Recebido: 25/04/2024
Aprovado: 04/06/2024

O DIREITO FUNDAMENTAL À EXPLICABILIDADE DA INTELIGÊNCIA ARTIFICIAL UTILIZADA EM DECISÕES ESTATAIS

THE FUNDAMENTAL RIGHT TO EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE APPLIED IN STATE DECISION-MAKING

Sérgio Rodrigo de Pádua[1]
Bruno Meneses Lorenzetto[2]

SUMÁRIO: Introdução. 1. Inteligência Artificial nas decisões na esfera pública e explicabilidade: elementos

1 Doutor e Mestre em Direitos Fundamentais e Democracia pelo UniBrasil - Centro Universitário Autônomo do Brasil (Curitiba). Professor de Direito no UniBrasil - Centro Universitário Autônomo do Brasil e no UniFatec - Centro Universi-tário de Tecnologia de Curitiba, lecionando as disciplinas de Direito Constitucional, Direito Administrativo, Direito Elei-toral, Teoria do Direito e Hermenêutica Jurídica. É pesquisador na área de Direito, Tecnologia e Inteligência Artifici-al. Analista Judiciário do Tribunal de Justiça do Estado do Paraná (TJPR). Membro da Associação Ibero-Americana de Direito e Inteligência Artificial (AID-IA).

2 Pós-doutorando em Direito pela Universidade Federal do Paraná (UFPR). Doutor em Direito pela UFPR na área de Direitos Humanos e Democracia. Mestre em Direito pela UFPR na área do Direito das Relações Sociais. Graduado em Direito pela Pontifícia Universidade Católica do Paraná (PUCPR). Visiting Scholar na Columbia Law School, Columbia University, Nova Iorque. Coordenador do Programa de Pós-Graduação em Direito (Direitos Fundamentais e Democracia) e Professor da Graduação do Centro Universitário Autônomo do Brasil (UniBrasil). Professor de Direito da PUCPR.

iniciais. 2. Interpretabilidade e explicabilidade nos marcos regulatórios da Inteligência Artificial. 3. A construção do direito à explicabilidade da Inteligência Artificial que auxilia decisões estatais. Conclusão. Referências.

RESUMO: O presente trabalho analisa o direito à explicabilidade da Inteligência Artificial utilizada como apoio às decisões estatais (da Administração Pública e do Judiciário). A pesquisa foi multidisciplinar, pois foram objeto de análise normas nacionais, documentos internacionais que vêm sendo produzidos sobre o tema (recomendações da OCDE e da UNESCO, normativas da União Europeia e Resolução da ONU), categorias teóricas do pensamento jurídico e de estudos científicos sobre os limites e as possibilidades da explicabilidade da IA. Este estudo se pauta em pesquisa qualitativa de métodos bibliográfico e descritivo-exploratório, mediante análise bibliográfica e documental inerente ao Direito e à Ciência da Computação. Os resultados demonstram: a) a necessária interseção entre a teoria da justificação das decisões estatais tomadas na esfera pública e as técnicas de inteligência artificial explicável (*xAI*), a fim de se garantir à *accountability* das decisões estatais auxiliadas por inteligência artificial; e b) a definição de direito fundamental à explicabilidade estrita das decisões estatais tomadas com auxílio de IA (na condição de um limite ético e normativo), de maneira de diferenciá-lo da explicabilidade em sentido amplo (engloba também a interpretabilidade dos processos computacionais necessários ao funcionamento da IA). As conclusões são: a) a explicabilidade não é um impedimento ao desenvolvimento científico da IA no campo das decisões públicas; b) a explicabilidade em sentido amplo pode ser subdividida em interpretabilidade e explicabilidade em sentido estrito; c) o direito fundamental à explicabilidade da IA aplicada às decisões estatais (na Administração Pública e no Judiciário) tem grande relevância no sistema de *accountability*.

PALAVRAS-CHAVE: Explicabilidade. Interpretabilidade. Inteligência Artificial. Decisões. Esfera Pública. Direito Fundamental.

ABSTRACT: This study analyzes the right to explainability of Artificial Intelligence (AI) systems used to support state decisions in Public Administration and the Judiciary. The research was multidisciplinary, encompassing national statutes, international documents (such as OECD and UNESCO recommendations, European Union regulations, and a UN Resolution), theoretical legal constructs, and scientific studies on the limits and possibilities of AI explainability. Employing qualitative research

methods, this study relied on bibliographic and descriptive-exploratory approaches through the analysis of documents pertinent to Law and Computer Science. The results demonstrate: a) the essential intersection between the theory of justification for state decisions in the public sphere and explainable artificial intelligence (*xAI*) techniques, aiming to ensure the accountability of state actions supported by AI; and b) the articulation of the fundamental right to strict explainability of state decisions facilitated by AI (within ethical and normative boundaries), distinguishing it from a broader concept of explainability that includes the interpretability of computational processes essential for AI operation. The conclusions affirm that: a) explainability does not hinder the scientific development of AI in making public decisions; b) broad-sense explainability can be divided into interpretability and strict-sense explainability; c) the fundamental right to explainability when AI is applied to state decisions (in Public Administration and the Judiciary) holds significance within the accountability framework.

KEYWORDS: Explainability. Interpretability. Artificial Intelligence. Decision-making. Public Sphere. Fundamental Right.

INTRODUÇÃO

O uso crescente de sistemas de Inteligência Artificial (IA) para auxiliar decisões públicas marca uma tendência significativa. Por exemplo, pesquisas da Fundação Getulio Vargas revelam que há sessenta e quatro iniciativas de IA nos tribunais brasileiros, abrangendo quarenta e sete tribunais (Salomão *et al.*, 2022). Um dos mais notáveis entre esses projetos é o Victor, implementado pelo Supremo Tribunal Federal, que influencia diretamente o trâmite da admissibilidade de recursos na corte (Hartmann Peixoto, 2020b).

Essa expansão da IA no setor público traz o desafio de atualizar e reinterpretar as teorias jurídicas tradicionais. Neste contexto, é urgente discutir as implicações das regulações advindas de organizações como a OCDE (2020), a UNESCO (2021), a União Europeia (2024) e as Nações Unidas (2024) sobre ética e regulação em IA. Essa discussão envolve integrar conceitos de responsabilidade (*accountability*) e a necessidade de desenvolver sistemas de IA que sejam explicáveis (*xAI*), estabelecendo um vínculo direto com os princípios jurídicos aplicáveis.

Neste caminho, este estudo tem como objetivos principais: a) explorar o potencial de explicabilidade de decisões públicas proferidas com o auxílio de Inteligência Artificial; e b) identificar padrões éticos e jurídicos que contribuam para a construção do direito à explicabilidade das decisões públicas (da Administração e do Judiciário) auxiliadas por Inteligência Artificial.

Mediante marco teórico interdisciplinar (com estudo de autores do Direito e da Ciência da Computação que se dedicam ao tema pesquisado), o texto foi desenvolvido em três capítulos. No primeiro capítulo são delineados aspectos iniciais do problema (com enfoque na proposta da OCDE e em conceitos técnicos que são premissas necessárias ao estudo), no segundo capítulo há um aprofundamento no debate global sobre a explicabilidade de sistemas de IA (UNESCO, União Europeia e ONU) e, ao final, no terceiro capítulo são definidos elementos caracterizadores do direito à explicabilidade.

Os resultados da pesquisa apontam para: a) a urgente necessidade de criar modelos de Inteligência Artificial para o setor público que incorporem a explicabilidade a partir de uma abordagem de transparência integrada ao *design*, respeitando os limites éticos e normativos que regem a integração da IA nas decisões dos agentes estatais; b) o reconhecimento de que existem barreiras técnicas que impactam o direito à explicabilidade das decisões públicas auxiliadas pela IA; e c) a definição do direito fundamental à explicabilidade das decisões públicas tomadas com o apoio de IA.

Feita a breve introdução, passa-se ao aprofundamento no tema.

1. INTELIGÊNCIA ARTIFICIAL NAS DECISÕES NA ESFERA PÚBLICA E EXPLICABILIDADE: ELEMENTOS INICIAIS

O direito proporciona estímulos para uma motivação/fundamentação adequada e racional das decisões públicas (judiciais ou da administração pública), estabelecendo limites para evitar excessos de racionalidade por parte do agente decisório (Alexy, 2005, p. 193-194). Nesse contexto, a transparência assume uma posição central na arquitetura das decisões, especialmente em políticas públicas onde é essencial a proatividade em explicar as razões das decisões tomadas, conforme incentivado pelo ordenamento jurídico (Thaler; Sunstein, 2019, posição 89-90).

Além disso, há um conjunto normativo coeso (Ávila, 2006, p. 49) que indica a importância da publicidade dos atos estatais (Alexy, 2011, p. 147), da Administração Pública e do Judiciário, como um valor fundamental no sistema jurídico-constitucional brasileiro. Esse bloco normativo demonstra a evolução do princípio da publicidade para o dever de transparência dos órgãos estatais. Portanto, a criação de um ambiente institucional favorável à responsabilização é impulsionada pela promoção de uma cultura de transparência (art. 5º, XXXIII, art. 37, *caput*, e art. 93, IX, da Constituição Federal, e art. 3º, IV, da Lei n. 12.527/2011). A partir disso, no enfoque aqui trabalhado, a transparência também envolve fornecer *acesso claro às informações e processos de tomada de decisão, explicando como as decisões são feitas e quais critérios e algoritmos (sequências de instruções lógicas) são aplicados.*

Nesse aspecto, verifica-se a importância dos Princípios de Asilomar, que foram desenvolvidos durante conferência realizada pelo *Future of Life Institute* (2017). Eles consistem em *vinte e três diretrizes que visam garantir que o desenvolvimento e a aplicação da Inteligência Artificial (IA) sejam realizados de forma benéfica e segura para a sociedade*. Ao introduzir a discussão sobre o assunto, os Princípios de Asilomar estabelecem que a participação de sistemas autônomos em decisões públicas (judiciais e da administração pública) deve ser acompanhada de uma explicação satisfatória, que possa ser auditada por uma autoridade humana competente (Future of Life Institute, 2017). Além disso, esses princípios destacam a importância da transparência em caso de falhas, estipulando que, se um sistema de IA causar danos, deve ser possível esclarecer as razões (Future of Life Institute, 2017). A partir dessa premissa, desde já se pode trabalhar com um conceito inicial de *explicabilidade como a capacidade de descrever e justificar as decisões tomadas por sistemas de IA, permitindo que sejam entendidas e confiáveis para os humanos*.

Por sua vez, a *Recomendação* n. 449 da OCDE (Organização para a Cooperação e Desenvolvimento Econômico) (2020), que contou com a adesão do Brasil em 21 de maio de 2019, estabelece *diretrizes para o desenvolvimento, a implementação e o uso responsável de sistemas de Inteligência Artificial (IA)*. Esta resolução enfatiza a importância da transparência, explicabilidade e *accountability* (responsabilização) no contexto da IA, com o objetivo de assegurar que os sistemas de IA sejam benéficos e seguros para as pessoas. A Recomendação n. 449 da OCDE enfatiza, especialmente na seção 1.3 (OCDE, 2020), a necessidade de transparência e explicabilidade nos sistemas de Inteligência Artificial.

Nesta parte, a seção 1.3 estabelece que os envolvidos com Inteligência Artificial devem se comprometer com a transparência e com a divulgação responsável dos sistemas de IA. Este compromisso inclui: a) *fomentar um entendimento amplo sobre os sistemas de IA* (inciso I); b) *aumentar a conscientização das partes interessadas sobre suas interações com esses sistemas* (inciso II); c) *permitir que indivíduos afetados compreendam os resultados gerados pela IA* (inciso III); e d) *possibilitar que aqueles adversamente afetados questionem os resultados de um sistema de IA*, fornecendo-lhes informações claras e acessíveis sobre os fatores e a lógica que fundamentam as previsões, recomendações ou decisões (inciso IV).

Adicionalmente, a seção 1.5 da Recomendação n. 449 da OCDE (2020) impõe um dever de *accountability*, exigindo que os atores da Inteligência Artificial sejam responsáveis pelo funcionamento adequado dos sistemas de IA e pelo cumprimento dos princípios de transparência e explicabilidade. Essa responsabilidade deve ser contextualizada de acordo com os papéis dos atores dentro da esfera da IA na sociedade e alinhada com o progresso corrente na Ciência da Computação.

Portanto, a Recomendação n. 449 da OCDE representou um passo significativo em direção ao aprimoramento futuro do marco normativo nacional para a regulação da IA. Isso inclui os Projetos de Lei n. 21/2020 e n. 2.338/2023 que tramitam no Senado Federal (2020 e 2023), que também têm especial enfoque nos direitos à transparência e à explicabilidade dos sistemas de IA. No entanto, a seção 1.5 da Recomendação n. 449 reconhece as limitações práticas na implementação desses direitos. Esta seção sugere que o "estado da arte" da tecnologia de Inteligência Artificial deve ser levado em conta, assegurando o nível mais alto de transparência e explicabilidade que as técnicas atuais permitirem. Simultaneamente, a seção 1.5 admite implicitamente o uso das tecnologias de IA atuais que podem apresentar limitações que resultam em um nível mais baixo de transparência nas predições, recomendações e decisões geradas por IA (OCDE, 2020), como, por exemplo, as *redes neurais (modelos computacionais inspirados no cérebro humano*, compostos por camadas de neurônios artificiais que aprendem padrões a partir dos dados analisados).

Neste ponto, apenas para fins de contextualização do leitor, as redes neurais são espécie de modelos de IA baseada em *deep learning* (aprendizado profundo), que se trata de técnica avançada de aprendizado de máquina (*machine learning*) utilizada para desenvolver sistemas que aprendem com os dados e que, sem necessidade de *programação explícita, melhoram seu desempenho em cenários complexos* (Russell; Norvig, 2013, p. 26-27).

Por sua vez, no âmbito normativo nacional, a Lei n. 13.709/2018 (Lei Geral de Proteção de Dados – LGPD) trata de *regras e princípios para garantir a proteção dos direitos de privacidade e segurança dos dados dos cidadãos*. O art. 6º da LGPD estabelece os princípios fundamentais para a proteção de dados no Brasil. Notavelmente, o inciso VI desse artigo define o princípio de transparência, garantindo aos titulares acesso a informações claras, precisas e facilmente acessíveis sobre como seus dados são tratados e quem são os agentes responsáveis, sempre considerando a preservação dos segredos comercial e industrial (art. 10, § 2º). Este conceito é reforçado pelo artigo 20, § 1º, que obriga os controladores a fornecer, quando solicitado, detalhes claros e adequados sobre os critérios e procedimentos utilizados em decisões automatizadas, também respeitando segredos comerciais e industriais.

Desse modo, a LGPD em seu artigo 20 detalha direitos específicos relativos a decisões automatizadas, incluindo: a) a possibilidade de revisão humana dessas decisões (*caput*); b) o direito a informações claras sobre os critérios e procedimentos utilizados nas decisões automatizadas, mantendo a confidencialidade necessária (§ 1º); e c) a prerrogativa de auditoria pela Autoridade Nacional de Proteção de Dados para investigar possíveis discriminações em decisões automatizadas protegidas por segredo comercial ou industrial (§ 2º).

Nesse contexto, uma estratégia para otimizar a aplicabilidade do art. 20 da LGPD em decisões de entes públicos apoiadas por sistemas de Inteligência Artificial pode ser desenvolvida para assegurar maior transparência. Essa estratégia encontra fundamento no art. 5º, XXXIII, no art. 37, *caput*, e no art. 93, IX, da Constituição Federal, que exigem que todas as decisões sejam públicas e devidamente fundamentadas, sob risco de nulidade.

Para reforçar a publicidade das decisões de entes públicos que utilizam IA (especialmente do Judiciário), a Resolução n. 332, de 21 de agosto de 2020, do Conselho Nacional de Justiça (CNJ) introduziu medidas importantes. Especificamente, o artigo 8º, VI, dessa Resolução estipula a necessidade de fornecer uma explicação satisfatória e auditável por uma autoridade humana para qualquer decisão proposta por um modelo de Inteligência Artificial em casos judiciais (Brasil, 2020a). Além disso, o artigo 19, *caput*, da mesma Resolução destaca que os sistemas que empregam IA como ferramenta auxiliar na elaboração de decisões judiciais devem priorizar, como critério fundamental, a clareza na explicação dos processos que levaram aos resultados obtidos (Brasil, 2020a). Finalmente, o artigo 25, *caput*, da Resolução determina a necessidade de total transparência na prestação de contas por qualquer sistema de IA judicial (Brasil, 2020a). Para uma leitura mais ampla, na forma tratada neste artigo, destaque-se que, na pendência de uma regulamentação clara sobre o uso de IA pelo Poder Executivo, os princípios gerais da Resolução n. 332/2020 do CNJ podem ser utilizados como balizas mínimas pelos órgãos da Administração Pública, mediante aplicação analógica (art. 4º da Lei de Introdução às Normas do Direito Brasileiro), haja vista a necessidade de concretização de proteção de direitos já previstos no regime constitucional.

Para essa finalidade as técnicas de *Inteligência Artificial Explicável* (*xAI*) (Deeks, 2020) visam tornar os sistemas de IA mais transparentes e compreensíveis, *fornecendo ou aumentando a explicabilidade dos sistemas de IA para facilitar o controle das decisões auxiliadas pela referida tecnologia*. Portanto, para abordar modelos de Inteligência Artificial Explicável (*xAI*) (Deeks, 2020, p. 1834) é necessário estabelecer as bases das normas aplicáveis que promovam a *accountability by design* (Wagner, 2020), que se refere *à incorporação de mecanismos de responsabilização diretamente no design dos sistemas de IA* (ou seja, *desde o projeto até a efetiva implementação*), assegurando que suas operações possam ser auditadas e os responsáveis identificados. Isso envolve aumentar a *transparência*, a interpretabilidade e a explicabilidade dos sistemas públicos baseados em IA, aproveitando também os conceitos propostos por recomendações e declarações de organizações internacionais, apesar de sua não obrigatoriedade nas relações jurídicas internas (*soft law*).

Nesse caminho, segundo Frank Pasquale (2017), a explicabilidade deve ser considerada como a quarta lei da robótica, de maneira a complementar

as três primeiras leis da robótica que decorrem da obra clássica "Eu, Robô", de Isaac Asimov (1950). Assim, para fins de contexto ao leitor, Asimov, em suas histórias de ficção científica, estabeleceu as Três Leis da Robótica que preveem que: 1) "um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano venha a ser ferido"; 2) "um robô deve obedecer às ordens dadas por seres humanos, exceto nos casos em que tais ordens entrem em conflito com a Primeira Lei"; e 3) "um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira ou com a Segunda Lei" (Asimov, 1950).

Desse modo, a melhoria na transparência das motivações das decisões públicas (Lorenzetto; Clève, 2015) é fundamental para a integridade do processo decisório em um ambiente democrático (Habermas, 2002, p. 79). Isso é reforçado por Alexy (2005, p. 195), que considera a racionalidade pública das decisões como intrínseca à própria democracia. Assim, a transparência dos atos estatais emerge como um dos pilares mais básicos para o desenvolvimento de um ambiente institucional propício à *accountability* algorítmica e ao debate público, elementos que são extremamente necessários para fundamentar adequadamente as decisões tomadas (Pasquale, 2019).

Estabelecidas essas bases, pode-se passar para o aprofundamento do direito à explicabilidade.

2. INTERPRETABILIDADE E EXPLICABILIDADE NOS MARCOS REGULATÓRIOS DA INTELIGÊNCIA ARTIFICIAL

A distinção entre os conceitos de explicabilidade e interpretabilidade é necessária para se entender seus impactos jurídicos e o alcance das regulações sobre o direito à explicabilidade. Nesse sentido, a UNESCO (2021) elaborou a "Recomendação sobre a Ética da Inteligência Artificial", que visa estabelecer padrões éticos e diretrizes para o desenvolvimento e uso de IA. Esta recomendação contempla a totalidade do ciclo de vida dos sistemas de IA, desde a pesquisa até o encerramento da utilização, e destaca a necessidade de padrões éticos em todas as fases (UNESCO, 2021).

Além disso, a seção 3, "b", do documento ressalta que a IA tem relevância até mesmo nas ciências sociais e humanas, impactando os conceitos científicos e criando uma nova base para a tomada de decisões (UNESCO, 2021).

Consequentemente, os modelos de IA destinados ao apoio às decisões públicas (judiciais e da Administração Pública) são contemplados por princípios estabelecidos na Recomendação da UNESCO (2021), como proporcionalidade, justiça, não discriminação, supervisão e determinação humana, transparência, explicabilidade e responsabilidade. Esses princípios

são interpretados de maneira a garantir a avaliação do impacto ético e a gestão ética, bem como a comunicação e informação eficaz.

Desse modo, a transparência permite às pessoas compreenderem como os sistemas de Inteligência Artificial são pesquisados, projetados, implementados e utilizados, levando-se em conta a sensibilidade de cada sistema para vida em sociedade. Isso engloba informações detalhadas sobre os fatores que influenciam previsões ou decisões. A abordagem da UNESCO (2021) para a transparência, embora não exija explicitamente a divulgação de códigos e bases de dados (*datasets*), destaca que a transparência é uma faceta necessária para fomentar a confiança humana em sistemas de IA. Por outro lado, em situações onde os direitos humanos estão em risco, a transparência pode incluir acesso ao código-fonte ou datasets de treinamento (UNESCO, 2021).

Nesse sentido, a *explicabilidade* foca na obrigação de tornar os resultados de sistemas de IA compreensíveis para os seres humanos, fornecendo informações que facilitem o entendimento das entradas (*inputs*), saídas (*outputs*) e dos processos algorítmicos fundamentais que levam aos resultados processados e influenciam nas decisões.

Dessa forma, explicabilidade e transparência são conceitos interligados, pois tanto os resultados quanto os processos que a eles conduzem devem ser claros e verificáveis dentro de seu contexto de aplicação. Nesse sentido, por exemplo, a Resolução n. 332/2020 do CNJ ressalta a importância da transparência na IA, assegurando o direito dos cidadãos a uma explicação satisfatória e auditável por autoridades humanas em relação a decisões sugeridas por modelos de IA, garantindo assim a sua *accountability*, especialmente no que tange às decisões judiciais (cuja aplicabilidade, por analogia, é possível para a Administração Pública).

Já no contexto das diretrizes estabelecidas pela UNESCO, observa-se que o nível de transparência e explicabilidade em sistemas de IA deve estar alinhado com o contexto de uso do sistema. Isso implica que, dentro dos limites de explicabilidade próprios de cada sistema de IA, deve haver um equilíbrio com outros princípios fundamentais, como segurança e privacidade.

Para demarcar as premissas deste estudo, pode-se conceituar a *interpretabilidade* de um modelo de IA como a característica que se refere ao nível de compreensão que os especialistas em computação (cientistas e engenheiros) têm dos processos e resultados algorítmicos e dos sistemas de Inteligência Artificial (Guidotti *et al*., 2019), e que pode variar conforme a técnica de IA empregada. Já a *explicabilidade* busca técnicas para tornar os processos e resultados da IA compreensíveis em linguagem natural ou de outra forma inteligível para humanos (Gilpin *et al*., 2018). Ou seja, enquanto a interpretabilidade pode não ser sempre plenamente alcançável em todos os

casos, a explicabilidade busca soluções técnicas para facilitar o entendimento humano, promovendo a acessibilidade às pessoas impactadas pela decisão.

Em suma, o *conceito amplo de explicabilidade* engloba tanto a *explicabilidade em sentido estrito* (que é um direito fundamental decorrente do regime constitucional brasileiro), destinada àqueles sem conhecimento técnico de IA, quanto a *interpretabilidade*, direcionada a profissionais da computação, direito e outras ciências sociais que possuam tal conhecimento técnico (Pádua, 2023a, p. 99).

Além disso, com base nas definições de proporcionalidade, justiça, supervisão humana, transparência e explicabilidade, destacadas na Recomendação da UNESCO (2021), pode-se avaliar os efeitos dos institutos da *responsabilidade* e da *accountability* no contexto de IA. *Responsabilidade* refere-se à conformidade com padrões legais e éticos internacionais em decisões e ações tomadas por sistemas de IA, atribuindo essa obrigação à pessoa ou entidade que desenvolve ou utiliza a IA. Quanto à *accountability*, é necessário haver mecanismos de auditoria e rastreabilidade robustos para os sistemas de IA e seus resultados, considerando também os aspectos técnicos e institucionais.

Nessa linha, a *Policy Area 1* (avaliação de impacto ético) da UNESCO (2021) enfatiza a importância de estabelecer requisitos claros de transparência e explicabilidade para os sistemas de IA, especialmente na tomada de decisões por entidades públicas (Judiciário e Administração Pública), abordando o comportamento desses sistemas (o que inclui os algoritmos e os dados envolvidos). Isso é vital, pois tanto as decisões do Judiciário como da Administração Pública envolvem interação direta com os usuários finais, como delineado na seção 52 da recomendação da UNESCO (2021).

Além disso, para assegurar a promoção da explicabilidade no contexto institucional, é importante destacar a *Policy Area 2*, que trata de governança e gestão ética, a qual determina que para manter a independência do Judiciário, sistemas de IA que auxiliam na decisão judicial devem fornecer mecanismos de monitoramento para órgãos de controle e fiscalização (UNESCO, 2021).

Ademais, a seção 70 da Recomendação (UNESCO, 2021) ressalta que o nível e o tipo de informação sobre os algoritmos e os resultados de um sistema de IA e a forma de explicação necessária podem variar conforme o público-alvo que solicita a explicação, sejam eles usuários finais, especialistas ou desenvolvedores.

Noutro enfoque, também é necessário avaliar a viabilidade da explicabilidade dentro do atual estado da arte em Inteligência Artificial. Dessa maneira, considerando-se as limitações tecnológicas atuais, muitos algoritmos de IA não são completamente explicáveis, e em alguns casos, a explicabilidade pode acrescentar custos significativos de implementação e processamento (Deeks, 2019). Assim, pode haver uma troca entre a precisão/qualidade de um sistema de IA e o seu nível de explicabilidade (Gilpin *et al.*,

2018), onde, frequentemente, um aumento na explicabilidade pode diminuir a acurácia do sistema de IA, enquanto uma maior acurácia pode levar a uma menor transparência operacional (O'Neil, 2016).

No contexto nacional, o desafio mencionado é reconhecido no artigo 19, *caput*, da Resolução n. 332/2020 do CNJ, que orienta que os sistemas de IA usados como ferramentas auxiliares na tomada de decisão devem priorizar a técnica que permite explicar os passos que levaram ao resultado. Esse princípio de exigir a explicação dos "passos que conduziram ao resultado" serve para guiar a intersecção entre a Teoria do Direito e a Ciência da Computação no desenvolvimento de sistemas de IA que auxiliam a tomada de decisões públicas.

Adicionalmente, devido ao risco de opacidade, existe uma considerável crítica quanto à aplicação de sistemas de IA no processo de tomada de decisões públicas. Por outro lado, em que pesem as críticas, o Direito pode ser chamado a formular o método de decisão mediante claros limites jurídicos (Pádua, 2023b, p. 420). Um caminho a ser explorado é a utilização do processamento de linguagem natural (NLP) para desenvolver justificativas racionais para as decisões, algo ainda não plenamente desenvolvido, mas que se encontra na fronteira da Teoria do Direito e Ciência da Computação, visando a explicabilidade das decisões influenciadas por sistemas de IA (Hartmann Peixoto; Bonat, 2023), direta ou indiretamente, nas decisões públicas.

Desse modo, nos modelos de Inteligência Artificial voltados para a construção de decisões públicas, a teoria da argumentação (Atienza, 2016, p. 192-219) pode oferecer um modelo teórico útil e valioso, considerando-se sua abordagem analítica e procedimental para a justificação das decisões (Alexy, 2005, p. 261-264). Isso pode ser a base para desenvolver modelos de IA que integrem explicabilidade desde o início do desenvolvimento do sistema até o processamento dos resultados (Licht, K.; Licht, J., 2020), ou que permitam, através de um sistema auxiliar, proporcionar explicabilidade para os resultados obtidos por meio de *machine learning* (Cassol da Silva, 2024) ou *deep learning*.

Nessa perspectiva, a seção 68 da Recomendação da UNESCO (2021) demonstra a necessidade de revisão e adaptação dos marcos regulatórios e legais para assegurar a responsabilização pelos conteúdos e resultados produzidos pela IA em todas as fases de seu ciclo de vida, e também estipula o dever de *accountability* das entidades públicas responsáveis.

Além disso, a avaliação de impacto da implementação do sistema de IA, conforme a seção 50 da Recomendação da UNESCO (2021), deve ser realizada de forma democrática, permitindo a participação dos cidadãos e avaliando os benefícios e riscos relacionados aos direitos humanos, ao meio ambiente, à ética e à sociedade, além de contemplar medidas de prevenção, mitigação e monitoramento de riscos. Este é um passo necessário para que

o setor público (Poder Judiciário e Administração Pública) realize uma autoavaliação dos sistemas de IA propostos, garantindo que os métodos utilizados estejam alinhados com as práticas adequadas. Nesse caminho, conforme a seção 51 da Recomendação da UNESCO (2021), devem ser estabelecidos mecanismos de supervisão voltados para a auditabilidade, a rastreabilidade e a *explicabilidade* desses sistemas.

A avaliação de sistemas públicos de IA deve ser um esforço multidisciplinar (Susskind, 2010), envolvendo as partes interessadas de diversas áreas, como atores do ecossistema de IA, representantes da sociedade civil, servidores públicos, investidores, fabricantes, engenheiros, advogados e usuários, conforme a seção 69 da Recomendação da UNESCO (2021). Além disso, essa fase deve ser multicultural, pluralista e inclusiva.

Sobre isso, a seção 77 da Recomendação da UNESCO (2021) destaca a importância de abordar a interoperabilidade dos bancos de dados relacionados aos sistemas do setor público (Pádua; Berberi, 2021, p. 240), permitindo compartilhar dados de qualidade em um espaço comum, que deve ser seguro e protegido.

Portanto, o cenário descrito até aqui aponta para a necessidade de criar um ambiente normativo e institucional que posicione a explicabilidade como peça central na união dos princípios de confiança, responsabilidade, *accountability* e justiça no âmbito da Inteligência Artificial utilizada como auxiliar à tomada de decisões públicas.

3. A CONSTRUÇÃO DO DIREITO À EXPLICABILIDADE DA INTELIGÊNCIA ARTIFICIAL QUE AUXILIA DECISÕES ESTATAIS

A partir das premissas anteriores para o uso da IA como auxiliar na tomada de decisões públicas, é necessário analisar o contexto cultural de cada agente regulatório envolvido globalmente, bem como considerar as complexas relações políticas e comerciais entre o Brasil e a Europa (o que demanda um afinamento mínimo entre os respectivos marcos regulatórios). Nesse cenário, destacam-se diretrizes supranacionais como *as Ethical Guidelines for Trustworthy AI* (Diretrizes Éticas para IA Confiável) da União Europeia (2019), que visam estabelecer o conceito de IA confiável por meio de três componentes essenciais ao longo do ciclo de vida do sistema: a) *licitude*, assegurando a conformidade com todas as leis e regulamentos aplicáveis; b) *ética*, aderindo a princípios e valores éticos; e c) *robustez técnica e social*.

Desse modo, a regulamentação europeia em relação à IA emergiu como uma evolução das regras da Convenção do Conselho da Europa para a Proteção de Pessoas Relativamente ao Tratamento Automatizado de Dados de Caráter Pessoal (União Europeia, 1981). Essa normativa,

alinhando-se com os artigos 7 e 8 da Carta dos Direitos Fundamentais da União Europeia, busca garantir proteção aprimorada às pessoas em relação ao processamento automatizado de seus dados pessoais, incluindo coleta e operações diversas em tais dados, visando à preservação da privacidade e da proteção de dados.

Por sua vez, o art. 14, n. 2, alínea "g", e o art. 15, n. 1, alínea "h", do Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia (2016b) estabelecem que o direito à explicabilidade pode ser deduzido do direito de acesso a informações úteis que permitem compreender a lógica computacional utilizada na produção de decisões automatizadas. No entanto, questões de "interesse público" podem limitar o direito de oposição ao tratamento automatizado de dados pessoais (União Europeia, 2016b). É importante notar que, desde que haja uma revisão humana nas decisões tomadas com o auxílio de sistemas de IA, a decisão não seria considerada como "exclusivamente com tratamento automatizado", o que poderia influenciar o direito de oposição, na forma do art. 22º, do GDPR (União Europeia, 2016). No mais, a Diretiva n. 2016/680 também incorporou princípios éticos para a IA confiável, limitando o uso de decisões totalmente automatizadas em situações que afetam significativamente os indivíduos e a sociedade (União Europeia, 2016a).

Na sequência, de acordo com o Regulamento sobre Inteligência Artificial (*AI Act*) da União Europeia (2024), sistemas de IA considerados de "risco elevado" (aqueles que podem afetar direitos fundamentais, tanto diretamente quanto indiretamente) estão sujeitos à obrigação de fornecer documentação detalhada que ateste a sua explicabilidade. Portanto, o *AI Act* estipula que o direito à explicabilidade está condicionado à disponibilização de documentação técnica abrangente, que deverá cobrir aspectos como: a) *descrição geral do sistema de IA*: seu propósito, responsáveis, interação com hardware ou software externo, atualizações, colocação no mercado e instruções de uso e instalação; b) *detalhes do desenvolvimento*: métodos e estágios de desenvolvimento, *design* do sistema, arquitetura e algoritmos, incluindo metodologias de treinamento e interpretação dos resultados; c) *informações de monitoramento e operação*: limites de desempenho, identificação de resultados não intencionais e riscos, e medidas de supervisão humana; d) *sistema de gestão de riscos e alterações realizadas durante o ciclo de vida do sistema*; e e) *avaliação de desempenho pós-comercialização*, que inclui acompanhamento após a implementação do sistema (União Europeia, 2024).

Esses requisitos demonstram a necessidade de transparência e responsabilização em sistemas de IA de alto risco (como os utilizados para auxílio à tomada de decisões públicas), em linha com o compromisso da

União Europeia com a proteção dos direitos fundamentais dos indivíduos no contexto da digitalização e da automação.

Na União Europeia, o princípio da explicabilidade é visto como fundamental para construir e manter a confiança dos usuários em sistemas de IA, por meio de processos transparentes que clarifiquem as capacidades e as finalidades desses sistemas, dentro dos limites do que é tecnicamente viável. Isso se reflete na obrigação de que as decisões públicas auxiliadas pelos sistemas de IA devem ser compreensíveis para as pessoas afetadas, permitindo contestações ou pedidos de revisão.

Entretanto, a normativa europeia reconhece que nem sempre é possível fornecer uma explicação completa sobre como um modelo de IA chegou a uma determinada saída (*output*) ou decisão, especialmente nos casos de algoritmos *black box*. Em tais situações, outras medidas de explicabilidade, como rastreabilidade, auditabilidade e comunicação transparente sobre as capacidades do sistema, podem ser necessárias para garantir que sejam respeitados os direitos fundamentais (União Europeia, 2019).

Para além dos avanços no contexto da regulação europeia, em novembro de 2023, o Brasil e outros 27 países assinaram a Declaração de Bletchley, um instrumento internacional que reconhece o potencial da IA para a humanidade e estabelece esforços regulatórios voltados para vários aspectos jurídicos, incluindo o reconhecimento do direito à explicabilidade (Reino Unido, 2023).

Em 21 de março de 2024, as Nações Unidas (ONU) aprovaram sua primeira Resolução sobre o tema, documento que destacou o direito à explicabilidade das decisões estatais auxiliadas pelos sistemas de Inteligência Artificial:

> 4. Convoca os Estados Membros e convida outros interessados a tomar medidas para cooperar e fornecer assistência aos países em desenvolvimento para acesso inclusivo e equitativo aos benefícios da transformação digital e sistemas de inteligência artificial seguros, protegidos e confiáveis, incluindo por meio de: (...)
> (k) Promovendo transparência, previsibilidade, confiabilidade e *compreensibilidade* (grifo nosso) ao longo do ciclo de vida de sistemas de inteligência artificial que tomam ou apoiam decisões que impactam os usuários finais, incluindo o fornecimento de notificações e *explicações* (grifo nosso), e promovendo supervisão humana, como, por exemplo, através da revisão de decisões automatizadas e processos relacionados ou, quando apropriado e relevante, alternativas de decisão humana ou reparação eficaz e responsabilidade para aqueles adversamente impactados por decisões automatizadas de sistemas de inteligência artificial (Nações Unidas, 2024, tradução nossa).

Após a análise dos diversos marcos normativos internacionais e supranacionais sobre o tema, percebe-se que o direito à explicabilidade atua em simbiose com direito fundamental à **autodeterminação informativa**, que, no plano infraconstitucional, é explicitado no art. 2º, II, da Lei n. 13.709/2018. Dessa forma, o direito fundamental à autodeterminação informativa reconhece que o processamento automatizado de dados ameaça a capacidade de decisão autônoma do indivíduo a respeito da oportunidade, da extensão e das consequências do fornecimento de seus dados pessoais para terceiros, ainda que o direito à autodeterminação informativa não seja ilimitado, frente à proporcional e justificada necessidade na utilização pública de parte dos dados (Mendes, 2022, p. 36). Isso ocorre porque os riscos do processamento automatizado de dados residem "mais na finalidade do processamento e nas possibilidades de processamento do que no tipo dos dados tratados" (Mendes, 2022, p. 38).

Sendo assim, como um alicerce para o direito fundamental à explicabilidade, deve-se verificar os três elementos que delineiam o direito à autodeterminação informativa: a) o poder de decisão atribuído ao indivíduo sobre a coleta e a utilização de seus dados pessoais (Mendes, 2022, p. 39-40); b) suporte fático abstrato típico de uma estrutura de princípio de direito público (Alexy, 2011, p. 305); e c) possibilidade de identificação dos dados pessoais tratados, de maneira que os registros de dados pessoais sejam protegido (Mendes, 2022, p. 40) conforme as possibilidades tecnicamente possíveis. Nesse contexto, veja-se que, no julgamento da Medida Cautelar na Ação Direta de Inconstitucionalidade 6387 (ADI 6387 MC), o Supremo Tribunal Federal reconheceu a autodeterminação informativa como um direito fundamental inerente ao sistema constitucional brasileiro (Brasil, 2020b).

A partir das premissas até aqui desenvolvidas, o direito à explicabilidade da Inteligência Artificial nas decisões estatais é caracterizado como um direito fundamental que assegura aos cidadãos transparência e controle sobre as decisões automatizadas que os afetam. Esse direito é essencial para garantir a *accountability* das decisões estatais e promover a justiça e a igualdade.

Desse modo, para avançar no âmbito constitucional brasileiro, é necessário estabelecer as balizas do direito fundamental à explicabilidade, na forma delineada nesse estudo, de maneira a possibilitar o reconhecimento de seus elementos como inerentes a um *direito fundamental por atribuição* (Alexy, 2012, p. 74) que pode ser *inferido a partir de outros direitos fundamentais* (conforme pressupostos já descritos neste estudo).

Nessa linha, a respeito da sua *titularidade*, todas as pessoas têm direito à explicabilidade, permitindo-lhes compreender a utilização da IA nas decisões estatais (em relação ao grau de extensão, à finalidade, aos riscos, aos benefícios, aos resultados e aos demais dados e às informações disponíveis), o que decorre do mais amplo escrutínio a que devem ser submetidas as

decisões estatais (Habermas, 2012, p. 61). Ou seja, o direito fundamental à explicabilidade tem a *universalidade* como uma de suas características, promovendo a igualdade e a justiça no acesso às informações sobre decisões auxiliadas estatais por IA.

Quanto ao *conteúdo*, o direito fundamental à explicabilidade inclui a obrigação dos órgãos públicos de fornecer informações claras e compreensíveis sobre os critérios e procedimentos utilizados por sistemas de IA, de maneira a garantir a transparência e a auditabilidade das decisões. Nessa linha, as entidades públicas, em especial a Administração Pública e o Judiciário, são responsáveis por garantir a explicabilidade, assegurando que as decisões automatizadas ou auxiliadas por IA sejam justificadas e possam ser questionadas pelos cidadãos.

A *normatividade* do direito à explicabilidade possui eficácia plena (imediata), de maneira que ele pode ser exigido judicialmente, assegurando que os cidadãos tenham acesso às informações necessárias para entender as decisões estatais assistidas por IA.

Sobre suas características de *essencialidade e fundamentalidade*, o direito fundamental à explicabilidade é necessário à proteção à dignidade humana e ao funcionamento do Estado Democrático de Direito na era da sociedade da informação (Floridi, 2014, p. 55), promovendo a confiança pública nas decisões estatais assistidas por IA.

Ademais, tratando-se de direito que é inerente à dignidade humana, o direito fundamental à explicabilidade garante que os cidadãos mantenham o controle sobre as informações que afetam suas vidas, o que qualifica sua *inalienabilidade* e sua *indisponibilidade* que são próprias da típica *eficácia vertical* voltada às relações estatais (considerado o recorte adotado pelo presente artigo). No mesmo sentido, a *imprescritibilidade* do direito à explicabilidade se delineia na medida em que ele não se perde pelo não uso ao longo do tempo, assegurando que os cidadãos possam exigir explicações sobre decisões automatizadas a qualquer momento.

No que toca às hipóteses de sua *limitação* em determinados casos (sigilo comercial e industrial, proteção à própria dignidade humana e segurança nacional, por exemplo), é possível determinar ao direito à explicabilidade a *relatividade* que é própria dos direitos fundamentais, uma vez que as eventuais restrições do direito à explicabilidade em favor da força normativa de outros direitos fundamentais devem ser justificadas através do método da proporcionalidade (Alexy, 2012, p. 95).

No que se refere aos demais aspectos, o *suporte fático em sentido amplo* (mediante a estrutura triádica de bem protegido, intervenção e restrição) para o direito à explicabilidade, inerente à sua proteção *prima facie* (Alexy, 2012, p. 308). Assim, o direito fundamental à explicabilidade é estruturado através da *dinâmica triádica* de bem protegido, intervenção e restrição, assegura

a transparência e a justiça nas decisões estatais automatizadas. Esse direito protege os cidadãos, permitindo-lhes compreender e questionar as decisões auxiliadas por IA que os afetam, enquanto reconhece a necessidade de restrições proporcionais para proteger outros direitos fundamentais. Já o *suporte fático em sentido estrito* do direito à explicabilidade abrange as situações específicas em que decisões são tomadas com o auxílio de IA. Dessa forma, as entidades estatais "que tratam na prática o domínio da norma como normativo, não sucumbem a nenhuma normatividade apócrifa do fáctico" (Müller, 2010, p. 59-60). Por exemplo, quando uma decisão judicial é influenciada por um sistema de IA, os cidadãos têm o direito de entender como dados foram usados, como foram utilizados e quais algoritmos e critérios foram aplicados para auxiliar na construção do modelo de IA ou na tomada de decisão.

Portanto, após a incursão nos diversos diplomas normativos que vêm influenciando na construção do direito à explicabilidade das decisões públicas auxiliadas por IA, bem como nos elementos que o caracterizam como um direito fundamental por atribuição, destaca-se que esse direito (Pádua, 2023a, p. 270) também engloba:

a) a oportunidade de conhecimento para o indivíduo cujos dados pessoais foram tratados no processo de modelagem da IA, para que ele possa entender como seus dados são utilizados;

b) o fornecimento de explicações compreensíveis sobre a lógica, etapas e processos do desenvolvimento de sistemas de IA que auxiliam a tomada de decisão por autoridades públicas (da Administração Pública e do Judiciário), sempre respeitando os limites técnicos e o máximo alcance viável conforme a tecnologia utilizada;

c) a possibilidade de profissionais técnicos (com conhecimento em Ciência da Computação) desafiarem o sistema de IA por meio de informações que permitam a interpretabilidade do sistema computacional.

Este entendimento promove um elo estreito entre o Direito e a Ciência da Computação, propiciando avanços científicos e jurídicos que beneficiam a sociedade como um todo, garantindo-se a maior eficácia possível ao art. 5º, XXXIII, ao art. 37, *caput*, e ao art. 93, IX, da Constituição Federal.

CONCLUSÃO

A questão da interpretabilidade e da explicabilidade das decisões públicas não é um desafio recente, já que historicamente envolve a transparência na fundamentação/motivação.

Desse modo, o direito à explicabilidade das decisões públicas (administrativas e judiciais) que utilizam o auxílio de Inteligência Artificial está amparado por um conjunto de normas constitucionais e

legais relacionadas à *accountability*. Isso se estende desde uma interpretação aprofundada do artigo 20 da LGPD até a formulação dos conceitos de interpretabilidade e explicabilidade com base em várias iniciativas internacionais e supranacionais, incluindo os Princípios de Asilomar, a Recomendação n. 449 da OCDE, a Recomendação da UNESCO, as normativas da União Europeia, a Declaração de Bletchley e a Resolução das Nações Unidas sobre o uso da IA. Além disso, a Resolução n. 332/2020 do CNJ é outro instrumento normativo que aborda a explicabilidade das decisões públicas apoiadas por IA, cujos princípios podem ser analogicamente aplicados para a Administração Pública, no intuito de se garantir direitos fundamentais.

Portanto, a explicabilidade não deve ser vista como um obstáculo ao desenvolvimento científico da IA jurídica (conforme o art. 5º, IX, e o art. 218, *caput* e § 1º, da Constituição Federal), mas sim como uma expressão do avanço tecnológico, um valor protegido pela Constituição, que se ajusta aos limites estabelecidos pelo estado atual da tecnologia.

Logo, é possível se reconhecer o direito fundamental à explicabilidade da IA utilizada no auxílio às decisões estatais (da Administração Pública e do Poder Judiciário), mediante características típicas de um direito fundamental por atribuição (na forma desenvolvida no último capítulo).

Além disso, deve-se definir claramente os conceitos de transparência e explicabilidade em um sentido amplo. Isso é particularmente importante porque a explicabilidade, em seu sentido mais abrangente, pode ser dividida em duas categorias: *interpretabilidade*, que diz respeito aos processos computacionais da IA e aos resultados gerados, facilitando a compreensão por parte dos profissionais da tecnologia; e *explicabilidade em sentido estrito*, que se refere ao desenvolvimento de modelos em linguagem natural ou outras formas de representação que sejam facilmente compreensíveis pelos destinatários humanos das decisões auxiliadas por IA.

Portanto, frente ao aumento do uso de técnicas de Inteligência Artificial para auxiliar na tomada de decisões na esfera pública (pela Administração e pelo Judiciário), o direito fundamental à explicabilidade dessas decisões mediadas por IA deve ser reconhecido.

REFERÊNCIAS

ALEXY, Robert. *Teoria da Argumentação Jurídica*: A Teoria do Discurso Racional como Teoria da Justificação Jurídica. 2. ed. Tradução: Zilda Hutchinson Schild Silva. São Paulo: Landy, 2005.

ALEXY, Robert. Teoria dos Direitos Fundamentais. Tradução: Virgílio Afonso da Silva. *Revista dos Tribunais*, São Paulo, 2. ed., 2011.

ASIMOV, Isaac. *Eu, Robô*. Tradução: Aline Storto Pereira. São Paulo: Aleph, 2015. E-book Kindle.

ÁVILA, Humberto Bergmann. *Teoria dos Princípios*: da definição à aplicação dos princípios jurídicos. 5. ed. rev. e amp. São Paulo: Malheiros Editores, 2006.

BRASIL. Conselho Nacional de Justiça. *Resolução n. 332*, de 21 de agosto de 2020. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429. Acesso em: 24 abr. 2024.

BRASIL. Senado Federal. *Projeto de Lei n. 21, de 2020*. Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências. Brasília, DF: Senado Federal, 3 fev. 2022. Disponível em: legis.senado.leg.br/sdleg-getter/documento?dm=9063365&ts=1656528542410&disposition=inline. Acesso em: 24 abr. 2024.

BRASIL. Senado Federal. *Projeto de Lei n. 2338, de 2023*. Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Senado Federal, 3 mai. 2023. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1713911560851& disposition=inline. Acesso em: 24 abr. 2024.

BRASIL. Supremo Tribunal Federal (Plenário). *Medida Cautelar em Ação Direta de Inconstitucionalidade 6387*. Defere medida cautelar para afastar os efeitos da Medida Provisória n. 954/2020 em razão da ofensa ao direito fundamental à autodeterminação informativa. Autor: Conselho Federal da Ordem dos Advogados do Brasil. Relator: Min. Gilmar Mendes, 7 de maio de 2020. Disponível em: https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=754357629. Acesso em: 15 abr. 2024.

CASSOL DA SILVA, Bruno. *Inteligência Artificial Explicável e Decisões Judiciais*: Experimentações com o método SHAP. (Dissertação de Mestrado em Direito) – Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, 14 mar. 2024, p. 196.

DEEKS, Ashley. *The Judicial Demand for Explainable Artificial Intelligence*. Columbia Law Review, v. 119, n. 7, p. 1829-1850, 2020.

FLORIDI, Luciano. *The Fourth Revolution*: How the Infosphere is Reshaping Human Reality. Oxford: Oxford University Press, 2014.

FUTURE OF LIFE INSTITUTE. *Asilomar AI Principles*. 2017. Disponível em: https://futureoflife.org/ai-principles. Acesso em: 5 ago. 2020.

GILPIN, Leilani H.; BAU, David; YUAN, Ben Z.; BAJWA, Ayesha; SPECTER, Michael; KAGAL, Lalana. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Itália, p. 80-89, 2018. Disponível em: https://doi.org/10.1109/DSAA.2018.00018.

GUIDOTTI, Riccardo; MONREALE, Anna; RUGGIERI, Salvatore; TURINI, Franco; GIANNOTTI, Fosca; PEDRESCHI, Dino. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, v. 51, n. 5, artigo 93, jan. 2019, 42 pages. Disponível em: https://doi.org/10.1145/3236009.

HABERMAS, Jürgen. *A Inclusão do Outro*. Tradução: George Sperber. São Paulo: Loyola, 2002.

HABERMAS, Jürgen. *Teoria do Agir Comunicativo, 1*: Racionalidade da Ação e Racionalidade Social. Tradução: Paulo Astor Soethe. São Paulo: WMF Martins Fontes, 2012.

HARTMANN PEIXOTO, Fabiano. *Inteligência Artificial e Direito*: Convergência Ética e Estratégica. Curitiba: Alteridade, 2020, v. 5

HARTMANN PEIXOTO, Fabiano. Projeto Victor: Relato do Desenvolvimento da Inteligência Artificial na Repercussão Geral do Supremo Tribunal Federal. *Revista Brasileira de Inteligência Artificial e Direito*, v. 1, n. 1, jan-abr. 2020.

HARTMANN PEIXOTO, Fabiano; BONAT, Debora. *GPTs e Direito*: impactos prováveis das IAs generativas nas atividades jurídicas brasileiras. Seqüência: Estudos jurídicos e políticos, v. 44, n. 93, p. 1-31, Florianópolis, 2023. Disponível em: 10.5007/2177-7055.2023.e94238.

LICHT, Karl de Fine; LICHT, Jenny de Fine. *Artificial intelligence, transparency, and public decision-making.* AI & Society, mar. 2020. Disponível em: https://doi.org/10.5007/2177-7055.2023.e94238.

LORENZETTO, Bruno Meneses; CLÈVE, Clèmerson Merlin. Constituição Federal, Controle Jurisdicional e Níveis de Escrutínio. *Direitos Fundamentais & Justiça*, v. 32, jul./set. 2015, p. 97-123.

MENDES, Laura Schertel Ferreira. In: DONEDA, Danilo; SARLET, Ingo Wolfgang; MENDES, Laura Schertel Ferreira. *Estudos Sobre Proteção de Dados*

Pessoais: Dados num Mundo em Transformação. São Paulo: Expressa, 2022. E-book Kindle.

MÜLLER, Friedrich. *Metodologia do Direito Constitucional*. 4. ed. São Paulo: Revista dos Tribunais, 2010.

NAÇÕES UNIDAS. *UN AI Resolution (A/78/L.49)*. 24 mar. 2021. Disponível em: https://undocs.org/Home/Mobile?FinalSymbol=A%2F78%2FL.49&Language=E&DeviceType=Desktop&LangRequested=False. Acesso em: 24 abr. 2024.

OCDE. Recommendation 0449, of 21 May 2019: *Recommendation of the Council on Artificial Intelligence*. Disponível em: https://legalinstruments.oecd.org/en/instruments /OECD-LEGAL-0449. Acesso em: 20 abr. 2024.

O'NEIL, Cathy. *Weapons of Math Destruction*: how Big Data increases inequality and threatens democracy. Nova Iorque: Crown, 2016. E-book Kindle.

PÁDUA, Sérgio Rodrigo de. *Da Jurisdição "Ex Machina" ao Juiz Ciborgue*: Inteligência Artificial e Interpretação do Direito. São Paulo: Thomson Reuters, 2023.

PÁDUA, Sérgio Rodrigo de. Inteligência artificial judicial e a representação do suporte fático hipotético. Suprema - *Revista de Estudos Constitucionais,* v. 3, n. 1, p. 415-438, Brasília, 2023. Disponível em: https://doi.org/10.53798/suprema.2023.v3.n1.a224.

PÁDUA, Sérgio Rodrigo de; BERBERI, Marco Antonio. Robô Processual: Inteligência Artificial, Atos Processuais e Regras Padrão. *Revista da AGU*, v. 20, n. 3, Brasília, 2021. Disponível em: https://doi.org/10.25109/2525-328X.v.20.n.03.2021.2744.

PASQUALE, Frank. *Data-Informed Duties in AI Development*. Columbia Law Review, v. 119, n. 7, p. 1917-1940, 2019.

PASQUALE, Frank. Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society. *Ohio State Law Journal*, v. 78, 2017.

REINO UNIDO. *The Bletchley Declaration by Countries Attending the AI Safety Summit*, 1-2 November 2023. Disponível em: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-

declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Acesso em: 22 abr. 2024.

RUSSELL, Stuart; NORVIG, Peter. *Inteligência Artificial*. Tradução de Regina Célia Simille. Rio de Janeiro: Elsevier, 2013.

SALOMÃO, Luis Felipe (coord.) *et al*. *Inteligência Artificial*: Tecnologia Aplicada à Gestão dos Conflitos no Âmbito do Poder Judiciário Brasileiro. 2. ed. FGV Conhecimento: Centro de Inovação, Administração e Pesquisa do Judiciário, 2022. Disponível em: https://ciapj.fgv.br/sites/ciapj.fgv.br/files/relatorio_ia_2fase.pdf. Acesso em: 24 abr. 2024.

SUSSKIND, Richard. *The End of Lawyers?* Rethinking the Nature of Legal Services. Oxford: Oxford University Press, 2010. E-book Kindle.

THALER, Richard H.; SUNSTEIN, Cass R. *Nudge*. Tradução: Ângelo Lessa. Rio de Janeiro: Objetiva, 2019. E-book Kindle.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. 24 nov. 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000380455. Acesso em: 20 mar. 2022. Acesso em: 23 abr. 2024.

UNIÃO EUROPEIA. *AI Act* (Anexos). Disponível em: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0004.02/DOC_1&format=PDF. Acesso em: 24 abr. 2024.

UNIÃO EUROPEIA. *Carta de Direitos Fundamentais da União Europeia*. 2000. Disponível em: https://www.europarl.europa.eu/charter/pdf/text_pt.pdf. Acesso em: 24 abr. 2024.

UNIÃO EUROPEIA. *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*. Disponível em: https://rm.coe.int/1680078b37. Acesso em: 24 abr. 2024.

UNIÃO EUROPEIA. *Diretiva (UE) 2016/680 do Parlamento Europeu e do Conselho*, de 27 de abril de 2016. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016L0680&from=HR. Acesso em: 5 jan. 2023.

UNIÃO EUROPEIA. *Ethics Guidelines for Trustworthy AI*. 8 abr. 2019. Disponível em: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. Acesso em: 24 abr. 2024.

UNIÃO EUROPEIA. *Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho*, de 27 de abril de 2016. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=PT. Acesso em: 24 abr. 2024.

WAGNER, Ben. Accountability by design in technology research. *Computer Law & Security Review*, v. 37, p. 105398, ISSN 0267-3649, Oxford, 2020. Disponível em: https://doi.org/10.1016/j.clsr.2020.105398.