Pontifícia Universidade Católica de São Paulo
PUC-SP

Letícia De Conti Serec

Transparência na Inteligência Artificial: mecanismos de enfrentamento nos projetos de regulação da Europa e do Brasil

Mestrado em Tecnologias da Inteligência e Design Digital

São Paulo
2023

Letícia De Conti Serec

Transparência na Inteligência Artificial: mecanismos de enfrentamento nos projetos de regulação da Europa e do Brasil

Dissertação apresentada à banca examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de Mestre em Tecnologias da Inteligência e Design Digital, sob a orientação da Profa. Dra. Dora Kaufman.

São Paulo

2023

Gerenciador de ficha catalográfica:

http://biblio2.pucsp.br/ficha/?_ga=2.154384056.1415767632.1628681585-1429258994.1628681585

Obs. Após inserir a ficha deletar este texto

Banca Examinadora

_____

_____

_____

_____

_____

À comunidade da Pontifícia Universidade
Católica de São Paulo pelo apoio permanente.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – nº 88887.662229/2022-00.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – nº 88887.662229/2022-00.

AGRADECIMENTOS

Expresso meus sinceros agradecimentos a todos que contribuíram para a realização desta dissertação. Em primeiro lugar, agradeço à minha orientadora, Dora Kaufman, por sua sabedoria, orientação e dedicação, que foram inestimáveis para o meu crescimento acadêmico e pessoal. Sou grata por todo o conhecimento transmitido e pelo estímulo constante em busca da excelência.

Agradeço aos meus pais, Maria Serec e Fernando Serec, e aos meus irmãos, Fernando e Beatriz, pelo apoio incondicional e pela força que sempre me deram nos momentos de desafios. A presença e o encorajamento de vocês foram fundamentais para a minha jornada.

Agradeço ao meu namorado, Gabriel, cujo amor, apoio e compreensão foram essenciais durante essa fase da minha vida. Sua presença trouxe leveza e tranquilidade, ajudando-me a enfrentar os desafios com mais serenidade.

Um agradecimento especial aos brilhantes professores do programa de mestrado da Tecnologias da Inteligência e Design Digital ("TIDD") e meus queridos colegas. Por fim, quero agradecer aos meus amigos; sua presença, apoio e palavras de incentivo foram fundamentais para que eu pudesse superar obstáculos e seguir adiante.

Agradeço a cada um de vocês por compartilharem essa jornada comigo. A todos que mencionei e a tantos outros que de alguma forma estiveram presentes em minha caminhada. Sem o apoio, o incentivo e o amor de vocês, esta dissertação não teria sido possível.

Sure, we humans can’t always truly explain our thought processes either—but we find ways to intuitively trust and gauge people. Will that also be possible with machines that think and make decisions differently from the way a human would? (KNIGHT, 2017)

RESUMO

O objetivo da presente dissertação é investigar nos documentos regulatórios sobre inteligência artificial (IA) no Brasil e na União Europeia se os dispositivos presentes relacionados aos princípios da transparência e da explicabilidade consideram características técnicas e sistêmicas inerentes a sistemas de IA baseados em aprendizado de máquina, mais especificamente em redes neurais profundas. Para isso, são estabelecidos marcos teóricos. O primeiro aborda a definição dos conceitos de IA, com ênfase no aprendizado de máquina e no aprendizado profundo, essenciais para uma avaliação adequada do impacto das regulamentações. O segundo trata da transparência como ferramenta para proteger o direito à igualdade, considerando a autonomia privada e o direito à informação. Por fim, o terceiro marco teórico avalia os documentos regulatórios no Brasil e na União Europeia, com enfoque nos princípios da transparência e da explicabilidade.

Palavras-chave: Regulamentação; Inteligência Artificial; Opacidade; Explicabilidade; Transparência.

ABSTRACT

The aim of this work is to investigate the regulatory documents on artificial intelligence (AI) in Brazil and the European Union (EU) to determine whether the provisions related to transparency and explainability principles consider the technical and systemic characteristics inherent in machine learning-based AI systems, specifically deep neural networks. To achieve this, three theoretical frameworks are established. The first framework addresses the definition of AI concepts, with a focus on machine learning and deep learning, which are essential for a proper assessment of the impact of regulations. The second framework examines transparency as a tool to protect the right to equality, considering private autonomy and the right to information. Finally, the third theoretical framework evaluates the regulatory documents in Brazil and the EU, with a specific focus on transparency and explainability principles.

Keywords: Regulation; Artificial Intelligence; Opacity; Explainability; Transparency.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

art.	Artigo
Anvisa	Agência Nacional de Vigilância Sanitária
CJSUBIA	Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil
CNPD	Conselho Nacional da Autoridade de Proteção de Dados
IA	Inteligência Artificial
LGPD	Lei Geral de Proteção de Dados
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PL	projeto de lei
UE	União Europeia

SUMÁRIO

1. INTRODUÇÃO

1.1. Apresentação e delimitação do objeto

A Inteligência Artificial (IA) é considerada a "tecnologia de propósito geral" do século XXI, de forma que tende a ter um impacto cada vez maior na vida em sociedade. As tecnologias não possuem as mesmas características; enquanto algumas adicionam valor incremental à sociedade, outras são consideradas disruptivas. Kaufman (2022) argumenta que as "tecnologias de propósito geral", como foram a máquina a vapor, a eletricidade e o computador, são aquelas que efetivamente têm a capacidade de moldar uma era e reorientar inovações nos mais diferentes setores.

A IA faz parte da vida diária, media nossa comunicação e sociabilidade, e está na essência dos modelos de negócio de plataformas e aplicações como Waze, Google research, Netflix e Spotify – nos primeiros dois casos, para a definição do melhor itinerário e a precisão de pesquisa, respectivamente; nos últimos dois casos, para recomendações de conteúdo ou público.

> A Siri, da Apple, e a Alexa, da Amazon são assistentes pessoais digitas inteligentes que nos ajudam a localizar informações úteis com acesso por meio de voz. Os algoritmos de inteligência artificial mediam as interações nas redes sociais, como a seleção do que será publicado no *feed* de notícias do Facebook. Eles estão igualmente presentes nos diagnósticos médicos, nos sistemas de vigilância na prevenção a fraudes, nas análises de crédito, nas contratações de RH, na gestão de investimentos, na indústria 4.0, no atendimento automatizado (*chatbot*); bem como nas estratégias de marketing, nas pesquisas, na tradução de idiomas, no jornalismo automatizado, nos carros autônomos, no comércio físico e virtual, nos canteiros de obras, nas perfurações de petróleo, na previsão de epidemias. Estamos na era da personalização, viabilizada pela extração das informações contidas nos dados que geramos em nossas movimentações *online* (KAUFMAN, 2022, p. 25).

Além de a IA estar presente no cotidiano da sociedade, há uma tendência de que a lógica de sua aplicação se torne dominante na criação de riqueza, atribuindo-se um valor econômico sem precedentes a esse tipo de tecnologia. Dessa forma, a sociedade atual está migrando para uma economia baseada em dados, haja vista que estes serão a matéria-prima estratégica de um novo modelo econômico (KAUFMAN, 2022, p. 22).

Atento a essa realidade, o Congresso Nacional, no dia 10 de fevereiro de 2022, promulgou a Emenda Constitucional 115 (BRASIL, 2022), que, ao acrescer o inciso LXXIX ao art. 5º da Constituição Federal, incluiu a proteção de dados pessoais no rol de direitos e garantias fundamentais dos indivíduos – nos termos do referido inciso, "[...] é assegurado, nos termos da lei, o direito à proteção dos dados pessoais, inclusive nos meios digitais".

O texto também fixou a competência privativa da União para legislar sobre proteção e tratamento de dados pessoais. Isso significa que a proteção de dados pessoais se tornou uma prioridade na legislação brasileira, o que traz mais segurança e transparência ao uso de informações de indivíduos e favorece os investimentos em tecnologia no País.

Em sua obra *Inteligência artificial – como os robôs estão mudando o mundo*, Kai-Fu Lee (2019) considera que a IA está em sua era de implementação, etapa que envolve três elementos cruciais para o sucesso de um algoritmo: *big data*, poder de computação e engenheiros de IA habilidosos. Nesse contexto, Kai-Fu Lee entende que dados são o aspecto central para esse sucesso: quanto maior o número de exemplos a que uma rede estiver exposta, mais precisamente ela poderá identificar padrões no mundo real.

> Os pesquisadores de IA de elite ainda têm um potencial de levar o campo a um nível superior, mas esses avanços ocorrem uma vez a cada várias décadas. Enquanto esperamos pelo próximo avanço, a crescente disponibilidade de dados será a força motriz por trás da modificação, causada pelo aprendizado profundo, de inúmeras indústrias ao reder do mundo (LEE, 2019, p. 27).

Antes vista como algo restrito a laboratórios de pesquisa acadêmica e filmes de ficção científica, a IA agora está no centro do discurso público. Isso se deve aos avanços teóricos recentes na área, que finalmente estão produzindo repercussões práticas que já estão mudando significativamente a vida de toda a humanidade (LEE, 2019).

> Hoje tudo isso mudou. Artigos sobre as mais recentes inovações de IA cobrem as páginas dos jornais. Conferências de negócios sobre como alavancar a IA para aumentar os lucros estão acontecendo quase todos a dias. E os governos do mundo todo estão lançando seus próprios planos nacionais para explorar a tecnologia. De repente, a IA está no centro do discurso público, e por boas razões. [...] IA já alimenta muitos de nossos aplicativos e sites favoritos, e nos próximos anos dirigirá nossos carros, gerenciará nossos portfólios, fabricará muito do que compramos e potencialmente tirará nossos empregos (LEE, 2019, p. 10).

Ainda que não exista uma legislação sobre IA vigente no Brasil, em maio de 2023 o presidente do Senado, Rodrigo Pacheco, apresentou o Projeto de Lei (PL) 2.338, que dispõe sobre a regulação do desenvolvimento e uso da IA no país, projeto subsidiado por relatório substitutivo apresentado por uma comissão de juristas.

No entanto, o processo brasileiro se iniciou em 2020, com o lançamento pelo Ministério da Ciência, Tecnologia e Inovações (“MCTI”) da “Estratégia Brasileira para a Transformação Digital”, que possuía, como um de seus objetivos, a criação de um marco legal para IA até 2022.

Em 2019, o PL 5.051/2019 foi apresentado ao Senado Federal pelo senador Styvenson Valentim. O referido projeto teve como objetivo estabelecer os princípios para o uso da IA no

Brasil. Posteriormente, foi proposto à Câmara dos Deputados, pelo deputado Eduardo Bismarck, o PL 21/2020, aprovado no dia 29 de setembro de 2021 com 413 votos a favor e 15 contra. O PL estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da IA no Brasil. Seguindo o processo legislativo ordinário, ele foi enviado para apreciação do Senado Federal. Na sequência, foi apresentado ao Senado Federal o PL 872/2021, de autoria do senador Veneziano Vital do Rêgo, que dispõe sobre marcos éticos e diretrizes para o desenvolvimento e uso da IA no Brasil.

Em 2022, diante da complexidade técnica do tema, o Presidente do Senado Federal, Rodrigo Pacheco, por meio do Ato nº 4, instituiu uma comissão de juristas ("CJSUBIA") responsável por analisar os três PLs existentes sobre o tema e elaborar um substitutivo a ser analisado pelo Senado (PLs 5.051/2019, 21/2020 e 872/2021). A comissão, formada por 18[1] juristas, tinha como integrantes nomes como Miriam Wimmer, atualmente Diretora da Autoridade Nacional de Proteção de Dados; Thiago Sombra, professor de Direito da Universidade de Brasília e autor do livro *Fundamentos da regulação da privacidade e proteção de dados pessoais* (2019); e Bruno Bioni, pesquisador na área de Direito e Tecnologia e membro do Conselho Nacional da Autoridade de Proteção de Dados – CNPD.

A comissão promoveu reuniões, seminários e audiências públicas divididas por eixos temáticos. Além disso, foram promovidos doze painéis temáticos pela comissão, que receberam 102 manifestações de entidades da sociedade civil organizada, consolidadas em um relatório pelos juristas (SENADO FEDERAL, 2022). A comissão entregou, em dezembro de 2022, seu relatório final ao Presidente do Senado, após 240 dias de trabalho. O relatório substitutivo apresentou 40 artigos que foram submetidos ao Senado como PL 2.338/2023.

No contexto internacional, a Comissão Europeia e os países-membros da comunidade europeia estão engajados na regulação da IA desde 2017, quando foi publicado um relatório do Parlamento Europeu abordando a responsabilidade civil, a ética, o impacto no mercado de trabalho e a privacidade relacionados à robótica e à IA.

Desde 2018, a Comissão Europeia lançou diversas iniciativas com o objetivo de estabelecer um arcabouço regulatório para a inteligência artificial. Vale destacar o observatório criado pela Comissão, denominado *AI Watch*, para monitorar e analisar a aplicação da IA no continente.

---

[1] Ricardo Villas Bôas Cueva (Presidente); Laura Schertel Ferreira Mendes (relatora); Ana de Oliveira Frazão; Bruno Ricardo Bioni; Danilo Cesar Maganhoto Doneda (*in memoriam*); Fabrício de Mota Alves; Miriam Wimmer; Wederson Advincula Siqueira; Claudia Lima Marques; Juliano Souza de Albuquerque Maranhão; Thiago Luís Santos Sombra; Georges Abboud; Frederico Quadros D'Almeida; Victor Marcel Pinheiro; Estela Aranha; Clara Iglesias Keller; Mariana Giorgetti Valente; Filipe José Medon Affonso.

> Em 2019, o Grupo de Especialistas de Alto Nível em Inteligência Artificial da Comissão Europeia publicou diretrizes de ética para uma IA confiável. A partir destas diretrizes, em fevereiro de 2020, a Comissão publicou um documento propondo uma estrutura base para as próximas fases da ação legislativa ("white paper"), que foi seguido de um processo de consulta pública que envolveu partes interessadas de vários setores, visando influenciar a redação da Proposta de Regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de inteligência artificial ("AI Act") (SALOMÃO et al., 2021, p. 15).

Em fevereiro de 2020, a Comissão Europeia lançou a chamada "Estratégia de IA" para acelerar o desenvolvimento e a implementação da IA na União Europeia (UE), bem como garantir a sua segurança e parâmetros éticos (COMISSÃO EUROPEIA, 2020). A Estratégia de IA inclui iniciativas como a criação de um ecossistema europeu de dados, o fomento a investimentos em pesquisa e inovação em IA e o estabelecimento de parcerias internacionais.

Em abril de 2021, a Comissão Europeia colocou em consulta pública uma proposta de regulação de IA ("AI Act") visando estabelecer regras claras para o desenvolvimento, a colocação no mercado e o uso de sistemas de IA na UE (COMISSÃO EUROPEIA, 2021). O objetivo da chamada "Proposta de Regulamento de Inteligência Artificial" é garantir que a IA seja usada de maneira segura, justa e ética, evitando-se riscos para a segurança e os direitos dos indivíduos.

O AI Act foi aprovado pelo Parlamento Europeu por uma maioria durante uma votação em 14 de junho de 2023, com 499 votos a favor, 28 votos contra e 93 abstenções. A próxima etapa do processo envolverá negociações detalhadas com os membros do Parlamento Europeu para resolver questões específicas e encontrar um consenso sobre os detalhes da legislação e garantir sua implementação adequada em toda a União Europeia.

Diante da análise documental e dos debates prévios ocorridos desde 2018 na UE, a Comissão Europeia adotou uma abordagem mais ampla e objetiva ao estabelecer obrigações para as aplicações de IA na Proposta de Regulamento de Inteligência Artificial, levando em conta as observações acerca do uso dessa tecnologia no continente. É válido ressaltar que o processo implementado pela UE ocorreu de maneira gradual e aberta a contribuições multidisciplinares.

Na contramão do processo de regulamentação da IA na UE, o Poder Legislativo brasileiro, ao propor os três PLs sobre o tema (PLs 5.051/2019, 21/2020 e 872/2021), não havia realizado um diagnóstico acerca do uso da IA no País, tampouco identificado quais são os problemas éticos reais associados ao tema. Ademais, não houve consideração da necessidade de estabelecer arranjos institucionais que garantam a aplicação dos PLs propostos.

No ordenamento brasileiro, há diferentes tipos de conselhos instituídos, tais como o Conselho Nacional de Justiça, o Conselho Nacional de Saúde, o Conselho Nacional de Educação, o Conselho Nacional do Ministério Público, o Conselho Nacional de Combate à Pirataria etc. Em comum, é possível notar que eles são essencialmente órgãos de produção de política pública, não figurando como órgãos do Poder Executivo central. Ademais, os conselhos geralmente reúnem atores de diferentes setores, buscando conferir à instituição à qual se vinculam algum grau de representação democrática e expertise setorial (DATA PRIVACY BRASIL, 2021). Existem preocupações jurídicas com os riscos e possíveis danos que a IA pode produzir sobre valores fundamentais de um Estado de Direito, tais como a igualdade e a dignidade da pessoa humana.

Considerando que a criação de uma regulação da IA é uma questão de grande complexidade técnica e que possui impacto significativo em diversas áreas de atuação do poder público e setores econômicos, é crucial que a eficácia desses projetos de lei seja submetida a um escrutínio público rigoroso. Tal processo é fundamental para assegurar um amplo debate para se alcançar uma regulação que seja capaz de proteger valores fundamentais de um Estado Democrático de Direito, tais como o direito à igualdade e o exercício da autonomia privada.

Esta dissertação tem como objetivo lançar luz sobre a existência de características intrínsecas aos sistemas de IA, particularmente as técnicas de aprendizado de máquina – relacionadas à opacidade –, que podem ter impacto significativo na concretização dos valores fundamentais de um Estado de Direito, salvaguardados pelo ordenamento jurídico. O objetivo, dessa forma, é investigar se as propostas de regulamentação da IA – PL 2.338/2023 e AI Act – contemplam devidamente essas características intrínsecas nos dispositivos legais que tratam de transparência e explicabilidade.

Tem-se como hipótese que a falta de consideração das características técnicas intrínsecas dos sistemas atuais de IA na redação dos projetos de lei em análise pode levar a lacunas na regulamentação da IA, principalmente no que diz respeito à concretização da transparência e da explicabilidade.

1.2. Justificativa

A justificativa para a escolha desse objeto de pesquisa parte de três aspectos primordiais: sua relevância, atualidade e a existência de lacunas na análise de projetos voltados para a regulamentação da IA. A relevância de analisar esse tema reside em duas questões fundamentais.

Em primeiro lugar, a velocidade acelerada do avanço da IA e as complexidades e consequências jurídicas que esse fenômeno acarreta. Outro aspecto relevante do tema está intrinsecamente ligado ao estudo da transparência e da explicabilidade dos sistemas de IA. Esses conceitos não apenas permeiam a maioria dos Frameworks e Guias de boas práticas, que propõem medidas para uma IA justa e confiável, mas também representam um tópico de extrema importância capaz de afetar aspectos jurídicos fundamentais, como a igualdade e a dignidade da pessoa humana. Além de ser crucial que a interação entre humanos e máquinas seja o mais clara possível para que as pessoas possam compreender os atributos que influenciaram o sistema a chegar a uma determinada decisão, é também fundamental para a imputação de responsabilidade em casos de sistemas que causaram danos.

À época do início desta pesquisa, o objeto único da dissertação poderia estar relacionado ao desenvolvimento de sistemas de IA, com reflexões sobre a Lei Geral de Proteção de Dados, no Brasil, e o *General Personal Data Protection Law*, na UE. No entanto, ao longo da pesquisa, tornou-se evidente a relevância do debate público sobre as iniciativas e os projetos de regulamentação da IA no Brasil e na Europa. O tema passou a ser amplamente discutido por políticos, juristas e estudiosos, culminando na publicação do AI Act e do PL 2.338/2023.

Dessa forma, a escolha de analisar tanto a proposta brasileira quanto a da UE, apesar da diferença radical nos processos e na forma como as propostas normativas são submetidas ao debate público, advém do fato de se acreditar no poder que a experiência europeia traz consigo.

A UE, como pioneira na iniciativa de regulamentar a IA e com uma tradição jurídica consolidada na elaboração de arcabouços legais robustos e bem estruturados, oferece uma perspectiva interessante para compreender os desafios e as soluções propostas. Além disso, sua influência no campo político global é indiscutível, e sua abordagem está sendo atentamente observada por países e regiões em todo o mundo.

Ademais, há estudos específicos que se dedicam a analisar a utilização da IA no Poder Judiciário e de que forma as decisões jurídicas estão sendo automatizadas por meio do uso de sistemas de IA nos tribunais em todo o território brasileiro. Também se destaca uma terceira categoria de dissertação, que investiga os impactos da IA em diversas áreas do Direito, como o mercado de capitais, o direito de família, entre outras. No entanto, o objetivo central desta dissertação não é a análise do uso da IA no campo de atuação do Direito, mas, sim, a compilação e análise dos documentos legais que regulamentam sistemas de IA.

Durante a revisão da literatura, também foram encontrados trabalhos que trataram de forma fragmentada alguns projetos de lei relacionados à regulamentação da IA no Brasil, assim como iniciativas regulatórias na UE. No entanto, uma análise detalhada dos principais

documentos legais que regulamentam a IA em ambos os países ainda não foi realizada de forma abrangente e comparativa.

A partir dessas constatações, pretende-se verificar se os dispositivos presentes no AI Act e no PL 2.338/2023, relacionados aos princípios da transparência e da explicabilidade, levam em consideração características intrínsecas técnicas e sistêmicas encontradas em sistemas de IA baseados em aprendizado de máquina, mais especificamente em redes neurais profundas.

1.3. Metodologia

Para a elaboração da presente dissertação foi adotada a abordagem exploratória, de natureza qualitativa, com ênfase na análise de documentos e estudos relativos à IA e ao Direito. A pesquisa bibliográfica e documental incluiu a análise da literatura acadêmica e relatórios de pesquisa sobre características técnicas inerentes da IA, bem como a análise documental de relatórios e diretrizes no campo do Direito. O recorte geográfico da dissertação foi centralizado na Europa e no Brasil, realizado em inglês e português. Os procedimentos metodológicos se desdobraram estrategicamente nos três caminhos descritos a seguir.

O primeiro marco teórico necessário ao desenvolvimento da pesquisa consistiu na conceituação técnica de IA. Foi realizada uma pesquisa bibliográfica abordando os conceitos-chave de IA, subcampo aprendizado de máquina, técnica de aprendizado de máquina baseado em redes neurais, bem como as características técnicas inerentes aos modelos de aprendizado de máquina e os possíveis impactos associados ao seu desenvolvimento. Esse aspecto foi crucial para o desenvolvimento desta dissertação, pois sem uma definição clara de IA e suas características técnicas inerentes não é possível avaliar adequadamente o impacto que as regulamentações têm na sociedade. Esta análise contemplou autores como Mahesh (2020), Mitchell (1997), Bochie et al. (2020), Alpaydin (2016) e Kaufman (2018; 2019; 2022).

No segundo marco teórico, foi realizada uma avaliação da transparência como um recurso fundamental para proteger o direito à igualdade. Nesse contexto, foram apresentados conceitos relacionados ao exercício da autonomia privada, bem como ao direito à informação, conforme previsto no art. 5º, XIV, da Constituição. A transparência também foi abordada na perspectiva da atividade econômica, levando em consideração o princípio da igualdade estabelecido no *caput* do art. 5º, da Constituição.

E por fim, foi realizada uma análise documental com aspectos descritivos e valorativos, e uma análise comparativa, que também combina descrições e valorações, na avaliação dos

documentos regulatórios no Brasil e na UE. Para isso, foi analisado o PL 2.338/2023 derivada do Relatório Final elaborado pela CJSUBIA. Na UE, foram analisados o AI Act da Comissão Europeia, de 21 de abril de 2021, e a minuta adjacente com alterações propostas pela consulta pública, pelos países-membros e pelo Parlamento Europeu. Em ambos os projetos foi dado enfoque específico aos dispositivos relacionados aos princípios da transparência e da explicabilidade.

1.4. Estrutura e organização da dissertação

Para o desenvolvimento do tema anteriormente exposto, a dissertação está dividida em quatro capítulos, cada qual dividido em subitens.

O capítulo dois expõe, em linhas gerais, os conceitos-chave da IA, destacando a falta de consenso em sua definição. Além disso, apresenta as definições adotadas pelos projetos regulatórios do AI Act e do PL 2.338/2023. Em seguida, explora-se o aprendizado de máquina, com ênfase em redes neurais profundas, abordando sua importância atual. Por fim, são discutidas as complexas características inerentes ao aprendizado de máquina, como a opacidade do desenvolvimento dos sistemas de IA, fato que tem implicações jurídicas relevantes e demanda atenção específica.

O capítulo três se dedica a explorar a transparência como ferramenta fundamental para viabilizar a disseminação do conhecimento. Através da transparência, é possível promover a divulgação de informações relevantes e proporcionar a concretização de direitos fundamentais, como o direito à igualdade, o exercício da autonomia privada e o direito de escolha. Ao analisar a importância da transparência no contexto da IA, esse capítulo destaca como a divulgação clara e acessível das informações sobre sistemas de IA pode empoderar os indivíduos e possibilitar maior confiabilidade em sistemas de IA.

No capítulo quatro, foi conduzida uma análise minuciosa dos projetos legislativos já mencionados – PL 2.338/2023 e AI Act –, buscando identificar os dispositivos legais que abordam os conceitos de “transparência” e “explicabilidade” ou que tratam dessas temáticas. A fim de avaliar as soluções propostas pelos projetos, foram examinados os dispositivos legais que abordam essas questões, permitindo uma análise crítica sobre como esses projetos estão lidando com a transparência e a explicabilidade nos sistemas de IA.

No último capítulo, houve uma comparação entre os artigos presentes no PL 2.338/2023 e no AI Act que abordam as medidas de governança relacionadas à transparência no desenvolvimento, comercialização e pós-comercialização de sistemas de IA. Além disso, foram

analisados os direitos dos titulares em relação à explicação dos sistemas. Nessa análise comparativa, buscou-se avaliar se os projetos regulatórios contemplaram adequadamente a existência de características intrínsecas aos sistemas de IA de aprendizado de máquina – relacionadas à opacidade – e se propõem soluções efetivas para lidar com os desafios relacionados à efetivação da transparência e da explicabilidade dos sistemas de IA.

2. CONCEITO DE INTELIGÊNCIA ARTIFICIAL

2.1. Conceito de IA

A história da IA teve grande influência de John McCarthy. Após concluir seu doutorado em 1951, ele partiu para Stanford e, posteriormente, para o *Dartmouth College*, que seria o berço oficial da IA. Lá, convenceu Marvin Minsky a se juntar a ele na organização de um seminário de dois meses no verão de 1956. O objetivo era reunir pesquisadores norte-americanos interessados na teoria de redes neurais e no estudo da inteligência. O termo "inteligência artificial" foi usado pela primeira vez no título desse evento: "Dartmouth Summer Research Project on Artificial Intelligence" (RUSSEL; NORVIG, 2009).

A proposta do seminário de Dartmouth enfatizou a necessidade de a IA ser estabelecida como um campo distinto, o que se deve ao fato de ela buscar reproduzir habilidades humanas, como criatividade e uso da linguagem, que não são abordadas por outros campos de conhecimento (RUSSEL; NORVIG, 2009).

A premissa do seminário de Dartmouth era que "todos os aspectos da aprendizagem ou qualquer outra característica da inteligência podem, em princípio, ser descritos tão precisamente de modo que uma máquina pode ser construída para simulá-la" (KAUFMAN, 2019). E o objetivo era descobrir como fazer com que as máquinas usassem abstrações de linguagem e resolvessem problemas do domínio humano. Embora o seminário não tenha alcançado seus objetivos, ele foi importante porque criou o campo de conhecimento da IA e colocou em contato alguns dos principais pesquisadores da época.

Existem distintas definições para IA. Russell e Norvig (2009) apresentam oito diferentes abordagens, sendo agrupadas em dois tipos: as que se relacionam com processos mentais e raciocínio, e as que se relacionam com comportamento.

Figura 1 – Definições distintas para IA

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al*., 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: RUSSEL; NORVIG, 2009.

O conceito desenvolvido propriamente por Russel e Norvig (2009, p. 1196) define IA como "agentes inteligentes capazes de perceber seu meio ambiente e de realizar ações com a expectativa de selecionar uma ação que maximize seu desempenho".

Já Taddeo e Floridi (2018) definem a IA como um sistema interativo com capacidade de operar de forma autônoma e apto à autoaprendizagem. A IA pode, então, ser definida em termos de recursos de modelos computacionais que se baseiam em uma arquitetura tecnológica.

Agrawal, Gans e Goldfarb (2019), por sua vez, definem IA como uma tecnologia de predição, em que as previsões são utilizadas como entradas para a tomada de decisões. Segundo esses autores, as ferramentas de IA são úteis para tornar as máquinas de previsão eficazes.

Para McCarthy (2007, p. 2), IA "é a ciência e engenharia de criar máquinas inteligentes, especialmente programas de computador inteligentes. Está relacionado à tarefa semelhante de usar computadores para entender a inteligência humana, mas a IA não precisa se limitar a métodos biologicamente observáveis".

Segundo Kaufman:

> A inteligência artificial refere-se a um campo de conhecimento associado à linguagem e à inteligência, ao raciocínio, à aprendizagem e à resolução de problemas. A IA propicia a simbiose entre o humano e a máquina ao acoplar sistemas inteligentes artificiais ao corpo humano (prótese cerebral, braço biônico, células artificiais, joelho inteligente e similares), e a interação entre homem e máquina como duas "espécies" distintas conectadas (homem-aplicativos, homem-algoritmos de IA) (2019, p. 19).

Cada definição oferece uma perspectiva única sobre o que é a IA e como ela pode ser aplicada. Mas a principal distinção é que cada definição enfatiza aspectos específicos da tecnologia, tendo sempre como perspectiva a interação fundamental entre humano e máquina.

Com base nas posições doutrinárias citadas anteriormente e mais bem desenvolvidas no item 2.2, pode-se compreender que a IA é o uso de algoritmos complexos e técnicas estatísticas, que permitem que as máquinas aprendam e melhorem a sua performance ao longo do tempo. Isso significa que as máquinas são capazes de aprender a partir de dados coletados, de forma a tomar decisões mais precisas e eficientes (por meio de técnicas de aprendizado de máquina, como o aprendizado profundo, que será abordado de forma mais detalhada no item 2.2).

No que se refere aos projetos em análise, como se observa nos quadros a seguir, ambas as abordagens para a regulamentação – PL 2.338/2023 e AI Act – se fundamentam na definição de IA como *fattispecie*[2] para a incidência das normas que comporiam o complexo regulatório das atividades de sistema de IA. Ou seja, a partir da proposição de uma definição, que abarque uma série de atividades relacionadas ao que se considera IA, essa definição servirá como base ou pressuposto fático para a aplicação de normas ou um conjunto de normas em relação à IA (BULGARELLI, 1985).

Nesse contexto, serão expostos os conceitos de IA conforme os documentos regulatórios do Brasil e, em seguida, da UE.

Quadro 1 – Conceitos de IA conforme os documentos regulatórios brasileiros

Projeto de Lei	Conceito	Artigo
PL 5.051/2019	Não foi apresentado conceito de IA.	Não aplicável
PL 872/2021	Não foi apresentado conceito de IA.	Não aplicável
PL 21/2020	Sistema de inteligência artificial: o sistema baseado em processo computacional que pode, para um determinado conjunto de objetivos definidos pelo homem, fazer previsões e recomendações ou tomar decisões que influenciam ambientes reais ou virtuais.	Art. 2º, inciso I

[2] No italiano, por exemplo, consagrou-se a expressão *fattispecie*, proposta inicialmente por Emílio Betti. Ele explica que "o termo [*fattispecie*] deriva do latim medieval e significa *facti species*, que, à letra, significa figura do fato. A denominação é preferível à outra, comumente usada, de 'fato jurídico', porque indica tanto o fato propriamente dito, como, conjuntamente, o estado de fato e de direito, em que o fato incide e se enquadra" (BETTI, 2008, p. 20, nota 2).

PL 2.338/2023	Sistema computacional, com graus diferentes de autonomia, desenhado para inferir como atingir um dado conjunto de objetivos, utilizando abordagens baseadas em aprendizagem de máquina e/ou lógica e representação do conhecimento, por meio de dados de entrada provenientes de máquinas ou humanos, com o objetivo de produzir previsões, recomendações ou decisões que possam influenciar o ambiente virtual ou real.	Art. 4º, inciso I

Fonte: Elaborado pela autora.

Os primeiros projetos de lei propostos no Brasil – PL 5.051/2019 e PL 872/2021 – não apresentaram o conceito de IA. Já o PL 21/2020 trouxe em seu art. 2º, I, o conceito de sistema de IA.

A CJSUBIA, responsável por subsidiar a elaboração do PL 2.338/2023, promoveu uma discussão no Painel 1, que abordou o tema "inteligência artificial e regulação: objeto a ser regulado e aspectos sociotécnicos". Esse painel culminou no debate acerca da relevância sobre a definição da IA no contexto regulatório brasileiro. O Relatório Final, no quadro 2, apresentado pela CJSUBIA, compilou treze manifestações de especialistas sobre a questão.

A partir da análise dessas treze manifestações, pode-se afirmar que não houve um consenso inclusive nas posições sobre qual seria a melhor abordagem para conceituar a IA em um documento regulatório. Houve sugestão de revisão[3] dos termos da proposta no PL 21/2020, a fim de garantir uma definição clara e precisa de IA. Ressaltou-se a importância do papel desempenhado pela Comissão na elaboração de uma conceituação sólida[4] e atualizada sobre a matéria, além da sugestão de que a definição de IA apresentada pelo substitutivo deve ser capaz

[3] De acordo com Raquel Lima Saraiva (SENADO FEDERAL, 2022, p. 350): "Os textos dos projetos de lei sobre IA em tramitação **não são exitosos em estabelecer uma definição funcional para fins legais**. A definição proposta no art. 2º do PL nº 21, de 2020, por exemplo, tem relação direta com uma parte do conceito proposto pela OCDE, mas inclui elementos que prejudicam sua precisão e compreensão. É o que ocorre, por exemplo, ao estabelecer como critério a capacidade ampla e pouco objetiva de – abro aspas: '[...] aprender a perceber e a interpretar o ambiente externo, bem como interagir com ele [...]'. Perceber como? De que forma? Interpretar como? De que forma? E também essa interação: de que forma se dá? Isso não está claro, e essa definição do objeto a ser regulado deve ser prontamente revisitada, avaliando-se inclusive os benefícios de incluir uma definição do tipo em primeiro lugar e quais elementos e enfoques devem tomar parte dela em caso afirmativo. Se a gente não sabe que evento a gente está regulando, como pensar mais à frente?".

[4] Segundo Sergio Paulo Gallindo (SENADO FEDERAL, 2022, p. 350): "Então, é importante que a lei dê essa segurança jurídica para todos nós. No entanto, a própria União Europeia vai classificar a IA, no seu anexo 1, como técnica [...]. Acho que precisamos debater um pouco mais, para ter mais assertividade sobre o que essa IA significa efetivamente e talvez fazer salvaguardas para não embotar a evolução da IA".

de abranger as diferentes técnicas da IA[5] e seus variados usos[6], garantindo uma interpretação uniforme e consistente da lei[7].

Os referidos debates resultaram no PL 2.338/2023, que alterou a definição proposta pelo PL 21/2020. A distinção conceitual sobre a IA para o PL 21/2020 e o PL 2.338/2023 está relacionada às definições específicas de sistema de IA presentes em cada um dos projetos de lei. A principal diferença entre as definições está na abrangência e no nível de autonomia dos sistemas de IA; enquanto o PL 21/2020 foca na capacidade do sistema em cumprir objetivos definidos pelo ser humano, o PL 2.338/2023 reconhece a presença de diferentes graus de autonomia e a utilização de abordagens avançadas, como aprendizado de máquina e representação do conhecimento, para atingir objetivos específicos.

O processo de definição da inteligência artificial na Europa foi consistente com o processo de proposição do AI Act. O primeiro conceito de IA foi vislumbrado no Livro Branco sobre IA, que desempenhou um papel importante ao estabelecer uma base inicial para a compreensão e discussão da IA na região.

Em seguida, com base nos debates e nas discussões internas realizadas por especialistas e parlamentares europeus, o AI Act define um sistema de inteligência artificial como um programa informático desenvolvido com uma ou várias técnicas e abordagens enumeradas em seu Anexo I.

Quadro 2 – Conceitos de IA conforme os documentos regulatórios europeus

Documento	Conceito	Artigo
Livro Branco sobre Inteligência Artificial	Os sistemas de inteligência artificial (IA) são sistemas de software (e eventualmente também de hardware) concebidos por seres humanos que, tendo recebido um objetivo complexo, atuam na dimensão física ou digital, percecionando o seu ambiente mediante a aquisição de dados, interpretando os dados estruturados ou não estruturados recolhidos, raciocinando sobre o conhecimento	Não aplicável

[5] O contexto latino-americano é diverso em termos de implementação, e nós detectamos a existência de sistemas automatizados no âmbito público que, ainda que não necessariamente contemplem elementos de aprendizagem de máquina, por exemplo, igualmente apresentam novas camadas de opacidade na tomada de decisões sobre políticas públicas e um amplo potencial de violação a direitos fundamentais que merecem atenção.

[6] De acordo com Diogo Cortiz (SENADO FEDERAL, 2022, p. 348): “E eu, nos meus textos, nas minhas aulas, eu sempre tento levar essa ideia de que a IA não é necessariamente uma tecnologia. Eu entendo a IA como uma área de conhecimento, que tem, inclusive, o seu nascimento paralelo com a própria ciência da computação. Então, quando a gente pensa em discutir os problemas da IA, a gente tem que ter essa ideia de que a gente está falando de uma área muito ampla, que ela vai se materializar com diferentes usos, usando diferentes técnicas”.

[7] Segundo Gabrielle Sarlet (SENADO FEDERAL, 2022, p. 348): “Notadamente, aqui todos os colegas disseram, de forma brilhante, que há diversos tipos de IA e, portanto, em razão dessa grande paleta de tecnologias, as quais nós chamamos de IA, até se reafirma a necessidade, talvez, de um acordo semântico principiológico, já no próprio substitutivo, para que nós possamos entender exatamente o que nós estamos chamando de IA”.

	ou processando as informações resultantes desses dados e decidindo as melhores ações a adotar para atingir o objetivo estabelecido. Os sistemas de IA podem utilizar regras simbólicas ou aprender um modelo numérico, bem como adaptar o seu comportamento mediante uma análise do modo como o ambiente foi afetado pelas suas ações anteriores[8].	
Estratégia de IA da Comissão Europeia	Não foi apresentado conceito de IA.	Não aplicável
Proposta de regulamento do Parlamento Europeu e do Conselho	Sistema de inteligência artificial (sistema de IA): um programa informático desenvolvido com uma ou várias das técnicas e abordagens enumeradas no Anexo I, e capaz de, tendo em vista um determinado conjunto de objetivos definidos por seres humanos, criar resultados, tais como conteúdos, previsões, recomendações ou decisões, que influenciam os ambientes com os quais interage.	Art. 3º, item 1

Fonte: Elaborado pela autora.

Portanto, o Livro Branco sobre IA estabeleceu as bases iniciais, destacando os elementos essenciais da IA, enquanto o AI Act refinou e aprimorou essa definição, levando em consideração as discussões e o amadurecimento do assunto. O AI Act definiu um sistema de IA como um programa informático desenvolvido com uma ou várias técnicas e abordagens enumeradas em seu Anexo I.

A abordagem assumida pelo AI Act, de enumerar outras técnicas e abordagens no Anexo I, é adotada por outros regulamentos brasileiros que tiveram sucesso na sua efetivação. Ela é baseada na criação de listas ou categorias que definem e classificam determinados elementos ou critérios. Um exemplo dessa abordagem é a lista de medicamentos cobertos pelo Sistema Único de Saúde no Brasil, conhecida como Relação Nacional de Medicamentos Essenciais, elaborada pelo Ministério da Saúde. Essa lista estabelece quais medicamentos devem ser disponibilizados de forma gratuita para a população, considerando critérios como eficácia, segurança e custo-benefício. A lista é atualizada periodicamente com base em evidências científicas e necessidades de saúde pública.

Outro exemplo é a lista de drogas ilícitas, regulamentada pela Agência Nacional de Vigilância Sanitária (Anvisa), denominada de “substâncias e medicamentos sujeitos a controle especial”. Essa lista classifica substâncias que possuem potencial de abuso ou risco à saúde

[8] Definição apresentada pelo Grupo de Peritos de Alto Nível da Comissão Europeia utilizada para efeitos do Livro Branco.

pública, como entorpecentes e psicotrópicos, estabelecendo restrições legais para sua produção, venda e uso.

Segundo o *Framework para Comitês de Ética em IA* e sobre *Governança da Inteligência Artificial em Organizações*[9], o AI Act emprega conceituação ampla e abordagem baseada em avaliação de riscos, e não em um conceito restrito de IA (MAHLER, 2021). Dessa forma, procura-se manter agnóstico aos meios e se concentra no impacto das soluções sobre o bem-estar das pessoas (FLORIDI et al., 2022). Com essa estratégia, tenta impedir que a maioria dos sistemas utilizados atualmente para balizar decisões automatizadas seja excluída de seu escopo de incidência, o que poderia ocorrer sob uma definição mais estrita.

Devido à diversidade de definições e à velocidade da inovação, a prerrogativa de arbitrar se algo é ou não IA é um desafio significativo. No Brasil, a CJSUBIA discutiu o tema e propôs um novo conceito de IA para o PL 2.338/2023, como descrito no Quadro 1. A proposta enfatizou a utilização de abordagens baseadas em aprendizado de máquina e lógica e representação do conhecimento para definir a IA. No entanto, é importante ressaltar que a IA inclui várias abordagens e técnicas que vão além das mencionadas no artigo.

Na União Europeia, os documentos regulatórios propõem uma abordagem complementar na definição de "sistema de IA" com uma lista de técnicas e abordagens enumeradas no Anexo I. Essa abordagem tem como objetivo aprimorar a definição da IA ao longo do tempo, acompanhando as mudanças e os avanços tecnológicos.

Uma abordagem possível para definir a IA, que não foi mencionada por nenhum dos projetos em análise, seria considerar a natureza da tarefa a ser executada, em vez da técnica utilizada. Nessa perspectiva, um sistema seria considerado IA quando substitui ações que requerem decisão humana.

No entanto, como se optou por definir a IA do ponto de vista regulatório, a prerrogativa de arbitrar se algo é ou não IA poderia ser atribuída às agências reguladoras setoriais, seguindo práticas comuns usadas em outras áreas reguladas, conforme citado anteriormente com a lista de medicamentos cobertos pelo SUS ou a lista de drogas ilícitas da Anvisa, na medida em que as agências poderiam manter suas listas de sistemas de IA atualizadas conforme necessário, proporcionando uma governança adaptada à evolução tecnológica e aos desafios emergentes.

[9] ZAVAGLIA COELHO, 2023.

2.2. O aprendizado de máquina e a técnica de redes neurais profundas

Neste tópico, serão explorados o subcampo da IA denominado aprendizado de máquina e a técnica de redes neurais profundas, abordando sua importância no cenário atual.

Em seus primórdios, o desafio do campo da IA foi capacitar as máquinas para exercer funções que envolvessem tarefas executadas pelos humanos intuitivamente[10], com relativo grau de subjetividade. Resolver tarefas que normalmente exigem inteligência humana, como raciocínio, compreensão de linguagem natural e tomada de decisões com precisão e eficiência, não era fácil. Porém, ao longo dos anos, várias tentativas foram pensadas para enfrentar esse desafio. Segundo Kaufman:

> Várias tentativas envolvendo linguagens formais, apoiadas em regras de inferência lógica, tiveram êxito limitado, sugerindo a necessidade de os sistemas gerarem seu próprio conhecimento extraindo padrões de dados, ou seja, "aprender" com os dados sem receber instruções explícitas. Esse processo é usualmente denominado "aprendizado de máquina" (*machine learning*), subcampo da IA criado em 1959 e hoje certamente o maior subcampo da IA em número de praticantes (2022, p. 196).

O subcampo aprendizado de máquina foi proposto pelo pesquisador Arthur Lee Samuel. Segundo ele, a aprendizagem de máquina é definida como o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados (MAHESH, 2020, p. 389)[11]. O aprendizado de máquina é usado para ensinar as máquinas a lidar com os dados de forma mais eficiente (MAHESH, 2020).

> Aprendizado de máquina é empregado em uma variedade de tarefas de computação, nas quais programar os algoritmos é difícil ou inviável. Esses modelos analíticos permitem que pesquisadores, cientistas de dados, engenheiros e analistas produzam decisões e resultados confiáveis e replicáveis, e revelem ideias ocultas em relacionamentos históricos e tendências de dado (KAUFMAN, 2018, p. 20).

---

[10] No original: "*Ironically, abstract and formal tasks that are among the most difficult mental undertakings for a human being are among the easiest for a computer. Computer shave long been able to defeat even the best human chess player but only recently have begun matching some of the abilities of average human beings to recognize objects or speech. A person's everyday life requires an immense amount of knowledge about the world. Much of this knowledge is subjective and intuitive, and therefore difficult to articulate in a formal way. Computers need to capture this same knowledge in order to behave in an intelligent way. One of the key challenges in artificial intelligence is how to get this informal knowledge into a computer*" (GOODFELLOW; BENGIO; COURVILLE, 2016, p. 2).

[11] No original: "*According to Arthur Samuel Machine learning is defined as the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel was famous for his checkers playing program. Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning*" (MAHESH, 2020, p. 389).

De acordo com Mitchell (1997)[12], esse subcampo da IA teve como foco responder à questão: como construir programas de computador capazes de melhorar automaticamente com a experiência? Segundo Kaufman (2022), um algoritmo de aprendizado de máquina é capaz de aprender com base em grandes volumes de dados (*big data*).

Mitchell (1997) estabelece que, para se ter um problema de aprendizagem bem definido, deve-se identificar três características: a classe de tarefas, a medida de desempenho a ser melhorada e a fonte da experiência.

> Por exemplo, um programa de computador que aprende a jogar damas pode **melhorar seu desempenho** medido por sua habilidade de vencer **na classe de tarefas** envolvendo jogos de damas, através da **experiência obtida** jogando jogos contra si mesmo (MITCHELL, 1997, p. 3 – grifos nossos)[13].

Os algoritmos tiram vantagem de possíveis padrões e estruturas presentes nos dados do problema para ajustar seus parâmetros e, como consequência, melhorar o seu desempenho na tarefa apresentada, conforme explicam Bochie et al. (2020). O caráter autônomo – atribuído por alguns especialistas, mas não consensual – dessa abordagem fica evidente, visto que os parâmetros do algoritmo são definidos pela interação do próprio algoritmo com o conjunto de dados (BOCHIE et al., 2020, p. 4).

Dessa forma, entende-se que o aprendizado de máquina explora o estudo e a construção de algoritmos que fazem previsões baseadas em dados, modelos elaborados a partir de entradas de amostras, evoluindo a partir do estudo do reconhecimento de padrões e da teoria de aprendizagem computacional na IA (KAUFMAN, 2018). A vantagem dos sistemas de aprendizado é que eles próprios estabelecem os algoritmos, i.e., adaptam-se automaticamente aos requisitos da tarefa (KAUFMAN, 2018).

A técnica de aprendizado de máquina denominada "redes neurais de aprendizado profundo" (*deep learning*) está presente na maior parte das implementações atuais de IA (KAUFMAN, 2019).

---

[12] No original: "*The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. In recent years many successful machine learning applications have been developed, ranging from data-mining programs that learn to detect fraudulent credit card transactions, to information-filtering systems that learn users' reading preferences, to autonomous vehicles that learn to drive on public highways. At the same time, there have been important advances in the theory and algorithms that form the foundations of this field*" (MITCHELL, 1997, p. XV).

[13] No original: "*For example, a computer program that learns to play checkers might improve its performance as measured by its abiliry to win at the class of tasks involving playing checkers games, through experience obtained by playing games against itself. In general, to have a well-defined learning problem, we must identity these three features: the class of tasks, the measure of performance to be improved, and the source of experience*" (MITCHELL, 1997, p. XV).

> Essa relativamente nova técnica de aprendizado de máquina, baseada fortemente em redes neurais de aprendizado profundo (*deep learning neural networks* – DLNN), tem sua inspiração no funcionamento do cérebro biológico. [...] DLNN estabelecem correlações não perceptíveis aos desenvolvedores humanos, cuja tendência é considerar apenas as correlações "mais fortes", embora as correlações "mais fracas", quando agrupadas, possam impactar sensivelmente a acurácia dos modelos (KAUFMAN, 2022, p. 197).

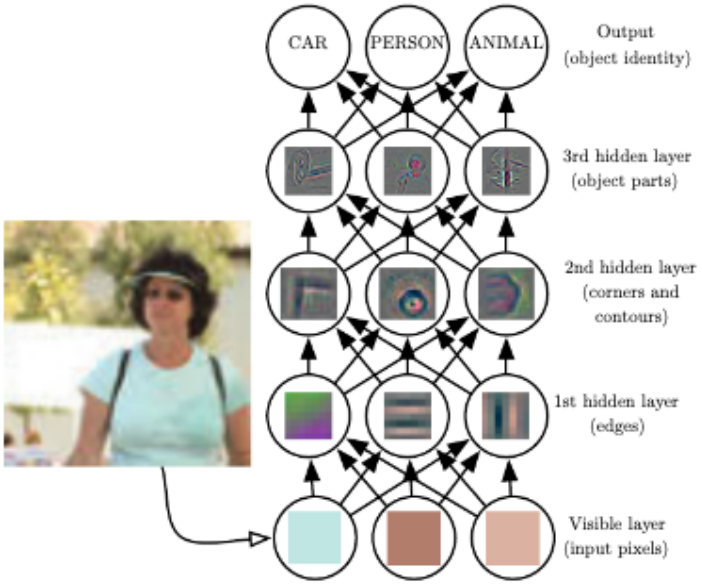
As redes neurais de aprendizado profundo são organizadas em diversas camadas. Conforme explicam Bochie et al. (2020), os neurônios nas redes neurais são normalmente organizados em grupos de unidades de processamento chamados de camadas, que, por sua vez, são organizadas em uma cadeia. A primeira e a última camada são chamadas de camada de entrada e de saída, enquanto as demais são as camadas ocultas (ou internas) (BOCHIE et al., 2020, p. 9).

Os dados de entrada são denominados de *inputs*, que são apresentados a uma camada visível (*visible layer*), assim chamada porque contém as variáveis observáveis, seguida de uma série de camadas ocultas contendo variáveis não observáveis e internas ao próprio modelo (KAUFMAN, 2022). Essa estrutura de camadas (*visible layer* e as demais *hidden layers*) codifica uma função matemática que mapeia conjuntos de valores de entrada (*inputs*) para valores de saída (*outputs*). Segundo Kaufman (2022, p. 197), "Em redes neurais profundas, os parâmetros aprendidos a partir de dados são chamados de *weights* (pesos); após a fase de treinamento (ou aprendizado), esses pesos compõem o algoritmo e passam a ser fixos".

Ocorre um processo de aprendizado em etapas, utilizando redes neurais artificiais, o qual é realizado por meio de técnicas que permitem que o sistema descubra automaticamente as representações necessárias para detectar funções ou classificações a partir dos dados brutos, ao contrário da construção manual de funções ou classificações.

Na imagem a seguir, pode-se observar a estrutura de uma rede neural profunda – a entrada de dados, as diversas camadas de processamento dos dados e, por fim, os dados que são "entregues" pelo sistema.

Figura 2 – Estrutura de uma rede neural profunda



Fonte: GOODFELLOW; BENGIO; COURVILLE, 2016[14].

No aprendizado profundo, o aumento no número de camadas ocultas permite que cada uma delas se especialize em identificar um conjunto de características em particular.

> As camadas mais próximas da camada de entrada são responsáveis por identificar as características mais primitivas, como arestas em uma imagem, enquanto as seguintes combinam essas informações para identificar padrões mais complexos, como animais ou semáforos na mesma imagem. São diversos os tipos de redes neurais existentes (BOCHIE et al., 2020, p. 4).

O processo no qual o algoritmo tem seus parâmetros ajustados com o objetivo de aprender a realizar uma tarefa é denominado de treinamento, que pode ocorrer de diferentes

[14] No original: "*Illustration of a deep learning model. It is difficult for a computer to understand the meaning of raw sensory input data, such as this image represented as a collection of pixel values. The function mapping from a set of pixels to an object identity is very complicated. Learning or evaluating this mapping seems insurmountable if tackled directly. Deep learning resolves this difficulty by breaking the desired complicated mapping into aseries of nested simple mappings, each described by a different layer of the model. The input is presented at the visible layer, so named because it contains the variables that we are able to observe. Then a series of hidden layers extracts increasingly abstract features from the image. These layers are called 'hidden' because their values are not given in the data; instead the model must determine which concepts are useful for explaining the relationships in the observed data. The images here are visualizations of the kind of feature represented by each hidden unit. Given the pixels, the first layer can easily identify edges, by comparing the brightness of neighboring pixels. Given the first hidden layer's description of the edges, the second hidden layer can easily search for corners and extended contours, which are recognizable as collections of edges. Given the second hidden layer's description of the image in terms of corners and contours, the third hidden layer can detect entire parts of specific objects, by finding specific collections of contours and corners. Finally, this description of the image in terms of the object parts it contains can be used to recognize the objects present in the image*" (GOODFELLOW; BENGIO; COURVILLE, 2016, p. 2).

maneiras, dependendo da construção dos dados em relação ao mapeamento entre entradas e saídas, conforme explicam Bochie et al. (2020). Os algoritmos de aprendizado de máquina podem ser divididos em supervisionado, não supervisionado e por reforço (BOCHIE et al., 2020, p. 4).

Quadro 3 – Abordagens dos algoritmos de aprendizado de máquina

Aprendizado supervisionado	**Aprendizado não supervisionado**	**Aprendizado por reforço**
Os algoritmos têm acesso a um conjunto de dados rotulados, ou seja, existem exemplos do mapeamento entre entradas e saídas. A presença dos rótulos possibilita que os algoritmos ajustem seus parâmetros para reproduzir as mesmas saídas caso entradas semelhantes sejam apresentadas. Em uma analogia ao aprendizado humano, o algoritmo de aprendizado supervisionado tem acesso às respostas corretas das perguntas de um teste e aprende com o acesso a essas respostas. As respostas corretas das perguntas do teste são análogas ao mapeamento entre entradas e saídas promovido pelo rotulamento dos dados.	O conjunto de dados carece de rótulos, não existindo um mapeamento entre entradas e saídas. Nesse cenário, os algoritmos buscam relações e características presentes no conjunto de dados que possam ser exploradas para classificar internamente os elementos. Essa classificação pode levar a grupos de dados que compartilhem características semelhantes ou a grupos de dados que possuam algum tipo de correlação. As relações inferidas são mensuradas por métricas que verificam se a classificação obtida é adequada, possibilitando que os algoritmos ajustem os seus parâmetros. Analogamente ao aprendizado humano, os algoritmos de aprendizado não supervisionado avaliam padrões, assim como um bebê observa o comportamento e as características que definem uma pessoa conhecida, por exemplo. Observando os padrões de comportamento e as características de uma pessoa qualquer, o bebê é capaz de associar aquele conjunto de entradas de dados à saída que determina se a pessoa é conhecida ou não. Não é necessário, nesse caso, que seja informado previamente ao bebê que ele conhece a pessoa.	Os algoritmos se baseiam em um modelo de recompensas e punições à medida que o modelo interage com o ambiente no qual está inserido. Assim, em vez de existir um mapeamento direto entre entradas e saídas, os resultados são obtidos a partir da realimentação (ciclo de *feedback*) entre o sistema de aprendizado e o ambiente. A cada interação, as ações disponíveis são apresentadas ao modelo no seu estado atual e, após a mudança de estado, recebem um sinal de reforço. De forma similar ao aprendizado humano, os algoritmos de aprendizado por reforço buscam instigar um conjunto de comportamentos desejados, como o uso correto de talheres por uma criança através de recompensas, como palavras de reforço ou um prêmio. O objetivo dessa abordagem é escolher ações que maximizem a recompensa a longo prazo (KAELBLING et al., 1996).

Fonte: BOCHIE et al., 2020, p. 3-4.

Assim como o aprendizado de máquina, as abordagens de aprendizado profundo podem ser categorizadas da mesma forma (ALOM et al., 2018[15]).

Figura 3 – Abordagens do aprendizado profundo

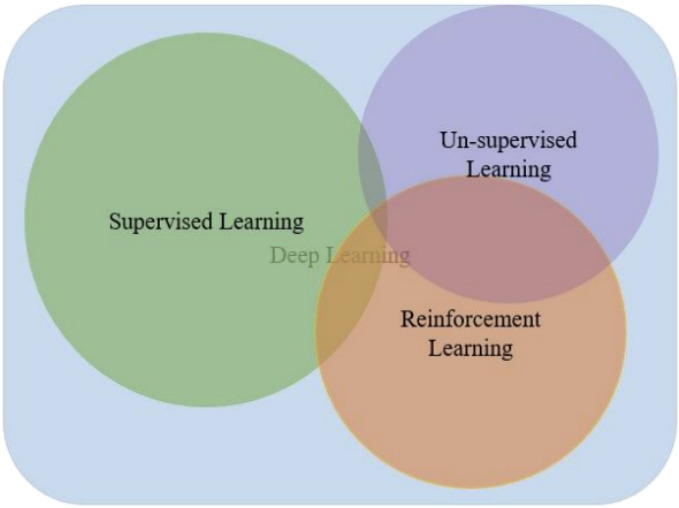


Fig. 2. Category of Deep Leaning approaches

Fonte: ALOM et al., 2018.

Além das categorias de algoritmos, é imprescindível, para o aprendizado de máquina, que haja capacidade de generalização, isto é, habilidade para reagir apropriadamente a situações às quais o algoritmo não teve acesso durante sua fase de aprendizado, conforme estabelecem Bochie et al. (2020). As amostras dos dados coletados para verificar essa capacidade são divididas em três conjuntos: treinamento, validação e teste (BOCHIE et al., 2020, p. 5).

Quadro 4 – Conjuntos de amostras de dados

Conjunto de treinamento	**Conjunto de validação**	**Conjunto de teste**
Contém as amostras que o algoritmo efetivamente utiliza para ajustar os seus parâmetros. Portanto, é durante o treinamento que os modelos de	O conjunto de validação é usado para acompanhar o progresso no aprendizado, servindo para medir o erro, ou o custo, associado à configuração atual do algoritmo.	O conjunto de teste contém amostras inéditas que não foram utilizadas no treinamento e é usado para avaliar o comportamento do modelo já ajustado. Portanto, a

[15] No original: "*Like machine learning, deep learning approaches can be categorized as follows: supervised, semi-supervised or partially supervised, and unsupervised. In addition, there is another category of learning called Reinforcement Learning (RL) or Deep RL (DRL) which are often discussed under the scope of semi supervised or sometimes under unsupervised learning approaches*". (ALOM et al., 2018, p. 3)

aprendizado acessam parte dos dados para ajustar seus parâmetros internos. O processo de treinamento propriamente dito pode diferir de acordo com os modelos de aprendizado adotados. No treinamento, o modelo de aprendizado passa por diversas etapas de atualização de parâmetros e verificação de desempenho.	O modelo utiliza o desempenho atual sobre os dados para alterar seus parâmetros, por exemplo, a acurácia para modelos de classificação, ou o Erro Médio Quadrático (*Mean Squared Error* – MSE) para modelos de regressão. A partir dessa medição, verifica-se se há margem para melhorar o algoritmo ou se o aprendizado deve ser encerrado.	avaliação do desempenho do algoritmo é feita a partir do conjunto de teste. É muito importante que o usuário do modelo apenas utilize o conjunto de teste após o ajuste completo do algoritmo feito nos conjuntos de treino e validação.

Fonte: BOCHIE et al., 2020, p. 3-4.

Dessa forma, pode-se concluir que a técnica de aprendizado de máquina baseado em redes neurais de aprendizado profundo permite que sistemas inteligentes aprendam a partir de grandes conjuntos de dados, também conhecidos como *big data* (REIS, 2021).

> O ressurgimento do interesse sobre essas técnicas foi impulsionado pela contínua evolução do hardware e do software que permitiu o uso de ferramentas que exigem processamento intensivo. No contexto atual, o aprendizado de máquina aparece como uma ferramenta alternativa aos algoritmos tradicionais, baseados em regras, que possibilita a análise de diferentes cenários em Redes Desafiadoras sem que as regras de análise sejam explicitamente programadas (BOCHIE et al., 2020, p. 5).

Alpaydin (2016) estabelece que, levando em consideração os recentes sucessos no aprendizado de máquina em vários domínios, pode-se afirmar que o que precisávamos não era de novos algoritmos, mas de muitos dados de exemplo e de poder computacional suficiente para executar os algoritmos nessa quantidade de dados. O que foi possível a partir de 2012, sendo que a IA começou efetivamente a ser adotada em larga escala a partir de 2016-2017.

2.3. Características técnicas inerentes à técnica de aprendizado de máquina baseado em redes neurais

A utilização da IA vem crescendo em diversas áreas, graças aos avanços no aprendizado de máquina baseado em redes neurais. Além disso, as aplicações de IA já têm um grande impacto em indivíduos, empresas e outras organizações em diversos setores, abrangendo bilhões de pessoas. No entanto, a discussão sobre transparência torna-se complexa diante desses

sistemas, os quais apresentam uma característica intrínseca relacionada à opacidade, situação denominada de "caixa preta"[16] na IA.

> A técnica de aprendizado de máquina que hoje melhor resolve esses desafios é o aprendizado profundo (*deep learning*), que introduz representações complexas, frequentemente referidas como "redes neurais profundas", expressas em termos de outras representações mais simples organizadas em diversas camadas. As entradas (*inputs*) são apresentadas a uma camada visível, assim chamada porque contém as variáveis observáveis, seguida de uma série de camadas ocultas contendo variáveis não observáveis e internas ao próprio modelo (origem do problema da não explicabilidade) (KAUFMAN, 2022, p. 197).

Kaufman (2022) destaca que a opacidade em sistemas de redes neurais de aprendizado profundo, ou falta de explicabilidade, decorre do desconhecimento de como os chamados "dados de entrada" (*inputs*) geraram os dados de saída (*output*) e como o sistema correlacionou as variáveis contidas nos dados de entrada e os pesos atribuídos (denominados de "parâmetros").

> Em redes neurais profundas, os parâmetros aprendidos a partir de dados são chamados de *weights* (pesos); após a fase de treinamento (ou aprendizado), esses pesos compõem o algoritmo e passam a ser fixos. No caso de uma imagem, em que os pixels são os dados de entrada, a saída do sistema reflete a soma das multiplicações de pesos pelos pixels de entrada. Cada camada processa o que supõe-se serem conceitos mais abstratos do que da camada anterior, gerando o nível de abstração requerido pela saída. Por exemplo, a saída pode ser dog *vs* cat, e a entrada pode ser a imagem (conjunto de pixels); cada camada mais "profunda" (mais próximo da saída) tem valores representando conceitos mais abstratos que ajudam, eventualmente, a concluir se é gato ou cachorro (KAUFMAN, 2022, p. 197).

As arquiteturas das redes neurais profundas são formadas por várias camadas (*layers*), e cada uma delas realiza representações mais abstratas do que na camada anterior, visando alcançar a abstração requerida pelos dados de saída (*output*). Esse processo de representações cada vez mais abstratas está no cerne do problema da não explicabilidade dos sistemas (KAUFMAN, 2022).

Como salientam Goodefellow, Bengio e Courville (2016), surge uma incompatibilidade entre a otimização matemática em dimensões elevadas e a capacidade do ser humano de raciocinar e interpretar semanticamente, fenômeno denominado pelos cientistas do campo de "problema da interpretabilidade". Segundo Ferrari (2018),

> [...] a opacidade dos *learners* é consequência da alta dimensionalidade de dados, da complexidade de código e da variabilidade da lógica de tomada de decisões. Por empregarem centenas ou milhares de regras, por suas predições estarem combinadas

[16] Pasquale define "caixa preta" como "the way scientific and technical work is made invisible by its own success. When a machine runs efficiently, when a matter of fact is settled, one focuses only on its inputs and outputs and not on its internal complexiy" (no original) (PASQUALE, 2015, p. 200).

> probabilisticamente de formas complexas, pela velocidade no processamento das informações, e pela multiplicidade de variáveis operacionais, parece estar além das capacidades humanas apreender boa parte – senão todas – as estruturas decisórias que empreguem a técnica de *machine learning*. Assim, o mero acesso ao código comunica muito pouco, remanescendo a dificuldade de compreender o processo decisório (FERRARI, 2018, p. 6).

Diante da incapacidade de compreender ou explicar os motivos pelos quais erros podem ocorrer, bem como o processo pelo qual as conclusões são alcançadas em sistemas, a falta de compreensão sobre como esses sistemas chegam às suas conclusões gera dúvidas legítimas sobre sua confiabilidade. Assim, a questão da confiança torna-se mais urgente à medida que se delega e confia na tomada de decisões automatizadas para proteger bens humanos significativos, como segurança e saúde (VON ESCHENBACH, 2021).

Especialistas estão empenhados em reduzir a opacidade. Como resultado, estão em andamento esforços científicos para gerar interpretações compreensíveis do funcionamento desses sistemas e um novo programa de pesquisa surgiu, denominado Inteligência Artificial Explicável (*XAI – Explainable Artificial Intelligence*).

XAI refere-se a um conjunto de métodos que, a fim de aumentar a transparência e a confiabilidade dos modelos opacos, busca desenvolver outros modelos para gerar explicação dos comportamentos dos modelos originais (ZEDNIK; BOELSEN, 2021).

O projeto tem como objetivo contornar o problema da caixa preta por meio de duas abordagens científicas: (i) métodos de aprendizado de máquina que sejam "inerentemente interpretáveis"; e (ii) técnicas analíticas para promover explicações após o sistema ter sido desenvolvido (*post hoc*). Independentemente da abordagem, o objetivo geral é superar a opacidade no aprendizado de máquina (ZEDNIK; BOELSEN, 2021). Até o momento, essas tentativas ainda não comprovaram eficácia, inclusive não existe um *benchmarking* para compará-las (COZMAN; KAUFMAN, 2022).

Diante do desejo de manter o potencial transformador do aprendizado de máquina baseado em redes neurais e, ao mesmo tempo, mitigar os efeitos negativos que a opacidade acarreta no problema da interpretabilidade, é crucial examinar os limites da transparência e explorar abordagens que visem proporcionar um nível viável de explicabilidade, a fim de aprimorar a compreensão desses sistemas.

2.4. Desvendando os aspectos interpretáveis no aprendizado de máquina

Os sistemas de IA de aprendizado de máquina baseado em redes neurais tornaram-se um fator estratégico nos processos decisórios, devido à sua capacidade de gerar resultados

preditivos com taxas altas de acurácia. No entanto, nas últimas décadas, surgiram preocupações em relação aos potenciais riscos e danos que esses sistemas podem causar.

No contexto desses sistemas, a interpretabilidade[17] é definida como a capacidade de explicar ou fornecer o significado, em termos compreensíveis para um ser humano, de como esse sistema opera (GUIDOTTI et al., 2018, p. 6). Embora existam barreiras em relação à interpretabilidade dos sistemas, é possível mitigar riscos e danos associados ao seu uso por meio de atributos específicos determinados em cada etapa de desenvolvimento do modelo.

Ao contrário de outras técnicas de IA, os desenvolvedores de aprendizado de máquina não especificam de fato como um problema de IA será resolvido, mas especificam os atributos sob as quais serão desenvolvidas (ZEDNIK, 2021).

Embora a transparência e a explicabilidade completa dos sistemas possam ser inviáveis, por conta da opacidade, é possível avaliar os sistemas por meio da escolha da base de dados de treinamento dos algoritmos do sistema, da seleção de um ambiente de aprendizado adequado para atribuir variáveis iniciais (hiperparâmetros) e da visualização e interpretação dos resultados.

Em relação à seleção das variáveis iniciais, também conhecidas como hiperparâmetros, é possível considerar cuidadosamente quais variáveis foram incluídas no modelo, levando-se em conta fatores como relevância e representatividade. Além disso, é possível avaliar a seleção da base de dados de treinamento dos algoritmos para garantir que os conjuntos de dados foram representativos e imparciais, evitando viés discriminatório ou distorção que possa afetar os resultados do sistema.

Um exemplo emblemático é o caso que ocorreu em 2019, com o algoritmo de concessão de crédito da Apple Card, em parceria com a Goldman Sachs, em que o banco foi investigado por suposta discriminação de gênero, pois clientes relataram que os homens recebiam limite de crédito maior que as mulheres. A reportagem do *The Washington Post* (TELFORD, 2019) revelou que o algoritmo concedia limites de crédito mais baixos para mulheres do que para homens, mesmo quando elas tinham históricos financeiros melhores. A Apple e a Goldman Sachs negaram qualquer viés de gênero em seu algoritmo e defenderam que o sistema era neutro em relação a esse aspecto. De acordo com Kaufman (2022), o viés pode ter origem nos seguintes contextos:

[17] Interpretar significa dar ou fornecer o significado ou explicar e apresentar em termos compreensíveis alguns conceitos.

Quadro 5 – Diferentes fontes de origem do viés

(a) na geração dos dados	A discriminação na produção de dados está presente tanto na predominância de usuários dos países desenvolvidos com mais acesso a tecnologias e às redes sociais, o que engendra uma base de dados enviesada pelo biotipo racial de pele clara, quanto na não desagregação dos dados por gênero e/ou o tratamento dado ao homem como "humano padrão".
(b) nas escolhas dos desenvolvedores	No desenvolvimento de um modelo de DLNN, a tarefa inicial dos cientistas da computação é identificar o problema a ser resolvido pelo sistema, em que situação e com qual objetivo o sistema será utilizado. O segundo passo é traduzir esse problema a ser resolvido em variáveis que possam ser observadas e manipuladas (*feature engineering process*). São eles que definem, por exemplo, quais termos de pesquisa serão usados para coletar os dados, o número de camadas ocultas e o número de nós em cada camada. Identificar a influência da subjetividade humana no projeto e na configuração do algoritmo de IA não é trivial, além de não ser possível eliminá-la mesmo se identificada (Hao, 2019). O Alan Turing Institute (Leslie, 2020) aponta como um dos problemas críticos que permitem que os vieses sistêmicos se infiltrem nos dados a postura dos desenvolvedores e designers de algoritmos, que não priorizam as ações para identificar e corrigir desequilíbrios potencialmente discriminatórios na representação demográfica e fenotípica. O instituto atribui esses vieses à complacência dos produtores de tecnologia, em geral, parte do grupo dominante e, logo, isentos dos efeitos adversos de resultados discriminatórios.
(c) na base de dados	O viés ocorre se os dados de referência forem menos diversificados demograficamente do que a população-alvo, ou seja, se a base de dados contiver poucos ou nenhum exemplo de uma determinada subpopulação por etnia e/ou gênero. A diferença entre ambientes controlados (laboratórios) e ambientes não controlados (mundo real), igualmente, tem o potencial de gerar resultados tendenciosos; nas ruas, por exemplo, as câmeras podem captar imagens em baixa resolução, o ângulo captado da face e a luminosidade podem dificultar a extração de características faciais ou mesmo distorcê-las, provocando erro no reconhecimento facial (Learned-Miller et al., 2020b).
(d) no processo de rotulagem dos dados	Criar uma base de dados de treinamento significa amostrar um mundo quase infinitamente complexo e variado, e fixá-lo em taxonomias compostas de classificações. [...] Manter a uniformidade na classificação manual de grandes conjuntos de dados é um desafio, que se torna quase inviável quando envolve classificar imagens de pessoas; são inúmeras as categorias classificatórias, incluindo raça, idade, nacionalidade, profissão, *status* econômico, comportamento, caráter e até mesmo moralidade. Estruturar uma taxonomia para classificar imagens de pessoas com a lógica utilizada para objetos gera inúmeras distorções e, consequentemente, vieses.
(e) nos dados de treinamento dos algoritmos	Considera-se que existe um enviesamento na base de dados quando o sistema exibe um erro sistemático no resultado ("enviesamento estatístico" ou "discriminação algorítmica"). Estritamente, qualquer conjunto de dados poderá ser imparcial para a execução de uma determinada tarefa, contudo, potencialmente existe o risco de que, se usado para uma tarefa distinta, seja tendencioso para essa segunda tarefa. Um sistema citado com frequência nos debates sobre discriminação algorítmica é o Compas.

Fonte: COZMAN; KAUFMAN, 2022, p. 200-205.

No caso mencionado, seria possível avaliar os atributos interpretáveis do sistema, conforme indicado por Kaufman (2022), em relação aos seguintes aspectos: (a) na geração dos dados; (b) nas escolhas dos desenvolvedores; (c) na base de dados; (d) no processo de rotulagem dos dados; e (e) nos dados de treinamento dos algoritmos.

3. A TRANSPARÊNCIA COMO MEIO DE PROTEÇÃO DO DIREITO À IGUALDADE

3.1. Transparência: resguardando o direito à igualdade

A transparência é um conceito que remonta ao latim medieval, com origem no termo *transparentia*, que é uma qualidade atribuída a um objeto transparente (ARRUDA, 2021). O termo se relaciona ao verbo transparecer, do latim *transparere*, formado pelos termos *trans*, que significa através, atravessar, e *parere*, que significa aparecer, deixar vir à luz, sendo transparente o objeto que permite a vinda ou a passagem da luz (ARRUDA, 2021).

Durante o período do Iluminismo, a transparência tornou-se a esperança de "descobrir" uma verdade singular e objetiva, capaz de superar as perspectivas individuais. Além de ser vista como essencial para estabelecer uma sociedade justa e harmônica (DASTON, 1992). Diversos historiadores reconhecem a importância de práticas iluministas como os primeiros contextos em que a transparência surgiu de forma moderna (CRARY, 1990; DASTON, 1992; DASTON; GALISON, 2007; HOOD, 2006)[18].

> Durante o Iluminismo, pensadores e filósofos da elite europeia enalteciam o conhecimento e a razão como uma forma de libertação, já que a compreensão sobre os fenômenos significava também um domínio maior sobre eles. O movimento marca o fim da idade média e do Estado absolutista, consequentemente influenciando os debates que viriam a conformar as democracias liberais da atualidade (MEIRELES, 2020, p. 8).

Além disso, a transparência possuía, nesse contexto, uma dimensão afetiva[19], relacionada ao medo de segredos e à crença de que a visibilidade pode levar ao controle sobre algo, sustentando a promessa da democracia liberal de que a abertura gera segurança (ANANNY; CRAWFORD, 2018). A transparência começou a se manifestar principalmente nas práticas relacionadas à ciência e à engenharia social, nas quais era vista como uma condição prévia para se alcançar a verdade e promover uma sociedade melhor.

O termo transparência foi incorporado, aos poucos, à ciência do direito. Segundo Arruda:

---

[18] No original: "*The hope to 'uncover' a singular truth was a hallmark of The Enlightenment, part of what Daston (1992: 607) calls the attempt to escape the idiosyncrasies of perspective: a 'transcendence of individual viewpoints in deliberation and action [that] seemed a precondition for a just and harmonious society'. Several historians point to early Enlightenment practices around scientific evidence and social engineering as sites where transparency first emerged in a modern form*" (ANANNY; CRAWFORD, 2018).

[19] Isso inclui uma dimensão afetiva, relacionada ao medo de segredos, à sensação de que ver algo pode levar ao controle sobre ele e à promessa da democracia liberal de que a abertura cria segurança (PHILLIPS, 2011).

> Considerando que a ciência jurídica se ocupa das relações jurídicas, sendo estas "situações fáticas de conduta humana que estão unidas umas às outras pela norma jurídica", a transparência jurídica passa a ser uma propriedade, uma qualidade a ser atribuída aos diversos meios onde se desenvolvem as relações jurídicas. Assim, haveremos de encontrar aplicação do princípio da transparência nas diversas áreas do direito, como, por exemplo, nas relações comerciais, de consumo e fiscais (2021, p. 42).

A transparência desempenha um papel crucial no direito privado, pois, quanto mais transparentes as relações jurídicas, maior conhecimento é permitido às partes, que passam a ficar mais seguras e confiantes nas relações (ARRUDA, 2021), permitindo o exercício da autonomia privada[20].

No art. 4º[21], *caput*, do Código de Defesa do Consumidor, a transparência tem como fim reequilibrar as relações de consumo, harmonizando e dando maior clareza às relações contratuais. Segundo Marques (2003, p. 139): "... significa 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, isto é, na fase negocial dos contratos de consumo".

Uma das formas concretas de expressão do princípio da transparência é reconhecida pelo direito à informação, conforme prevê o art. 5º, XIV[22], da Constituição Federal, que assegura que os cidadãos tenham acesso a informações relevantes e necessárias para participar ativamente da sociedade e fiscalizar as ações dos poderes públicos.

O dever de informar também está presente no direito privado, podendo ser compreendido como a prática de tornar disponíveis informações e conhecimentos relevantes, de forma clara e acessível, para que possam ser compreendidos e avaliados por todas as partes interessadas e envolvidas nas relações comerciais.

---

[20] "O princípio da autonomia põe em relevo as possibilidades de ser e atuar e a responsabilidade pelas consequências da conduta humana. Autonomia pode, aqui, ser compreendida como poder, reconhecido ou concedido pelo ordenamento estatal a um indivíduo ou a um grupo, de determinar vicissitudes jurídicas, como consequência de comportamentos – em qualquer medida – livremente assumidos.1 Àqueles que desenvolvem, titularizam o domínio ou fazem uso de sistemas de IA cabe exercer controle eficaz sobre eles. Afinal, os sistemas de IA não devem comprometer a autonomia dos seres humanos de estabelecer, dentro da licitude, seus próprios padrões comportamentais. Ante a inegável vulnerabilidade da pessoa diante das máquinas inteligentes, necessário buscar soluções para preservar, proteger e promover a autonomia sob dupla perspectiva: a autodeterminação nas questões individuais (ou seja, na construção da pessoalidade) e autonomia na convivência com os outros humanos" (LIMA; SÁ, 2020).

[21] "Art. 4º A Política Nacional das Relações de Consumo tem por objetivo o atendimento das necessidades dos consumidores, o respeito à sua dignidade, saúde e segurança, a proteção de seus interesses econômicos, a melhoria da sua qualidade de vida, bem como a transparência e harmonia das relações de consumo, atendidos os seguintes princípios: [...]."

[22] "Art. 5º 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: [...] XIV – é assegurado a todos o acesso à informação e resguardado o sigilo da fonte, quando necessário ao exercício profissional."

> O princípio da transparência é concretizado através dos deveres informativos decorrentes da boa-fé, é o direcionamento de condutas, é a pedra angular no exercício da autonomia contratual, apresentando com a devida clareza os pormenores que cercam aquela relação contratual. Assim, é possível compreender o quão profundo e imprescindível é o dever de informar, uma conduta ativa imposta ao fornecedor; o consumidor é o detentor do direito subjetivo de informação (artigo 6º III CDC). Característico de tal direito é o fato de a pessoa já ter alguma noção prévia sobre a existência de uma informação requerida, mas não conhece os detalhes ou sua abrangência. Pode a informação consistir em dever principal como em dever acessório, instrumental ou anexo na relação de consumo (SALOMÃO, 2012).

Presente também na intersecção entre direito e economia, a transparência vinculada ao dever de informar é fundamental para a teoria da assimetria da informação[23], a fim de promover a divulgação de informações relevantes, permitir que as partes envolvidas tomem decisões informadas e evitar a exploração de informações privilegiadas.

Antes de prosseguir com a análise, é importante considerar que a transparência não pode ser vista como "um estado final preciso em que tudo é claro e aparente" (ANANNY; CRAWFORD, 2018), mas, sim, como um meio para atingir uma finalidade. Embora a transparência esteja presente em diversas abordagens no âmbito jurídico, antes de se analisar sua aplicação no âmbito tecnológico, é imprescindível considerar qual finalidade se almeja resguardar.

Para esse ponto, é interessante tratar a transparência sob a perspectiva da atividade econômica. Nesse sentido, focaremos na transparência associada ao dever de informar, conforme apresentado anteriormente. Essa forma de transparência envolve a disponibilização clara e acessível de informações relevantes, de modo a permitir que todas as partes interessadas compreendam e avaliem adequadamente a situação em questão.

A dispersão do conhecimento – fornecimento de informações igualitárias e de qualidade, na perspectiva da atividade econômica – pode ser compreendida como um requisito indispensável para a garantia da concretização do princípio da igualdade diante de um processo de integração econômica equilibrada, que nada mais é do que a efetivação da igualdade material[24] entre os agentes econômicos. Nesse caso, a igualdade material é o valor

---

[23] "Podemos então entender que a assimetria de informação pode ser representada por uma assimetria ou desequilíbrio de conteúdos de repertório. Tendo por certo que tal desequilíbrio pode resultar em desequilíbrio da capacidade de barganha e sobreposição de interesses numa relação entre dois sujeitos em torno de bens e/ou valores" (GABAN, 2002, p. 130).

[24] "Igualdade material quer, aqui significar efetiva, e não meramente formal, de oportunidades [...] isso só pode ocorrer com a difusão forçada do conhecimento econômico entre os indivíduos, que, por sua vez, só pode ser assegurada através de uma garantia firme de existência de concorrência."

constitucional[25] que está se protegendo por meio da transparência, compreendida como difusão forçada do conhecimento no contexto econômico.

A importância da igualdade no Estado Democrático de Direito brasileiro se expressa na abertura do art. 5º e está concretizada no *caput* desse mesmo artigo na Constituição Federal. Os requisitos para se atingir a referida igualdade material entre agentes econômicos, segundo Salomão (2015), são a dispersão do conhecimento (informações igualitária e de qualidade) e a concorrência. Esta dissertação se concentrará no tópico da disseminação do conhecimento.

A dispersão do conhecimento, de acordo com Salomão (2015), está relacionada à interação entre valores e conhecimento na sociedade, especialmente no campo econômico.

> Essa definição entre valores e conhecimento da sociedade é bastante cristalina no campo econômico. Como visto, a proteção da concorrência leva à descoberta da verdadeira utilidade dos produtos e das melhores opções para o consumidor. O valor "concorrencial" influi, portanto, duplamente sobre a realidade – primeiro modelando-a, e em seguida permitindo seu conhecimento. [...] A regra jurídica, aí, é eminentemente instrumental. A afirmação da concorrência como valor fundamental (modelagem) garante a liberdade de escolha e informação mais abundante possível para o consumidor. Ele, então sozinho, descobrirá a solução mais adequada para suas necessidades (SALOMÃO, 2012, p. 200).

Dessa forma, a disseminação do conhecimento é relevante, pois permite que os agentes econômicos tomem decisões embasadas em informações adequadas. Nesse contexto, a transparência é ferramenta essencial viabilizada por meio da disseminação do conhecimento, pois promove a divulgação de informações relevantes e possibilita a concretização do direito à igualdade, permitindo o exercício da autonomia privada.

3.2. Transparência e explicabilidade nos sistemas de IA

De acordo com a Organização para a Cooperação e Desenvolvimento Econômico (OCDE)[26], deve haver transparência e divulgação responsável em torno dos sistemas de IA para garantir que as pessoas entendam os resultados baseados nesses sistemas e, assim, possam desafiá-los.

---

[25] Preâmbulo da Constituição Federal: "Nós, representantes do povo brasileiro, reunidos em Assembléia Nacional Constituinte para instituir um Estado Democrático, destinado a assegurar o exercício dos direitos sociais e individuais, a liberdade, a segurança, o bem-estar, o desenvolvimento, a igualdade e a justiça como valores supremos de uma sociedade fraterna, pluralista e sem preconceitos, fundada na harmonia social e comprometida, na ordem interna e internacional, com a solução pacífica das controvérsias, promulgamos, sob a proteção de Deus, a seguinte Constituição da República Federativa do Brasil".

[26] Os princípios da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) sobre Inteligência Artificial (AI) podem ser encontrados no seguinte link: https://www.oecd.org/going-digital/ai/principles/.

Essa lógica repousa em uma suposição epistemológica de que a verdade tem uma correspondência com um fato (DAVID, 2015). Desse modo, a premissa dessa transparência nos sistemas de IA é que, quanto mais informações forem conhecidas sobre o funcionamento interno de um sistema, maior seria a sua capacidade de ser compreendido pelos seus observadores e usuários.

De acordo com Turilli e Floridi (2009), o primeiro aspecto importante a ser levantado é que, para que a informação seja considerada transparente, ela deve ser acessível e compreensível. Nesse ponto reside a diferença conceitual entre a mera disponibilização de informações sobre o funcionamento do sistema, entendido como a transparência, e a compreensão de fato sobre o funcionamento dele, entendido como a explicabilidade.

Algoritmos só podem ser considerados explicáveis se um humano puder articular o modelo treinado ou a lógica de uma determinada decisão, por exemplo, explicando a influência (quantificada) de determinados *inputs* ou atributos (FLORIDI et al., 2016, p. 6).

Um segundo aspecto que merece atenção é que a discussão sobre a transparência e a explicabilidade se torna ainda mais complexa diante dos sistemas algorítmicos de redes neurais profundas, os quais apresentam características intrínsecas que impossibilitam a efetividade desses conceitos, conforme destacado no Capítulo 1.

Os esforços para tornar os algoritmos transparentes enfrentam o grande desafio de tornar os complexos processos de tomada de decisão acessíveis e compreensíveis.

> "'Algoritmos' são opacos no sentido de que se alguém é um receptor da saída do algoritmo (a decisão de classificação da classe), raramente se tem qualquer senso concreto de como ou porque uma determinada classificação foi obtida a partir de entradas" (Burrell, 2016: 1). Tanto as entradas quanto as saídas podem ser desconhecidas. A opacidade dos algoritmos de aprendizagem da máquina é um produto da alta dimensionalidade dos dados, do código complexo e da lógica mutável da tomada de decisões (Burrell, 2016). Matthias (2004: 179) sugere que a aprendizagem de máquinas pode produzir *out-puts* para os quais "o próprio treinador humano é incapaz de fornecer uma representação algorítmica" (FLORIDI et al., 2016, p. 6).

Devido às características intrínsecas que tornam difícil entender e avaliar como esses sistemas tomam suas decisões, existe uma dificuldade de se gerar transparência e explicabilidade integral em relação ao seu funcionamento, apresentando um desafio significativo para compreendê-los adequadamente.

> Essa técnica de inteligência artificial amplamente usada oferece diversos benefícios, mas tem limitações. O mais prudente é que seus em seus resultados. Em primeiro lugar, porque são técnicas estatísticas de probabilidade, logo, possuem grau de incerteza intrínseco, e, em segundo, por conta da opacidade de seu funcionamento (como confiar plenamente em algo que não se domina, não se compreende?), além

> das várias limitações técnicas. A inteligência artificial implementada atualmente em larga escala deve ser encarada como parceira dos profissionais humanos nos processos de decisão, e não soberana, ou seja, capaz de contribuir para aumentar a inteligência humana especializada, e não substituí-la (KAUFMAN, 2022, p. 43).

A situação ainda é agravada pelo fato de os sistemas que utilizam a técnica de redes neurais profundas estarem se tornando cada vez mais precisos para a realização de previsões complexas em diversos campos de aplicação, graças à sua velocidade e sofisticação. No entanto, ao contrário de sugestões em serviços de entretenimento ou navegação, como Netflix ou Waze, a opacidade desses sistemas na medicina pode levar à resistência por parte dos profissionais de saúde em adotá-los.

E essa resistência dos profissionais de saúde em aceitar plenamente os resultados apurados pelos sistemas de IA surge devido à falta de compreensão sobre como esses resultados são obtidos. A falta de transparência levanta uma questão ética crucial diante das limitações dessa técnica: os profissionais de saúde não podem considerar os resultados da IA como soberanos, pois são responsáveis por tomar decisões informadas e fornecer cuidados de qualidade, e, para isso, precisam confiar em um processo e metodologia os quais possam compreender e justificar.

Esse impasse tem criado questões éticas, especialmente quando as consequências dessas decisões são significativas, como diagnósticos médicos incorretos ou acidentes causados por veículos autônomos. Conforme destaca Arbix (2020), a discussão sobre os sistemas de IA prioriza a importância de explicar e entender o funcionamento desses sistemas, diante da necessidade de se avaliar a confiabilidade na tomada de decisões.

> Qualquer máquina é concebida para funcionar e antes de ser comercializada é submetida a uma série de testes, em que são apurados indicadores confiáveis, com percentuais de acerto em níveis aceitáveis para considerá-la aprovada. No caso de máquinas de inteligência artificial, o processo é bem mais complexo, por conta da opacidade (caixa-preta). A chamada interpretabilidade do sistema de IA é a tentativa de entender e determinar qual grau de confiança atribuir ao resultado obtido (KAUFMAN, 2022, p. 42).

Cientistas estão empenhados em reduzir a opacidade, ou seja, serem capazes de explicar como o sistema chegou a determinada previsão e, preferencialmente, de forma que o usuário compreenda e possa até identificar os erros cometidos pelos algoritmos (KAUFMAN, 2022). No entanto, ainda não existem soluções eficazes para reduzir a opacidade dos sistemas de aprendizado de máquina baseado em redes neurais.

As redes neurais são reconhecidas como modelos estatísticos de probabilidade que incorporam a variável da incerteza intrinsecamente. Devido a essa natureza complexa, as

decisões e os resultados obtidos não podem ser explicados em sua plenitude. No entanto, é fundamental focalizar no que é viável, considerando a importância ética de fornecer informações sobre o processo de tomada de decisão dos sistemas de IA.

3.3. Transparência: o pilar fundamental para a confiança no desenvolvimento da IA

Muller (2021), em seu artigo "We need to talk about artificial intelligence", publicado no *World Economic Forum*, destaca que começa a se formar um consenso em torno do impacto que a IA terá na humanidade ao afirmar que "considerações éticas, como o viés da IA (por raça, gênero ou outros critérios) e a transparência algorítmica (clareza sobre as regras e métodos pelos quais as máquinas tomam decisões) já impactam negativamente a sociedade por meio das tecnologias que usamos diariamente". Nesse cenário, o autor acrescenta que está aumentando a demanda para que a sociedade civil, o setor público e o setor privado trabalhem em mecanismos para a construção de responsabilidades e confiança no desenvolvimento da IA.

Diante da rápida disseminação da IA, é importante considerar que essa tecnologia apresenta tanto benefícios quanto riscos à sociedade, conforme atestam os debates, entre outros, sobre a utilização de programas de pontuação de risco na justiça criminal, de sistemas de reconhecimento facial na segurança pública e de carros autônomos.

Erick Trickey (2018) aponta que, no mês de novembro de 2017, onze professores e pesquisadores da Universidade de Harvard e do Massachusetts Institute of Technology (MIT) publicaram uma carta aberta ao Poder Legislativo do estado de Massachusetts, nos Estados Unidos, solicitando que não fosse aprovada uma disposição específica constante de um projeto de lei de justiça criminal. Essa disposição obrigaria os tribunais estaduais a utilizar programas de pontuação de risco para determinar a fiança dos réus. Os especialistas recomendaram que o estado revisasse os programas de pontuação de risco a fim de identificar potenciais vieses negativos relacionado raça ou gênero ocultos e considerasse a possibilidade de construir um programa próprio, através de um processo aberto e público (BAVITZ, 2017).

A falta de transparência e explicabilidade em sistemas algorítmicos suscita preocupações significativas, uma vez que os usuários afetados não teriam a oportunidade de compreender como esses algoritmos foram desenvolvidos, como chegaram a determinados cálculos ou se o sistema cometeu vieses negativos discriminatórios ao atribuir informações incorretas a uma decisão em particular.

Segundo Trickey (2018), alguns modelos de pontuação de risco usam fatores que podem reforçar indicadores de discriminação racial, tais como nível de educação, isolamento social ou

instabilidade habitacional de um réu. A opacidade desses sistemas dificulta a detecção de possíveis erros, vieses negativos e injustiças, impossibilitando que os usuários afetados exerçam seu direito de conhecer e compreender o processo que afeta suas vidas.

Enquanto não houver uma solução técnica para reduzir a opacidade dos sistemas, conforme discutido no Capítulo 1, o caminho será buscar os atributos que possam contribuir com o máximo de transparência e explicabilidade. Essa busca por atributos que contribuam para a transparência e a explicabilidade dos sistemas de IA pode representar uma solução para auxiliar na detecção de possíveis erros, vieses negativos e injustiças.

Embora a lógica por trás do uso desses modelos seja alocar recursos adequados e reduzir o viés, Hao (2019) argumenta que esses sistemas podem amplificar preconceitos e gerar dados tendenciosos. Isso decorre do fato de que as ferramentas de avaliação de risco, baseadas em IA, são frequentemente impulsionadas por algoritmos treinados com dados históricos de crimes (HAO, 2019).

> Agora, populações que historicamente foram desproporcionalmente visadas pela aplicação da lei – especialmente comunidades de baixa renda e minorias – correm o risco de receber pontuações altas de reincidência. Como resultado, o algoritmo pode ampliar e perpetuar vieses incorporados e gerar ainda mais dados contaminados para alimentar um ciclo vicioso. Como a maioria dos algoritmos de avaliação de risco são proprietários, também é impossível interrogar suas decisões ou responsabilizá-los. O debate sobre essas ferramentas ainda está em andamento. Em julho passado, mais de 100 organizações de direitos civis e comunitárias, incluindo ACLU e NAACP, assinaram uma declaração contra o uso de avaliação de risco. Ao mesmo tempo, cada vez mais jurisdições e estados, incluindo a Califórnia, recorreram a elas em um esforço desesperado para consertar suas prisões sobrecarregadas (HAO, 2019, *on-line*).

Em 19 de março de 2018, um carro autônomo provocou uma morte por atropelamento de pedestre em Tempe, no Arizona (G1, 2018). Quem seria o culpado quando um carro autônomo provoca um acidente é apenas uma das muitas questões incômodas que os debates sobre IA levantam. Uma situação como essa pode suscitar inúmeros questionamentos: o motorista que estivesse dentro carro – e que, em tese, poderia assumir o controle sobre o veículo – deveria ser culpado pelo acidente? Quem deveria assumir a responsabilidade por falhas no sistema autônomo? O que o acidente significaria em termos éticos e jurídicos? Conforme preleciona Roberto:

> Seria necessário averiguar, na prática, se houve ação ou omissão voluntária, negligência ou imprudência nos termos do Código Civil – por parte de algum ser humano. 5 Nesse caso, o motorista "reserva", um dos polos humanos da interação homem-máquina, poderia ser culpado: será que poderia ter intervindo antes do acidente, e só não o fez por negligência ou imperícia? O teste de subsunção nesse caso já é conhecido do direito há séculos, por mais que os fatos sejam novos, e não cabe aos nossos propósitos delimitá-lo aqui. A responsabilidade subjetiva também poderia

> recair sobre os próprios produtores ou outros envolvidos na cadeia de produção do veículo, naturalmente. Seria o caso de peças montadas com imperícia, falta de vistorias legal ou tecnicamente necessárias, etc. (2020, p. 132).

O website "Moral Machine", do MIT[27], criado por Iyad Rahwan, professor associado do *Media Lab* daquela mesma universidade, entrevistou inúmeras pessoas em todo o mundo para questioná-las a respeito de dilemas éticos que envolvessem acidentes provocados por carros autônomos: qual seria a melhor decisão tomada pelo sistema – adotar uma trajetória que provocasse a morte de dois passageiros ou a que provocasse a morte de cinco pedestres?

Diante da complexidade em determinar a responsabilidade por decisões indevidas tomadas por sistemas de IA, definições sobre a responsabilidade ainda permanecem objeto de intenso debate.

[27] Disponível em: https://www.moralmachine.net/. Acesso em: 19 jan. 2023.

4. ANÁLISE DA TRANSPARÊNCIA E DA EXPLICABILIDADE NOS PROJETOS DE REGULAMENTAÇÃO DE IA

O crescente uso de tecnologias de IA tem gerado preocupações em relação à responsabilidade legal por falhas sistêmicas em decisões automatizadas, equívocos em diagnósticos médicos, discriminação e fraude financeira (KAUFMAN, 2019). Embora a transparência por si só não garanta uma maior segurança dos sistemas, a explicação de seu funcionamento e a interpretabilidade sobre os sistemas podem assegurar a concretização do direito à igualdade, permitindo o exercício da autonomia privada e o direito de escolha ao permitir que os indivíduos tomem decisões informadas e tenham maior confiança nesses sistemas.

Em contrapartida, no contexto do aprendizado de máquina baseado em redes neurais, a lógica da transparência como pilar fundamental e exclusivo para gerar igualdade entre as partes pode não ser factível devido à natureza da técnica. Essa ideia foi abordada no Capítulo 2, no qual foram discutidos os desafios relacionados à transparência diante dessa tecnologia.

No contexto em que a transparência é limitada, pode haver preocupações legítimas sobre a violação do direito à igualdade, uma vez que as partes afetadas não possuem conhecimento completo sobre o funcionamento dos sistemas IA. Essas preocupações fortalecem a discussão atual sobre a regulamentação do tema. Neste capítulo, serão analisadas as soluções propostas tanto no PL 2.338/2023 quanto no AI Act em relação à transparência e à explicabilidade desses sistemas.

4.1. Brasil: transparência como conceito norteador do PL 2.338/2023

Neste tópico, será realizada uma análise do processo de trabalho da CJSUBIA, que deu origem ao PL 2.338/2023, que, atualmente, está em discussão no Senado Federal.

A CJSUBIA – constituída pelo Presidente do Senado Federal, senador Rodrigo Pacheco, em fevereiro de 2022, com a tarefa de elaborar um substitutivo aos projetos de lei apresentados na Câmara do Deputados e no Senado Federal –, discutiu a importância da transparência como princípio norteador na redação desse projeto. Essa discussão foi apresentada em três frentes: durante as audiências públicas; na organização do seminário internacional; e no recebimento das contribuições escritas como parte da consulta pública.

As audiências públicas foram realizadas pela Comissão, em formato multissetorial, com o objetivo de coletar visões do setor público, academia, indústria e terceiro setor. Dentre os

doze painéis, discutiu-se no Painel 9[28] o tema "direitos e deveres: transparência e explicabilidade; revisão e o direito à intervenção humana; correção de vieses", e as principais ideias debatidas nesse painel serão expostas a seguir.

De acordo com Renato Leite Monteiro e Tainá Junquilho, o direito à explicabilidade está associado a capacitar indivíduos e sociedade com informações relevantes sobre processos automatizados, permitindo-lhes compreender seu funcionamento e tomar decisões a seu respeito. Eles entendem que esse conceito não deve se confundir com a publicização de códigos-fonte. Tainá Junquilho (SENADO FEDERAL, 2022) acrescentou que cada cultura pode determinar o tipo de explicação necessária, não sendo imprescindível compreender todos os caminhos percorridos pelo algoritmo.

Bruno Jorge Soares (SENADO FEDERAL, 2022) argumentou que a exigência de uma explicabilidade absoluta poderá comprometer os benefícios da tecnologia, que foi desenvolvida para ampliar as capacidades humanas. Ele defende a necessidade de equilíbrio nesse aspecto. Já Caroline Tauk (SENADO FEDERAL, 2022) mencionou que os modelos opacos não permitem uma explicação completa; além disso, Dora Kaufman (SENADO FEDERAL, 2022) afirmou que parte da opacidade nos resultados da técnica de aprendizado de máquina baseado em redes neurais é inevitável, pois é intrínseca à natureza da técnica.

Dora Kaufman (SENADO FEDERAL, 2022) também destacou que "as propostas de regulamentação mundo afora contemplam uma modalidade de auditoria, mas [...] elas não equacionam propriamente como concretizá-las". Nesse sentido, Virgílio Almeida (SENADO FEDERAL, 2022) defendeu ser necessário "estabelecer práticas para auditoria e regras para tornar os sistemas mais transparentes".

Em complementação às audiências públicas, a Comissão abriu prazo para o envio de contribuições escritas[29] para subsidiar os trabalhos, através da Nota Informativa Senado Federal – 2022-08682.

[28] Estiveram presentes o Dr. Renato Leite Monteiro, do Data Privacy Brasil e do Twitter; a Dra. Dora Kaufmann, professora da Pontifícia Universidade Católica de São Paulo (PUC-SP); a Dra. Caroline Tauk, juíza federal no Rio de Janeiro; o Dr. Bruno Jorge, representante da Agência Brasileira de Desenvolvimento Industrial (ABDI); o Dr. Virgílio Almeida, professor da Universidade Federal de Minas Gerais (UFMG); a Dra. Tainá Aguiar Junquilho, professora da Universidade de Brasília (UnB) e do Instituto Brasiliense de Direito Público (IDP); a Dra. Ana Paula Bialer, advogada e sócia fundadora da Bialer & Falsetti Associados.

[29] Foram recebidas 102 manifestações de entidades representantes da sociedade civil, de órgãos governamentais, da academia, do setor privado, além de contribuições individuais.

Houve também realização de Seminário Internacional, organizado por painéis. O Painel 3[30] debateu sobre "transparência, viés e devido processo na tomada de decisão automatizada". A seguir, serão apresentadas as principais concepções discutidas acerca desse tópico.

Courtney Lang[31] (SENADO FEDERAL, 2022) destaca que é "importante notar que a transparência não é um fim, e, sim, um meio pelo qual podemos lidar com a responsabilidade e o empoderamento de sistema". Essa seria uma prerrogativa que podemos usar como administração de risco em conjunto com outras ações. Mireille Hildebrandt (SENADO FEDERAL, 2022), por sua vez, entende que há transparência vinculada ao entendimento do funcionamento do sistema para que possa haver responsabilização e mitigação de danos. Afirma, ainda, que é importante pensarmos sobre transparência, nos sistemas de IA, para que as pessoas sejam capazes de desafiar, contestar esses sistemas, a fim de se transmitir uma segurança robusta e crescente. Além disso, apresenta a ressalva de que isso não quer dizer que nós temos que saber sobre todos os intrincados funcionamentos internos do sistema, mas devemos nos imbuir de uma supervisão correta sobre eles.

As contribuições citadas sobre a transparência na discussão da IA ficaram, em sua maioria, relativamente vagas e generalistas, deixando em aberto a questão de como efetivamente abordar esse tema na legislação. Embora tenham sido mencionados pontos relevantes sobre a distinção entre a publicização de códigos-fonte e a transparência em si, não houve explanação aprofundada sobre como essa publicização poderia beneficiar ou não a obtenção de sistemas mais transparentes. Além disso, a questão da opacidade dos sistemas de IA, decorrente das características das técnicas de aprendizado de máquina baseado em redes neurais, não foi objeto de uma discussão abrangente sobre as possíveis soluções viáveis para alcançar um nível satisfatório de transparência, sem esbarrar nesses desafios. Embora os depoimentos reafirmem a importância da transparência, falta clareza quanto ao seu conteúdo específico e à forma de concretizá-la.

As manifestações nesse contexto foram muito enfáticas sobre a importância de documentação técnica detalhada e também sobre o monitoramento no mercado, conforme mencionou Pam Dixon (SENADO FEDERAL, 2022). Além disso, o que fica claro das posições internacionais é que as obrigações de transparência devem ser compatíveis com o nível de risco

[30] Contou com apresentações do Dr. Anupam Chander, professor de Direito e Tecnologia na Georgetown Law; do Dr. Marc Rotenberg, presidente e fundador do Center for AI and Digital Policy; da Dra. Bojana Bellamy, presidente do Centre for Information Policy Leadership (CIPL); da Dra. Carly Kind, diretora da Ada Lovelace Coordenação de Comissões Especiais, Temporárias e Parlamentares de Inquérito 72 Institute; e da Dra. Nanjira Sambuli, membra do Programa de Tecnologias e Assuntos Internacionais do Carnegie Endowment for International Peace e do Ford Global Fellowship.

[31] Dra. Courtney Lang, do Conselho da Indústria de Tecnologia da Informação.

do sistema, conforme mencionou Courtney Lang (SENADO FEDERAL, 2022) – quanto maior o risco, mais medidas de transparência devem ser adotadas.

A ênfase nas manifestações internacionais sobre a importância de documentação técnica detalhada é um aspecto relevante a ser considerado. Importante observar que o AI Act, conforme será analisado nos itens subsequentes, na redação. Diante desse contexto, surge a questão de saber se documentar o processo de desenvolvimento e/ou uso de um sistema de IA seria uma resposta adequada à demanda por transparência.

4.1.1. Brasil: previsões sobre transparência no Projeto de Lei 2.338/2023

O PL 2.338/2023, que possui um total de nove capítulos, incorpora medidas de transparência, em grande parte vinculadas às medidas de explicabilidade, nos Capítulos I, II, III e IV. Dessa forma, foi realizada uma busca no PL pelos termos "transparência" e "explicabilidade" e encontrou-se um total de sete resultados relacionados ao termo "transparência", sendo que cinco desses resultados estão presentes em artigos específicos da lei – arts. 3º, 18, 19 e 24 –, e os outros dois foram encontrados no texto de "justificação".

Em relação ao termo "explicabilidade" ou "explicação", foram identificados cinco resultados, todos eles presentes em artigos da lei – arts. 3º, 5º, 8º, 18 e 20 –, dos quais dois também fazem menção ao termo "transparência". Além dos sete artigos explicitamente relacionados aos termos supramencionados, os arts. 7º, 13 e 22 foram incluídos nesta análise devido à abordagem implícita em relação aos conceitos "transparência" e "explicabilidade" em seu conteúdo.

Essa distribuição pode ser observada no Quadro 5:

Quadro 6 – Proposta de Regulamento Brasileira

Capítulos	Menção explícita aos termos	Menção implícita aos termos
Capítulo I – Disposições Preliminares	Artigo 3º, inciso VI	
Capítulo II – Dos Direitos	Artigo 5º, inciso II	
		Artigo 7º
	Artigo 8º	
Capítulo III – Da Categorização dos Riscos		Artigo 13
	Artigo 18	

Capítulo IV – Da Governança dos Sistemas de Inteligência Artificial	Artigo 19	
	Artigo 20	
		Artigo 22
	Artigo 24	
Capítulo V – Da Responsabilidade Civil	Não foram encontrados resultados	
Capítulo VI – Códigos de Boas Práticas e de Governança	Não foram encontrados resultados	
Capítulo VII – Da Comunicação de Incidentes Graves	Não foram encontrados resultados	
Capítulo VIII – Da Supervisão e Fiscalização	Não foram encontrados resultados	
Capítulo IX – Das Disposições Finais	Não foram encontrados resultados	

Fonte: Elaborado pela autora.

O conceito de transparência assume formas e desempenha papéis diferentes em cada um desses capítulos. Inicialmente, ele é introduzido como um princípios orientadores do PL 2.338/2023 (Capítulo I). Em seguida, desempenha um papel distinto ao estar relacionado aos direitos das pessoas afetadas por sistemas de IA. Posteriormente, assume uma abordagem mais enfática ao ser integrado aos requisitos para categorização de riscos desses sistemas (Capítulo II). Além disso, o conceito faz parte das diretrizes de governança estabelecidas para os agentes de IA que fornecem ou operam sistemas de alto risco (Capítulo III). Por fim, o conceito de transparência se destaca como ponto central na avaliação de impacto algorítmico (Capítulo IV).

A título de introdução, o PL 2.338/2023 no Capítulo 1, art. 3º, VI, estabelece os princípios que devem ser seguidos para o desenvolvimento, a implementação e o uso de sistemas de IA. Nesse contexto, são mencionados explicitamente os princípios de transparência, explicabilidade, inteligibilidade e auditabilidade como diretrizes para esse desenvolvimento.

> Art. 3º O desenvolvimento, a implementação e o uso de sistemas de inteligência artificial observarão a boa-fé e os seguintes princípios:
> [...]
> VI – transparência, explicabilidade, inteligibilidade e auditabilidade;
> [...]
> IX – rastreabilidade das decisões durante o ciclo de vida de sistemas de inteligência artificial como meio de prestação de contas e atribuição de responsabilidades a uma pessoa natural ou jurídica.

Além disso, o inciso IX do mesmo artigo apresenta o conceito de rastreabilidade das decisões tomadas ao longo do ciclo de vida dos sistemas de IA. Essa rastreabilidade tem como objetivo possibilitar a prestação de contas e a atribuição de responsabilidades a pessoas físicas ou jurídicas. Com base nessa explicação, podemos considerar que o inciso IX representa uma medida de transparência, em uma abordagem indireta do princípio.

Já os arts. 5º, 7º e 8º, presentes no Capítulo II, estabelecem os direitos de acesso às informações sobre a compreensão das decisões tomadas por esses sistemas de IA ao usuário afetado. Tais direitos estão organizados em duas categorias: a) o direito de receber informações claras e adequadas, previamente à contratação ou utilização do sistema, conforme estabelece o art. 7º; e b) os direitos a serem exercidos pelas pessoas afetadas por sistemas de IA que poderão solicitar explicação sobre a decisão a qualquer tempo, conforme o art. 5º, I e II, e o art. 8º. Um maior detalhamento sobre essa análise será realizado no Capítulo 4.

O direito à explicação e à informação está intrinsecamente relacionado ao princípio da transparência, uma vez que esta pressupõe a disponibilização acessível de informações relevantes e compreensíveis, bem como a capacidade de fornecer explicações adequadas para compreensão dos processos, conforme já mencionado de forma mais aprofundada no Capítulo 2.

O conceito de transparência também aparece no Capítulo III do PL 2.338/2023, em "Categorização dos riscos dos sistemas de IA". Antes de um sistema ser disponibilizado no mercado ou utilizado, será necessário que o fornecedor o submeta a uma avaliação preliminar para determinar seu nível de risco, conforme estabelecido no art. 13 do PL 2.338/2023. O projeto se baseia, de forma semelhante ao projeto da UE, na categorização dos sistemas de IA com base no risco, que pode ser classificado como baixo, alto ou excessivo.

> Art. 13. Previamente a sua colocação no mercado ou utilização em serviço, todo sistema de inteligência artificial passará por avaliação preliminar realizada pelo fornecedor para classificação de seu grau de risco, cujo registro considerará os critérios previstos neste capítulo.
> § 1º Os fornecedores de sistemas de inteligência artificial de propósito geral incluirão em sua avaliação preliminar as finalidades ou aplicações indicadas, nos termos do art. 17 desta lei.

Durante essa avaliação, o fornecedor deve incluir as finalidades e as aplicações do sistema; além disso, o registro e a documentação dessa avaliação devem ser mantidos por ele, para fins de responsabilização e prestação de contas, mesmo nos casos em que os sistemas não sejam classificados como de alto risco.

Ao exigir o fornecimento de uma descrição clara sobre as finalidades e aplicações do sistema, em relação ao propósito e aos objetivos do sistema, o legislador pretende exigir a transparência para facilitar que usuários e partes interessadas compreendam como o sistema será utilizado e como ele pode afetar a vida e as atividades dos usuários.

Nesse contexto, a falta de transparência e de explicabilidade é critério explícito para a classificação de risco do sistema, conforme estabelecido no art. 18.

> Art. 18. Caberá à autoridade competente atualizar a lista dos sistemas de inteligência artificial de risco excessivo ou de alto risco, identificando novas hipóteses, com base em, pelo menos, um dos seguintes critérios:
> [...]
> VII – baixo grau de transparência, explicabilidade e auditabilidade do sistema de inteligência artificial, que dificulte o seu controle ou supervisão.

No entanto, devido à abrangência do art. 18, surgem algumas questões importantes. Uma delas é: quem será a autoridade responsável por realizar a fiscalização desses documentos? Outro aspecto que não fica claro é o que se considera como "baixo grau" de transparência e explicabilidade. Além disso, surge a questão de qual padrão de mercado será adotado para essa avaliação de transparência e explicabilidade.

Uma faceta essencial da transparência inclui a apresentação de documentos, pois estes são a base fundamental para a ocorrência de avaliação sobre o funcionamento de sistemas, processos ou ações de investigação, e podem ser o mecanismo adequado para fornecer a evidência necessária para avaliar e analisar de forma objetiva o sistema em questão.

As medidas de transparência também aparecem de forma clara no art. 19 (Capítulo IV), que estabelece em seus incisos medidas para governança dos sistemas de IA. Conforme prescreve esse artigo, os agentes de IA estabelecerão estruturas de governança e processos internos aptos a garantir a segurança dos sistemas e o atendimento dos direitos de pessoas afetadas. Referido artigo estabelece um rol com seis medidas de governança:

> Art. 19. [...]
> I – medidas de transparência quanto ao emprego de sistemas de inteligência artificial na interação com pessoas naturais, o que inclui o uso de interfaces ser humano-máquina adequadas e suficientemente claras e informativas;
> II – transparência quanto às medidas de governança adotadas no desenvolvimento e emprego do sistema de inteligência artificial pela organização;
> III – medidas de gestão de dados adequadas para a mitigação e prevenção de potenciais vieses discriminatórios;
> IV – legitimação do tratamento de dados conforme a legislação de proteção de dados, inclusive por meio da adoção de medidas de privacidade desde a concepção e por padrão e da adoção de técnicas que minimizem o uso de dados pessoais;
> V – adoção de parâmetros adequados de separação e organização dos dados para treinamento, teste e validação dos resultados do sistema;
> VI – adoção de medidas adequadas de segurança da informação desde a concepção até a operação do sistema.

Os incisos I e II desse artigo estabelecem especificamente que as medidas de transparência deverão ser suficientemente claras e informativas quanto ao emprego de sistemas de IA na interação com pessoas naturais e quanto às medidas de governança adotadas no desenvolvimento e emprego desses sistemas.

Além do rol com seis medidas de governança estabelecidas pelo art. 19, o art. 20 determina medidas adicionais que deverão ser adotadas pelos agentes de IA que forneçam ou operem sistemas de alto risco.

Com atenção especial ao inciso I, que descreve medida relacionada à geração de transparência dos sistemas de alto risco ao estabelecer que os agentes de IA deverão documentar os detalhes a respeito do funcionamento do sistema e das decisões envolvidas em sua construção, implementação e uso, considerando todas as etapas relevantes no ciclo de vida do sistema, tais como design, desenvolvimento, avaliação, operação e descontinuação do sistema. Discussão mais aprofundada sobre este tema poderá ser encontrada no item 3.5.

As medidas de transparência também estão presentes na avaliação de impacto algorítmico prevista no art. 22. Segundo esse artigo, é obrigação dos agentes de IA, sempre que o sistema for considerado como de alto risco pela avaliação preliminar, realizar uma avaliação de impacto algorítmico.

> Art. 22. A avaliação de impacto algorítmico de sistemas de inteligência artificial é obrigação dos agentes de inteligência artificial, sempre que o sistema for considerado como de alto risco pela avaliação preliminar.
> Parágrafo único. A autoridade competente será notificada sobre o sistema de alto risco, mediante o compartilhamento das avaliações preliminar e de impacto algorítmico.

Ao examinar minuciosamente o art. 22, emerge a necessidade de questionar como será garantido que a avaliação de impacto tenha sido conduzida de maneira adequada. Quem será responsável por estabelecer os parâmetros para essa avaliação de impacto? Será incumbência das agências setoriais definir e exigir uma padronização desses impactos?

Segundo o art. 24, § 1º, a avaliação de impacto deverá registrar ao menos oito recomendações descritas nas alíneas *a*[32] a *i*, que vão desde descrever os riscos conhecidos e

[32] "Art. 24. [...] § 1º A avaliação de impacto considerará e registrará, ao menos: a) riscos conhecidos e previsíveis associados ao sistema de inteligência artificial à época em que foi desenvolvido, bem como os riscos que podem razoavelmente dele se esperar; b) benefícios associados ao sistema de inteligência artificial; c) probabilidade de consequências adversas, incluindo o número de pessoas potencialmente impactadas; d) gravidade das consequências adversas, incluindo o esforço necessário para mitigá-las; e) lógica de funcionamento do sistema de inteligência artificial; f) processo e resultado de testes e avaliações e medidas de mitigação realizadas para verificação de possíveis impactos a direitos, com especial destaque para potenciais impactos discriminatórios; g) treinamento e ações de conscientização dos riscos associados ao sistema de inteligência artificial; h) medidas de mitigação e indicação e justificação do risco residual do sistema de inteligência artificial, acompanhado de testes

previsíveis associados ao sistema até estabelecer medidas de transparência ao público a respeito dos riscos residuais, principalmente quando envolver alto grau de nocividade ou periculosidade à saúde ou segurança dos usuários.

Contudo, diante da complexidade inerente à execução da avaliação de impacto estabelecida no art. 24, suscitam-se sobre a relação entre os esforços empregados na realização dessa avaliação e os benefícios resultantes em termos de transparência.

4.2. Europa: transparência como conceito norteador no AI Act

Neste tópico, será realizada uma análise aprofundada da primeira versão da proposta de regulamentação do Parlamento Europeu sobre IA, datada de 21 de abril de 2021, juntamente com o texto compilado contendo as alterações em relação ao texto original, atualizado em 8 de maio de 2023.

A proposta europeia de 2021 estabelece nas exposições de motivos uma metodologia de análise de riscos para sistemas de IA considerados de "risco elevado", que serão assim classificados quando criarem riscos significativos para a saúde e a segurança, ou para os direitos fundamentais das pessoas. A proposta segue essa abordagem baseada no risco e impõe encargos regulamentares apenas quando é provável que um sistema de IA represente riscos elevados para os direitos fundamentais e a segurança.

Esses sistemas de IA terão de cumprir um conjunto de requisitos obrigatórios para uma IA de confiança e seguir procedimentos de avaliação da conformidade antes de poderem ser colocados no mercado da UE. A proposta busca garantir a proteção dos direitos fundamentais afetados pelo uso da IA com o objetivo de proteger direitos como dignidade humana, privacidade, proteção de dados, não discriminação e igualdade de gênero.

Por exemplo, os sistemas de IA utilizados no contexto da manutenção da ordem pública serão considerados de "risco elevado", e a exatidão, a confiabilidade e a transparência desses sistemas são especialmente importantes para evitar impactos adversos, manter a confiança do público e garantir responsabilidade e recursos eficazes, além de serem cruciais para respeitar os direitos fundamentais das pessoas envolvidas. No caso dos sistemas de IA que não são de risco elevado, apenas são impostas obrigações de transparência bastante limitadas, por

---

de controle de qualidade frequentes; i) medidas de transparência ao público, especialmente aos potenciais usuários do sistema, a respeito dos riscos residuais, principalmente quando envolver alto grau de nocividade ou periculosidade à saúde ou segurança dos usuários, nos termos dos artigos 9º e 10 da Lei nº 8.078, de 11 de setembro de 1990 (Código de Defesa do Consumidor)."

exemplo, no que diz respeito à prestação de informações para sinalizar a utilização de um sistema de IA quando este interage com seres humanos.

Em relação à opacidade de determinados sistemas de IA, que pode 61orna-los incompreensíveis ou muito complexos para os usuários, será necessário garantir um grau específico de transparência para que os usuários sejam capazes de interpretar os resultados do sistema e usá-lo adequadamente. Portanto, a proposta estabeleceu que esses sistemas devem ser acompanhados de documentação robusta, instruções de uso e informações concisas e claras, incluindo as que dispõem sobre possíveis riscos aos direitos fundamentais e discriminação, quando aplicável.

A confidencialidade das informações e dos códigos-fonte é garantida às autoridades públicas e aos organismos notificados que necessitam acessá-los para verificar o cumprimento das obrigações. A proposta busca garantir a proteção dos direitos fundamentais afetados pelo uso da IA, como dignidade humana, privacidade, proteção de dados, não discriminação e igualdade de gênero.

4.2.1. Europa: previsões sobre transparência no AI Act

Na versão inaugural da proposta de regulação sobre IA de 2021, em conjunto com o texto compilado que incorpora as modificações em relação ao conteúdo original, foi identificado um total de treze artigos que abordam a temática da transparência e/ou da explicabilidade.

Os artigos foram selecionados por apresentarem menção explícita ou implícita desses conceitos. Seis dos artigos analisados tratam explicitamente da transparência e/ou da explicabilidade (arts. 1º, 10, 13, 52, 68c e 82b), enquanto os outros sete abordam esses termos de forma implícita (arts. 9º, 11, 12, 16, 17, 61 e 64). Além disso, é importante ressaltar que a proposta europeia (“AI Act”) está dividida em onze títulos, sendo que em seis deles são encontradas referências explícitas ou implícitas sobre transparência. Essa temática é especialmente enfatizada no título referente aos Sistemas de IA de Risco Elevado (Título III), e essa distribuição pode ser observada no quadro a seguir:

Quadro 7 – Proposta de Regulamento do Parlamento Europeu

Título	Menção explícita aos termos	Menção implícita aos termos
Âmbito de Aplicação e Definições	Artigo 1º	

(Título I)		
Práticas de IA Proibidas (Título II)	Não foram encontrados resultados	
Sistemas de IA de Risco Elevado (Título III)	Capítulo 2 – Requisitos Aplicáveis a Sistemas de IA de Risco Elevado	
	Artigo 10	Artigo 9º
	Artigo 13	Artigo 11
		Artigo 12
	Capítulo 3 – Obrigações dos Fornecedores e Utilizadores de Sistemas de IA de Risco Elevado e de Outras Partes	
	Não foram encontrados resultados	Artigo 16
		Artigo 17
Obrigações de Transparência Aplicáveis a Determinados Sistemas de IA (Título IV)	Artigo 52	
Medidas de Apoio à Inovação (Título V)	Não foram encontrados resultados	
Governação e Execução (Títulos VI, VII e VIII)	Título VIII – Capítulo 3 – Execução Artigo 68c	Título VIII – Capítulo 1 – Acompanhamento Pós-Comercialização Artigo 61
Códigos de Conduta (Título IX)	Não foram encontrados resultados	Artigo 64
Disposições Finais (Títulos X, XI e XII)	Artigo 82b	Não foram encontrados resultados

Fonte: Elaborado pela autora.

O art. 1º[33], apresentado no Título I, informa que um dos objetos estabelecidos pelo regulamento será a definição das regras de transparência para certos sistemas de IA. Na UE, o uso do termo "transparência" nesse artigo não tem caráter principiológico, mas, sim, informativo, pois foi incluído entre os cinco objetos da norma[34] que a norma visa atender. A

[33] "Artigo 1.º Objeto O presente regulamento estabelece: [...] d) Regras de transparência harmonizadas determinados sistemas de IA."

[34] "Artigo 1.º Objeto O presente regulamento estabelece: a) Regras harmonizadas para a colocação no mercado, a colocação em serviço e a utilização de sistemas de inteligência artificial ('sistemas de IA') na União; b) Proibições de certas práticas de inteligência artificial; c) Requisitos específicos para sistemas de IA de risco elevado e obrigações para os operadores desses sistemas; d) **Regras de transparência harmonizadas para sistemas de IA** concebidos para interagir com pessoas singulares, sistemas de reconhecimento de emoções e sistemas de categorização biométrica, bem como para sistemas de IA usados para gerar ou manipular conteúdos de imagem, áudio ou vídeo; e) Regras relativas à fiscalização e vigilância do mercado."

alteração proposta[35] de uma versão para outra consiste em não especificar quais sistemas serão regulamentados, adotando uma nomenclatura mais genérica de determinados sistemas de IA.

No Título III, que aborda sistemas de IA de alto risco, mais especificamente no Capítulo 2, que trata dos requisitos aplicáveis aos sistemas de IA de alto risco, o art. 13 aborda a transparência e a prestação de informações aos usuários desses sistemas. Referido artigo estabelece de forma detalhada as diretrizes a serem seguidas para o cumprimento do requisito de transparência e fornecimento de informações sobre o sistema de IA de alto risco. Diante da relevância e do grau de detalhamento desse artigo, o item 2.1.3 se dedicará a analisá-lo de maneira minuciosa.

Não houve nenhuma proposta de alteração em relação ao art. 13 no texto de 8 de maio de 2023, mas foi sugerida inclusão de trecho específico sobre transparência no art. 10 (Título III, Capítulo 2). Adicionou-se ao art. 10, item 2, uma disposição que aborda a transparência na governança dos dados. O item 2 estabelece que os conjuntos de dados de treinamento, validação e teste devem estar sujeitos a práticas adequadas de governança e gestão de dados. No texto original, não havia menção específica a práticas de transparência, mas foi incluído no item (aa)[36] a adoção da "transparência quanto ao propósito original da coleta de dados".

No que se refere ao Título III, Capítulo 2, os arts. 9º, 11 e 12 dispõem sobre os requisitos aplicáveis a sistemas de IA de risco elevado. Para tanto, o art. 9º estabelece que deve ser criado, implantado, documentado e mantido um sistema de gestão de riscos em relação a sistemas de IA de risco elevado. No item 2 desse artigo é descrito de forma detalhada o que deve compreender essa documentação e quais são as etapas dessa análise. O art. 11, de forma complementar, estabelece que uma documentação técnica do sistema de IA de risco elevado deve ser elaborada com base em elementos previstos no Anexo IV (irá versar sobre disposições relativas à documentação técnica).

---

[35] Texto inicial: "d) Regras de transparência harmonizadas para sistemas de IA concebidos para interagir com pessoas singulares, sistemas de reconhecimento de emoções e sistemas de categorização biométrica, bem como para sistemas de IA usados para gerar ou manipular conteúdos de imagem, áudio ou vídeo. Proposta de alteração: d) Regras de transparência harmonizadas para sistemas de IA".

[36] "*Article 10 – Data and data governance 1. High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 as far as this is technically feasible according to the specific market segment or scope of application. Techniques that do not require labelled input data, such as unsupervised learning and reinforcement learning shall be developed on the basis of data sets such as for testing and verification that meet the quality criteria referred to in paragraphs 2 to 5. 2. Training, validation and testing data sets shall be subject to data governance appropriate for the intended purpose of the AI system. Those practices shall concern in particular, (a) the relevant design choices; (aa) transparency as regards the original purpose of data collection*."

E, por fim, preleciona o art. 12 que devem ser implementadas medidas para a manutenção dos registros atualizados enquanto o sistema estiver em funcionamento, a fim de que se possa avaliar se o sistema ainda está adequado à finalidade prevista na sua concepção, ao longo do seu ciclo de vida.

As exigências de produção de documentação técnica e os procedimentos de avaliação propostos na regulamentação estão intrinsecamente relacionados aos princípios da transparência e da explicabilidade. Ao exigir que os desenvolvedores documentem adequadamente o funcionamento, os algoritmos utilizados, os conjuntos de dados empregados e as etapas do processo de desenvolvimento, o regulador busca proporcionar maior entendimento sobre como esses sistemas tomam decisões ou realizam tarefas específicas para, por exemplo, ser viável verificar se existe alguma faceta discriminatória nesses sistemas.

Apesar de a proposta enfatizar a produção de documentação técnica e os procedimentos de avaliação, é necessário ponderar a exequibilidade e a clareza das diretrizes propostas referentes às informações que devem ser incluídas nessa documentação. Nesse contexto, uma análise minuciosa se faz relevante para avaliar se as organizações desenvolvedoras e usuárias possuem a capacidade de implementar e compreender de forma adequada tais diretrizes. Assim, torna-se crucial investigar se essas diretrizes são alcançáveis e compreensíveis, visando assegurar a qualidade e a utilidade da documentação, aspecto que será abordado de forma detalhada no Capítulo 5.

No Capítulo 13 do mesmo título são descritas as obrigações dos fornecedores e utilizadores dos sistemas de IA de alto risco. Além dos documentos e registros estabelecidos nos artigos anteriores, é obrigação do fornecedor criar um sistema de gestão da qualidade que assegure a conformidade do sistema de IA que se está fornecendo. O projeto faz distinção entre fornecedor, distribuidor e utilizador. O fornecedor é o organismo que desenvolve um sistema de IA ou que possui um sistema de IA desenvolvido com vista à sua colocação no mercado. Já o distribuidor é uma pessoa, distinta do fornecedor e do utilizador, que disponibiliza um sistema de IA no mercado. Por sua vez, o utilizador é quem utiliza, sob a sua autoridade, um sistema de IA, salvo se este for usado no âmbito de uma atividade pessoal de caráter não profissional.

São elencados no art. 17 treze aspectos que devem ser levados em conta na documentação, sob a forma de políticas, procedimentos ou instruções escritas. O aspecto que merece atenção pode ser observado no item (a) do art. 17, que aborda a adoção de procedimentos de avaliação da conformidade e de gestão de modificações do sistema de IA de risco elevado. Essa abordagem visa garantir que o desenvolvimento e a utilização desses sistemas sejam realizados de maneira responsável, levando em consideração a conformidade

com os requisitos regulatórios e a implementação de medidas para lidar com possíveis mudanças no sistema ao longo do tempo.

Além disso, a regulamentação busca verificar se os sistemas são capazes de fornecer explicações claras e compreensíveis sobre suas decisões ou comportamentos ao estabelecer requisitos de testes, auditorias e avaliações independentes.

No Título IV, que diz respeito às obrigações adicionais de transparência aplicáveis a determinados sistemas de IA, o art. 52[37] estabelece que os sistemas de IA projetados para interagir com indivíduos devem garantir que eles sejam informados de que estão interagindo com um sistema de IA. Em relação ao item 2[38], são estabelecidas diretrizes específicas para sistemas de reconhecimento de emoções e de biometria, exigindo que as pessoas expostas sejam informadas sobre o funcionamento do sistema com o qual estão interagindo.

No caso do item 3[39] do mesmo artigo, o foco está nos sistemas de IA que geram ou manipulam conteúdos de imagem, áudio ou vídeo que sejam significativamente semelhantes a pessoas, objetos ou locais reais, sendo necessário divulgar que o conteúdo foi gerado ou manipulado artificialmente. Não houve nenhuma proposta de alteração em relação ao art. 52 no texto de 8 de maio de 2023.

Além da documentação exigida antes da comercialização, mencionada no Título III, Capítulo 2, o art. 61 (Título VIII, Capítulo 2) estabelece que deve ser implementado um plano de acompanhamento pós-comercialização aplicável para os sistemas de IA de risco elevado. Esse plano deve ser documentado e precisa conter informação sobre o desempenho dos sistemas ao longo de sua vida útil, bem como permitir ao fornecedor avaliar sua conformidade. Nesse mesmo contexto, o art. 64 estabelece que deve ser garantido às autoridades de fiscalização o acesso aos conjuntos de dados de treino, validação e teste utilizados pelo fornecedor.

---

[37] "Artigo 52.º Obrigações de transparência aplicáveis a determinados sistemas de inteligência artificial 1. Os fornecedores devem assegurar que os sistemas de IA destinados a interagir com pessoas singulares sejam concebidos e desenvolvidos de maneira que as pessoas singulares sejam informadas de que estão a interagir com um sistema de IA, salvo se tal se revelar óbvio dadas as circunstâncias e o contexto de utilização. Esta obrigação não se aplica a sistemas de IA legalmente autorizados para detetar, prevenir, investigar e reprimir infrações penais, salvo se esses sistemas estiverem disponíveis ao público para denunciar uma infração penal."

[38] "2. Os utilizadores de um sistema de reconhecimento de emoções ou de um sistema de categorização biométrica devem informar sobre o funcionamento do sistema as pessoas a ele expostas. Esta obrigação não se aplica a sistemas de IA usados para categorização biométrica que sejam legalmente autorizados para detetar, prevenir e investigar infrações penais."

[39] "3. Os utilizadores de um sistema de IA que gera ou manipula conteúdos de imagem, áudio ou vídeo que sejam consideravelmente semelhantes a pessoas, objetos, locais ou outras entidades ou acontecimentos reais e que, falsamente, pareçam ser autênticos e verdadeiros a uma pessoa ('falsificação profunda') devem divulgar que o conteúdo foi gerado ou manipulado artificialmente. Contudo, o primeiro parágrafo não se aplica se a utilização for legalmente autorizada para detetar, prevenir, investigar e reprimir infrações penais ou for necessária para exercer o direito à liberdade de expressão e o direito à liberdade das artes e das ciências consagrados na Carta dos Direitos Fundamentais da UE, desde que salvaguarde adequadamente os direitos e as liberdades de terceiros."

Afora as alterações apontadas anteriormente, o texto compilado que incorporou as modificações propostas inclui os arts. 82b e 68c, que tratam da transparência e do "direito à explicação de tomada de decisão individual", respectivamente.

Foi proposta a inclusão do art. 82b ao projeto de lei, que estabelece que a Comissão Europeia desenvolverá, em consulta com o órgão fiscalizador do Regulamento de IA[40], diretrizes sobre a implementação prática do Regulamento sobre IA. Essas diretrizes tratarão, em particular, da implementação prática das obrigações de transparência estabelecidas no art. 52.

Também foi incluído o art. 68c[41], que aborda o "direito à explicação da tomada de decisão individual". Esse artigo estabelece que qualquer pessoa afetada por uma decisão tomada por meio de um sistema de IA de alto risco, que tenha efeitos legais ou impacte negativamente a saúde, a segurança, os direitos fundamentais, o bem-estar socioeconômico ou quaisquer outros direitos decorrentes das obrigações estabelecidas no Regulamento sobre IA, terá o direito de solicitar ao provedor uma explicação clara sobre o papel do sistema de IA no processo de tomada de decisão, os principais parâmetros da decisão tomada e os dados de entrada relacionados, conforme estipulado no art. 13.

[40] AI Office.

[41] "*Article 68c A right to explanation of individual decision-making Any affected person subject to a decision which is taken by the deployer on the basis of the output from an high-risk AI system which produces legal effects or similarly significantly affects him or her in a way that they consider to adversely impact their health, safety, fundamental rights, socio-economic well-being or any other of the rights deriving from the obligations laid down in this Regulation, shall have the right to request from the deployer clear and meaningful explanation pursuant to Article 13(1) on the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the related input data*."

5. BRASIL E EUROPA: ANÁLISE COMPARATIVA DOS PROJETOS DE REGULAMENTAÇÃO DE IA

Neste tópico, serão contrapostos os artigos presentes no PL 2.338/2023 e no AI Act que versam sobre (i) as medidas de governança relacionadas à transparência aplicadas ao desenvolvimento, à comercialização e à pós-comercialização de sistemas de IA; e (ii) os direitos dos titulares em relação à explicação dos sistemas. Será feita uma análise comparativa das semelhanças e diferenças entre essas duas propostas, com o propósito de entender as abordagens adotadas por cada uma delas.

Posteriormente, será realizada a análise dos principais artigos referenciados no AI Act e no PL 2.338/2023, a fim de se verificar se foi dada a devida consideração à opacidade inerente à técnica de aprendizado de máquina baseado em redes neurais, a qual pode limitar a transparência e dificultar a explicação e interpretação completa das decisões tomadas por esses sistemas.

5.1. Das medidas de governança dos sistemas de IA

Existem iniciativas isoladas denominadas “Responsible AI”, lideradas por instituições como ONU, OCDE, fóruns empresariais, governos, que têm como objetivo definir boas práticas e viabilizar técnicas de explicação de algoritmos, de auditabilidade, de prestação de contas ou, ainda, de mitigação de vieses negativos, reunindo essas estratégias por meio de *frameworks* de governança para o design, o desenvolvimento e a implantação de soluções de IA (ZAVAGLIA, 2023).

Esse movimento se propõe a lidar com a complexidade da tecnologia por meio da elaboração de documentação do processo de desenvolvimento e avaliação da performance dos modelos.

> Isso permite interpretar os detalhes da construção das soluções e avaliar como se comportam em determinadas situações e quais seus resultados (*outcomes*), o que, inclusive, permite uma nova linha de pesquisa e desenvolvimento (P&D) por meio de uma engenharia reversa para analisar as variáveis explicativas mais pontuadas, o que aumenta a transparência e melhor entendimento sobre os ajustes necessários e as escolhas para mitigar impactos negativos (ZAVAGLIA, 2023, p.13).

Segundo Zavaglia (2023), o caminho para a transparência e genuíno sentido de explicabilidade não compreende o *black box*, o funcionamento das redes neurais e suas milhares

de correlações matemáticas e estatísticas, mas as etapas, os dados, as escolhas, as variáveis mais importantes e os resultados esperados (testes) e alcançados (monitoramento).

No entanto, a opacidade das redes neurais estabelece limites para a governança, uma vez que nem todas as correlações matemáticas e estatísticas podem ser totalmente compreendidas. Apesar disso, a busca por uma governança informada e transparente continua sendo crucial para identificar e mitigar, por exemplo, possíveis fontes de viés discriminatório e outros problemas éticos associados à utilização de IA em diversas áreas, incluindo saúde, justiça e segurança pública.

Diante desse cenário, é imprescindível o desenvolvimento de sistemas a partir de uma estrutura de governança adequada e de um programa de gerenciamento de riscos capaz de integrar os desafios técnicos com os aspectos sociais.

> Por isso, é preciso analisar todas as etapas e impactos antes da colocação no mercado de um produto ou serviço suportado por IA, não apenas o tipo de tecnologia ou a visão de negócios. Organizações públicas e privadas devem analisar a tecnologia existente (*narrow AI*) e suas características para implantar programas de gestão de risco que funcionem de verdade no contexto atual, com aplicabilidade prática para os desafios relacionados a dilema éticos e jurídicos, sem perder a noção de futuro e de todo o potencial dessa tecnologia (ZAVAGLIA, 2023, p.13).

O *Policy Paper* sobre *Regulação de Inteligência Artificial no Brasil*[42] destacou que, atualmente, não existem procedimentos de documentação padronizados destinados a divulgar as características de desempenho de sistemas de IA. Diante dessa lacuna, é ressaltada a necessidade de adoção de procedimentos com o intuito de esclarecer pontos-chave do funcionamento do algoritmo, bem como os usos pretendidos para determinado modelo de aprendizado de máquina baseado em redes neurais.

Importante notar, no entanto, que a governança da IA e a regulamentação da IA desempenham papéis complementares no avanço tecnológico. Enquanto a governança se baseia em boas práticas e princípios voluntários para orientar o desenvolvimento ético da IA, a regulamentação impõe orientações e dispositivos vinculantes que visam garantir o uso responsável e seguro da tecnologia.

Enquanto a governança busca fomentar boas práticas, a regulamentação da IA assume um papel estruturado e normativo. Nesse sentido, a ausência de procedimentos padronizados para divulgar as características de desempenho e documentação de desenvolvimento dos

[42] Contribuição do Centro de Tecnologia e Sociedade (CTS) – Fundação Getulio Vargas (FGV Direito Rio) à Consulta Pública do Ministério da Ciência Tecnologia Inovações e Comunicações – MCTIC sobre a Estratégia Brasileira de Inteligência Artificial. Disponível em: file:///Users/leticia.serec/Downloads/policypaperiaegoverno.pdf.

sistemas de IA destaca a necessidade de adoção de direcionamento para o fornecimento de documentação detalhada sobre o funcionamento e o uso pretendido do sistema.

De forma a responder à necessidade de promover a transparência no desenvolvimento e uso de sistemas de IA, por meio de procedimentos de documentação padronizados, tanto o AI Act quanto o PL 2.338/2023 propõem um arcabouço regulatório para sistematizar as medidas de governança que devem ser adotadas antes, durante e após a comercialização desses sistemas. O Quadro 8, a seguir, apresenta os treze artigos encontrados no AI Act e no PL 2.338/2023 referentes à classificação dos sistemas de IA e às medidas de governança: (i) no AI Act, foram encontrados sete artigos no Título III, intitulado "Sistemas de IA de Risco Elevado", Capítulos 2 e 3, além de um artigo adicional estar presente no Título VIII, intitulado "Acompanhamento Pós-Comercialização, Partilha de Informações, Fiscalização do Mercado", Capítulo 1; (ii) no PL 2.338/2023, foram encontrados cinco artigos relacionados a esse tópico no Capítulo IV, que trata especificamente da "Governança dos Sistemas de Inteligência Artificial".

Quadro 8 – Artigos referentes à classificação dos sistemas de IA e medidas de governança

<table>
<tr><th colspan="2">Proposta de Regulamento do Parlamento Europeu</th><th colspan="2">Proposta de Regulamento Brasileira</th></tr>
<tr><td rowspan="2">Sistemas de IA de Risco Elevado (Título III)</td><td>Capítulo 2 – Requisitos Aplicáveis a Sistemas de IA de Risco Elevado
Artigos 9º, 10, 11, 12 e 13</td><td rowspan="2">Capítulo IV – Da Governança dos Sistemas de Inteligência Artificial</td><td rowspan="2">Artigos 20, 22 e 24</td></tr>
<tr><td>Capítulo 3 – Obrigações dos Fornecedores e Utilizadores de Sistemas de IA de Risco Elevado e de Outras Partes
Artigos 16 e 17</td></tr>
<tr><td>Título VIII</td><td>Capítulo 1 – Acompanhamento Pós-Comercialização
Artigo 61</td><td>Artigos 13 e 19</td></tr>
</table>

Fonte: Elaborado pela autora.

Antes de prosseguir com a análise sobre governança, é importante observar que o regulamento europeu segue uma abordagem baseada no risco e diferencia as utilizações de IA entre as que criam: (i) um risco inaceitável; (ii) um risco elevado; e (iii) um risco baixo ou mínimo. O PL brasileiro, por sua vez, adotou abordagem similar com uma nomenclatura um pouco distinta sobre o uso e a implementação de sistemas de IA: (i) risco excessivo; (ii) alto risco; e (iii) risco baixo. A lei não faz menção expressa à última nomenclatura, adotando o item iii no caso de o sistema de IA não ser classificado como de risco excessivo ou alto. A fim de

fornecer melhor entendimento sobre as classificações de risco propostas, apresenta-se a seguir um quadro comparativo com a conceituação das categorias de risco para as duas propostas em questão.

Quadro 9 – Conceituação das categorias de risco

AI Act	
i) Um risco inaceitável	O Título II inclui todos os sistemas de IA cuja utilização seja considerada inaceitável por violar os valores[43] da União, como, por exemplo, os direitos fundamentais.
ii) Um risco elevado	O sistema de IA é considerado de risco elevado quando estejam satisfeitas ambas as condições que se seguem: a) o sistema de IA destina-se a ser utilizado como um componente de segurança de um produto ou é, ele próprio, um produto abrangido pela legislação de harmonização da União enumerada no Anexo II; b) nos termos da legislação de harmonização da União enumerada no Anexo II, o produto cujo componente de segurança é o sistema de IA, ou o próprio sistema de IA enquanto produto deve ser sujeito a uma avaliação da conformidade por terceiros com vista à colocação no mercado ou à colocação em serviço (art. 6º, item 1). Além dos sistemas de IA de risco elevado referidos anteriormente, os sistemas de IA referidos no Anexo III são também considerados de risco elevado (art. 6º, item 2).
iii) Um risco baixo ou mínimo	Sistemas que não foram considerados de risco inaceitável ou elevado.
PL 2.338/2023	
i) Risco excessivo	Sistemas de IA que empreguem técnicas subliminares que tenham por objetivo ou por efeito induzir a pessoa natural a se comportar de forma prejudicial ou perigosa à sua saúde ou segurança ou contra os fundamentos desta lei; Sistemas de IA que explorem quaisquer vulnerabilidades de grupos específicos de pessoas naturais, tais como associadas à sua idade ou deficiência física ou mental, de modo a induzi-las a se comportar de maneira prejudicial à sua saúde ou segurança ou contra os fundamentos desta lei; Sistemas de IA utilizados pelo poder público, para avaliar, classificar ou ranquear as pessoas naturais, com base no seu comportamento social ou em atributos da sua personalidade, por meio de pontuação universal, para o acesso a bens e serviços e políticas públicas, de forma ilegítima ou desproporcional (art. 14, I, II e III).
ii) Alto risco	O art. 17[44] descreve as quatorze finalidades específicas que determinam se um sistema de IA deve ser considerado de alto risco, por exemplo, sistemas de IA para aplicações

[43] As proibições abrangem práticas com potencial significativo para manipular as pessoas por meio de técnicas subliminares que lhes passam despercebidas ou explorar as vulnerabilidades de grupos específicos, como as crianças ou as pessoas com deficiência, para distorcer substancialmente o seu comportamento de uma forma que seja suscetível de causar danos psicológicos ou físicos a essa ou a outra pessoa (COMISSÃO EUROPEIA, 2020).

[44] "Art. 17. São considerados sistemas de inteligência artificial de alto risco aqueles utilizados para as seguintes finalidades: I – aplicação como dispositivos de segurança na gestão e no funcionamento de infraestruturas críticas, tais como controle de trânsito e redes de abastecimento de água e de eletricidade; II – educação e formação profissional, incluindo sistemas de determinação de acesso a instituições de ensino e de formação profissional ou para avaliação e monitoramento de estudantes; III – recrutamento, triagem, filtragem, avaliação de candidatos, tomada de decisões sobre promoções ou cessações de relações contratuais de trabalho, repartição de tarefas e controle e avaliação do desempenho e do comportamento das pessoas afetadas por tais aplicações de inteligência

	na área da saúde, inclusive as destinadas a auxiliar diagnósticos e procedimentos médicos e em sistemas biométricos de identificação.

Fonte: Elaborado pela autora.

A principal diferença entre a metodologia de classificação de risco europeia e a brasileira está na abordagem adotada para determinar quais sistemas de IA são considerados de alto risco. No caso da metodologia europeia, um sistema de IA é classificado como de risco elevado quando atende a duas condições principais elencadas no artigo, além dos sistemas de IA listados no Anexo III como de risco elevado.

Já a metodologia brasileira abrange uma variedade de finalidades específicas que indicam o alto risco associado ao uso desses sistemas em determinados setores, como saúde e identificação biométrica. Segundo manifestação em apoio ao PL 2.338/2023 da Coalizão Direitos na Rede[45], essa abordagem é relevante pois:

> A partir de uma abordagem regulatória baseada em riscos e em direitos (*risks and rights-based approach*), o texto do PL cria uma regulação assimétrica dos agentes regulados, com obrigações mais ou menos fortes de acordo com o nível de risco do sistema de IA, o que será determinado a partir de uma avaliação preliminar. Assim o PL estabelece direitos e medidas de governança básicos deflagrados por toda ferramenta de IA, mas também cria certos direitos e obrigações específicos para os casos potencialmente mais arriscados. Ao mesmo tempo, o projeto define que as medidas de governança dos sistemas de IA devem ser aplicadas ao longo de todo o seu ciclo de vida (desde a concepção até o seu encerramento/descontinuação) (DIREITOS NA REDE, 2023).

artificial nas áreas de emprego, gestão de trabalhadores e acesso ao emprego por conta própria; IV – avaliação de critérios de acesso, eligibilidade, concessão, revisão, redução ou revogação de serviços privados e públicos que sejam considerados essenciais, incluindo sistemas utilizados para avaliar a elegibilidade de pessoas naturais quanto a prestações de serviços públicos de assistência e de seguridade; Coordenação de Comissões Especiais, Temporárias e Parlamentares de Inquérito V – avaliação da capacidade de endividamento das pessoas naturais ou estabelecimento de sua classificação de crédito; VI – envio ou estabelecimento de prioridades para serviços de resposta a emergências, incluindo bombeiros e assistência médica; VII – administração da justiça, incluindo sistemas que auxiliem autoridades judiciárias na investigação dos fatos e na aplicação da lei; VIII – veículos autônomos, quando seu uso puder gerar riscos à integridade física de pessoas; IX – aplicações na área da saúde, inclusive as destinadas a auxiliar diagnósticos e procedimentos médicos; X – sistemas biométricos de identificação; XI – investigação criminal e segurança pública, em especial para avaliações individuais de riscos pelas autoridades competentes, a fim de determinar o risco de uma pessoa cometer infrações ou de reincidir, ou o risco para potenciais vítimas de infrações penais ou para avaliar os traços de personalidade e as características ou o comportamento criminal passado de pessoas singulares ou grupos; XII – estudo analítico de crimes relativos a pessoas naturais, permitindo às autoridades policiais pesquisar grandes conjuntos de dados complexos, relacionados ou não relacionados, disponíveis em diferentes fontes de dados ou em diferentes formatos de dados, no intuito de identificar padrões desconhecidos ou descobrir relações escondidas nos dados; XIII – investigação por autoridades administrativas para avaliar a credibilidade dos elementos de prova no decurso da investigação ou repressão de infrações, para prever a ocorrência ou a recorrência de uma infração real ou potencial com base na definição de perfis de pessoas singulares; XIV – gestão da migração e controle de fronteiras."

[45] DIREITOS NA REDE. Carta de apoio ao PL nº 2338/23/2023. Disponível em: https://direitosnarede.org.br/2023/06/14/carta-de-apoio-ao-pl-2338-2023/. Acesso em: 14 jun. 2023.

Retomando a análise, no caso do PL 2.338/2023, todos os sistemas, independentemente de sua classificação de risco, deverão adotar as medidas estabelecidas nos arts. 13 e 19 como medidas de governança e transparência envolvidas na categorização dos riscos e sua documentação. Essas medidas de governança se aplicam durante todo o ciclo de vida de todos os sistemas de IA, independente de sua classificação.

Nesse contexto, o art. 13 do PL 2.338/2023 estabelece que todo sistema de IA deve passar por uma avaliação preliminar realizada pelo fornecedor antes de ser colocado no mercado. Essa avaliação tem como objetivo classificar o nível de risco do sistema e deve incluir as finalidades ou aplicações desses sistemas. Além disso, é necessário registrar e documentar essa avaliação para fins de prestação de contas caso o sistema de IA não seja classificado como de alto risco.

> Avaliações, relatórios e diagnósticos de impacto são instrumentos que têm ganhado cada vez mais importância em uma sociedade na qual as ações humanas e empresariais podem provocar riscos de difícil ou impossível reparação. Atualmente, no Brasil existem ao menos três avaliações de impacto setoriais definidas por lei, sendo elas (i) a avaliação de impacto ambiental, (ii) a avaliação de impacto regulatório e (iii) o relatório de impacto à proteção de dados (RIPD). No campo da IA, a Avaliação de Impacto de Inteligência Artificial (AIIA) é vista como um instrumento de governança que possibilita ao desenvolvedor ou aplicador da tecnologia identificar e reduzir possíveis riscos que determinado sistema de IA possa causar aos direitos e liberdades fundamentais (LAPIN, 2022).

O art. 19 do PL 2.338/2023 institui que os agentes de IA devem estabelecer estruturas de governança e processos internos para garantir a segurança dos sistemas e o respeito aos direitos das pessoas afetadas. Isso inclui as seguintes observâncias:

- medidas de transparência na interação com os usuários desses sistemas;
- transparência quanto às medidas de governança adequadas para evitar discriminação;
- conformidade com a legislação de proteção de dados;
- separação e organização adequada dos dados para treinamento, teste e validação;
- adoção de medidas de segurança da informação desde a concepção até o encerramento do sistema.

O PL 2.338/2023 estabelece nos arts. 20 e 24 as medidas de governança e os processos internos que deverão ser adotados pelos agentes de IA que forneçam ou operem sistemas de alto risco. São eles:

- documentação a respeito do funcionamento do sistema e das decisões envolvidas em sua construção, implementação e uso, considerando todas as etapas relevantes no ciclo de vida do sistema;
- adoção de medidas técnicas para viabilizar a explicabilidade dos resultados dos sistemas de IA;
- avaliação de impacto algorítmico de sistemas de IA.

O AI Act não apresentou qualquer previsão relativa à necessidade de documentação ou medidas de governança para sistemas de baixo risco; apenas os sistemas de IA classificados como de risco elevado devem cumprir os requisitos previstos. E estabelece que, antes de serem disponibilizados e comercializados no mercado, esses sistemas deverão contar com os seguintes documentos:

- documentação e estruturação de sistema de gestão de riscos em relação a sistemas de IA de risco elevado (art. 9º);
- documentação técnica elaborada antes da colocação no mercado ou colocação em serviço desse sistema (art. 11 e Anexo IV);
- instruções de utilização que incluam informações concisas, completas, corretas e claras que sejam pertinentes, acessíveis e compreensíveis para os utilizadores (art. 13);
- registro automático de eventos enquanto o sistema de IA de risco elevado estiver em funcionamento (art. 12);
- criação de sistema de gestão da qualidade que assegure a conformidade com o presente regulamento (art. 17);
- criação e documentação de um sistema de acompanhamento pós-comercialização (art. 61).

Conclui-se, dessa forma, que a distinção entre os projetos está na ênfase mais focalizada do AI Act na documentação e governança dos sistemas de IA de alto risco, enquanto a proposta brasileira busca abordar a governança e a transparência de forma mais abrangente, incluindo todos os sistemas de IA, independentemente de sua classificação de risco.

Em relação ao aspecto convergente de ambos os projetos, foi estabelecido o dever de informar uma lista extensa de informações detalhadas sobre a concepção do sistema de IA de alto risco. De forma alinhada ao posicionamento apresentado pelo *Policy Paper* sobre

Regulação de Inteligência Artificial no Brasil produzido pela Fundação Getulio Vargas (FGV Direito Rio, 2022), destaca-se que uma das abordagens propostas por pesquisadores da área é a utilização de "cartões de modelo", documentos curtos e objetivos com a finalidade de detalhar o modelo proposto.

> A recomendação é de que os modelos postos em produção sejam acompanhados dessa documentação detalhando o contexto no qual o algoritmo pretende atuar, suas características de desempenho, procedimentos de avaliação de desempenho, bem como outras informações relevantes, como métricas que capturam considerações sobre vieses, equidade e inclusão. Os denominados "cartões de modelo" devem ser utilizados conjuntamente às já conhecidas "Fichas Técnicas para Bases de Dados" (*Datasheets for Datasets*), que revelam detalhes acerca do conjunto de dados utilizado para treinar e testar os modelos de *machine learning*. Enquanto as "fichas" focam nas características dos dados utilizados para alimentar o modelo, os "cartões" focam nas características de treinamento do modelo, como o tipo de modelo, os usos pretendidos, as informações sobre atributos para os quais o desempenho do modelo pode variar e as medidas de desempenho do modelo. [...] Esses procedimentos têm a finalidade de aumentar a transparência sobre como determinado sistema de IA funciona e, assim, minimizar o uso de algoritmos em contextos ou finalidades para os quais aquele modelo não é adequado (CTS-FGV Rio, 2023, p. 17).

No AI Act, essas obrigações de informação estão definidas no Anexo IV, que inclui detalhes sobre o propósito do sistema, suas características técnicas, métodos utilizados, capacidades e limitações, além de informações sobre dados utilizados e potenciais impactos nos direitos fundamentais. No caso do PL 2.338/2023, essas obrigações estão previstas no art. 20, o que envolve especificações técnicas, métodos utilizados, critérios e parâmetros adotados, bem como informações sobre os dados utilizados e as etapas de treinamento do sistema.

5.1.1. Análise dos artigos sob a perspectiva da opacidade

Diante da relevância da classificação dos sistemas de IA e das medidas de governança para alcançar a transparência e explicabilidade destes, este tópico busca verificar se a opacidade inerente às técnicas de aprendizado de máquina baseado em redes neurais foi devidamente considerada na redação dos principais artigos mencionados no AI Act e no PL 2.338/2023, sendo que o foco será direcionado para a avaliação do Anexo IV do AI Act e do art. 20 do PL 2.338/2023.

No PL 2.338/2023 será analisado de forma detalhada o art. 20, I, II, IV e V, que versa sobre as medidas de governança que devem ser adotadas pelos agentes de IA diante de sistemas de alto risco.

De acordo com o que foi defendido no Capítulo 2, a opacidade dos sistemas em análise apresenta desafios significativos em termos de compreensão de como eles chegam nos seus

resultados. O inciso I do art. 20[46] é um dos trechos mais relevantes sobre documentação apresentado no projeto, pois exige que sejam documentadas todas as etapas relevantes no ciclo de vida do sistema, as fases sobre o seu funcionamento e as decisões envolvidas em sua construção, implementação e uso. Além disso, um aspecto de extrema importância é que o inciso estabelece explicitamente a necessidade de a documentação ser em formato adequado ao processo de desenvolvimento e à tecnologia usada. Isso porque deverão ser levadas em conta informações relevantes que podem variar de um sistema para outro e impactar o tipo de avaliação e o grau de transparência viável de ser fornecido, observando-se as características intrínsecas a cada tipo de sistema, conforme tópico abordado no Capítulo 2.

Já em relação ao inciso II do art. 20[47], a redação apresenta falta de clareza ao mencionar os "registros automáticos", não especificando exatamente sua natureza, além de não ficar evidente como esses registros permitirão avaliar a acurácia e a robustez do sistema em questão. O texto também não leva em consideração as particularidades e os desafios que podem surgir relacionados à tecnologia e ao sistema em uso. No caso de um sistema de IA de aprendizado de máquina baseado em redes neurais, por exemplo, para avaliação dos efeitos adversos, é necessário considerar as limitações decorrentes da opacidade do sistema, uma questão que não foi abordada no texto.

Na alínea *a* do inciso IV do art. 20 são estabelecidos critérios relacionados à avaliação dos dados para evitar a incorporação de vieses negativos sociais estruturais que possam ser perpetuados e ampliados pela tecnologia. Importante notar que o texto levou em conta um aspecto relevante do desenvolvimento desse sistema, que é a necessidade da gestão dos dados que serão usados nesse sistema, aspecto que, conforme mencionado no Capítulo 2, item 2.4, pode ser compartilhado, pois é um atributo transparente sobre sistema. Além disso, o compartilhamento dessa informação tende a facilitar a identificação e a correção de possíveis vieses negativos gerados pelo sistema.

O art. 20, V, estabelece a adoção de medidas técnicas para viabilizar a explicabilidade. O dispositivo menciona a adoção de três medidas distintas em um mesmo inciso. A primeira delas trata da implementação de medidas técnicas para garantir a explicabilidade dos resultados

[46] "Art. 20. Além das medidas indicadas no art. 19, os agentes de inteligência artificial que forneçam ou operem sistemas de alto risco adotarão as seguintes medidas de governança e processos internos: I – documentação, no formato adequado ao processo de desenvolvimento e à tecnologia usada, a respeito do funcionamento do sistema e das decisões envolvidas em sua construção, implementação e uso, considerando todas as etapas relevantes no ciclo de vida do sistema, tais como estágio de design, de desenvolvimento, de avaliação, de operação e de descontinuação do sistema."

[47] "Art. 20. [...] II – uso de ferramentas de registro automático da operação do sistema, de modo a permitir a avaliação de sua acurácia e robustez e a apurar potenciais discriminatórios, e implementação das medidas de mitigação de riscos adotadas, com especial atenção para efeitos adversos."

dos sistemas de IA. O texto sugere que a adoção de medidas técnicas é uma abordagem viável para promover a explicabilidade dos resultados do sistema. Nem sempre a simples divulgação das informações sobre como o sistema chegou a determinada conclusão é a solução mais adequada, portanto, o uso da expressão "medidas técnicas" no inciso pode ser uma abordagem apropriada, pois abrange diferentes possibilidades para promover uma maior compreensão do sistema. No entanto, o texto não deixa claro que tipo de abordagens seriam essas, o que ficaria aberto para interpretação.

A segunda medida envolve fornecer informações gerais sobre o funcionamento do modelo de IA utilizado, aspecto que, tecnicamente, conforme mencionado no Capítulo 2, item 2.4, pode ser viável para o desenvolvedor disponibilizar essas informações, pois se referem a dados intrínsecos à estruturação e criação do sistema.

Por fim, há a medida de fornecer informações que permitam a interpretação dos resultados concretos produzidos pelo sistema. Nesse ponto, não foi considerada a opacidade presente em algumas técnicas de IA, o que pode impedir que o agente de IA consiga fornecer uma explicação precisa para cada resultado gerado pelo sistema. Em outras palavras, pode haver situações em que a natureza das técnicas utilizadas torne tecnicamente difícil oferecer uma explicação completa para cada resultado produzido pelo sistema.

Em relação ao AI Act, o art. 11 direciona a documentação técnica ao Anexo IV, no qual são especificados os elementos que devem ser incluídos nessa documentação.

A redação do item 1[48] estabelece dois aspectos centrais que devem ser considerados na produção da documentação sobre o sistema de IA: a natureza das tecnologias de IA e os riscos associados a sistemas de alto risco. O enfoque dessa análise será sobre a natureza das tecnologias de IA, pois o texto acertadamente destaca a importância de considerar essa questão

[48] "A documentação técnica referida no Artigo 11 deve conter, no mínimo, as seguintes informações, conforme aplicáveis ao sistema de IA relevante. Uma descrição geral do sistema de IA, incluindo: (a) seu propósito pretendido, o nome do provedor e a versão do sistema refletindo sua relação com versões anteriores e, quando aplicável, mais recentes, na sucessão de revisões; (aa) a natureza dos dados prováveis ou pretendidos a serem processados pelo sistema e, no caso de dados pessoais, as categorias de pessoas naturais e grupos prováveis ou pretendidos a serem afetados; (b) como o sistema de IA pode interagir ou ser utilizado para interagir com hardware ou software, incluindo outros sistemas de IA, que não fazem parte do próprio sistema de IA, quando aplicável; (c) as versões do software ou firmware relevantes e, quando aplicável, informações para o implantador sobre quaisquer requisitos relacionados à atualização de versão; (d) a descrição das várias configurações e variantes do sistema de IA que se destinam a ser colocadas no mercado ou colocadas em serviço; (e) a descrição do hardware em que o sistema de IA se destina a ser executado; (f) quando o sistema de IA é um componente de produtos, fotografias ou ilustrações mostrando características externas, marcação e disposição interna desses produtos; (fa) a descrição da interface do implantador; (g) instruções de uso para o implantador, de acordo com o Artigo 13(2) e (3), bem como 14(4)(e), e, quando aplicável, instruções de instalação; (ga) uma descrição detalhada e facilmente compreensível do principal objetivo ou objetivos de otimização do sistema; (gb) uma descrição detalhada e facilmente compreensível da saída esperada do sistema e da qualidade esperada da saída; (gc) instruções detalhadas e facilmente compreensíveis para interpretar a saída do sistema; (gd) exemplos de cenários para os quais o sistema não deve ser usado."

na produção da documentação. Nesse sentido, é importante reconhecer que a documentação de um sistema de IA baseado em aprendizado de máquina da técnica de redes neurais profundas será diferente da documentação de um sistema que utiliza outra técnica em que a opacidade não é uma realidade. A opacidade desses sistemas apresenta desafios adicionais na produção de documentação, uma vez que a explicação detalhada de como uma decisão foi alcançada pode ser inviável de ser atendida. Portanto, será essencial que o operador do sistema adapte a documentação de acordo com a tecnologia empregada, considerando a presença ou a ausência de opacidade, para fornecer informações adequadas e relevantes para a compreensão e avaliação do sistema.

O item 2 do Anexo IV do AI Act, [49] em questão se refere ao detalhamento dos métodos e etapas utilizados no desenvolvimento do sistema de IA. Essa abordagem é relevante, pois fornece informações essenciais para compreender a lógica subjacente à concepção do sistema. Compreender as etapas do desenvolvimento permite uma análise mais aprofundada sobre o funcionamento do sistema de IA, possibilitando a busca de esclarecimentos, quando necessário, para uma melhor compreensão dos resultados obtidos. No entanto, é importante reconhecer que a transparência pode ser limitada em determinadas técnicas, o que requer uma compreensão cuidadosa dessas limitações ao interpretar os resultados.

O tópico (b) do item 2[50] do Anexo IV do AI Act é uma das seções mais técnicas do AI Act, no qual são listadas as informações que devem ser incluídas na documentação técnica sobre a concepção do produto. Um adendo deve ser feito em relação ao "detalhamento dos componentes e interfaces" dos algoritmos e estruturas de dados, incluindo informações de como eles se relacionam entre si, pois é importante observar que certos sistemas de IA podem ser opacos e utilizar técnicas que dificultam a divulgação dessas informações. Além disso, não há um consenso doutrinário claro sobre quais informações devem ser fornecidas para garantir a transparência quanto às decisões tomadas pelo produto ou serviço. Isso significa que os operadores de IA podem enfrentar desafios significativos ao tentar fornecer essas informações, sem necessariamente esclarecer de forma abrangente sistemas complexos e opacos.

---

[49] "Uma descrição detalhada dos elementos do sistema de IA e do processo de desenvolvimento, incluindo: (a) os métodos e etapas realizados para o desenvolvimento do sistema de IA, incluindo, quando relevante, o uso de sistemas pré-treinados ou ferramentas fornecidas por terceiros e como esses foram usados, integrados ou modificados pelo provedor."

[50] "(b) uma descrição da arquitetura, especificações de design, algoritmos e estruturas de dados, incluindo um detalhamento de seus componentes e interfaces, como eles se relacionam entre si e como eles fornecem o processamento geral ou a lógica do sistema de IA; as principais escolhas de design, incluindo a justificativa e suposições feitas, também em relação a pessoas ou grupos de pessoas em que o sistema se destina a ser usado; as principais escolhas de classificação; para o que o sistema é projetado para otimizar e a relevância dos diferentes parâmetros; as decisões sobre qualquer possível compensação feita em relação às soluções técnicas adotadas para cumprir os requisitos estabelecidos no Título III, Capítulo 2."

Também no item 2, tópico (d), destaca-se o fornecimento de informações detalhadas sobre a metodologia, a técnica de treinamento e o conjunto de dados utilizados. Essas informações podem ser benéficas para a avaliação do sistema, pois a qualidade dos registros utilizados pode fornecer uma compreensão das saídas dos sistemas, levando em consideração as limitações resultantes da transparência, dependendo da técnica do sistema em análise. No entanto, é importante reconhecer que a transparência pode ser limitada em determinadas técnicas, o que requer uma compreensão cuidadosa dessas limitações ao interpretar os resultados. Além disso, ao analisar essas informações de maneira mais detalhada, se poderá avaliar a presença de possíveis vieses negativos discriminatórios e identificar dados que possam causar danos se não forem devidamente segmentados. Portanto, o fornecimento transparente dessas informações pode contribuir para uma análise mais precisa e uma abordagem mais justa e segura na utilização dos sistemas.

5.2. Direito dos titulares à explicação

A norma referente ao direito à explicação não é inovadora no Brasil, pois a Lei 12.414/2011 (Lei de Cadastro Positivo) contemplou, entre os direitos do cadastrado, o de solicitar ao consulente a revisão de decisão realizada exclusivamente por meios automatizados (art. 5º, VI). O direito do titular dos dados pessoais à explicação das decisões tomadas unicamente com base em tratamento automatizado (direito à explicação) também já foi disciplinado na ordem jurídica nacional pela Lei Geral de Proteção de Dados (LGDP).

A inclusão desse direito no conjunto de leis relativas à proteção de dados pessoais expandiu o âmbito de aplicação. Conforme prelecionam Lima e Sá (2020), a definição "decisões tomadas unicamente com base em tratamento automatizado" revela a amplitude da expressão, que não se limita aos casos de decisão por sistemas de IA. Mesmo que o art. 20 da LGPD não faça menção expressa à IA, não há como deixar de visualizar, pelo menos, o início da regulação de seu uso.

> O direito à explicação, nos moldes do art. 20, é uma consequência do princípio da transparência, previsto no art. 6º, VI da LGPD. O *caput* do art. 20 confere ao titular dos dados o direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais, desde que afetem seus interesses. Estão aí incluídas as decisões destinadas a definir o perfil pessoal, profissional, de consumo e de crédito ou os aspectos da personalidade do titular (LIMA; SÁ, 2020, p. 24).

A disposição contida no art. 20 da LGPD pode se revelar limitada, uma vez que restringe o direito à explicação apenas às decisões totalmente automatizadas.

> Assim, as decisões que forem o resultado simultâneo da automação e da decisão humana não são alcançadas pela previsão normativa. Não há como ignorar, nos dias atuais, os processos decisórios complexos nos quais algumas fases são automatizadas e outras são implementadas com decisões puramente humanas. Tais processos decisórios são merecedores de igual cobertura legal (LIMA; SÁ, 2020, p. 24).

Além disso, no § 1º do art. 20 da LGPD é atribuído ao controlador o dever de fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.

> Conquanto já seja um avanço legislativo, os parâmetros legais para o exercício do direito à explicação podem não ser suficientes para assegurar a autonomia informativa do titular dos dados pessoais e para concretizar a principiologia sistematizada no art. 6º, em especial, os princípios do livre acesso (inc. IV) da transparência (inc. VI) e da não discriminação (inc. IX) (LIMA; SÁ, 2020, p. 24).

Diante da discussão já estabelecida sobre o direito à explicação previsto na LGPD, o PL 2.338/2023 estabelece uma série de disposições específicas para abordar esse tema. Assim, esse PL deve ser à luz da Constituição Federal e dos direitos fundamentais dos indivíduos.

> A existência de múltiplos microssistemas jurídicos (consumidor, idoso, pessoa com deficiência, criança e adolescente, entre outros) traz o desafio peculiar ao intérprete de aplicar o direito, segundo uma visão constitucionalizada e não fragmentada do sistema jurídico: Na mesma direção, a proliferação de estatutos protetivos no Brasil (Código de Proteção e Defesa do Consumidor, Estatuto da Criança e do Adolescente, Estatuto do Idoso etc.) não deve ser encarada como a oportunidade para a construção de novos "guetos" doutrinários, ancorados em "lógicas próprias", permeadas por princípios e conceitos setoriais. A unidade do ordenamento centrado sobre a Constituição da República impõe que as normas especiais se insiram no sistema jurídico unitário, atendendo aos conceitos gerais que o embasam, evitando-se o desenvolvimento de uma terminologia setorial e, consequentemente, de uma hermenêutica setorial, distinta daquela aplicada à ordem jurídica em sua totalidade. Aos chamados Estatutos não compete, portanto, reescrever as noções fundantes do sistema jurídico ou desenhar seus princípios próprios e autônomos, mas realizar o projeto constitucional em dado campo específico, sempre atendendo à necessidade de preservação do caráter sistêmico da ordem jurídica (SCHREIBER; KONDER, 2016, p. 45).

Neste sentido, será realizada uma análise comparativa entre o PL 2.338/2023 e o AI Act, com o objetivo de examinar e comparar suas abordagens em relação ao direito dos titulares à explicação.

O Quadro 10, a seguir, apresenta os cinco artigos encontrados no AI Act e no PL 2.338/2023 referentes ao direito dos titulares à explicação: (i) no AI Act foram encontrados dois artigos nos Títulos IV e VIII – intitulados "Obrigações de transparência aplicáveis a determinados sistemas de Inteligência Artificial" e "Acompanhamento pós-comercialização,

partilha de informações, fiscalização do mercado"; (ii) no PL 2.338/2023 foram encontrados três artigos relacionados a esse tópico no Capítulo II, que trata especificamente "Dos Direitos".

Quadro 10 – Artigos referentes ao direito dos titulares à explicação

Proposta de Regulamento do Parlamento Europeu		**Proposta de Regulamento Brasileira**	
Título IV e Título VIII	Capítulo 3 – Execução Artigos 52 e 68 (c)	Capítulo II	Artigo 5º, 7º e 8º

Fonte: elaboração da autora.

O direito à explicação é contemplado em ambos os projetos, sendo essa distinção o objeto de análise subsequente.

De forma mais detalhada, o PL 2.338/2023 possui um capítulo dedicado aos direitos das pessoas afetadas por sistemas de IA, com destaque para os arts. 5º, 7º e 8º.

O art. 5º do PL 2.338/2023 estabelece que pessoas afetadas por sistemas de IA têm direito à informação prévia sobre suas interações com esses sistemas. Já o art. 7º preleciona que pessoas afetadas por sistemas de IA têm o direito de receber informações claras e adequadas antes de contratar ou usar o sistema.

O direito à informação está vinculado à disponibilização de informações de forma compreensível, clara e adequada.

> As dificuldades técnicas derivam de alguns modelos computacionais que apresentam opacidade e, portanto, não permitem uma explicação completa e transparente durante todo o ciclo de vida do sistema. No que concerne aos aspectos práticos, algumas técnicas de explicação usam equações complexas e de baixa compreensão pelo usuário leigo. Além disso, a explicação completa pode não ser útil e nem necessária para estes usuários (CTS-FGV Rio, 2023, p. 17).

Essa conexão suscita questionamentos relevantes, por exemplo: o que seria considerado uma forma compreensível? Essa compreensão terá como objetivo quem, os usuários ou os auditores? Da mesma forma, o que seriam exatamente "informações claras e adequadas"? Além disso, quem será responsável por arbitrar o que é ou não compreensível, claro e adequado?

> Arya et al. identificaram que há uma lacuna entre o que a comunidade técnica está produzindo sobre transparência e o que os reguladores e a sociedade como um todo exigem desses sistemas. Uma razão para essa lacuna é a falta de uma definição precisa de como essas informações devem ser fornecidas, algo que se deve especialmente ao fato de que pessoas diferentes em ambientes diversos podem exigir diferentes tipos de explicações (LAPIN, 2022, *online*).

O art. 8º do PL 2.338/2023, por sua vez, versa que as pessoas afetadas por sistemas de IA têm o direito de solicitar explicações sobre as decisões, previsões ou recomendações. O artigo é abrangente, pois não está limitado a sistema de alto risco e aos efeitos negativos que o sistema pode causar a determinadas áreas, pois pode ser aplicado a todas as decisões advindas de sistemas de IA. Essas informações incluem a natureza automatizada das interações e decisões, a descrição geral do sistema, a identificação dos operadores, o papel do sistema e dos humanos envolvidos, os dados pessoais utilizados, as medidas de segurança, a não discriminação e a confiabilidade, entre outras.

Em relação ao direito de solicitar explicações sobre as decisões, previsões ou recomendações, é importante considerar a viabilidade técnica de discriminar todas essas informações. É relevante que se reflita se, tecnicamente, é possível fornecer uma descrição minuciosa e abrangente de todos os aspectos relevantes do sistema de IA. Além disso, deve-se levar em conta o custo associado ao cumprimento dessas exigências de documentação e compreensão do que elas representam para os resultados gerados pelo sistema.

> A inteligibilidade de sistemas de IA não deve consistir necessariamente em uma descrição precisa e detalhada de como os algoritmos funcionam. Tal forma de fornecimento de informações pode levar, em vários contextos, a um excedente informacional que pode ser inútil ou até prejudicial, levando ao que Ananny e Crawford chamam de "opacidade estratégica" (LAPIN, 2022, *online*).

Já o AI Act, em seu art. 68 (c), estabelece o direito à explicação para pessoas afetadas por uma decisão automatizada por um sistema de IA de alto risco, que tenha efeitos significantes na sua saúde, segurança, direitos fundamentais ou bem-estar socioeconômico. Nesse caso, o titular poderá pedir uma explicação sobre o papel do sistema de IA na decisão, os principais parâmetros da decisão tomada e os dados de entrada relacionados.

É importante notar que o conceito presente na proposta introduz certa subjetividade ao utilizar o trecho "efeitos significantes na sua saúde, segurança, direitos fundamentais ou bem-estar socioeconômico". Diante desse aspecto, torna-se crucial avaliar como os desenvolvedores, usuários e fornecedores irão abordar essa questão ao implementar efetivamente esse critério na prática, pois a interpretação e a aplicação do direito à explicação podem variar entre as partes envolvidas.

> Explicações dirigidas a pessoas ou grupos sobre os quais possam repercutir efeitos, que não se restringem aos jurídicos, a seus interesses advindos do uso de sistema de IA. A linguagem e o formato da explicação devem levar em conta as cinco questões[51]

[51] "(1) Quem é o destinatário das informações? Um regulador? Um advogado? Um indivíduo ou comunidade afetada pelo sistema? Uma organização da sociedade civil? Um especialista na área específica em que um sistema

> apresentadas anteriormente, com especial atenção ao nível de alfabetização digital do potencial receptor e o conhecimento que tem sobre IA e outras tecnologias digitais, bem como as particularidades relativas à origem regional, cultural, social da pessoa ou do grupo e o nível de especialização no campo em que o sistema é aplicado. A explicação também deve ser suficiente para que seu destinatário possa contestar e prestar contas da produção do sistema ou da decisão que ele influenciou, inclusive através do exercício do direito da pessoa de reclamar a uma autoridade supervisora (direito à transparência/explicabilidade) (LAPIN, 2022).

O art. 52 estabelece obrigações de transparência para determinados sistemas de IA[52] em que os fornecedores devem garantir que eles informem claramente que estão sendo utilizados, a menos que isso seja óbvio nas circunstâncias.

O aspecto que tanto o PL 2.338/2023 quanto o AI Act reconhecem e estabelecem é o direito dos titulares à explicação em relação aos sistemas de IA. O projeto brasileiro, contudo, demonstrou um maior destaque a esse assunto, dedicando um capítulo específico e três artigos extensos para abordar a questão. Já no AI Act, vale notar que o direito à explicação foi incorporado apenas na segunda versão da redação do projeto. Essa diferença ressalta a importância atribuída pelo projeto brasileiro à explicabilidade e à transparência na utilização de IA, enfatizando a necessidade de se fornecer informações claras e compreensíveis aos titulares dos dados.

5.2.1. Análise dos artigos sob a perspectiva da opacidade

está sendo implantado? Um engenheiro/cientista de computação? Disso dependerá uma compreensão do tipo de informação a ser prestada e seu grau de complexidade/tecnicidade, e é importante que certo nível de acesso a informações seja garantido a todos esses atores (2) Como as informações devem ser fornecidas? É necessário que o sistema seja auditado ou uma explicação é suficiente? No caso deste último, as informações podem ser transmitidas usando vocabulário especializado ou de uma forma que um leigo deva entender? (3) Para que fins? É para montar um desafio legal? Para consertar um bug? Para avaliar a imparcialidade ou o viés de uma saída? Para avaliar as razões do resultado de um sistema que alega que um paciente tem câncer? (4) Que tipo de informações são necessárias para atingir o objetivo pretendido? É suficiente avaliar o conjunto de dados de treinamento, informações gerais sobre o funcionamento do sistema ou, indo além do próprio sistema, seus gastos energéticos ou as escolhas econômicas e políticas que levaram ao desenvolvimento da aplicação? É importante saber como os dados foram coletados, tais como diretamente por humanos, sensores automatizados ou ambos? Além disso, é necessário saber como o sistema como um todo ou como ele conseguiu uma decisão específica, como, por exemplo, por que ele recusou crédito a um indivíduo? (5) Que tipo de sistema de IA está sendo avaliado? Seria uma aplicação de reconhecimento facial? De scoring de crédito? Ou um sistema de personalização de conteúdo?" (LAPIN, 2022).

[52] "Sistemas de IA projetados para interagir com pessoas, sistemas de reconhecimento de emoções ou categorização biométrica e sistemas que manipula conteúdos de imagem, áudio ou vídeo que sejam consideravelmente semelhantes a pessoas, objetos, locais ou outras entidades ou acontecimentos reais" (COMISSÃO EUROPEIA, 2020).

Anteriormente, foi evidenciada a importância do direito à explicação dos sistemas de IA para assegurar a transparência e a explicabilidade desses sistemas. Nesse contexto, este tópico tem como objetivo examinar se a opacidade inerente às técnicas de aprendizado de máquina baseado em redes neurais foi adequadamente considerada na formulação dos principais artigos mencionados no AI Act e no PL 2.338/2023 sobre direito à explicação.

Em relação ao PL 2.338/2023, serão analisados os arts. 5º, 7º e 8º, que se referem ao Capítulo II dos direitos.

Art. 5º Pessoas afetadas por sistemas de inteligência artificial têm os seguintes direitos, a serem exercidos na forma e nas condições descritas neste Capítulo: I – direito à informação prévia quanto às suas interações com sistemas de inteligência artificial; II – direito à explicação sobre a decisão, recomendação ou previsão tomada por sistemas de inteligência artificial.	Art. 7º Pessoas afetadas por sistemas de inteligência artificial têm o direito de receber, previamente à contratação ou utilização do sistema, informações claras e adequadas quanto aos seguintes aspectos: I – caráter automatizado da interação e da decisão em processos ou produtos que afetem a pessoa; II – descrição geral do sistema, tipos de decisões, recomendações ou previsões que se destina a fazer e consequências de sua utilização para a pessoa; III – identificação dos operadores do sistema de inteligência artificial e medidas de governança adotadas no desenvolvimento e emprego do sistema pela organização; IV – papel do sistema de inteligência artificial e dos humanos envolvidos no processo de tomada de decisão, previsão ou recomendação; V – categorias de dados pessoais utilizados no contexto do funcionamento do sistema de inteligência artificial; VI – medidas de segurança, de não discriminação e de confiabilidade adotadas, incluindo acurácia, precisão e cobertura; e VII – outras informações definidas em regulamento.	Art. 8º A pessoa afetada por sistema de inteligência artificial poderá solicitar explicação sobre a decisão, previsão ou recomendação, com informações a respeito dos critérios e dos procedimentos utilizados, assim como sobre os principais fatores que afetam tal previsão ou decisão específica, incluindo informações sobre: I – a racionalidade e a lógica do sistema, o significado e as consequências previstas de tal decisão para a pessoa afetada; II – o grau e o nível de contribuição do sistema de inteligência artificial para a tomada de decisões; III – os dados processados e a sua fonte, os critérios para a tomada de decisão e, quando apropriado, a sua ponderação, aplicados à situação da pessoa afetada; IV – os mecanismos por meio dos quais a pessoa pode contestar a decisão; e V – a possibilidade de solicitar intervenção humana, nos termos desta Lei.

Em relação ao inciso I do art. 5º, é factível a implementação dessa medida, uma vez que é razoável que o usuário tenha conhecimento de que está interagindo com um sistema de IA, a fim de tomar melhores decisões. Quanto ao inciso II, conforme mencionado nos subitens do Capítulo 2, sobre sistemas de IA de técnica de redes neurais profundas, existem limitações em relação ao tipo de decisão que pode ser completamente interpretada.

Dessa forma, ao generalizar sem mencionar possíveis considerações sobre a opacidade dos sistemas, o inciso II pode inviabilizar o detalhamento de como o direito à explicação poderá ser exercido. É relevante que se estabeleça em quais situações se pode exigir transparência completa do sistema e quais seriam as outras situações excepcionais em que explicações pontuais já seriam suficientes para solucionar a questão, diante da limitação dessa técnica.

As informações presentes no art. 7º são capazes de abordar, de certa forma, as lacunas levantadas em relação ao art. 5º, pois enumera as informações que devem ser fornecidas aos usuários antes de utilizarem o sistema em questão, evitando a possibilidade de solicitação de informações inviáveis ou muito custosas de serem fornecidas. Além disso, o artigo não deixa espaço para interpretações que se desviem dos sete incisos que descrevem quais informações podem ser solicitadas.

Por fim, o art. 8º exemplifica os tipos de informações que podem ser solicitadas para obter uma explicação sobre a decisão, previsão ou recomendação. Os cinco critérios mencionados levam em consideração aspectos que são passíveis de explicação, pois dizem respeito à estrutura do sistema, aos critérios e procedimentos utilizados em seu desenvolvimento, à racionalidade e lógica do sistema e aos mecanismos para contestar a decisão. No entanto, é importante destacar que o artigo não enumera os tópicos de forma exaustiva, o que significa que qualquer tipo de informação adicional pode ser solicitado para obtenção de uma explicação mais clara sobre a decisão, previsão ou recomendação.

Tudo isso dependerá do caso concreto e das respostas que de fato são relevantes para prover explicação adequada da situação que se visa esclarecer.

> Esse esforço deve ser direcionado de modo a garantir o que Frank Pasquale chama de "transparência qualificada" (*qualified transparency*). Na maioria das vezes, não é ter acesso ao código de um sistema que nos ajudará a resolver um problema relacionado ao funcionamento ou ao ambiente (práticas comerciais abusivas, custos ambientais) que circunda um sistema de IA, mas, em vez disso, ter acesso a "revelações limitadoras, a fim de respeitar todos os interesses envolvidos em uma determinada informação" (LAPIN, 2022, *online*).

No caso do AI Act, o enfoque de análise é o art. 68c:

> Artigo 68c Qualquer pessoa afetada sujeita a uma decisão tomada pelo implantador com base na saída de um sistema de IA de alto risco que produza efeitos jurídicos ou o afete significativamente de maneira semelhante, de uma forma que considere ter um impacto adverso em sua saúde, segurança, direitos fundamentais, bem-estar socioeconômico ou qualquer outro dos direitos decorrentes das obrigações previstas no presente regulamento, têm o direito de solicitar ao responsável pela instalação uma explicação clara e fundamentada nos termos do artigo 13º, nº 1, sobre a papel do sistema de IA no processo de tomada de decisão, os principais parâmetros da decisão tomada e os dados de entrada relacionados (PROPOSAL FOR AI ACT, 2023, *online*).

Em primeira análise, são estabelecidos os tipos de informações que podem ser requisitadas para o exercício do direito de explicação pelo titular. São elas: (i) questionar o papel do sistema de IA no procedimento de tomada de decisão; (ii) os principais parâmetros da decisão tomada; e (iii) questionar os dados de entrada relacionados.

A primeira conclusão que se pode aferir é que o direito à explicação estabelecido pelo AI Act restringe os casos em que pode ser exercido e as explicações que podem ser fornecidas apenas em relação a aspectos que, conforme pontuado no Capítulo 2, são viáveis de serem fornecidos pois se referem a atributos dos quais o desenvolvedor tem conhecimento.

Um segundo aspecto se refere a quem pode requisitar o exercício desse direito. Nesse caso, são apenas pessoas afetadas por uma decisão proveniente de um sistema de IA de alto risco que produza efeitos legais ou afete de forma significativa os direitos resguardados pelo arcabouço europeu.

> O as críticas decorrem do fato de que as pessoas geralmente são informadas apenas das decisões finais tomadas por IA, sejam concessões de empréstimos, admissão em universidades ou preços de seguros, mas ao mesmo tempo não têm ideia de como ou por que as decisões são tomadas. A questão da explicabilidade surge quando os humanos são afetados adversamente pela previsão feita por um sistema de IA, ou em casos extremos, são prejudicados por decisões baseadas em IA. O caso de lesões ou mortes resultantes de veículos autônomos é comumente citado (FLORIDI, 2022, p. 57).

Dessa forma, conclui-se que o regulador europeu optou por limitar a aplicação desse artigo apenas às decisões resultantes de sistemas de IA de alto risco que tenham afetado significativamente a vida das pessoas. Isso significa que apenas nessas situações esse direito à explicação será concedido.

6. CONCLUSÃO

Os sistemas de IA estão sendo amplamente adotados em diversos setores econômicos, o que resulta em uma nova dinâmica na interação entre seres humanos e máquinas. Na atualidade, grande parte das interações humanas envolve algum tipo de automação baseada em IA. Portanto, compreender e estudar esse campo e refletir sobre as implicações jurídicas desses avanços tecnológicos é fundamental para acompanhar as transformações e os desafios que surgem nessa relação cada vez mais integrada entre humanos e tecnologia.

De forma paralela à imensa relevância da IA na sociedade contemporânea, surge um movimento global em busca de regulamentar essa seara. A proliferação dos sistemas de IA em diversos setores econômicos e a nova dinâmica na interação entre seres humanos e máquinas têm despertado preocupações em relação aos possíveis riscos e danos que essa tecnologia pode causar. Assim, a análise dos projetos de regulamentação se tornou uma necessidade premente.

A complexidade de redigir sobre o presente tópico reside no fato de os projetos de regulamentação estarem ainda em tramitação. Tanto na Europa quanto no Brasil, desde 2021, essas discussões têm passado por inúmeras alterações, transformando-se em um *moving target*[53] difícil de acompanhar. As propostas de regulamentação têm sido objeto de amplos debates e revisões, resultando em uma dinâmica que exige constante atualização.

Outro aspecto que contribui para a complexidade da análise é a inexistência de julgados específicos sobre o tema. Em muitos casos, as discussões e as propostas de regulamentação ainda estão sendo debatidas, o que impede a disponibilidade de decisões judiciais que poderiam servir como base para análises comparativas ou embasar conclusões mais sólidas.

Embora ainda não haja promulgação de leis sobre o tema, foi possível estabelecer uma análise dos elementos considerados na redação dessas propostas, fornecendo subsídios para futuras discussões e debates sobre os desafios e as implicações da transparência e da explicabilidade nos sistemas de IA.

Tanto a transparência quando a explicabilidade, são valores fundamentais no desenvolvimento da IA, pois por meio delas é possível garantir que os direitos de igualdade sejam respeitados, fornecendo às pessoas que interagem com o sistema o poder de escolha e a capacidade de compreender como as decisões afetam seu a dia a dia. Dessa forma, a transparência e a explicabilidade são ferramentas que promovem confiança e viabilizam a

[53] "Something that is always changing, making it difficult to count, describe, achieve, etc." Em tradução livre, é algo que está sempre mudando, sendo difícil contá-lo, descrevê-lo, alcançá-lo etc. Disponível em: https://dictionary.cambridge.org/pt/dicionario/ingles/moving-targe.

concretização dos direitos à igualdade e à autonomia privada na interação com os sistemas de IA.

Embora esses conceitos sejam de extrema relevância quando aplicados aos sistemas de IA, especialmente os de aprendizado de máquina baseado em redes neurais, deve-se considerar os limites impostos por características intrínsecas a esses sistemas. A opacidade pode tornar inviável a compreensão de como um determinado resultado foi alcançado, além de tornar desafiadora a aplicação de abordagens tradicionais de transparência e explicabilidade. Diante dessa realidade, pesquisadores e cientistas estão buscando novas abordagens e técnicas para solucionar essa questão.

Nesse contexto, a dissertação teve como objetivo investigar e responder se a opacidade, característica inerente dos sistemas de IA de aprendizado de máquina baseado em redes neurais, foi considerada na redação do PL 2.338/2023 e do AI Act.

A partir dessa análise, pode-se concluir que a transparência é considerada um dos pilares centrais tanto do PL 2.338/2023 quanto do AI Act, sendo amplamente reconhecida em diversos mecanismos propostos nos projetos analisados.

Nesse contexto, os principais mecanismos identificados nos projetos que visam instrumentalizar a transparência são: a implementação de medidas de governança e o direito à explicação. Ambos com o propósito de promover a transparência e proporcionar maior confiança no desenvolvimento e uso seguro da IA.

A análise dos dispositivos referentes às medidas de governança e ao direito à explicação revela que eles representam um universo restrito de análise para conclusões sobre a consideração da opacidade nos projetos de regulamentação.

Além disso, em relação às medidas de governança, foram identificadas instruções claras e bem descritas sobre as informações a serem registradas sobre o sistema em desenvolvimento. Nesse aspecto, podem ser citados o art. 20, I, IV e V, do PL 2.338/2023 e o Anexo IV, itens I e II, (b) e (d), do AI Act.

No entanto, a questão da opacidade e os desafios associados ao fornecimento de informações técnicas sobre os sistemas não foram abordados de forma explícita e aberta nos projetos, conforme se pode observar no art. 20, II, do PL 2.338/2023 e no Anexo IV, item II, (b) e (d), do AI Act. Essa falta de consideração explícita poderá comprometer a aplicação de algumas dessas medidas de governança, pois em alguns casos pode ser inviável fornecer informações técnicas detalhadas sobre a constituição do sistema, devido à opacidade.

No âmbito do PL 2.338/2023, o direito à explicação foi abordado de forma mais extensa. O projeto de lei levou em consideração a importância desse direito ao estabelecer de forma

específica no art. 8º, que contempla aspectos passíveis de explicação relacionados à estrutura do sistema, critérios e procedimentos utilizados em seu desenvolvimento, racionalidade e lógica do sistema, bem como mecanismos para contestar as decisões tomadas por ele.

No entanto, é importante observar que os artigos analisados tanto no AI Act quanto no PL sobre direito à explicabilidade não abordam explicitamente a questão da opacidade ao tratar desse tema. Embora o direito à explicação seja reconhecido e regulamentado nessas legislações, a discussão sobre a opacidade dos sistemas de IA não foi abordada de maneira direta. Portanto, a falta de menção específica à opacidade nos artigos analisados requer considerações em abordagens futuras na busca de soluções que promovam uma maior compreensão dos sistemas de IA e a mitigação dos riscos associados.

Dessa forma, conclui-se que é importante que as propostas de regulamentação abordem de forma mais aberta e transparente os desafios relacionados à opacidade, para que seja possível encontrar soluções adequadas e realistas para a promoção da transparência nos sistemas de IA.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, A.; GANS, J.; GOLDFARB, A. **Máquinas preditivas**: a simples economia da inteligência artificial. Trad. Wendy Campos. Rio de Janeiro: Elsevier, 2019.

ALOM, Z. et al. **The history began from AlexNet**: a comprehensive survey on deep learning approaches. Cornell University, 2018. Disponível em: https://arxiv.org/pdf/1803.01164.pdf. Acesso em: 18 abr. 2023.

ALPAYDIN, E. **Introduction to machine learning**. The MIT Press, 2016. Disponível em: https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf. Acesso em: 18 abr. 2023.

ANANNY, M.; CRAWFORD, K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. **New Media & Society**, v. 20, n. 3, p. 973-989, 2018. Disponível em: https://doi.org/10.1177/1461444816676645. Acesso em: 30 jun. 2023.

ARBIX, G. A transparência no centro da construção de uma IA ética. **Novos Estudos CEBRAP**, v. 39, n. 2, p. 395-413, 2020. Disponível em: https://www.scielo.br/j/nec/a/pD9k5gtHpXwsgFcsMC5gbJg/?lang=pt. Acesso em: 18 abr. 2023.

ARRUDA, Carmen Silva Lima de. O princípio da transparência. **Revista do Direito da Administração Pública**, n. 1, p. 39-111, jan.-jun. 2021.

ARYA, Vijay et al. **One explanation does not fit all**: a toolkit and taxonomy of AI explainability techniques. Arvix, 2019. Disponível em: https://arxiv.org/abs/1909.03012. Acesso em: 26 jun. 2022.

BAVITZ, Christopher. **An open letter to the members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System**. November 9, 2017. Disponível em: https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9. Acesso em: 23 jun. 2023.

BETTI, Emílio. **Teoria do negócio jurídico**. Trad. Servanda Editora. Campinas: Servanda, 2008.

BOCHIE, K. et al. **Aprendizado profundo em redes desafiadoras**: conceitos e aplicações. S.l.: Sociedade Brasileira de Computação, 2020.

BRASIL. Câmara dos Deputados. **Projeto de Lei 21/2020**. Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências. Brasília: Câmara dos Deputados, 2020. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340. Acesso em: 30 abr. 2022.

BRASIL. Decreto-lei 9.854, de 25 de junho de 2019. Institui o Plano Nacional de Internet das Coisas e dispõe sobre a Câmara de Gestão e Acompanhamento do Desenvolvimento de Sistemas de Comunicação Máquina a Máquina e Internet das Coisas. **Diário Oficial da**

União, Brasília, 2019. Disponível em: https://www.in.gov.br/en/web/dou/-/decreto-n-9854-de-25-de-junho-de-2019-173021041. Acesso em: 26 ago. 2022.

BRASIL. Lei 8.078, de 11 de setembro de 1990. **Código de Defesa do Consumidor**. Dispõe sobre a proteção do consumidor e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/L8078.htm. Acesso em: 25 ago. 2022.

BRASIL. Lei 10.406, de 10 de janeiro de 2002. Institui o Código Civil. **Diário Oficial da União**: seção 1, Brasília, DF, 11 jan. 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Acesso em: 23 ago. 2022.

BRASIL. Ministério da Ciência, Tecnologia, Inovações e Comunicações – MCTIC. **Estratégia Brasileira de Inteligência Artificial**. 2019. Disponível em: http://participa.br/profile/estrategia-brasileira-de-inteligencia-artificial. Acesso em: 9 fev. 2022.

BRASIL. Ministério da Ciência, Tecnologia, Inovações e Comunicações – MCTIC. **Chamada de Propostas para a Criação de Centros de Pesquisas Aplicadas em Inteligência Artificial**. 2020. Disponível em: https://antigo.mctic.gov.br/mctic/opencms/tecnologia/inteligencia_artificial/Centros_Pesquisas_Aplicadas/CentrosPesquisasAplicadas_IA.html. Acesso em: 9 fev. 2022.

BRASIL. Senado Federal. **Comissão conclui texto sobre regulação da inteligência artificial no Brasil**. [*on-line*] Brasília, 6 dez. 2022. Disponível em: https://www12.senado.leg.br/noticias/materias/2022/12/06/comissao-conclui-texto-sobre-regulacao-da-inteligencia-artificial-no-brasil. Acesso em: 9 abr. 2023.

BRASIL. Senado Federal. **Projeto de Lei 5.051/2019**. Estabelece os princípios para o uso da Inteligência Artificial no Brasil. Brasília: Senado Federal, 2019. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/138790. Acesso em: 9 fev. 2022.

BRASIL. Senado Federal. **Projeto de Lei 5.691/2019.** Institui a Política Nacional de Inteligência Artificial. Brasília: Senado Federal, 2019. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/139586. Acesso em: 9 fev. 2022.

BRASIL. Senado Federal. **Projeto de Lei 872/2021.** Dispõe sobre o uso da Inteligência Artificial. Brasília: Senado Federal, 2021. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=8940096&ts=1656530049680&disposition=inline. Acesso em: 9 fev. 2022.

BRASIL. Senado Federal. **Relatório final – comissão de juristas responsável por subsidiar elaboração de Substitutivo sobre inteligência artificial no Brasil**. Brasília, 2022. Disponível em: file:///C:/Users/danielagen/Downloads/DOC-Relat%C3%B3rio%20Legislativo%20-%20SF225155204039-20221206.pdf. Acesso em: 18 abr. 2023.

BULGARELLI, Waldirio. **Teoria jurídica da empresa**: análise jurídica da empresarialidade. São Paulo: RT, 1985.

BURRELL, J. How the machine ‘thinks’: understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, 2016. Disponível em: https://doi.org/10.1177/2053951715622512. Acesso em: 30 jun. 2023.

CHRISTENSEN, L. T.; CHENEY, G. Peering into transparency: challenging ideals, proxies, and organizational practices. **Communication Theory**, v. 25, n. 1, p. 70-90, 2015.

COMISSÃO EUROPEIA. **Comissão adota orientações práticas para garantir a livre circulação dos trabalhadores essenciais** [Comunicado à imprensa]. 2020. Disponível em: https://ec.europa.eu/commission/presscorner/detail/pt/ip_20_273. Acesso em: 12 mar. 2023.

COMISSÃO EUROPEIA. **Livro branco sobre inteligência artificial – uma abordagem europeia para a excelência e a confiança**. 2020. Disponível em: https://op.europa.eu/pt/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1. Acesso em: 1º mar. 2023.

COMISSÃO EUROPEIA. **Proposal for a Regulation on a European approach for Artificial Intelligence**. Bruxelas, 2021. Disponível em: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Acesso em: 22 maio 2022.

COMISSÃO EUROPEIA. **Proposta de Regulamento do Parlamento Europeu e do Conselho relativo à criação do programa Europa Digital e à sua execução** (*Digital Europe Programme*). 2021. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/HTML/?uri=CELEX:52021PC0206&from=EN. Acesso em: 12 mar. 2023.

COZMAN, F. G.; KAUFMAN, D. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. **Revista USP**, n. 135, p. 195-210, 2022. Disponível em: https://www.revistas.usp.br/revusp/article/view/206235/189877. Acesso em: 17 abr. 2023.

CRARY, J. **Techniques of the observer**. Cambridge, MA: The MIT Press, 1990.

DASTON, L. Objectivity and the escape from perspective. **Social Studies of Science**, v. 22, n. 4, p. 597-618, 1992.

DASTON, L.; GALISON, P. **Objectivity**. Brooklyn, NY: Zone Books, 2007.

DATA PRIVACY BRASIL. **Nota técnica sobre o PL 21/2021**: o que muda na proteção de dados pessoais dos brasileiros. São Paulo, set. 2021. Disponível em: https://www.dataprivacybr.org/wp-content/uploads/2021/09/dpbr_notatecnica_pl21.pdf. Acesso em: 9 abr. 2023.

DAVID, M. **The correspondence theory of truth**. Stanford Encyclopedia of Philosophy, 2015. Disponível em: http://plato.stanford.edu/entries/truth-correspondence/. Acesso em: 23 jun. 2023.

DIREITOS NA REDE. **Carta de apoio ao PL nº 2338/23/2023**. Disponível em: https://direitosnarede.org.br/2023/06/14/carta-de-apoio-ao-pl-2338-2023/. Acesso em: 14 jun. 2023.

EUROPEAN PARLIAMENT. **Proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts**. Disponível em: https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_EN.html#_section1. Acesso em: 23 jun. 2023.

EUROPEAN PARLIAMENT. **Relatório que contém recomendações à Comissão sobre disposições de Direito Civil sobre Robótic**. Parlamento Europeu, 2017. Disponível em: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_PT.html. Acesso em: 10 abr. 2023.

FERRARI, I. ***Accountability*** **de algoritmos: a falácia do acesso ao código e caminhos para uma explicabilidade efetiva**. Inteligência Artificial: 3º Grupo de Pesquisa do ITS. Instituto de Tecnologia e Sociedade do Rio, 2018. Disponível em: https://itsrio.org/wpcontent/uploads/2019/03/Isabela-Ferrari.pdf. Acesso em: 9 abr. 2023.

FLORIDI, Luciano et al. **capAI – A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act**. 2022. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091.

FGV DIREITO RIO. **Policy Paper – Regulação de Inteligência Artificial no Brasil**. Agosto de 2022. Disponível em: https://direitorio.fgv.br/sites/default/files/2022-08/policypaperiaegoverno.pdf. Acesso em: 19 jul. 2023.

FRIEDMAN, B.; NISSENBAUM, H. Bias in computer systems. **ACM Transactions on Information Systems**, v. 14, n. 3, p. 330-347, jul. 1996.

G1. Carro autônomo da Uber atropela e mata mulher nos EUA. **Auto Esporte**, 19 mar. 2018. Disponível em: https://autoesporte.globo.com/videos/noticia/2018/03/carro-autonomo-da-uber-atropela-e-mata-mulher-nos-eua.ghtml. Acesso em: 19 jan. 2023.

GABAN, E. M. Regulação econômica e assimetria de informação. **Revista do IBRAC**, São Paulo, v. 9. n. 5, 2002.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. MIT Press, 2016.

GUIDOTTI, R. et al. Local rule-based explanations of Black Box Decision Systems. **ArXiv**, 1805.10820, 2018.

HAO, K. AI is sending people to jail—and getting it wrong. **MIT Technology Review**, 2019. Disponível em: https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/. Acesso em: 9 abr. 2023.

HARWELL, D. **The accent gap: how Amazon's and Google's smart speakers leave certain voices behind**. Jul. 2018. Disponível em: https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/. Acesso em: 9 abr. 2023.

HILDEBRANDT, M. Who needs stories if you can get thedata? ISPs in the era of big number crunching. **Philosophy & Technology**, v. 24, n. 4, p. 371-390, 2011.

HOOD, C. Transparency in historical perspective. *In*: HOOD, C.; HEALD, D. (eds.). **Transparency**: the key to better governance? Oxford: Oxford University Press, 2006.

KAUFMAN, D. **A inteligência artificial irá suplantar a inteligência humana?** Barueri, SP: Estação das Letras e Cores, 2019.

KAUFMAN, D. **Desmistificando a inteligência artificial**. São Paulo: Autêntica, 2022.

KAUFMAN, D. Equipes interdisciplinares: não basta "juntar campos", tem que construir pontes. **Época Negócios**, 2021.

KAUFMAN, D. Inteligência Artificial e os desafios éticos: a restrita aplicabilidade dos princípios gerais para nortear o ecossistema de IA. **PAULUS: Revista de Comunicação da FAPCOM**, v. 5, n. 9, 2021. Disponível em: https://doi.org/10.31657/rcp.v5i9.453. Acesso em: 17 abr. 2023.

KAUFMAN, D. O protagonismo dos algoritmos da Inteligência Artificial: observações sobre a sociedade de dados. **Teccogs: Revista Digital de Tecnologias Cognitivas**, São Paulo: TIDD PUC-SP, n. 17, p. 44-58, jan.-jun. 2018.

LAPIN. **Contribuição à Comissão de Juristas do Senado Federal responsável por subsidiar a elaboração de minuta de substitutivo para o Marco Regulatório da Inteligência Artificial no Brasil**. Junho de 2022. Disponível em: https://lapin.org.br/2022/07/05/contribuicao-a-comissao-de-juristas-do-senado-federal-responsavel-por-subsidiar-a-elaboracao-de-minuta-de-substitutivo-para-o-marco-regulatorio-da-inteligencia-artificial-no-brasil/. Acesso em: 5 jul. 2023.

LAVANCHY, M. **Why Amazon's sexist AI recruiting tool is better than a human**. 2018. Disponível em: https://www.imd.org/research-knowledge/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/. Acesso em:

LEE, K-F. **Inteligência artificial**. Trad. Marcelo Barbão. Rio de Janeiro: Editora Globo, 2019.

LIMA, Taisa Maria Macena de; SÁ, Maria de Fátima Freire de. Inteligência artificial e Lei Geral de Proteção de Dados Pessoais: o direito à explicação nas decisões automatizadas. **Revista Brasileira de Direito Constitucional**, [S.l.], v. 34, n. 2, p. 189-208, 2020. Disponível em: https://rbdcivil.ibdcivil.org.br/rbdc/article/view/584/425. Acesso em: 23 jun. 2023.

MAHESH, B. Machine Learning Algorithms—A Review. **International Journal of Science and Research**, n. 9, p. 381-386, 2020.

MAHLER, Tobias. Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal. **Nordic Yearbook of Law and Informatics 2020-2021: Law in the Era of Artificial Intelligence**, p. 245-267, 2022. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4001444.

MARQUES, Cláudia Lima. **Contratos no Código de Defesa do Consumidor**: o novo regime das relações contratuais. 4. ed. São Paulo: RT, 2003.

MCCARTHY, J. **What is Artificial Intelligence?** 2007. Disponível em: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf. Acesso em: 13 mar. 2023.

MEIRELES, Adriana Veloso. **Algoritmos, privacidade e democracia ou como o privado nunca foi tão político como no século XXI**. 181f. Tese (Doutorado em Ciência Política) – Universidade de Brasília, Brasília, 2020.

MIT MEDIA LAB. MIT Media Lab to participate in $27 million initiative on AI ethics and governance. **MIT News**, [S.l.], 25 jun. 2018. Disponível em: https://robotics.mit.edu/mit-media-lab-participate-27-million-initiative-ai-ethics-and-governance. Acesso em: 19 jan. 2023.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997.

MITTELSTADT, B. D.; ALLO, P.; TADDEO, M.; WACHTER, S.; FLORIDI, L. **The ethics of algorithms**: mapping the debate. Big Data & Society, 2016.

MULLER, R. We need to talk about artificial intelligence. **World Economic Forum**, 2021. Disponível em: https://www.weforum.org/agenda/2021/02/we-need-to-talk-about-artificial-intelligence/. Acesso em: 9 abr. 2023.

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

PHILLIPS, J. W. P. Secrecy and transparency: an interview with Samuel Weber. **Theory, Culture & Society**, v. 28, n. 7-8, p. 158-172, 2011.

RAHWAN, I. **Moral machine**. [S.l.], [2016-2023?]. Disponível em: https://www.moralmachine.net/. Acesso em: 19 jan. 2023.

REIS, P. **Inteligência artificial na advocacia: desafios e possibilidades**. Dissertação (Mestrado em Direito) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2021. Disponível em: https://sapientia.pucsp.br/handle/handle/25749. Acesso em: 18 abr. 2023.

ROBERTO, Enrico. Responsabilidade civil pelo uso de sistemas de inteligência artificial: em busca de um novo paradigma. **Internet & Sociedade**, n. 1, v. 1, fev. 2020. Disponível em: https://revista.internetlab.org.br/responsabilidade-civil-pelo-uso-de-sistemas-de-inteligencia-artificial-em-busca-de-um-novo-paradigma-2/. Acesso em: 5 jun. 2021.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: a modern approach. Global edition. Harlow, UK: Pearson Education, 2009.

RUSSELL, S. J.; NORVIG, P.; DAVIS, E. **Artificial Intelligence**: a modern approach. 3. ed. Upper Saddle River, NJ: Prentice Hall, 2010.

SALOMÃO FILHO, Calixto. **Teoria crítico-estruturalista do direito comercial**. São Paulo: Marcial Pons, 2015.

SALOMÃO, L. F. et al. **Nota técnica sobre o Projeto de Lei 21/2020**. Rio de Janeiro, 2021. Disponível em: https://conhecimento.fgv.br/sites/default/files/2022-08/publicacoes/notatecnica.pdf. Acesso em: 9 abr. 2023.

SCHREIBER, Anderson; KONDER, Carlos Nelson. Uma agenda para o direito civil-constitucional. **Revista Brasileira de Direito Civil – RBDCivil**, v. 10, p. 22-23, out.-dez. 2016. Disponível em: https://rbdcivil. ibdcivil.org.br/rbdc/article/view/42/36. Acesso em: 29 jan. 2020.

SILBERG, J.; MANYIKA, J. Notes from the AI frontier: Tackling bias in AI (and in humans). **McKinsey & Company**, jun. 2019. Disponível em: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-tackling-bias-in-ai-and-in-humans. Acesso em: 9 abr. 2023.

TADDEO, M.; FLORIDI, L. How AI can be a force for good. **Science**, v. 361, n. 6404, August 2018. Disponível em: https://philarchive.org/archive/TADHAC. Acesso em: 25 ago. 2022.

TELFORD, T. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. **The Washington Post**, 11 nov. 2019. Disponível em: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. Acesso em: 9 abr. 2023.

TRICKEY, E. **Morality in the Machines**. Disponível em: https://hlt.staging.tri.be/feature/morality-in-the-machines/. Acesso em: 9 abr. 2023.

TURILLI, M.; FLORIDI, L. The ethics of informationtransparency. **Ethics and Information Technology**, v. 11, n. 2, p. 105-112, 2009.

VON ESCHENBACH, Warren J. Transparency and the black box problem: why we do not trust AI. **Philosophy & Technology**, v. 34, n. 4, p. 1607+, Dec. 2021. Disponível em: link.gale.com/apps/doc/A686026313/AONE?u=anon~dd101449&sid=googleScholar&xid=2a811f38. Acesso em: 30 jun. 2023.

ZAVAGLIA COELHO, Alexandre et al. **Governança da Inteligência Artificial em organizações**: framework para Comitês de Ética em IA – versão 1.0. São Paulo: FGV Direito SP, 2023.

ZEDNIK, C. Solving the black box problem: a normative framework for explainable Artificial Intelligence. **Philos. Technol.**, v. 34, p. 265-288, 2021. Disponível em: https://doi.org/10.1007/s13347-019-00382-7. Acesso em: 30 jun. 2023.

ZEDNIK, Carlos; BOELSEN, Hannes (ed.). AISB 2021 symposium proceedings: overcoming opacity in machine learning. **Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour**, 2021. Disponível em: https://aisb.org.uk/wp-content/uploads/2021/04/AISB21_Opacity_Proceedings.pdf. Acesso em: 30 jun. 2023.