AS AVALIAÇÕES DE IMPACTO COMO INSTRUMENTOS DE INTELIGIBILIDADE ALGORÍTMICA E GARANTIA DE DIREITOS FUNDAMENTAIS NA REGULAÇÃO DE INTELIGÊNCIA ARTIFICIAL

Luis Henrique de Menezes Acioly[1]

Isabelle Brito Bezerra Mendes[2]

João Araújo Monteiro Neto[3]

Resumo

No contexto da discussão acerca da ética aplicada ao uso de sistemas de inteligência artificial, destaca-se a necessidade de transparência no processo decisório que envolva ou afete a vida humana. Tais algoritmos de aprendizado apresentam um potencial discriminatório que afeta diretamente direitos fundamentais, de forma que há uma movimentação de diversos atores para definição de parâmetros inerentes a uma IA responsável. O presente estudo parte dessa premissa para buscar fornecer subsídios à aplicação prática da transparência em sistema de IA, a partir da contextualização do conceito de inteligibilidade nos instrumentos de avaliação de impacto dessas tecnologias e de processos que envolvam dados pessoais. Para tanto, o presente estudo valeu-se da metodologia de revisão de literatura, a partir de pesquisa bibliográfica de abordagem qualitativa, para: contextualizar a discussão acerca sobre vetores éticos da IA ao panorama sobre discriminação algorítmica; abordar a conceituação de opacidade, transparência, inteligibilidade e explicabilidade, fazendo correlação com o movimento de IA Explicável; e analisar a estrutura dos instrumentos de avaliação de impacto para correlacioná-los à proteção de direitos fundamentais. Obteve-se como resultado principal que o conceito de explicabilidade como a interface entre sistema inteligente e ser humano, para ser atrelado a um instrumento de avaliação de impacto, precisa partir da premissa de que há uma gama de abordagens interativas, o que se convenciona denominar de multicamadas, viabilizando uma governança social dos algoritmos e promovendo direitos fundamentais no contexto tecnológico.

Palavras-chave: Inteligência Artificial. Transparência Algorítmica. Inteligibilidade. Avaliações de Impacto. IA Explicável.

---

[1] Graduado em Direito pelo Centro Universitário Ruy Barbosa (UniRuy). Pesquisador junto ao Grupo de Estudos em Tecnologia, Informação e Sociedade da UNIFOR – GETIS/CNPq. Vice-Presidente do Laboratório de Inovação e Direitos Digitais da UFBA - LABID². Pesquisador em Grupo de Pesquisa "Conversas Civilísticas" da UFBA. Treinee junto ao Chezzi Advogados. Orcid: https://orcid.org/0000-0002-1615-6048. Currículo Lattes: http://lattes.cnpq.br/0485009486625100. Contato: acioly10@gmail.com.

[2] Mestranda em Direito Constitucional pela Universidade Federal do Ceará. Pós-Graduanda em Proteção de Dados e Governança Digital pela UNIFOR. Pesquisadora junto ao Grupo de Estudos em Tecnologia, Informação e Sociedade da UNIFOR – GETIS/CNPq. Advogada. Tax Consultant na EY. Orcid: https://orcid.org/0000-0002-0573-616X. Currículo Lattes: http://lattes.cnpq.br/3286725229723354. Contato: isabellemendes06@gmail.com.

[3] PhD em Direito pela Universidade de Kent no Reino Unido. Mestre em Direito Constitucional pela Universidade de Fortaleza. Professor do Centro de Ciências Jurídicas da Universidade de Fortaleza. Coordenador do Grupo de Estudos em Tecnologia, Informação e Sociedade da UNIFOR – GETIS/CNPq. Advogado especializado em Proteção de Dados e Privacidade, Presidente da Comissão de Direito Digital da OAB/CE. Orcid: https://orcid.org/0000-0002-0690-2449. Currículo Lattes: http://lattes.cnpq.br/4255484163600547. Contato: joaoneto@unifor.br.

Abstract

In the context of the discussion on ethics applied to the use of artificial intelligence systems, the need for transparency in the decision-making process that involves or affects human life stands out. These learning algorithms have a discriminatory potential that directly affects fundamental rights, so that there is a transmission of different aspects to define parameters inherent to a responsible AI. The present study departs from this proposal to seek to provide support for the practical application of transparency in an AI system, based on the contextualization of the concept of intelligibility in instruments for evaluating the impact of these technologies and processes that involve personal data. To this end, the present study uses the literature review methodology, based on bibliographical research with a qualitative approach, to: contextualize the discussion about ethical vectors of AI to the panorama of algorithmic discrimination; Address the conceptualization of opacity, transparency, intelligibility and explainability, demonstrating with the Explainable AI movement; and analyze the structure of impact assessment instruments to correlate them with the protection of fundamental rights. The main result was that the concept of explainability as an interface between an intelligent system and a human being, to be linked to an impact assessment instrument, needs to be based on the location that there is a range of interactive approaches, which is conventionally called multilayer, enabling social governance of algorithms and promoting fundamental rights in the technological context.

Keywords: Artificial intelligence. Algorithmic Transparency. Intelligibility. Impact Assessment. Explanable AI.

1 INTRODUÇÃO

A utilização de sistemas de Inteligência Artificial (IA) na contemporaneidade tem trazido à tona uma gama de discussões no tocante à ética, aos riscos e à transparência de tais sistemas, especialmente quando se constata o poder decisório que os algoritmos detêm sobre a vida humana. A título de exemplo, são utilizados algoritmos de IA em relações de consumo para personalização de ofertas, decidindo sobre o quanto determinada pessoa deve pagar em razão de suas características individuais e/ou sociais, bem como para o direcionamento de publicidade personalizada, com base em um perfil de consumidor, afetando frontalmente a autonomia negocial do indivíduo (MEDON; FALEIROS, 2021; ACIOLY, 2022). A forma de decisão do algoritmo na realidade socioeconômica fundada em dados é tida como ativo financeiro e mercadológico das organizações, de forma que tem sido mantida sob o pálio da opacidade do segredo industrial.

O funcionamento de sistema de IA para tomada de decisões que envolvam a vida humana necessita, simbioticamente, de dados pessoais, seja para serem *inputs* a serem tratados pela cadeia de comando e geraram os resultados programados, seja para alimentar o sistema de aprendizagem de máquina e treinar a acurácia desse sistema (LACERDA, 2021). A forma como tais algoritmos operam com dados pessoais, bem como a medida de funcionamento da tomada de decisão é objeto de discussões acerca da necessidade de transparência e viabilidade para compreensão humana, tendo em vista o potencial discriminatório dessas ferramentas.

A título de exemplificação, Bettega (2021) aponta casos de discriminação algorítmica: (i) em 2018, a *Amazon* desistiu de utilizar um sistema de IA de recrutamento de colaboradores quando detectado que esse havia sido maculado com um viés discriminatório contra mulheres, privilegiando-se a contratação de homens, pois havia sido alimentado com dados do estado recente da empresa (há até 10 anos), de predomínio da figura masculina; (ii) após lançamento de cartão de crédito, a *Apple* foi posta em investigação, pois seu sistema de IA relacionava a questão de gênero ao limite oferecido, privilegiando-se homens com maiores linhas de crédito; (iii) em 2019, o Google admitiu uma falha em seu algoritmo de pesquisa que associava a busca por profissionais mulheres ou por lésbicas a conteúdo pornográfico.

Nesse contexto, a previsibilidade dos riscos se torna um importante aliado na conjectura necessária ao combate às diversas formas de discriminação algorítmica e proteção de direitos fundamentais na relação do ser humano com a inteligência artificial. Há, por conseguinte, uma variedade de instrumentos que viabilizam a compreensão dos riscos atinentes à inteligência artificial, inclusive em contextos regulatórios. Põe-se em foco o Relatório de Impacto à Proteção de Dados Pessoais (RIPD), previsto na Lei Geral de Proteção de Dados e a Avaliação de Impacto Algorítmico (AIA), presente no Projeto de Lei n. 2.338, de 2023.

O presente estudo parte dessas premissas para traçar o objetivo de compreender os elementos estruturais desses instrumentos e correlacioná-los à sua função de cognição de riscos em uso de dados pessoais e em inteligência artificial. Busca-se demonstrar mais um prisma de proteção de direitos fundamentais no contexto tecnológico a partir da necessidade de tornar a IA explicável – *Explainable AI* –, construindo premissas para o alcance da diretriz da transparência. Nesse contexto, aborda-se a discussão acerca de princípios éticos para o uso de IA, a discussão sobre a coesa conceituação sobre inteligibilidade do sistema de IA, e sistematização dos conceitos desses instrumentos e seu espectro de funcionalidade.

Para tanto, procedeu-se, metodologicamente, a partir da revisão crítica da literatura de referência no tema, à materialização em uma pesquisa bibliográfica e documental de cunho descritivo, natureza qualitativa e caráter exploratório, em que se preconizou o diálogo entre autores das diferentes áreas das ciências sociais. Empreendeu-se aqui uma revisão literária, materializada por meio de uma pesquisa bibliográfica, cuja coleta de dados se deu por livros, dissertações e artigos, repositados em bases de dados eletrônicas – *Scientific Electronic Library Online* (Scielo), *Index Law Journals* e Google Acadêmico –, tendo como descritores: cidadania virtual; Avaliação de Impacto; discriminação algorítmica; Transparência Algorítmica; Opacidade; Explicabilidade; inteligência artificial; aprendizagem de máquina; direitos fundamentais.

O presente constructo se materializa a partir de três capítulos de desenvolvimento e considerações finais, além desta introdução. O capítulo 2 apresenta o conceito de discriminação algorítmica e contextualiza a discussão acerca da implementação de vetores éticos ao desenvolvimento de inteligência artificial. O capítulo 3 aborda as questões que envolvem os conceitos de opacidade, transparência e explicabilidade, fazendo correlação com o movimento de IA Explicável. O capítulo 4 apresenta os elementos que compõem a gestão de risco na proposta legislativa de regulação de IA no Brasil e analisa a estrutura dos instrumentos de avaliação de impacto para correlacioná-los à proteção de direitos fundamentais. Ao cabo, são tecidas as considerações finais.

2 DISCRIMINAÇÃO ALGORÍTMICA E VETORES ÉTICOS PARA O DESENVOLVIMENTO DE INTELIGÊNCIA ARTIFICIAL

No contexto de uma sociedade hiperconectada, o uso de Inteligência Artificial se mostra cada dia mais acentuado, posto que sua popularidade se relaciona diretamente à abundância de recursos tecnológicos que servem de base para sua estrutura (BIGONHA, 2017), bem como à maior disponibilidade de dados pessoais que alimentam a aprendizagem de máquina (HOFFMANN-RIEM, 2020). A utilização de redes neurais artificiais em conjunto com a crescente exponencial de *big data* têm conduzido a uma conjuntura que torna a inteligência artificial um mecanismo de tomada de decisões sobre aspectos centrais da vida humana (LEE, 2019).

A utilização de sistemas baseados em IA, centrados em algoritmos, para tomada de decisões que afetem a vida humana pode, contudo, reverberar vieses discriminatórios em diversas formas, pois, ao mesmo tempo presente o risco de erros de programação ou vieses dos próprios programadores, podem apresentar resultados estatisticamente corretos, mas pautados em generalizações, as quais desconsideram ambiguidades e fatores subjetivos do indivíduo afetado (MENDES; MATTIUZZO, 2019). A discussão sobre o potencial discriminatórios da inteligência artificial tem conduzido a elaboração de diretrizes éticas, que orientam o desenvolvimento e aplicação dessa tecnologia.

2.1 Contornos gerais sobre Inteligência Artificial e Discriminação Algorítmica

Da importância de se falar em riscos discriminatórios em sistemas de inteligência artificial nasce o questionamento do que se considera uma "discriminação". Sem embargo da discussão sociojurídico que envolve a conceituação de uma discriminação, o presente estudo se filia à conceituação trazida por Duarte (2021) para quem a discriminação algorítmica decorre da construção social de práticas historicamente contingentes, pautadas na subjetivação de grupos de pessoas por fatores como a cor da pele, origem étnico-racial, gênero, orientação sexual, ou qualquer forma de estigmatização.

O conceito de discriminação algorítmica se afasta da semântica deferida a preconceito e estereótipo na medida em que, embora possa se fundar em alguns desses elementos, está efetivamente associado a um comportamento (DUARTE, 2021), uma ação concreta. Essa ação pode decorrer da subjetividade do programador ou na aplicação da IA, mas também pode ser inserida em um panorama mais amplo, a partir da identificação da abordagem de aprendizado implementada na máquina (BAROCA; SELBST, 2016).

A complexidade dos riscos inerentes à massiva utilização de sistemas de IA é descortinada na medida em que os algoritmos possuem a capacidade de aprendizagem com o meio social em que são inseridos, maculado com diversas formas de discriminação e preconceitos (DUARTE, 2021), assim como pelo fato de realizarem correlações para além da mera probabilidade, utilizando-se de redes complexas, de forma análoga ao processamento de informações por um cérebro, o que se convencionou denominar de redes neurais artificiais – *deep neural networks* (ACIOLY, 2022).

Alguns eventos têm levantado a discussão sobre o potencial discriminatório decorrente do aprendizado de máquina. Magrani (2019) cita o caso do sistema inteligente da Microsoft, denominado "Tay", que, dotado de uma capacidade de aprendizado profundo, ao ser inserido na rede social "Twitter", passou a interagir através de conversações com usuários e aprender com as experiências acumuladas. Com menos de vinte e quatro horas de funcionamento, a "Tay" passou a reproduzir conteúdo em textos que, além de violar os Termos de Uso da plataforma, tinham conteúdo racista, sexistas e antissemita (MAGRANI, 2019).

A complexidade desses sistemas é reconhecida, inclusive, na conceituação de Inteligência Artificial pelo Grupo Independente de Peritos de Alto Nível sobre a Inteligência Artificial (GPAN IA) da Comissão Europeia:

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

Aprendizado de máquina ("*machine learning*") diz respeito ao termo cunhado por Arthur Samuel em 1959, para definir o campo de estudo da ciência da computação que dá à máquina a habilidade de aprender sem ser explicitamente programado (SAMUEL, 1959; GABRIEL, 2022). O aprendizado de máquina constitui a forma como a árvore de decisões é estruturada em um dado sistema inteligente com vistas à formação de um resultado, constituindo critério para sua configuração como inteligência artificial (GABRIEL, 2022).

As formas de *machine learning* podem variar de acordo com a estruturação dos algoritmos de aprendizagem, sendo geralmente categorizado em: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço (GABRIEL, 2022; RUSSELL; NORVIG, 2013). O aprendizado supervisionado se desenvolve a partir de métodos de regressão e classificação, casos em que são fornecidos, além do *input*, um padrão de resultado esperado ("*target*"), classificando a memória do sistema na idealização do resultado e aplicação desse padrão (BARTNECK *et al*., 2021). O aprendizado não supervisionado ("*free learning*") é desenvolvido quando não há resultado esperado para determinados *inputs*, focando no potencial de agrupamento de dados ("*clusters*") para a formação de padrões (GABRIEL, 2022). A seu

turno, o aprendizado por reforço ("*reinforcement learning*") se desenvolve a partir de interações – seja por recompensas ou punições – que estimulam e reforçam a aprendizagem em um sistema de "tentativa e erro" (GABRIEL, 2022).

As redes neurais, ou neurônios artificiais, subsidiam o processo de *deep learning*, constituindo ferramentas de concretização de uma frente multicamada da inteligência artificial (RUSSELL; NORVIG, 2013). A estrutura em camadas de encadeamento lógico se torna imbricada o suficiente elaborar novas formas de interação dos "neurônios", para além da programação original, o que se convencionou denominar *deep neural network* (ACIOLY, 2022; GABRIEL, 2022). A atuação desses algoritmos se torna extremamente complexa, que apresentam uma questão de opacidade, a partir da inescrutabilidade dos caminhos que conduziram a tomada de decisão da inteligência artificial (SCHIPPERS, 2018).

Para esse contexto, Barocas e Selbst (2016) trazem à luz a inserção de potenciais discriminatórios na estruturação de um sistema inteligente, a partir de: (i) definição da variável-alvo ("*target*") e dos rótulos de classe; (ii) treinamento do sistema com dados; (iii) seleção de recursos, ou critérios; (iv) *proxies*; e (v) mascaramento. A seleção da variável-alvo pode representar um contexto discriminatório, na medida em que o "*target*" escolhido pode, ao estabelecer um caminho de ordenação dos algoritmos, representar valores do programador, conscientes ou não, que resultem em um injusto resultado (BAROCAS; SELBST, 2016).

O fomento de dados pessoais nos sistemas de IA ("*dataset*") com a finalidade de treinamento e os critérios estabelecidos para a construção do aprendizado, principalmente em sistemas de aprendizado não supervisionado, semelhantemente podem conduzir a um panorama discriminatório, tendo em vista não necessariamente o alvo, mas o caminho de aprendizado traçado para se chegar a um resultado (SCHIPPERS, 2018). A remoção de informações potencialmente discriminatórias, como raça e gênero, nas variáveis de entrada, no escopo da rotulagem ou ainda no "*dataset*", não garante que o aprendizado não se tornará discriminatório (SCHIPPERS, 2018), tendo em vista que o processo de aprendizagem pode gerar inferências a partir da conjugação de dados intermediários – os chamados *proxies* – levando a resultados carregados de premissas enviesadas (BAROCAS; SELBST, 2016).

Da conjugação das diversas formas de discriminação e dos inúmeros caminhos pelos quais ela pode acontecer a partir do uso de dados pessoais em sistemas de inteligência artificial, nasce a necessidade de adoção de mecanismos técnico-jurídicos capazes de estabelecer parâmetros mínimos para sua utilização e desenvolvimento.

2.2 Diretrizes Éticas sobre Inteligência Artificial

A dificuldade de compreensão dos algoritmos, aliada à ausência de transparência em seus modelos formativos, é criticada por Pasquale (2015), apontando para uma sociedade de segredos, denominando-a de *Black Box Society*. Diante dessa opacidade, *one way mirror*[4], é impossível compreender se os resultados foram justos ou se atenderam exclusivamente a interesses econômicos ou escusos dos programadores (PASQUALE, 2015).

A denominação "*Black Box Society*" se consolida a partir da difusão generalística de algoritmos predominantemente opacos, notadamente em aplicação nos setores financeiro e mercado securitário. Essa noção ganha relevo quando se questiona a forma de regulação de setores da economia cujos procedimentos são de conhecimento social limitado ou inexistente (PASQUALE, 2015). Insta ressaltar que essa parcela de mercado está estritamente relacionada ao acesso a direitos fundamentais como propriedade e autonomia negocial, como face do livre desenvolvimento da personalidade humana.

Levantando questões éticas para o uso de IA, Mulholland e Frajhof (2021) assinalam que a opacidade das decisões tomadas por algoritmos impede a avaliação de legitimidade de legalidade do código, bem como que a ausência de atividade humana nesse processo abre o dilema acerca da imputabilidade e responsabilidade civil em casos de danos.

A ética no contexto da regulamentação do uso de IA pressupõe a existência de vetores axiológicos de plena aplicabilidade, tais quais formulados pela *Fairness, Accountability and Transparency in Machine Learning Organization* (FAT-ML), a saber: (i) Responsabilidade; (ii) Explicabilidade; (iii) Precisão; (iv) Auditabilidade; e (v) Justiça (DUARTE, 2021). No campo doutrinário brasileiro, Mullholand e Frajhof (2021) elencam como princípios a beneficência, a não-maleficência, a autonomia e justiça, atrelados às diretrizes de justiça, acurácia e inteligibilidade. Na lição das autoras (MULHOLLAND;

[4] O conceito de "*one way mirror*" se relaciona com a figura dos espelhos unidirecionais utilizados em estabelecimentos policiais americanos, que permite a vigilância do ambiente por quem está atrás do espelho e reflete a imagem para aqueles que estão internos à sala. Cuidou o autor (PASQUALE, 2015) de realizar essa analogia para demonstrar o estado de supervigilância ocasionado pela opacidade dos algoritmos presentes no cotidiano social, permitindo a coleta massiva de informações pessoais dos cidadãos por grandes *players* do mercado e por governos.

FRAJHOF, 2021), a aplicabilidade desses vetores se relaciona a tomada de medidas preventivas e formativas:

> O reconhecimento desses princípios representa, resumidamente, a adoção de medidas que (i) impeçam a aplicação de sistemas de IA que violem o princípio da igualdade de tratamento; (ii) permitam reconhecer que os insumos utilizados pela IA e os resultados que advém de seu tratamento sejam precisos; e (iii) proporcionem à pessoa humana o conhecimento dos processos de decisão tomados pela IA (MULHOLLAND; FRAJHOF, 2021, p. 73).

Há ademais, uma gama de instituições e entidades que emitiram guidelines para o desenvolvimento de inteligência artificial. À guisa de exemplo, Mulholland e Gomes (2021) sintetizam a existência de documentos idealizados por instituições privadas transnacionais como a *Microsoft* e a *IBM*, bem como produzidos por entidades supranacionais, como a Conferência de Asilomar e a *Acess Now* e a *Association for Computing Machinery* ("*ACM*"), que estruturam suas diretrizes éticas no entorno da responsabilidade e precaução.

Pondo em foco instituições governamentais ou que tenham em consideração a reunião de entes estatais para a formulação de orientações, de caráter vinculativo ou não, pondera-se a *Recommendation of the Council on Artificial Intelligence*, da Organização para a Cooperação e Desenvolvimento Econômico (OCDE, 2019) e as Orientações Éticas para uma IA de Confiança, do Grupo Independente de Peritos de Alto Nível sobre a Inteligência Artificial da Comissão Europeia (COMISSÃO EUROPEIA, 2019).

A Organização para a Cooperação e Desenvolvimento Econômico ("*OCDE*") promoveu uma Recomendação a seu conselho, em que foram estabelecidos os princípios de (OCDE, 2019): (i) crescimento inclusivo, desenvolvimento sustentável e bem-estar, isto é, os agentes de IA devem realizar uma administração responsável para a produção de resultados benéficos para as pessoas e para o planeta; (ii) justiça e valores centrados no ser humano, isto é, os agentes de IA devem respeitar o estado de direito, os direitos humanos e os valores democráticos durante todo o ciclo de vida da IA; (iii) transparência e explicabilidade, isto é, os agentes de IA devem fornecer informações significativas, adequadas ao contexto e consistentes com o estado da arte do tema; (iv) robustez, segurança e precaução, em vista a promover o gerenciamento de riscos e fortalecer a segurança integral do sistema; e (v) *accountability*, que assume feição de responsabilidade, virtude de responsabilizar-se pelo bom funcionamento do sistema de IA (OCDE, 2019).

A seu turno, o Grupo Independente de Peritos de Alto Nível sobre a Inteligência Artificial erigiu os seguintes ditames éticos para o desenvolvimento de uma IA confiável: (i) respeito à autonomia humana; (ii) prevenção de danos; (iii) equidade; (iv) explicabilidade (COMISSÃO EUROPEIA, 2019). Desse panorama de diretrizes éticas para o desenvolvimento de inteligência artificial se extrai os contornos para uma discussão sobre regulação de inteligência artificial, intervenção estatal para proteção de direitos humanos. O caráter flexível desses modais deontológicos defere-lhes uma função de suporte à formulação de estruturas regulatórias para a concretização de direitos fundamentais, contextualizando-os à sociedade da informação.

3 OPACIDADE, INTELIGIBILIDADE E EXPLICABILIDADE: CONCEITOS EM DEBATE

Tem-se a explicabilidade dos sistemas inteligentes como supedâneo ao controle social dos algoritmos e escrutínio público dos riscos discriminatórios decorrentes do aprendizado de máquina (MITTELSTADT *et al.*, 2016). O mascaramento de vieses a partir do argumento de neutralidade tecnológica (BAROCAS; SELBST, 2016) tem deslocado um centro de poder para os desenvolvedores do sistema de IA e grandes corporações do mercado tecnológico (PASQUALE, 2015). O desconhecimento estratégico que circunscreve a opacidade das redes neurais artificiais tem sido utilizado politicamente para evitar responsabilidades dos agentes de IA (BUCHER, 2018), postergando ou mitigando discussões sobre resultados discriminatórios.

A opacidade dos algoritmos, a partir de sua contextualização como antônimo à transparência, conduz à imperiosa a adoção de mecanismos específicos de efetivação, uma vez que se depende da compreensibilidade para a constatação do atendimento às demais diretrizes principiológicas. Nos contornos da Lei Geral de Proteção de Dados, o princípio da transparência significa a "garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial" (BRASIL, 2018). O Projeto de Lei n. 2.338, de 2023, não apresentou uma definição para transparência, fazendo apenas menção na fileira de princípios para o desenvolvimento, implementação e uso de sistemas de inteligência artificial, ao lado dos princípios de explicabilidade, inteligibilidade e auditabilidade.

Desse contexto, toma-se como imprescindível afastar a concepção de transparência para a mera abertura do código-fonte (KROLL *et al.*, 2017; DUARTE, 2021), mas deve-se trazer à luz a compreensão de uma dimensão material deste princípio, consubstanciada na inteligibilidade do processo de decisão algorítmica, isto é, a possibilidade de ele ser compreensível ao homem-médio (FRAZÃO; GOETTENAUER, 2021). Esse é o resultado da combinação de uma série de fatores relacionados aos sistemas algorítmicos, dos quais os dados pessoais são apenas parte, ainda que relevante e significativa (FRAZÃO; GOETTENAUER, 2021). Nesse ponto, Pereira (2021) expõe o tema sob o prisma de IA aplicada às plataformas de interação social:

> A inteligibilidade de sistemas de ML [*machine learning*] não deve consistir necessariamente em uma descrição precisa e detalhada de como os algoritmos funcionam, especialmente porque tal forma de fornecimento de informações pode ser inútil para o usuário final de uma plataforma de mídia social que recebe desinformação e que pretende aprender mais sobre como as informações são direcionadas para sua conta (PEREIRA, 2021, p. 97).

Em breve síntese, Angelov *et al.* (2021) apresentam a taxonomia da discussão sobre transparência algorítmica, trazendo um panorama conceitual ao qual o presente estudo se filia. Segundo os autores (ANGELOV *et al*., 2021), transparência é um estado em potencial de um sistema de inteligência artificial ser compreendido, sendo o oposto de "blackboxes", enquanto interpretabilidade é a capacidade desse sistema fornecer interpretações em termos que possam ser compreendidos pelos seres humanos. Explicabilidade seria, assim, a noção de explicação como interface entre ser humano e sistema de inteligência artificial (ANGELOV *et al*., 2021).

Por sua vez, a inteligibilidade deve ser entendida como compreensibilidade racional, e pode ser categorizada conforme diferentes parâmetros. A título de comparação, Rudin (2019). destaca que explicabilidade é a possibilidade de o sistema ser inteligível através de explicações adicionais, não nativas ao próprio código, enquanto a interpretabilidade é a compreensibilidade intrínseca, inerente, a despeito de qualquer explicação adicional.

Quanto ao momento em que a inteligibilidade é implementada, têm-se: (1) *pre-model*; (2) *in-model*; e (3) *post-model* (PEREIRA, 2021) Importa ressaltar que a implementação apriorística de mecanismos de inteligibilidade – *pre-model* – tem o condão de melhor atender aos princípios de transparência, segurança e não-discriminação que rege o tratamento de dados

pessoais em âmbito brasileiro, uma vez que se coaduna com a própria noção de *privacy-by-design*[5].

Quanto à amplitude da inteligibilidade, Carvalho, Pereira e Cardoso (2019) sistematizam em três espécies: (i) transparência algorítmica, (ii) a interpretabilidade global e a (iii) interpretabilidade local. Especificamente quanto à transparência algorítmica, esta, por permitir compreender como o algoritmo aprende com os dados pessoais e que tipos de relações podem ser extraídas de tal operação, de forma abstrata e não individual (PEREIRA, 2021), atende ao desiderato de prestação de contas (KROLL *et al*., 2017), inclinando-se à existência normativa do direito à explicação de decisões automatizadas na LGPD. Não se busca a compreensão do funcionamento do sistema, mas somente da estrutura do algoritmo de aprendizado (PEREIRA, 2021). Contextualizando essa discussão, Pereira (2021) aponta que a redação da Lei Geral de Proteção de Dados Pessoais ("Lei n. 13.709/2018") apresenta um "grau considerável de incerteza sobre como e para que fins as informações sobre os critérios e procedimentos utilizados para uma decisão automatizada devem ser fornecidas".

Ademais, a interpretabilidade global tem sua funcionalidade direcionada à descrição do comportamento do modelo, o que incluir a compreensão das categorias de dados que são necessárias e a descrição detalhada do funcionamento do (CONFALONIERI *et al*., 2020; PHILLIPS *et al*., 2021). A interpretabilidade local, por sua vez, descreve na ótica individual como um sistema chegou a dado resultado, a partir de um *input* específico (CONFALONIERI *et al*., 2020; PHILLIPS *et al*., 2021). A utilidade das abordagens de inteligibilidade varia de acordo com a necessidade a qual se destinam (PEREIRA, 2021), ponderando os interesses que melhor se adequam ao caso concreto, a partir de uma dinâmica de sopesamento (KROLL *et al*., 2017).

A inteligibilidade do modelo de aprendizado de máquina, inclusive, também gera discussões no âmbito do *design thinking*. Na visão de Ruiz e Quaresma (2021), no contexto da IA explicável, *XAI - Explainable AI*, há a necessidade de procedimentos adaptativos ao contexto, ou seja, sistemas que construam modelos explicativos contextuais para a espécie de fenômeno a ser explicado. Enxergar uma explicação como uma linha de raciocínio representa,

[5] Como elucida, Bioni (2021), *privacy-by-design* é a ideia de que a proteção de dados pessoais deve orientar a concepção de um produto ou serviço, aplicando desde então tecnologias que facilitem o controle de fluxo e a proteção de dados pessoais. A Lei Geral de Proteção de Dados (LGPD) estabelece que medidas de segurança e proteção de dados pessoais devem ser observadas desde a concepção do produto ou serviço até a sua execução (art. 46, § 2º).

principalmente, compreender tal explicação como tão somente traços da forma como as regras de inferência são utilizadas pelo sistema para tomar determinada decisão (CONFALONIERI *et al*., 2020). É necessário fazer uma tradução de um sistema aos limites cognitivos humanos.

Phillips *et al.* (2021) trazem à luz um conjunto principiológico para reger a *Explanable AI* (XAI): (i) o princípio da explicação, que concerne afirmar que um sistema de IA deve fornecer evidências, suporte; ou raciocínio para cada decisão tomada pelo sistema; (ii) o princípio da significação, que atribui sentido à explicação fornecida pela IA, a qual deve ser compreensível e significativa para seus usuários, ajustável às necessidades de cada grupo de pessoas; (iii) o princípio da precisão, segundo o qual a explicação fornecida deve ser precisa e fidedigna com os processos efetivamente realizados no sistema; e (iv) o princípio do limite de conhecimento, através do qual se consigna o reconhecimento de que os sistemas de IA devem identificar casos para os quais não foram concebidos para operar e, portanto, as suas respostas podem não ser confiáveis.

O movimento de *Explanable AI* (*XAI*) preconiza não somente o embate às "*blackbloxes*", mas também a geração de uma Inteligência Artificial responsável, capaz de produzir modelos de transparência por design, desde a sua concepção (ADADI; BERRADA, 2018). A XAI faz parte de uma nova geração de tecnologias de IA chamada de terceira onda de IA, que tem por objetivo gerar algoritmos que possam se explicar com precisão concepção (ADADI; BERRADA, 2018).

Para um sistema de geração de explicações, Confalonieri *et al*. (2020) apontam um conjunto de parâmetros que podem reger a interface entre ser humano e inteligência artificial: (i) causal, isto é, que buscam fornecer uma relação causal entre os dados de entrada e saída no sistema de IA; (ii) contrafactual, isto é, as informações contrastantes podem gerar evidências empíricas que são melhor interpretadas pelos seres humanos; (iii) social, através do qual se busca uma transferência interativa de conhecimento, na qual a informação é adaptada de acordo com a formação e o nível de especialização do destinatário; (iv) seletivo, de forma que o conteúdo informativo das explicações deve ser selecionado de acordo com a formação e as necessidades do usuário, sem a necessidade de exposição massiva da causa completa de um evento; (v) semântica, isto é, a representação formal e o raciocínio podem implementar várias formas de manipulação do conhecimento, como abstração e refinamento, exercendo um papel importante na seleção de informações e adaptação social da explicação; (vi) interativo, pois as explicações devem ser interativas, permitindo ao operador do sistema de IA revisar e consolidar

alguns conhecimentos prévios, aprimorando o sistema de explicação com as evidências coletadas nesse processo.

Nessa conjectura de fatores, a inteligibilidade, nesse ponto, pode ser implementada de forma apriorística – *pre-model* – por meio de instrumentos de transparência e prestação de contas em um contexto de regulação e governança, especialmente quanto aos dados pessoais que alimentam a aprendizagem de máquina e as informações pessoais utilizadas como *inputs* para geração de resultados. Nesse ponto, destacam-se as avaliações de impacto – como o Relatório de Impacto à Proteção de Dados Pessoais (RIPD) e a Avaliação de Impacto Algorítmico (AIA) – como documentos voltados à responsabilização e prestação de contas.

Estruturalmente, é importante ressaltar que, ainda que ocorra a anonimização de dados pessoais que sirvam de subsídio para o *machine learning* do sistema de IA, que em tese desconfiguraria a pessoalidade do dado, aplica-se o ecossistema protetivo da LGPD quando há uma relação de causa e efeito que o mero tratamento de dados possa exercer sobre o indivíduo, tal como ocorre com algoritmos que mineram dados anonimizados, valendo-se para tanto da Teoria Consequencialista do dado pessoal (BIONI, 2021).

Deve-se ter em mente que a avaliação do desempenho de sistemas de IA diz respeito à sua robustez e desenvolvimento ético, isto é, à capacidade do sistema inteligente de, no caso concreto, agir com precisão e especificidade, ao passo que garanta a segurança, justeza e não discriminação do sistema (FLORIDI *et al*., 2022). A inteligibilidade, nesse contexto, atua como forma de garantir a observância das métricas de robustez e desenvolvimento ético, desde a sua concepção, de maneira que se viabilize um sistema de auditoria e prestação de contas na governança social dos algoritmos (KROLL *et al*., 2017).

O avanço na viabilização prática do princípio ético da transparência em sistemas de inteligência artificial tem o condão de conduzir a um ambiente de responsabilidade e prevenção contra o potencial discriminatório dos algoritmos de aprendizado, ponderando sua utilização no tratamento automatizado de dados pessoais.

4 AS AVALIAÇÕES DE IMPACTO COMO INSTRUMENTOS DE COGNIÇÃO DE RISCOS E PROTEÇÃO DE DIREITOS FUNDAMENTAIS

No ambiente regulatório, algumas iniciativas legislativas têm ganhado destaque por buscarem o estabelecimento de normas a nível federal, em caráter geral, sobre inteligência artificial. Pode-se citar o Projeto de Lei n. 5.051, de 2019, o PL n. 21, de 2020, e o PL n. 872, de 2021, que por um lapso relevante de tema foram os principais projetos de lei federal que versaram sobre o tema (MOURÃO, 2023). Após a formulação da minuta de um substitutivo a essas propostas legislativas por uma comissão de juristas, a CJUSBIA, o Projeto de Lei n. 2.338, de 2023, assumiu a dianteira nesse cenário, com robustos mecanismos normativos para consagração de uma Inteligência Artificial Responsável.

A minuta do Substitutivo – e, consequentemente, do PL 2.338, de 2023 – adotou uma modelagem baseada em riscos, tendo em vista endereçar obrigações específicas a partir de mecanismos de cognição de riscos e governança em IA (MENDONÇA JÚNIOR; NUNES, 2023). Importa, na hipótese, conhecer o tratamento destinado ao risco nesse projeto de lei para posteriormente correlacioná-lo à possibilidade de utilização dos instrumentos de cognição de risco como meios de concretização do vetor de transparência.

4.1 Tratamento destinado ao risco no Projeto de Lei n. 2.338 de 2023

O Projeto de Lei n. 2.338, de 2023prevê uma estratificação dos riscos que o sistema de IA, que considera não somente os riscos sociais, mas também aqueles inerentes à segurança e qualidade. Para tanto apontou os estratos de: (i) risco excessivo; (ii) alto risco e (iii) demais riscos. À luz do art. 14 da propositura legislativa, há vedação expressa aos usos de sistema de inteligência artificial para: (i) induzir a pessoa natural a se comportar de forma prejudicial ou perigosa à sua saúde ou segurança a partir de técnicas subliminares; (ii) explorar vulnerabilidades de grupos específicos de pessoas naturais, de forma a induzi-las a se comportar de forma prejudicial ou perigosa à sua saúde ou segurança; e (iii) avaliar, classificar ou ranquear as pessoas naturais, pelo Poder Público, com base em atributos de sua personalidade ou comportamento social, por meio de pontuação universal, para o acesso a bens, serviços ou políticas públicas (BRASIL, 2023).

Considerou-se como risco excessivo o uso de sistemas de identificação biométrica à distância, de forma contínua em espaços acessíveis ao público, no âmbito das atividades de

segurança pública, para o qual o art. 15 da Proposta de Lei trouxe como critérios a previsão em lei federal específica, para atividade geral, ou a autorização judicial para a persecução penal direcionada a pessoa individualizada, e restringiu aos casos de: (i) persecução de crimes passíveis de pena máxima de reclusão superior a dois anos; (ii) busca de vítimas de crimes ou pessoas desaparecidas; e (iii) crimes em flagrante delito (BRASIL, 2023). De todo modo, o art. 16 do PL n. 2.338, de 2023, defere à autoridade competente o papel de regulamentar os sistemas de IA de risco excessivo (BRASIL, 2023).

O Projeto de Lei n. 2.338, de 2023, estabelece como IA de alto risco o sistema utilizado para as finalidades dispostas em seu art. 17 (BRASIL, 2023). Essa lista, contudo, poderá ser alargada pela autoridade competente, após consulta ao órgão regulador setorial respectivo, com base nos critérios dispostos no art. 18 da Propositura Legislativa: (i) implementação em larga escala; (ii) possibilidade de o sistema impactar negativamente o exercício de direitos ou liberdades; (iii) alto potencial danoso de ordem moral ou material, ou discriminação; (iv) possibilidade de afetar pessoas de um grupo vulnerável específico; (v) possibilidade de os resultados prejudiciais do sistema serem irreversíveis ou de difícil reversão; (vi) constatação de um sistema de IA similar ter causado danos; (vii) baixo grau de transparência, explicabilidade ou auditabilidade, que dificulte o seu controle ou supervisão; (viii) alto nível de identificabilidade dos titulares de dados pessoais; (ix) existência de expectativa razoável do afetado quanto aos seus dados pessoais, tal como a expectativa de confidencialidade (BRASIL, 2023).

O Projeto de Lei estabelece a obrigatoriedade de o agente de inteligência artificial realizar uma análise preliminar, que resulte em um relatório, em que se identifique se o sistema se insere em alguma categoria de alto risco ou risco excessivo, de forma prévia à sua inserção no mercado ou utilização em serviço, na forma de seu art. 13 (BRASIL, 2023). A estratificação do sistema de IA como de alto risco resulta na obrigação de adotar as medidas de governança previstas no art. 20 da proposta de lei e elaborar uma Avaliação de Impacto Algorítmico, para além das disposições gerais sobre governança previstas no art. 19 do projeto (BRASIL, 2023), denotando uma carga suplementar de medidas de gestão do impacto do sistema inteligente, proporcional ao risco promovido, sob pena de incorrer em penalidade administrativa.

4.2 Avaliações de Impacto e sistema de explicação multicamada

O presente estudo põe em foco a Avaliação de Impacto Algorítmico e o Relatório de Impacto à Proteção de Dados, que tem sua função diretamente ligado à cognição de riscos para fins de governança e regulação em tecnologia. Uma avaliação de impacto é um processo específico para avaliar e documentar os impactos de um dado projeto ou empreendimento em determinadas áreas ou a partir de determinadas abordagens, e atribuir responsabilidades na mitigação desses impactos (RAAB, 2020). Incluindo a avaliação de impacto em um contexto dialógico, Metcalf *et al*. (2021) asseveram que estes instrumentos têm evoluído como mecanismos de responsabilização e prestação de contas, demonstrando que a conceituação e gestão dos impactos se fundam numa construção social, através de organizações e contestações legais, políticas e epistêmicas.

A Avaliação de Impacto Algorítmico (AIA) é considerada um instrumento mais amplo do que o Relatório de Impacto à Proteção de Dados, na medida em que não se esgota nos aspectos inerentes aos dados pessoais, mas relaciona-se com a própria programação algorítmica e com o aprendizado de máquina (LEMOS *et al*., 2023). Pode-se contextualizar a Avaliação de Impacto Algorítmico como sendo um instrumento de governança com objetivo de delinear a responsabilização, identificar possíveis danos oriundos da atividade de sistemas algorítmicos e dispor medidas de mitigação desses danos (METCALF *et al*., 2021). Selbst (2021) sintetiza as funções precípuas do AIA como instrumento de regulação em: (i) exigir que agentes de IA realizem uma avaliação dos impactos sociais de seus sistemas de forma prévia à sua inserção em mercado, ou utilização em serviço; e (ii) criar documentações sobre a avaliação realizada e os seus resultados para apoiar o desenvolvimento de políticas em termos de IA.

No panorama do Projeto de Lei n. 2.338, de 2023, a Avaliação de Impacto Algorítmico é disposta em seus art. 24, que prevê a obrigatoriedade de ela, ao menos, registrar: (i) os riscos conhecidos e previsíveis à época da elaboração do sistema inteligente; (ii) os benefícios associados ao sistema; (iii) a probabilidade de efeitos adversos e a quantidade de pessoas possivelmente impactadas; (iv) a gravidade das consequências e as medidas de mitigação; (v) a lógica do funcionamento do sistema inteligente; (vi) o processo e resultado dos testes e avaliações para verificação de possíveis impactos a direitos; (vii) o treinamento e ações de conscientização dos riscos associados ao sistema; (viii) as medidas de mitigação e justificação do risco residual; e (ix) as medidas de transparência pública, especialmente quanto aos possíveis usuários desse sistema (BRASIL, 2023).

A abordagem baseada em risco no cenário da proteção de dados pessoais também dispõe de instrumento de prestação de contas apto a ser aplicado no horizonte do tratamento automatizado de dados. Põe-se em foco o Relatório de Impacto à Proteção de Dados, como instrumento de governança de dados pessoais, *compliance* e *accountability* (GOMES, 2020). No cenário brasileiro, a LGPD não estabeleceu um rol taxativo, nem enumerou situações às quais será designado a elaboração de um Relatório de Impacto à Proteção de Dados, deferindo um papel central à Autoridade Nacional de Proteção de Dados - ANPD na definição de tais circunstâncias[6] (MACHADO; MENDES, 2020). Indo-se além, a definição conceitual do RIDP na LGPD defere um papel preponderante do Controlador na sua elaboração, na medida em que o restringe ao tratamento de dados pessoais que podem gerar riscos às liberdades civis e aos direitos fundamentais, mediante identificação prévia do próprio controlador (GOMES, 2020).

O RIPD é, à luz do conceito normativo da LGPD, o documento de descrição de atividades realizadas pelo controlador que possam gerar riscos às liberdades civis e aos direitos fundamentais dos titulares de dados, bem como as medidas de salvaguarda e mitigação. No que se refere à estrutura desse instrumento, a LGPD pouco disciplinou, apenas enumerando a necessidade de demonstração de: (i) a descrição dos tipos de dados tratados; (ii) a metodologia utilizada para o tratamento e para a garantia da segurança das informações; e (iii) a análise do controlador com relação a medidas, salvaguardas e mecanismos de mitigação proporcionais ao risco encontrado (BRASIL, 2018). A LGPD, entretanto, mencionou expressamente a possibilidade de a ANPD solicitar a elaboração de um Relatório de Impacto à Proteção de Dados na situação de utilização da hipótese legal do art. 7º, inciso IX, isto é, o Legítimo Interesse do controlador ou de terceiro, e determinar em outra ocasião, à sua discricionariedade (MORAIS JÚNIOR, 2023).

Nesse contexto, o RIPD, além de realizar a gestão de riscos associados à proteção de dados pessoais, também sinaliza para a novas perspectivas de gestão de dados, contribuindo de forma mais completa e eficiente na construção de processos de conformidade ética (GOMES, 2019a). A própria noção de riscos a direitos fundamentais pressupõe uma análise multiprismática, levando-se em conta além de riscos de segurança da informação, as lacunas

[6] LGPD: "Art. 55-J Compete à ANPD: (…) XIII - editar regulamentos e procedimentos sobre proteção de dados pessoais e privacidade, bem como sobre relatórios de impacto à proteção de dados pessoais para os casos em que o tratamento representar alto risco à garantia dos princípios gerais de proteção de dados pessoais previstos nesta Lei" (BRASIL, 2018).

jurídicas e diretrizes éticas do sistema em que se insere. Nesse vetor, Gomes (GOMES, 2019b) aponta:

> A ideia do relatório de impacto é refletir uma avaliação de impacto, cuja base regulatória é a identificação de riscos, que pode ser realizada para propósitos diferentes, como: avaliar o impacto de incidentes de segurança; avaliar o impacto de novas tecnologias; avaliar o impacto de novos produtos que podem gerar riscos aos direitos dos titulares de dados etc (GOMES, 2019b, p. 179).

Lacerda (2021) aponta a necessidade de analisar com atenção os impactos da inteligência artificial no âmbito da liberdade e da privacidade, evitando danos ou imputando responsabilidades pela violação a estes direitos fundamentais. A essa feita, utilização do Relatório de Impacto à Proteção de Dados como mecanismo de prestação de contas do processamento de dados por IA e o manejo de Avaliação de Impacto Algorítmico se coaduna diretamente com a necessidade de inteligibilidade do sistema, apontando para os riscos à proteção de dados, e para os riscos jurídicos e sociais imanentes à tomada de decisão sobre aspectos da vida humana, bem como para o funcionamento em abstrato do processo algorítmico de decisão, de forma prévia, descortinando ainda as medidas de mitigação a tais ameaças, e de prevenção à discriminação algorítmica.

Nessa perspectiva, a multiplicidade de camadas sobre as explicações no contexto da tomada de decisão algorítmica pode ser compatibilizada com as funções institucionais desses documentos. Kaminski e Malgieri (2020) correlacionam o processo de condução de uma avaliação de impacto no contexto da proteção de dados como sendo uma ocasião de percepção dos riscos e consequente explicação a partir de níveis de interação com os possíveis afetados. Edwards e Veale (2018) sugerem que há possibilidade de tais documentos serem utilizados para fornecimento de explicações centradas no modelo e explicações centradas no sujeito, se coadunando com as formas de abordagem de explicabilidade, dispostas em tópico específico do presente estudo.

Visualiza-se um nexo entre a governança sistêmica de inteligência artificial e os direitos dos possíveis afetados, a partir de um processo complexo de avaliação dos impactos, conduzido com suporte de diversos atores e em níveis diversos de fluxo de informação (KAMINSKI; MALGIERI, 2020). Essa noção se conjuga com a necessidade de tornar tais instrumentos públicos. Embora o PL 2.338/2023 traga uma norma específica que preveja a

divulgação da Avaliação de Impacto Algorítmico[7], a LGPD nada dispõe acerca da publicização do Relatório de Impacto à Proteção de Dados. Kaminski e Malgieri (2020) se posicionam a favor da divulgação da Avaliação de Impacto à Proteção de Dados (nome dado ao RIPD no texto normativo do Regulamento Geral de Proteção de Dados da União Europeia), tanto por orientação instrutiva do *Article 29 Data Protection Working Party* ("Grupo de Trabalho do Artigo 29" ou "A 29 WP"), quanto para cumprimento da necessária viabilização do exercício de direitos pelos titulares, afetados pelo sistema de IA:

> Moreover, there may be an argument for disclosure of group- or location-based explanations to individuals as part of the GDPR's individual transparency regime. That is, even if DPIAs are not required to be made public, and even if companies decide not to disclose to the public what they discover about the impact of algorithmic decision-making on particular groups, they may nonetheless have to do so to impacted individuals under Article 22. We understand the GDPR to suggest a connection between required DPIA analysis of systemic risks of unfairness and discrimination, and the individual rights to contestation, to express one's view, and to human intervention. That is, for a person to be able to effectively invoke her right to contest an algorithmic decision, she may need to know whether she is being treated the same or differently as other similarly situated individuals[8] (KAMINSKI; MALGIERI, 2020, p. 25).

A utilização de tais documentos no contexto de uma IA explicável, tendo em vista a formulação de uma interface de explicação calcada na gestão responsável do sistema inteligente, pode gerar benefícios para a proteção de direitos fundamentais e prevenção à discriminação algorítmica. O desenvolvimento de uma avaliação de impacto atrelada à consignação de salvaguardas por meio da explicabilidade estão, assim, intrinsecamente relacionadas com o direito de contestação de decisão automatizada (KAMINSKI; MALGIERI, 2020), e pode ser o melhor meio para superar a "falácia da transparência" através de um ciclo

[7] PL 2.338/2023: "Art. 26. Garantidos os segredos industrial e comercial, as conclusões da avaliação de impacto serão públicas, contendo ao menos as seguintes informações: I – descrição da finalidade pretendida para a qual o sistema será utilizado, assim como de seu contexto de uso e escopo territorial e temporal; II – medidas de mitigação dos riscos, bem como o seu patamar residual, uma vez implementada tais medidas; e III – descrição da participação de diferentes segmentos afetados, caso tenha ocorrido, nos termos do § 3º do art. 24 desta Lei" (BRASIL, 2023).

[8] Em tradução livre, entende-se que: "Além disso, pode haver um argumento a favor da divulgação de explicações baseadas em grupo ou localização a indivíduos como parte do regime de transparência individual do RGPD. Ou seja, mesmo que as DPIAs não sejam obrigadas a ser tornadas públicas, e mesmo que as empresas decidam não divulgar ao público o que descobrem sobre o impacto da tomada de decisões algorítmicas em grupos específicos, podem, no entanto, ter de o fazer aos indivíduos afetados sob Artigo 22. Entendemos que o RGPD sugere uma ligação entre a análise exigida pela DPIA dos riscos sistémicos de injustiça e discriminação e os direitos individuais à contestação, à expressão da própria opinião e à intervenção humana. Ou seja, para uma pessoa poder para invocar efetivamente o seu direito de contestar uma decisão algorítmica, ela pode precisar de saber se está a ser tratada da mesma forma ou de forma diferente de outros indivíduos em situação semelhante".

virtuoso de auditoria algorítmica e detecção e mitigação contínua de efeitos não desejados (EDWARDS; VEALE, 2018).

Um contexto de explicabilidade em multicamadas, considerando diversas abordagens tem o potencial de viabilizar um panorama de governança social dos algoritmos (DONEDA; ALMEIDA, 2016), promovendo uma IA mais justa e responsável. As métricas de precaução e prevenção podem dar sentido à formulação dessas avaliações de impacto (BIONI; LUCIANO, 2021), tendo em vista a explicabilidade como proteção de especiais situações, tal qual o perfilamento de consumidores e acesso à crédito (ACIOLY; 2022). A atribuição de valor às abordagens de explicação é um importante vetor de proteção de direitos fundamentais no cenário tecnológico, essencialmente quando contextualizada em um ambiente regulatório pautado em uma abordagem baseada em risco.

5 CONSIDERAÇÕES FINAIS

Ao longo desse estudo, buscou-se analisar os diversos sentido denotados à transparência substancial em matéria de inteligência artificial para proteção de direitos fundamentais. Para tanto, consignou-se que o enfrentamento da discriminação algorítmica não se restringe às medidas correlacionadas à prevenção de vieses ínsitos aos programadores, nem às ações que envolvem a restrição de acesso a dados potencialmente discriminatórios, tendo em vista que essa discriminação pode decorrer de questões relacionadas à abordagem do aprendizado de máquina e à dinâmica da tomada de decisão algorítmica.

Da conjugação de diversas formas de discriminação algorítmica e dos inúmeros meios pelos quais ela pode decorrer, nasce a necessidade de estabelecimento de mecanismos técnico-jurídicos capazes de orientar o desenvolvimento e utilização da Inteligência Artificial de forma responsável. Há, nesse sentido, um plexo de vetores éticos para o cenário da IA, ao redor do mundo, formulados por diversos atores. Desse complexo se extraem os contornos para uma discussão sobre regulação da IA, atribuindo objetivos para uma intervenção estatal com desígnio maior de proteção de direitos humanos. Esses modais deontológicos, por terem uma estrutura flexível, concretizada casuisticamente, tem em seu favor deferida uma função de suporte à formulação de estruturas regulatórias centradas na proteção de direitos fundamentais.

Dentre esses modais, a transparência se destaca por ser um *meta-princípio*, que serve à concretização dos seus pares. O presente estudo se afasta da mera compreensão de que a liberação do código-fonte da programação algorítmica seria suficiente para atender à necessidade de transparência da IA. O avanço na viabilização prática do vetor ético da transparência em sistemas de inteligência artificial tem o condão de conduzir a um ecossistema de IA responsável, viabilizando a efetividade da prevenção e precaução contra o potencial discriminatório dos algoritmos de aprendizado, ponderando a utilização de mecanismos de tratamento automatizado de dados pessoais.

Em um ambiente regulatório pautado na abordagem baseada em risco, a funcionalidade das avaliações de impacto, essencialmente algorítmica (AIA) e à proteção de dados pessoais (AIPD), demonstra compatibilidade com a necessidade de atribuição prática para a explicabilidade dos sistemas de inteligência artificial, considerando-se os riscos técnico-jurídicos que são consignados pela opacidade dos algoritmos de aprendizagem. A estratégia regulatória de estratificação de riscos conduz a uma carga positiva de obrigações, proporcional ao risco atribuído no caso concreto, impondo uma estrutura de governança capaz de garantir a proteção de direitos fundamentais dos possíveis afetados pelo sistema de IA.

A partir da discussão empreendida, observou-se que o conceito de explicabilidade, isto é, a interface entre sistema inteligente e ser humano, para ser atrelado a um instrumento de prestação de contas, precisa partir da premissa de que há uma gama de abordagens interativas, o que se convenciona denominar de multicamadas, viabilizando uma governança social dos algoritmos. Essa gama de abordagens deve considerar parâmetros específicos, que tragam individualização da interação, e consideração das nuances que envolvem o interlocutor, suas limitações de compreensão e contexto social.

A proteção de direitos fundamentais ante o potencial discriminatório da inteligência artificial deve ser pautada no vetor de prevenção, posto que a cognição de riscos enseja o empoderamento dos titulares de dados pessoais possivelmente afetados pela decisão algorítmica, viabilizando o exercício de seus direitos, especialmente o de contestação. A inserção dessa discussão no contexto social pode ensejar a mudança de prisma na relação ser humano e inteligência artificial, atribuindo uma dinâmica dialógica e o papel central ao ser humano, movimento que convencionou-se denominar *human-centered-design*. Para tanto, o presente estudo chega à conclusão da necessidade de aprimoramentos regulatórios que

disciplinem os usos e estruturação das avaliações de impacto aqui discutidas, para dar subsídios à interação dialógica para uma explicação significativa.

Deve-se ter em vista a atribuição de parâmetros normativos coerentes e práticos para atribuir sentido à explicabilidade, que considerem as premissas aqui dispostas. Ademais, no contexto de abordagem baseada em risco, é imperativa adoção de parâmetros de regulação de abordagem preventiva, de forma tal que o modelo prévio de inteligibilidade tem o condão de aprimorar a gestão dos riscos de tratamentos automatizados de dados pessoais. Por fim, é possível consignar que há compatibilidade funcional entre as avaliações de impacto e a explicabilidade da IA ainda que se utilize dados anonimizados, face ao caráter consequencialista e protetivo das normas de proteção de dados e da eminente regulação da Inteligência Artificial no Brasil.

REFERÊNCIAS

ACIOLY, Luis Henrique de Menezes. Reflexões sobre a Utilização da Inteligência Artificial e Algoritmos nas Relações Consumeristas à Luz da Lei Geral de Proteção de Dados. In: *Revista* **Brasileira de Direito Comercial**: Concorrência, Empresarial e do Consumidor. Porto Alegre: Lex Magister. v. 48, ago./set. 2022, p. 165-187.

ADADI, Amina; BERRADA, Mohammed; Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). In: **IEEE Acess**, v. 6, p. 52138–52160, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8466590>. Acesso em: 14 out. 2023.

ANGELOV, Plamen; SOARES, Eduardo; JIANG, Richard; ARNOLD, Nicholas; ATKINSON, Peter. Explainable artificial intelligence: an analytical review. In: **WIREs Data Mining and Knowledge Discovery**, v. 11, n. 5, e1424. Disponível em: <https://doi.org/10.1002/widm.1424>. Acesso em: 14 out. 2023.

BAROCAS, Solon; SELBST, Andrew. Big Data's Disparate Impact. In: **California Law Review**, v. 104, 2016, p. 671-732. Disponível em: <http://dx.doi.org/10.2139/ssrn.2477899>. Acesso em: 28 jun. 2023.

BARTNECK, Christoph; LUTGE, Christoph; WAGNER, Alan; WELSH, Sean. **An Introduction to Ethics in Robotics and AI**. Cham: Springer Switzerland, 2021.

BETTEGA, E. O que fazer sobre o viés algorítmico baseado em gênero?. In: BARBOSA, B.; TRESCA, L.; LAUSCHNER, T. (orgs.). **Governança da Internet e Gênero**: tendências e desafios. [s.l.]:CGI.Br., 2021.p. 125-133.

BIGONHA, Carolina. Inteligência Artificial em Perspectiva. In: **Panorama Setorial da Internet**, a. 10, n. 2, jul./out. 2018, p. 1-9.

BIONI, Bruno Ricardo. **Proteção de Dados Pessoais**: a função e os limites do consentimento. 3. ed. Rio de Janeiro: Forense, 2021.

BIONI, Bruno; LUCIANO, Maria. O Princípio da Precaução na Regulação de Inteligência Artificial: Seriam as Leis de Proteção de Dados o seu Portal de Entrada?. In: BIONI, Bruno. (org.). **Proteção de dados**: contexto, narrativas e elementos fundantes. São Paulo: B. R. Bioni Sociedade Individual de Advocacia, 2021 p. 281-313.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais. Diário Oficial da União: seção 1, Brasília, DF, ano 155, n. 157, p. 59-64, 15 ago. 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 05 mar. 2023.

BRASIL. **Projeto de Lei nº 2338, de 2023**. Dispõe sobre o uso da Inteligência Artificial. Senado Federal. Brasília, 2023. Disponível em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>. Acesso em: 11 jun. 2023.

BUCHER, Taina. **If… then**: algorithmic power and politics. New York: Oxford University Press, 2018.

CARVALHO, D.V.; PEREIRA, E.M.; CARDOSO, J.S. Machine Learning Interpretability:A Survey on Methods and Metrics. **Electronics,** n. 8, 832, 2019. Disponível em: https://doi.org/10.3390/electronics8080832. Acesso em: 28 out. 2022.

COMISSÃO EUROPEIA. **A definition of AI***:* Main capabilities and scientific disciplines. 2019. Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 29 out. 2022.

CONFALONIERI, Roberto; COBA, Ludovik; WAGNER, Benedikt; BESOLD, Tarek. A historical perspective of explainable Artificial Intelligence. In: **WIREs Data Mining and Knowledge Discovery**, v. 11, n. 1, e1391, 2021. Disponível em: <https://doi.org/10.1002/widm.1391>. Acesso em: 14 out. 2023.

DONEDA, Danilo; ALMEIDA, Virgílio. What is Alghorithm Governance?. In: **IEEE Internet Computing**, v. 20, n. 4. jul./ago., 2016, p. 60-63.

DUARTE, Alan. **A Antidiscriminação no contexto da inteligência artificial**: possibilidades de governança mediante a normatização de algoritmos**.** Fortaleza: Mucuripe, 2021.

EDWARDS, Lilian; VEALE, Michael. Enslaving the Algorithm: From a ‘Right to an Explanation’ to a ‘Right to Better Decisions’?. In: **IEEE Security & Privacy**, v. 16, n. 3, p. 46-54, 2018. Disponível em: <http://dx.doi.org/10.2139/ssrn.3052831>. Acesso em: 14 out. 2023.

FLORIDI, Luciano; HOLWEG, Matthias; TADDEO, Mariarosa; SILVA, Javier Amaya; MOKANDER, Jakob; WEN, Yuni. **capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act**. 23, mar. 2022. Disponível em: <https://dx.doi.org/10.2139/ssrn.4064091>. Acesso em: 15 jul. 2023.

FRAZÃO, Ana; GOETTENAUER, Carlos. Black box e o direito face à opacidade algorítmica. In: BARBOSA, M. M.; BRAGA NETTO, F.; SILVA, M.C.; FALEIROS JÚNIOR, J. L. M. (coords.). **Direito Digital e Inteligência Artificial**: Diálogos entre Brasil e Europa. Indaiatuba: Foco, 2021 p. 27-42.

GABRIEL, Martha. **Inteligência Artificial**: Do Zero ao Metaverso. São Paulo: Atlas, 2022.

GOMES, Maria Cecília O. Entre o método e a complexidade: compreendendo a noção de risco na LGPD. In: PALHARES, Felipe (Coord.). **Temas atuais de proteção de dados**. São Paulo: Thomson Reuters Brasil, 2020, p. 245-271.

GOMES, Maria Cecília O. Para além de uma "obrigação legal": o que a metodologia de benefícios e riscos nos ensina sobre o relatório de impacto à proteção de dados. In: LIMA, A. P.; HISSA, C.; SALDANHA, P. M. (orgs.). **Direito Digital**: Debates Contemporâneos**.** São Paulo: Revista dos Tribunais, 2019a, p. 141-153.

GOMES, Maria Cecília O. Relatório de impacto à proteção de dados: Uma breve análise da sua definição e papel na LGPD. In: **Revista do Advogado**. v. 39, n. 144, p. 174–183, nov., 2019b.

HOFFMANN-RIEM, Wolfgang. **Teoria Geral do Direito Digital**. Trad. Italo Fuhrmann. Rio de Janeiro: Forense, 2020.

KAMINSKI, Margot; MALGIERI, Gianclaudio. Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. In: **U of Colorado Law Legal Studies Research Paper**. Nº. 19-28, 2020. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3456224>. Acesso em 24 jun. 2023.

KROLL, Joshua; HUEY, Joanna; BAROCAS, Solon; FELTEN, Edward; REIDENBERG, Joel; ROBINSON, David; YU, Harlan. Accountable Algorithms. In: **University of Pennsylvania Law Review**, v. 165, 2017, p. 633-706. Disponível em: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/. Acesso em: 15 jul. 2023.

LACERDA, Bruno Torquato Zampier. A função do direito frente à inteligência artificial. In: BARBOSA, M. M.; BRAGA NETTO, F.; SILVA, M.C.; FALEIROS JÚNIOR, J. L. M. (coords.). **Direito Digital e Inteligência Artificial**: Diálogos entre Brasil e Europa. Indaiatuba: Foco, 2021 p. 81-93.

LEE, Kai-Fu. **Inteligência artificial**: como os robôs estão mudando o mundo, a forma como amamos, nos relacionamos, trabalhamos e vivemos. Trad. Marcelo Barbão. Rio de Janeiro: Globo Livros, 2019.

LEMOS, Alessandra; BUARQUE, Gabriela; SOARES, Ingrid; MULIN, Victor; CHIAVONE, Tayrone. **Avaliação de Impacto Algorítmico para a proteção dos direitos fundamentais**: Relatório. Brasília: Laboratório de Políticas Públicas e Internet, 2023.

MACHADO, Diego Carvalho; MENDES, Laura Schertel. Tecnologias de Perfilamento e dados agregados de geolocalização no combate à Covid-19 no Brasil: Uma análise dos riscos individuais e coletivos à luz da LGPD. In: **Revista Brasileira de Direitos Fundamentais & Justiça**, v. 14, n. 1, p. 105–148, 2020. Disponível em: <https://doi.org/10.30899/dfj.v0i0.1020>. Acesso em: 24 set. 2023.

MAGRANI, Eduardo. New perspectives on ethics and the laws of artificial intelligence. In: **Internet Policy Review**, v. 8, n. 3, 2019. Disponível em: <https://doi.org/10.14763/2019.3.1420>. Acesso em 08 jul. 2023.

MENDES, Laura Schertel; MATTIUZZO, Marcela. Discriminação algorítmica: conceito, fundamento legal e tipologia. In: **Revista de Direito Público**, v. 16, n. 90. Porto Alegre, nov./dez. 2019, p. 39-64.

MEDON, Filipe; FALEIROS JÚNIOR, José Luiz de Moura. Discriminação algorítmica de preços, Perfilização e Responsabilidade civil nas relações de consumo. In: **Revista Direito e Responsabilidade**. Coimbra, a. 3. 2021, p. 94-969. Disponível em: https://revistadireitoresponsabilidade.pt/2021/discriminacao-algoritmica-de-precos-perfilizacao-e-responsabilidade-civil-nas-relacoes-de-consumo-jose-luiz-de-moura-faleiros-junior-filipe-medon/. Acesso em: 28 abr. 2022.

MENDONÇA JÚNIOR, Claudio do Nascimento; NUNES, Dierle José Coelho. Desafios e Oportunidades para a Regulação da Inteligência Artificial: A necessidade de compreensão e mitigação dos riscos da IA. In: **Revista Contemporânea**, v. 3, n. 7, 2023, p. 7753-7785. Disponível em: <https://doi.org/10.56083/RCV3N7-024>. Acesso em: 09 set. 2023.

METCALF, Jacob; MOSS, Emanuel; WATKINS, Elizabeth Anne; SINGH, Ranjit; ELISH, Madeleine Clare. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In: **FAccT '21:** Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, mar. 2021, p. 735-746. Disponível em: <https://doi.org/10.1145/3442188.3445935>. Acesso em: 17 set. 2023.

MITTELSTADT, Brent Daniel; ALLO, Patrick; TADDEO, Mariarosaria; WACHTER, Sandra; FLORIDI, Luciano. The ethics of algorithms: Mapping the debate. In: **Big Data & Society**. v. 3, n. 2, jul./dez., 2016, p. 1-21. Disponível em: <https://doi.org/10.1177/2053951716679679>. Acesso em: 23 jul. 2023.

MORAIS JÚNIOR, Ricardo Antonio Maia de. **Accountability e direito fundamental à proteção de dados pessoais enquanto limites ao uso da inteligência artificial na relação de emprego**. 2023, 161 f. Dissertação (Mestrado em Direito) – Faculdade de Direito, Universidade Federal do Ceará, Programa de Pós-Graduação em Direito, Fortaleza, 2023.

MOURÃO, Licurgo. Regulação da Inteligência Artificial no Brasil. In: CAMINO, Geraldo Costa da (coord*.).* **Intellegentiae Artificialis, Imperium et citivatem**. Madrid: Alma Mater, 2023, p. 73-90.

MULHOLLAND, Caitlin; FRAJHOF, Isabella. Entre as leis da robótica e a ética: regulação para o adequado desenvolvimento da inteligência artificial. In: BARBOSA, M. M.; BRAGA NETTO, F.; SILVA, M.C.; FALEIROS JÚNIOR, J. L. M. (coords.). **Direito Digital e Inteligência Artificial**: Diálogos entre Brasil e Europa. Indaiatuba: Foco, 2021, p. 65-80.

MULHOLLAND, Caitlin; GOMES, Rodrigo Dias de Pinho. Inteligência Artificial e Seus Principais Desafios para os Programas de Compliance e as Políticas de Proteção de Dados. In: CUEVA, Ricardo Villas Bôas; FRAZÃO, Ana (coords). **Compliance e Política de Proteção de Dados**. São Paulo: Thomson Reuters Brasil, 2021. p. 161-180.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449)**. Paris: OECD, 2019.

PASQUALE, Frank. **The Black Box Society.** Cambridge: Havard Unity Press, 2015.

PEREIRA, J. R. L. de. Transparência pela cooperação: como a regulação responsiva pode auxiliar na promoção de sistemas de machine-learning inteligíveis. **Revista de Direito Setorial e Regulatório**, v. 7, nº 1, p. 194-223, mai.-jun. 2021.

PHILLIPS, P. Jonathon; HAHN, Carina A.; FONTANA, Peter C.; YATES, Amy N.; GREENE, Kristen; BRONIATOWSKI, David A.; PRZYBOCKI, Mark**. Four Principle of Explainable Artificial Intelligence**. Gaithersburg: National Institute of Standards and Technology (NIST), 2021.

RAAB, Charles. Information privacy, impact assessment, and the place of ethics. In: **Computer Law & Security Review**, v. 37, jul. 2020, p. 1-16. Disponível em: <https://doi.org/10.1016/j.clsr.2020.105404>. Acesso em: 17 set. 2023.

RUDIN, Cynthia. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. In: **Nature machine intelligence**, v. 1, n. 5, 2019, p. 206-215. Disponível em: <https://doi.org/10.1038%2Fs42256-019-0048-x>. Acesso em: 23 jul. 2023.

RUIZ, Cinthia; QUARESMA, Manuela. Inovação com Dados: a experiência do usuário com Sistemas baseados em Inteligência Artificial. In: **Human Factors in Design**. v.10, n 20, p. 23-44, dezembro 2021. Disponível em: <https://doi.org/10.5965/2316796310202021023>. Acesso em: 28 out. 2022.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**. 3. ed. Trad. Regina Célia Simille de Macedo. Rio de Janeiro: Elsevier, 2013.

SAMUEL, Arthur. Some Studies in Machine Learning Using the Game of Checkers. In: **IBM Journal of Research and Development**. v3. N. 3, jul. 1959, p. 206-226. Disponível em: <https://citeseerx.ist.psu.edu/doc/10.1.1.368.2254>. Acesso em: 09 jul. 2023.

SELBST, Andrew. An institutional view of algorithmic impact assessment. In: **Havard Jornal of Law & Technology**, Massachusetts, v. 35, n. 1, p. 117-173, 2021. Disponível em: <https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf>. Acesso em: 05 mar. 2023.

SCHIPPERS, Laurianne-Marie. **Algoritmos que discriminam:** uma análise jurídica da discriminação no âmbito das decisões automatizadas e seus mitigadores. 2018. 57 f. Trabalho de Conclusão de Curso (Bacharelado em Direito) – Escola de Direito de São Paulo, Fundação Getúlio Vargas, São Paulo, 2018.