação da prática de IA é a EBA, ou, *Ethics-Based Auditing*, uma forma de governança para as organizações projetarem e implantarem sistemas éticos. A auditoria é um mecanismo de avaliação de concepção, de risco e de impacto, com base nos princípios e normas relevantes. Essas auditorias são abordagens recentes e ainda não padronizadas, mas elas se dividem em três modalidades: auditorias de funcionalidade, que focam no porquê da decisão; auditorias de código que revisam o código-fonte e auditorias de impacto, para investigar os efeitos das decisões do algoritmo. A EBA é responsável por fiscalizar os sistemas independente da responsabilidade de fiscalização pelo auditado na gestão diária de seus sistemas. É uma forma potente de interação entre as organizações e o setor específico de fiscalização, pois permite monitoração e contestação de forma consistente, equilibrando os conflitos de interesses (MOKANDER; FLORIDI, 2021).

O processo de EBA deve ser contínuo, dialético, estratégico e orientado para o design (MOKANDER; FLORIDI, 2021). Cada um desses pontos tem uma aplicabilidade prática. A auditoria baseada em ética não é um processo único, não deve seguir os percursos comuns de uma auditoria como um "evento" específico, mas como um mecanismo contínuo de monitoramento e avaliação. Deve-se documentar continuamente o desempenho do sistema para prevenir potenciais inadequações. É interessante pensar que esse procedimento requer, inclusive, que o sistema seja alimentado ocasionalmente com dados-testes, num ambiente supervisionado, para entender de que maneira o ADMS está tomando decisões.

Os dados utilizados nesses testes devem ser suficientemente realistas, porque as falhas no design do ADMS só se manifestam com dados que apresentam problemas reais. Hartmann et al. (2019) defende monitoração do modelo a ser utilizado em todas as fases, tanto de treinamento quanto de implementação, para que a estabilidade do sistema possa ser validada. O tratamento de dados experimentais nessas fases é relevante, dentro dos limites da opacidade do sistema, para diagnosticar comportamentos potencialmente discriminatórios. A equipe envolvida nessas operações deve conceber e executar a avaliação em um ambiente diverso, porque as decisões tomadas pelo sistema precisam ser interpretadas por seres humanos e a diversidade de indivíduos pode complexificar o diagnóstico sobre o sistema. Métricas diversas devem ser desenvolvidas para analisar as categorias testadas, com o auxílio das perspectivas diferentes dos atores envolvidos no processo de avaliação. Ainda há a possibilidade de que sejam conduzidos testes

conscientemente controversos, ou "testes de adversidade", com a função deliberada de tentar quebrar o sistema para encontrar as suas vulnerabilidades.

Há benefícios das auditorias internas que vale destacar. Embora as auditorias externas sejam conduzidas por especialistas que não fazem parte da organização e, portanto, podem evitar conflitos de interesse, as auditorias internas possuem vantagens. Uma delas se refere ao acesso mais facilitado aos ADMS, mesmo com as políticas empresariais de segredo industrial e comercial. O acesso que os auditores internos possuem para avaliar o sistema, de forma direta, tem implicações na qualidade da avaliação (RAJI et al., 2020).

Como destacado em outro capítulo, a caixa preta dos sistemas de inteligência artificial tem relação com as leis de informação que resguardam às empresas o direito de não macular suas informações sigilosas (PASQUALE, 2016). O código-fonte de um ADMS, no cenário de competitividade empresarial, não pode ser divulgado sem que isso cause prejuízos para essa organização. Embora se destaque que a transparência em IA não precisa revelar segredos industriais, alguns objetos de tratamento requerem um escrutínio sobre o código-fonte. As auditorias internas são aconselháveis mesmo em um cenário em que haja regulação setorial e auditorias externas.

As auditorias internas podem avaliar quão bem determinado ADMS irá operar no mundo real, com testes que visam garantir a aderência do sistema aos padrões. O importante no processo de auditoria interna é que seja duradouro e holístico e que não adote o modelo pós-implantação para gerir os sistemas de IA (MOKANDER; FLORIDI, 2021). Para um modelo de auditoria adequado, deve-se proceder com uma auditoria pré-implantação em todo o processo de desenvolvimento da IA, para promover EBA. Medidas de mitigação implementadas apenas pós-implantação e dano é algo desaconselhável (RAJI et al., 2020).

Ressalta-se que, no caso de auditorias externas, no geral, as auditorias são pós-implantação, para avaliar possíveis impactos. As auditorias internas podem fiscalizar os modelos desde o início de seu design, pelo nível de acesso ao sistema interno da empresa. Para esse modelo de auditoria, deve-se cobrar relatórios mais rigorosos, já que a transparência per si conta com a auditabilidade da própria empresa. O processo de desenvolvimento dos sistemas automatizados de tomada de revisão pode apresentar, na medida do possível, engenharia auditável, que gerem relatórios de desenvolvimento

sérios e com rigor. A auditoria interna, nesse caso, é um braço do princípio da transparência na regulação em IA (RAJI et al., 2020).

A EBA não é um mecanismo de substituição dos mecanismos tradicionais de governança e *accountability*. A auditoria baseada em ética deve ser entendida como uma forma de complementar outras ferramentas, como a supervisão humana e a regulamentação. A auditoria, nesse sentido, é responsável por atender os pressupostos éticos sem incorrer em leis tão rígidas que sufoquem a possibilidade de inovação nos ramos que utilizam inteligência artificial. A EBA é uma forma de fortalecer a infraestrutura ética na sociedade, mas não substitui outros mecanismos importantes para proteção dos direitos humanos (MOKANDER; FLORIDI, 2021).

Há que se considerar, ainda, que algumas restrições econômicas podem reduzir a eficácia dessas autorias, assim como a viabilidade de que esse modelo seja utilizado por algumas organizações. O custo de auditoria é justificável diante do benefício levado para a empresa, mas as empresas menores podem ter impactos desproporcionais ao precisar deslocar uma equipe mais robusta para garantir a adequação ética nas auditorias. Os custos, assim como os benefícios, de utilização de modelos de auditoria baseados em ética devem ser divididos entre vários setores. As empresas que não puderem realizar auditoria contínua de como satisfatório, pelo tamanho da empresa e capital disponível, deve ser capaz de acessar recursos de regulação setorial (MOKANDER et al., 2021).

Em 2020, o Facebook criou um *Oversight Board* para decidir sobre a publicação de conteúdos no Instagram e no Facebook. Mais precisamente, para avaliar pedidos de revisão acerca de conteúdos bloqueados pela plataforma, no escopo inicial de sua jurisdição. O OB é um Conselho de Supervisão criado pelo Facebook, mas não subordinado à sua estrutura hierárquica, mantendo a condição de um órgão independente. O conselho é composto por membros de diversas regiões e com perfis muito diferentes. O Facebook decidiu criar esse modelo experimental diante das práticas controversas que cometeu, no que se refere à privacidade dos usuários e à existência de conteúdos que violam os direitos humanos em sua plataforma. Há duas considerações relevantes sobre o Oversight Board: ele foi criado como uma organização autônoma, os conselheiros têm mandatos fixos e contam com um orçamento de 130 milhões de dólares (LEMOS, 2020).

O Conselho não está subordinado ao Facebook e julgará de forma independe cada caso, sem avaliar as recomendações feitas pelo Facebook em qualquer situação. A ideia

é que o OB tenha autoridade para decidir se o Facebook deve, ou não, remover conteúdos considerado com impróprios, discriminatórios ou inadequados. A característica central desse princípio da autoridade, nesse caso, é que as decisões do conselho são vinculantes e o Facebook deve acatá-las imediatamente, sem pedidos de revisão. Acessoriamente, o comitê pode recomendar que as políticas de conteúdo mudem, em face de algum caso emblemático. Com base no princípio da transparência, o comitê do Conselho deve divulgar publicamente as explicações sobre suas decisões. As decisões, assim como suas justificativas, estarão disponíveis no site do comitê. Em qualquer caso que o comitê for solicitado a se manifestar para uma decisão, deve tornar públicas as razões desse veredicto. Por fim, relatórios anuais do trabalho desse Conselho serão publicizados em seu site.

O Comitê Independente do Facebook analisa casos emblemáticos em que decisões foram tomadas, seja para manter as decisões ou modificá-las. No campo da regulação em inteligência artificial, estima-se que é possível alargar o escopo de atuação de um *Oversight Board*, para julgar decisões tomadas pela empresa desde a concepção do ADMS. O mecanismo de *Oversight Board* preserva algumas características importantes para a auditação em inteligência artificial: atores com perfis diversos envolvidos com a análise, autonomia para acessar informações da empresa, transparência para prestar contas à sociedade sobre o tratamento dos dados dos titulares e independência para julgar concepções *ex-ante* e impactos *ex-post*. Outra característica relevante é a possibilidade de que o Conselho de Supervisão indique medidas que a empresa deve seguir para não incorrer em outras avaliações inadequadas.

Um Conselho de Supervisão para auditorias baseadas em ética em IA seria uma forma de resolver problemas destacados ao longo dos capítulos, em que o comitê preserva a autonomia para julgar os casos, por não estar subordinado à hierarquia da empresa. Como lembrado por Mokander et al. (2020), a responsabilidade de garantir ADMS éticos é das organizações que desenvolvem e operam esses sistemas, portanto, mecanismos de governança devem garantir que o poder de decisão sirva para garantir e preservar direitos e não para fugir das sanções. Ao mesmo tempo, delegar toda a capacidade de decidir para atores internos pode macular os princípios da transparência e da impessoalidade. Se há um mecanismo de auditoria independente, vinculado à empresa mas não subordinado a ela, esse deve ser capaz de decidir e indicar formas das organizações prevenirem os erros, preferencialmente, ou mitigá-los.

A abordagem regulatória que parece mais adequada, inclusive no que se refere a utilizar *Oversight Board* como um mecanismo de auditoria, pode ser baseada no risco. Ressalta-se quatro níveis de risco que um ADMS pode representar e cada um desses riscos devem ser avaliados em seus contextos sociotécnicos específicos. Os quatro níveis seriam: risco inaceitável, alto risco, risco mínimo ou limitado. Aderir aos códigos de conduta é uma responsabilidade coletiva e irrevogável, mas a forma como cada organização deve responder para garantir a transparência no tratamento dos dados pode ser diferente (MOKANDER; FLORIDI, 2021).

Por exemplo, um sistema que não lida diretamente com dados sensíveis, ou que utilizam dados para previsão e mitigação de danos ambientais, pode não estar submetido a mesma abordagem que um sistema que trata dados para estabelecer critérios de atendimento em hospitais. Ou seja, o escopo de atuação da própria organização, que implica na concepção dos sistemas, pode ser um critério de avaliação do risco. Nos sistemas que forem avaliados como risco inaceitável, é aconselhável que a proposta seja banida (FRAZÃO, 2021). Ou que, num cenário ideal, seja possível colocar uma camada de intervenção humana diversa após a decisão do algoritmo. Os sistemas de alto risco devem ser submetidos a avaliações de conformidade ética obrigatórias, *ex-ante* e fornecerem relatórios rígidos e até mesmo públicos sobre as decisões do sistema (MOKANDER et al., 2021).

Mokander e Floridi (2021) argumentam que não é possível medir completamente o risco que um sistema possui, mas é possível ter uma ideia de quais impactos um algoritmo pode causar, em face de algumas métricas. Brown, Davidovic e Hasan (2021) desenvolveram métricas para avaliar o desempenho específico de um algoritmo, como um passo inicial para definição do risco que determinado algoritmo apresenta. Para os autores, deve-se estabelecer uma avaliação das partes interessadas com o uso de determinado sistema, para identificar de que forma as partes podem ser impactadas. Isso deve ocorrer porque grupos e indivíduos diferentes são impactados de formas diversas com o tratamento de dados e decisões automatizadas. Portanto, essas métricas devem ser capazes de medir os riscos para identificar qual eixo regulatório cabe melhor para o caso.

Os Oversight Boards são aconselháveis para organizações de maior porte ou que tratem dados para fins sensíveis, cumulativamente ou não. As empresas que automatizam decisões com potenciais discriminatórios, com base até mesmo na revisão de outros casos de vieses relatados na área, devem constituir um mecanismo de supervisão mais rígido e

robusto. Cabe a essas empresas auditar o ADMS desde o princípio, realizar testes com os algoritmos, tentar quebrar o sistema para identificar eventuais falhas e relatar periodicamente os impactos do sistema na prática.

É possível, via Comitê de Supervisão, fazer isso sem quebrar os sigilos comerciais e industriais, para garantir a prestação de contas aos titulares dos dados sobre como o sistema procede. Esses comitês podem ser acionados para esclarecimento sobre algum caso, por um terceiro interessado ou por um desenvolvedor de sistemas da empresa que encontrou problemáticas na utilização do sistema. O comitê pode revisar os casos e decidir quais as medidas que a empresa deve tomar. A equipe, como enfatizado, deve ser diversa, especializada e compreender os cenários sociotécnicos de utilização do ADMS profundamente.

Considerações finais

O objetivo desse trabalho era analisar os mecanismos de regulação em inteligência artificial e sistemas automatizados de tomadas de decisão. A ideia do trabalho foi situar as previsões sobre a transparência e a explicabilidade na Lei Geral de Proteção de Dados Brasileira, com o intuito de pensar se as figuras previstas pela LGPD eram suficientes para propor um modelo de regulação em sistemas automatizados. Dada a discussão sobre a literatura e reflexões trazidas ao longo do capítulo sobre a necessidade de uma regulação robusta em IA, entende-se que a previsão sobre decisões automatizadas da LGPD, e sua responsabilização, não é o suficiente para fiscalizar, supervisionar e agir sobre um tema que tem cadeias de decisões tão complexas.

A figura do *Data Protection Officer*, prevista na LGPD como o encarregado, tem função de mediar a comunicação entre o controlador, os titulares de dados e a agência reguladora. O encarregado é o responsável por proceder em face dos pedidos de explicação feito pelos titulares, mediando as informações, embora a LGPD tenha previsão de que o controlador deve fornecer todas as informações solicitadas pelo titular dos dados, ressalvadas as possibilidades de segredo comercial e industrial. A hipótese que norteava essa pesquisa sugeria que a figura do DPO não é suficiente para a revisão de decisões automatizadas, em face da complexidade dos sistemas e da limitação em reconhecer potenciais discriminatórios em casos mais robustos. Como demonstrado ao longo do

trabalho, uma equipe diversa e especializada parece mais adequada para compor um mecanismo de fiscalização e avaliação das decisões dos sistemas automatizados.

Concentrar o poder de revisão em um indivíduo, mesmo que esse encarregado tenha grande conhecimento técnico e especializado, subestima a complexidade da tomada de decisão em sistemas automatizados. Embora esse modelo possa funcionar no escopo e limites da LGPD, a regulação em inteligência artificial deve ser mais robusta. Sugeriu-se, como hipótese, que um mecanismo de *Oversight Board* pudesse qualificar a atuação de profissionais envolvidos com o tratamento de dados e daqueles com a responsabilidade de dar prosseguimento aos requerimentos de explicação dos titulares. O objetivo geral desse trabalho, portanto, parece ter sido alcançado. Os Conselhos de Supervisão, ou mesmo as auditorias setoriais, tem maior aderência à complexidade técnica dos ADMS.

A Lei Geral de Proteção de Dados fornece uma porta de entrada importante para a regulação em IA, em fase da previsão expressa sobre o direito à explicação em casos de decisões automatizadas, mesmo que a expressão "explicação" não apareça, ela está presente como desenvolvimento do princípio da informação e da transparência. Em comparação com o GDPR, que não prevê direito à explicação dessa forma, a LGPD tem passos que abrem margens relevantes para a discussão sobre regulação no Brasil, com precedentes que envolvem o direito à explicação e alguns mecanismos de regulação passíveis de serem empregados também em inteligência artificial.

É impossível oferecer respostas taxativas para a regulação em inteligência artificial, porque o campo carece de reflexão constante acerca de quais mecanismos podem possibilitar um ambiente ético e de proteção dos direitos humanos e sociais. Apesar disso, é possível indicar caminhos possíveis e abrir debates sobre a validade de alguns instrumentos nesse campo. A regulação deve estar acompanhada desses instrumentos, destaca-se aqui as auditorias como mecanismos de avaliação dos sistemas automatizados desde o seu design.

As auditorias externas, feitas por terceiros que não possuem nenhuma relação com a empresa, tem vantagens em termos de conflito de interesse, mas desvantagens pela dificuldade de acesso a dados mais complexos resguardados por segredo comercial. As auditorias internas, por outro lado, podem facilitar o acesso à informação, mas a estrutura hierárquica da empresa pode influenciar a responsabilização e até mesmo a opacidade, de forma consciente, em caso de materiais discriminatórios, perigosos ou inadequados. Uma

auditoria independente, mas com acesso irrestrito aos dados e capacidade de decidir acerca das ações da empresa com os sistemas automatizados têm grande vantagem, com o mecanismo de *Oversight Board*.

Esse Comitê deve ter uma equipe diversa e especializada, para avaliar os sistemas automatizados de tomada de decisão em diversos cenários, não apenas quando solicitado pelo titular de dados o seu direito à informação. Porém, nos casos em que o titular de dado requerer o direito à explicação, o Conselho pode avaliar e decidir sobre o caso, publicizar informações sobre o tratamento e sobre a sua decisão. Pode, ainda, indicar à empresa como proceder em diante. Um mecanismo como esse requer um dispêndio financeiro significativo, então a sua existência deve estar atrelada ao risco dos fins do sistema automatizado, do escopo de atuação da empresa e do seu porte.

Nos casos em que o risco sobre os dados for mínimo ou limitado, a autorregulação setorial pode permitir que as organizações dividam os custos das operações de prestação de contas. O mais importante é que as empresas e organizações assumam a responsabilidade de gerir os seus sistemas automatizados de tomada de decisão de forma ética, seguindo os princípios de *accountability* e com aplicações práticas para oferecer para os indivíduos o direito de saber, adequadamente e segundo as suas limitações de conhecimento técnico, como o tratamento dos seus dados interfere nas decisões que afetam as suas vidas.

Para pesquisas futuras, entende-se que é necessário aprofundar os esforços em torno das regulações setoriais ou multissetoriais. A partir desse trabalho, com a conclusão de que os custos de manutenção de uma estrutura de supervisão de ADMS *ex-ante* e *ex-post* pode ser muito custoso para pequenas empresas, entende-se que outras vias podem ser mais adequadas, que considerem o risco, o escopo de atuação da empresa, a finalidade dos seus sistemas algoritmos e o porte dessa empresa. Apesar dessas diretrizes, trabalhos futuros devem se debruçar mais longamente como como é possível operacionalizar isso no Brasil. Para as grandes empresas, sugere-se os Conselhos de Supervisão, ainda que observados os critérios descritos acima. Com base nisso, sugere-se que trabalhos futuros aprofundem nessas formas de regulação. A regulação deve ser entendida em sentido amplo, com esferas diversas de atuação: a regulação via dispositivos jurídicos pode posicionar normas gerais e específicas, indicar mecanismos de avaliação de risco e instrumentos de regulação setoriais ou multissetoriais. Artigos futuros devem se debruçar sobre "o como" dessa regulação.

Referências:

BIONI, B. Proteção de dados pessoais: a função e os limites do consentimento. Rio de Janeiro: Forense, 2019.

BIONI, B.; LUCIANO, M. O princípio da precaução na regulação de Inteligência Artificial: seriam as leis de proteção de dados o seu portal de entrada? In: FRAZÃO, A.; MULHOLLAND, C. (coords.) Inteligência artificial e direito: ética, regulação e responsabilidade. [livro eletrônico] 2ª ed, rev, São Paulo: Thomson Reuters Brasil, 2020.

BRASIL. Lei nº 13.709, de agosto de 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm Acesso em: 03 maio 2022.

BROWN, S.; DAVIDOVIC, J.; HASAN, A. The algorithm audit: Scoring the algorithms that score us. Big Data & Society, v. 8, n. 1, p. 205395172098386, jan. 2021. Disponível em: https://doi.org/10.1177/2053951720983865. Acesso em: 3 maio 2022.

BURRELL, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, v. 3, n. 1, p. 205395171562251, 5 jan. 2016. Disponível em: https://doi.org/10.1177/2053951715622512. Acesso em: 3 maio 2022.

CAMPISI, N. From Inherent Racial Bias to Inocrrect Data - The Problema With Current Credit Scoring Models. Forbes, 26 fev 2021. Disponível em: https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/ Acesso em: 03 maio 2022.

DONEDA, D. Da privacidade à proteção de dados pessoais: elementos da formação da Lei Geral de Proteção de Dados. [livro eletrônico] 2ª ed, São Paulo: Thomson Reuters Brasil, 2020.

FLORIDI, L; COWLS, J. A Unified Framework of Five Principles for AI in Society. Issue 1, 23 jun. 2019. Disponível em: https://doi.org/10.1162/99608f92.8cd550d1. Acesso em: 3 maio 2022.

FLORIDI, L.; TADDEO, M. What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, v. 374, n. 2083, p. 20160360, 28 dez. 2016. Disponível em: https://doi.org/10.1098/rsta.2016.0360. Acesso em: 3 maio 2022.

FRAZÃO, A. Decisões algorítmicas e direito à explicação. Jota, 24 nov. 2021. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/decisoes-algoritmicas-e-direito-a-explicacao-24112021 Acesso em: 28 abr 2022.

GUTIERREZ, A. É possível confiar em um sistema de Inteligência Artificial? Práticas em torno da melhoria da sua confiança, segurança e evidências de Accountability. In: FRAZÃO, A.; MULHOLLAND, C. (coords.) Inteligência artificial e direito: ética, regulação e responsabilidade. [livro eletrônico] 2ª ed, rev, São Paulo: Thomson Reuters Brasil, 2020.

HARTMANN, I.A. et al. Policy Paper Regulação de Inteligência Artificial no Brasil. FGV Direito Rio, 2020. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/30078/PolicyPaperIAeGoverno.pdf?sequence=1&isAllowed=y Acesso em 03 mai. 2022.

LEMOS, R. O Oversight Board do Facebook. ITS Rio, 12 mai 2020. Disponível em: https://itsrio.org/pt/artigos/o-oversight-board-do-facebook/ Acesso em: 03 mai 2022.

LESSIG, L. Code version 2.0. New York: Basic Books, 2006.

MÖKANDER, J. et al. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Science and Engineering Ethics, v. 27, n. 4, 6 jul. 2021. Disponível em: https://doi.org/10.1007/s11948-021-00319-4. Acesso em: 3 maio 2022.

MÖKANDER, J.; FLORIDI, L. Ethics-Based Auditing to Develop Trustworthy AI. Minds and Machines, v. 31, n. 2, p. 323-327, 19 fev. 2021. Disponível em: https://doi.org/10.1007/s11023-021-09557-8. Acesso em: 3 maio 2022.

MORLEY, J et al. Operationalising AI ethics: barriers, enablers and next steps. AI & SOCIETY, 15 nov. 2021. Disponível em: https://doi.org/10.1007/s00146-021-01308-8. Acesso em: 3 maio 2022.

PASQUALE, Frank. Black box society: The secret algorithms that control money and information. [S. l.: s. n.], 2016. 311 p. ISBN 9780674970847.

RAJI, I.D. et al. Closing the AI accountability gap. In: FAT* '20: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, Barcelona Spain. FAT* '20: Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM, 2020. Disponível em: https://doi.org/10.1145/3351095.3372873. Acesso em: 3 maio 2022.

SCHERER, M.U. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. Harvard Jounal of Law & Technlogy, v. 29, n. 2, Spring 2016.

VEGA, I.S. Inteligência artificial e tomada de decisão - a necessidade de agentes externos. In: FRAZÃO, A.; MULHOLLAND, C. (coords.) Inteligência artificial e direito: ética, regulação e responsabilidade. [livro eletrônico] 2ª ed, rev, São Paulo: Thomson Reuters Brasil, 2020.

WORLD ECONOMIC FORUM. AI Governance A holistic Approach to Implement Ethics into AI. Switzerland, 2019. Disponível em: https://weforum.my.salesforce.com/sfc/p/#b0000000GycE/a/0X000000cPl1/i.8ZWL2HIR_kAnvckyqVA.nVVgrWIS4LCM1ueGy.gBc Acesso em: 02 mai. 2022.