SISTEMAS DE INTELIGÊNCIA ARTIFICIAL E AVALIAÇÕES DE IMPACTO PARA DIREITOS HUMANOS*

*ARTIFICIAL INTELLIGENCE SYSTEMS AND HUMAN RIGHTS IMPACT ASSESSMENTS***

Sergio Marcos Carvalho de Ávila Negri[1]

Joana de Souza Machado[2]

Carolina Fiorini Ramos Giovanini[3]

Nathan Pascoalini Ribeiro Batista[4]

Resumo: A partir de uma abordagem exploratória, este trabalho analisa estratégias de regulação de sistemas de inteligência artificial, com foco em modelos mais recentes baseados na classificação de riscos. A pesquisa investiga a hipótese de que o modelo regulatório de sistemas de inteligência artificial centrado na classificação de riscos pode considerar de modo insuficiente o impacto diferenciado das tecnologias para direitos humanos de grupos vulneráveis, aprofundando desigualdades e violações já existentes, especialmente no contexto do Sul Global, marcado pela colonialidade do poder. Em conclusão, apontamos que, embora as avaliações de impacto sejam importantes instrumentos em termos de *accountability* e de estrutura de governança de sistemas de inteligência artificial, devem necessariamente considerar a existência de riscos inaceitáveis. Caso contrário, os processos regulatórios que nelas se apoiem poderão legitimar práticas violadoras de direitos humanos.

* Artigo submetido em 19/12/2022 e aprovado para publicação em 15/12/2023. Artigo publicado em formato antecipado ("Ahead of Print") em 13/07/2023.
** O trabalho apresenta resultados parciais da pesquisa realizada no âmbito do Projeto "Inovação e Direito na Inteligência Artificial: mapeamento normativo e análise de impacto para o exercício de direitos fundamentais" (CNPq Universal).
1Professor do Departamento de Direito Privado e do corpo permanente do PPGD em Direito e Inovação da Universidade Federal de Juiz de Fora (UFJF). Doutor e Mestre em Direito Civil pela UERJ, com especialização junto à Università di Camerino (Itália). Coordenador do NEAPID e do Projeto acima nomeado. E-mail: smcnegri@yahoo.com ORCID: https://orcid.org/0000-0003-2156-3518.
2 Professora do Departamento de Direito Público Material e do corpo permanente do PPGD em Direito e Inovação da Universidade Federal de Juiz de Fora (UFJF). Doutora e Mestre em Teoria do Estado e Direito Constitucional pela PUC-Rio, com estágio doutoral junto à Harvard Law School (EUA). Coordenadora do LAVID e integrante do projeto acima nomeado. E-mail: joana.machado@ufjf.br. ORCID: https://orcid.org/0000-0001-6467-2357.
3 Mestranda em Direito e Inovação no PPGD da Universidade Federal de Juiz de Fora (UFJF). Pesquisadora do NEAPID e integrante do projeto acima nomeado. E-mail: carolina.giovanini@direito.ufjf.br. ORCID: https://orcid.org/0000-0003-1961-7417.
4 Mestrando em Direito e Inovação no PPGD da Universidade Federal de Juiz de Fora (UFJF). Membro do Núcleo de Coordenação da Rede de Pesquisa em Governança da Internet. Pesquisador do NEAPID e integrante do projeto acima nomeado. E-mail: nathanprb18@gmail.com. ORCID: https://orcid.org/0000-0003-1798-9356.

Palavras-chave: Inteligência artificial; Direitos humanos e Inovação; Regulação; Discriminação; Avaliações de impacto; Classificação de riscos.

Abstract: From an exploratory approach, this work analyzes regulation strategies of artificial intelligence systems, focusing on more recent models based on risk classification. The research investigates the hypothesis that the regulatory model of artificial intelligence systems centered on risk classification may insufficiently consider the differentiated impact of technologies on vulnerable groups human rights, deepening existing inequalities and violations, especially in the global south context, marked by the coloniality of power. In conclusion, we point out that although impact assessments are important instruments in terms of accountability and a governance structure for artificial intelligence systems, they must necessarily consider the existence of unacceptable risks. Otherwise, the regulatory processes that rely on them may legitimize practices that violate human rights.

Keywords: Artificial Intelligence; Human Rights and Innovation; Regulation; Discrimination; Impact Assessments; Risk Assessment.

Introdução

Sistemas de inteligência artificial (IA)[5] estão cada vez mais presentes em nosso dia a dia, sob a promessa de mudanças positivas, contínuas e de amplo alcance social: processos seletivos, avaliações de crédito, *chatbots* de atendimento, carros autônomos e assistentes pessoais são alguns exemplos de aplicações da tecnologia. Conforme aponta Bigonha (2018), a crescente popularização de tecnologias de inteligência artificial está relacionada (i) ao barateamento da infraestrutura para processamento; (ii) aos avanços no desenvolvimento de algoritmos; (iii) à maior disponibilidade de dados; (iv) à disponibilidade de tecnologias em código aberto; e (v) à maior conectividade.

Simultaneamente, a disseminação do uso de tais sistemas em atividades cotidianas levanta uma série de preocupações relacionadas aos eventuais impactos para direitos humanos, uma vez que a lógica algorítmica passa a mediar cada vez mais as interações humanas, incidindo, inclusive, em processos decisórios que potencialmente retroalimentam opressões estruturais da sociedade (Silva, 2022).

Em resposta à popularização de sistemas de IA, surgem movimentos regulatórios,

[5] Desde a introdução do trabalho, é importante situar a diferença entre os termos "inteligência artificial" e "sistemas de inteligência artificial", a ser mais explorada no item 1. Enquanto o primeiro refere-se a uma área de estudo, composta por subcampos que incluem linguagem natural, aprendizagem de máquinas, redes neurais e robótica (Cerka, Grigiene, Sirbikyte, 2015), o segundo termo faz referência aos sistemas que utilizam as abordagens técnicas de inteligência artificial.

inicialmente centrados na regulação pela ética, direcionados à construção de um quadro normativo sobre o desenvolvimento, comercialização e uso da inteligência artificial, como é o caso de diversas Estratégias Nacionais voltadas para a inteligência artificial. É possível, no entanto, identificar uma mudança no eixo regulatório que passa a apontar para uma abordagem baseada em risco. O gerenciamento de riscos demanda que sejam tomadas medidas relacionadas à *accountability* desses sistemas de IA, com a implementação de rotinas e procedimentos voltados a uma maior transparência.

Face a esse contexto, o presente trabalho busca responder ao seguinte problema: considerando a abordagem regulatória baseada em risco e a possibilidade de um sistema de IA produzir resultados discriminatórios, quais são os desafios na implementação de ferramentas de accountability, como avaliações de impacto, para a garantia de direitos das pessoas afetadas por tais sistemas?

O trabalho investiga a hipótese de que o modelo regulatório centrado na classificação dos riscos por vezes negligencia o impacto diferenciado das tecnologias, como a do reconhecimento facial, para grupos de vulnerabilidade politicamente induzida. Analisa-se de modo mais específico a armadilha de se transplantar esse modelo para o contexto do sul global, marcado pela colonialidade do poder.

Em termos metodológicos, este trabalho se apoia em uma abordagem exploratória, considerando o incipiente contexto regulatório ao qual os sistemas de inteligência artificial estão submetidos. Desse modo, busca-se uma familiaridade com o problema de pesquisa desenvolvido, com intuito de formulação de hipóteses mais robustas posteriormente (Gil, 2018). O presente estudo adota a revisão sistemática de bibliografia como a principal técnica de pesquisa, com intuito de compreender o estado da arte acerca do contexto regulatório de sistemas de IA e de como ferramentas de *accountability*, a exemplo de avaliações de impacto para direitos humanos, podem ser implementadas durante o processo de desenvolvimento de sistemas de inteligência artificial de modo a mitigar os riscos às pessoas-alvo de tais sistemas. Por meio, ainda, de análise documental, realiza-se coleta de dados sobre disposições relacionadas a discriminações e a vieses algorítmicos nas propostas regulatórias em tramitação no Brasil.

Para tanto, o trabalho explora, em primeiro lugar, o plano tecnológico de funcionamento de sistemas de inteligência artificial e as possíveis definições a serem adotadas para identificação de tais tecnologias. Posteriormente, no âmbito ético e social, são detalhados os impactos

discriminatórios para grupos historicamente marcados por vulnerabilidades, buscando-se demonstrar que o uso de sistemas de inteligência artificial pode aprofundar desigualdades e violações já existentes, impactando tais grupos de modo diferenciado. Por fim, a partir de uma perspectiva normativa, as estratégias de regulação da inteligência artificial são apresentadas, com enfoque para a elaboração de avaliações de impacto a partir de uma abordagem centrada em direitos humanos, compreendidos não como categoria universal, mas como ferramenta política de luta por reconhecimento de sujeitos concretos (Pires, 2017; Herrera Flores, 2009).

1. Definições em disputa

Apesar de o termo "inteligência artificial" ter sido cunhado por John McCarthy (1965) na conferência Dartmouth Summer Research Project on Artificial Intelligence, atualmente, não há um consenso sobre a sua definição. É essencial explorar as definições em disputa, uma vez que, a partir do conceito adotado, os limites de aplicação de normas previstas em legislações nacionais e regras de governança criadas por instrumentos privados ou guias orientativos em âmbito global serão definidos, criando-se uma verdadeira moldura normativa.

Em primeiro lugar, é importante diferenciar os termos "inteligência artificial" e "sistemas de inteligência artificial". Enquanto o primeiro refere-se a uma área de estudo, composta por subcampos que incluem linguagem natural, aprendizagem de máquinas, redes neurais e robótica (Cerka, Grigiene, Sirbikyte, 2015), o segundo termo faz referência aos sistemas que utilizam as abordagens técnicas de inteligência artificial.

Russel e Norvig (2013) visualizam oito tipos diferentes de definições para o termo "Inteligência Artificial". Os autores dividem as abordagens em quatro categorias: (i) pensando como um humano; (ii) pensando racionalmente; (ii) agindo como seres humanos; e (iv) agindo racionalmente. Nesse sentido, observa-se que as definições se dividem em perspectivas sobre processos de pensamento, raciocínio e comportamento, além de utilizarem como referencial o desempenho humano ou a racionalidade, respectivamente.

De acordo com Steibel, Vicente e Jesus (2019), a expressão "inteligência artificial" tem sido definida como a habilidade de um sistema de interpretar corretamente dados externos, aprender a partir desses dados e usar o aprendizado para alcançar objetivos e tarefas específicas por meio da adaptação flexível. Nesse sentido, os autores pontuam que a IA utiliza informações externas como inputs para identificar regras e modelos.

Silva (2022) estabelece uma diferenciação entre a abordagem simbólico-dedutiva e a abordagem conexionista-indutiva. Na primeira, o ambiente calculatório recebe dados para análise e uma série de instruções para obter saídas (*outputs*) classificados ou operacionalizados de acordo com os objetivos declarados; por outro lado, na abordagem conexionista-indutiva, os algoritmos recebem dados de treinamento (*inputs*) e exemplos de resultados (*outputs*) para realizar correlações.

Ainda, destaca-se que Alexandre Quaresma (2020) entende que são inteligências artificiais todos os sistemas cibernético-informacionais e de computação que simulam ou tentam reconstituir artificialmente os diversos tipos de comportamentos das entidades biológicas humanas do mundo natural, no que se refere às ações e condutas consideradas por elas – entidades – como inteligentes. Contudo, o autor ressalta que a própria expressão "inteligência artificial" contém um equívoco, qual seja, uma ambiguidade conceitual que decorre do fato de que esses sistemas não surgem artificialmente por si, mas sim da própria inteligência humana.

Nesse sentido, Quaresma (2020) esclarece que os sistemas de inteligência artificial seguem comandos pré-estabelecidos, ou seja, regras fixas de conduta e de comportamento formuladas por seres humanos. Sendo assim, a inteligência ali presente é, na verdade, uma inteligência humana, uma vez que segue parâmetros pré-estabelecidos por humanos. Conforme aponta Negri (2020), a indefinição ontológica e jurídica que marca o desenvolvimento do campo da inteligência artificial faz com que construções baseadas em metáforas sejam utilizadas, por exemplo, para se pensar em sistemas de inteligência artificial a partir de uma retórica antropomórfica, como se fossem pessoas. No atual contexto de criação de definições em legislações nacionais[6], é essencial ter cautela quando conceitos criados, inicialmente, a partir de metáforas e de analogias são utilizados em seu sentido literal. Além disso, é necessário tomar cuidado com definições pautadas em generalizações abstratas e reduções unitárias, indiferentes às distintas abordagens técnicas presentes no campo da inteligência artificial, caso contrário, os

[6] No âmbito da União Europeia, a Comissão Europeia apresentou a Proposta de Regulamento sobre a Inteligência Artificial ("Artificial Intelligence Act"). Atualmente, a proposta encontra-se na fase de "trílogo", na qual são realizadas discussões entre a Comissão Europeia, o Conselho e o Parlamento para negociação de uma versão final do texto antes de sua aprovação. Faz-se necessário destacar que a definição do termo "sistemas de inteligência artificial" é um dos debates em questão (UNIÃO EUROPEIA. Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206). Bruxelas: Comissão Europeia, 2021). No contexto nacional, Projeto de Lei 21/20- aprovado na Câmara dos Deputados e atualmente em tramitação no senado Federal - apresenta definição e abordagens técnicas que caracterizam sistemas de inteligência artificial, assim como o Projeto de Lei 2.338/23, apresentado no Senado Federal.

limites de aplicação das normas acerca do tema poderão ser inadequadamente esvaziados.

Assim, os sistemas de inteligência artificial já fazem parte de uma série de atividades cotidianas e o uso de tais sistemas pode levantar riscos e impactos para direitos humanos. É necessário, portanto, que as definições a serem adotadas sejam suficientemente abrangentes para alcançar diferentes usos e aplicações de inteligência artificial, fazendo com que a moldura normativa de proteção promova uma tutela efetiva de direitos humanos, cujo sentido, pensado a partir das dimensões intercultural e política, acesse os diversos corpos, (r)existências humanas, nos distintos territórios (Pires, 2017; Herrera Flores, 2009).

2. Impactos do uso de sistemas de inteligência artificial para grupos politicamente vulnerabilizados

O desenvolvimento e o uso de sistemas de inteligência artificial também levantam preocupações acerca da exacerbação de desigualdade e do fortalecimento de estigmas e discursos discriminatórios já vivenciados - historicamente - por grupos marginalizados. Nesse sentido, busca- se demonstrar que tais grupos historicamente submetidos a opressões e violências, por vezes mascaradas pelo discurso hegemônico abstrato de direitos humanos (Douzinas 2009), estão expostos a um risco maior de violações concretas de direitos humanos na medida em que o desenvolvimento de sistemas de inteligência artificial avança.

Silva (2022) pontua a existência de uma narrativa de que no "ciberespaço" ou em ambientes "virtuais" os marcadores sociais, por vezes reduzidos a pautas identitárias, como raça, gênero, classe ou nacionalidade, perderiam a sua relevância/força epistêmica. De acordo com o autor, essa narrativa teria se amparado, sobretudo, em três contextos: (i) os ambientes digitais ainda eram informacionalmente escassos, com poucas possibilidades de comunicação além da textualidade; (ii) os pesquisadores advindos de populações minorizadas nos países de diáspora africana ainda eram poucos e ignorados; (iii) o tecnoliberalismo em consolidação gerava a pretensão de neutralidade das plataformas e mídias.

Nesse contexto, Silva (2022) aponta que, no debate sobre tecnologias digitais, é necessário ir além da linguagem textual - que produz o discurso discriminatório explícito em textos e imagens

- e discutir também as manifestações construídas e expressas na infraestrutura ou *back*

end (por exemplo, nos algoritmos) ou através da interface (como símbolos, imagens, voz, textos e representações gráficas).

Para Silva (2022), essa não é uma questão unidirecional, pois a estrutura técnico-algorítmica pode facilitar manifestações de racismo e, simultaneamente, tais manifestações podem ser fonte e conteúdo para aspectos da estrutura técnica, uma vez que a circulação e engajamento de conteúdos discriminatórios são convertidos em métricas e faturamento para plataformas.

Em outras palavras, a estrutura técnico-algorítmica pode atuar como mais um elemento constitutivo do dispositivo de racialidade/biopoder (Carneiro, 2005), e os diversos elementos (práticas heterogêneas, discursos, o dito e não dito) se realinham e retroalimentam em função de um objetivo estratégico, como o da colonialidade.

A colonialidade do ser, na contribuição de Maldonado-Torres (2007), diz respeito à reprodução de hierarquias de raça, gênero e geopolítica, que foram inventadas ou instrumentalizadas como ferramentas de controle colonial. Em linha semelhante, para Quijano (2005), a colonialidade do poder nomeia a continuidade de padrões estabelecidos de poder entre colonizadores e colonizados; os resquícios contemporâneos dessas relações; e como esse poder molda a nossa compreensão da cultura, trabalho, intersubjetividade e produção de conhecimento.

A colonialidade pode ser também verificada em estruturas digitais, na forma de imaginações socioculturais, sistemas de conhecimento e nas escolhas sobre o desenho e o uso das tecnologias, situadas no tempo e espaço, influenciadas por instituições, sistemas, valores que dizem muito sobre a permanência do passado. Tecnologias emergentes, como os sistemas algorítmicos, estão diretamente sujeitas à continuidade de padrões estabelecidos de poder e é fundamental que a consideremos no discurso atual sobre justiça, responsabilidade e transparência, mobilizado na agenda das inovações tecnológicas (Mohamed, Png, Isaac, 2020).

É essencial reconhecer que pessoas com identidades diferentes podem passar por experiências distintas de marginalização, estigmatização e discriminação, atravessadas e por vezes interseccionadas por marcadores sociais como raça, gênero, orientação sexual e identidade regional. Conforme aponta Cortiz (2020), os sistemas de inteligência artificial necessitam de um algoritmo de treinamento, que pode ser compreendido como uma "receita de bolo" que utiliza dados de treinamento como "ingredientes" para produção de um modelo. Assim, um projeto de inteligência artificial envolve dois algoritmos: (i) o algoritmo de treinamento, que não faz juízo de valor ou apresenta vieses; e (ii) o modelo treinado, que é a saída do algoritmo de treinamento e

que será - de fato - utilizado em produção, mas que, a depender dos dados utilizados no treinamento, poderá demonstrar comportamentos enviesados.

É possível notar que o comportamento de um sistema de inteligência artificial reflete os padrões dos dados de treinamento. Por tal razão, é essencial avaliar a base de dados utilizada no momento de aprendizagem, verificando, por exemplo, se tal base é representativa e conta com dados atualizados e verídicos. No entanto, também é importante notar que, para além de questões técnicas, aspectos sociais, históricos, culturais e geográficos também influenciam o desenvolvimento de sistemas de inteligência artificial.

Por exemplo, Cortiz (2022) destaca que a maioria das ferramentas utilizadas no Sul Global foi desenvolvida por empresas do Norte, além disso, os conjuntos de dados mais populares são centrados nos Estados Unidos e na Europa ocidental, fazendo com que aspectos culturais específicos de outras localidades sejam desconsiderados. Na mesma direção, Tomasev et al (2021) apontam que é importante ir além das abordagens puramente computacionais para avaliar plenamente os aspectos sócio-técnicos do desenvolvimento e implementação de tecnologias digitais.

No que se refere aos impactos para a população LGBTQIAPN+[7], Tomasev et al (2021) apontam que, em razão da opressão histórica e os desafios contemporâneos enfrentados, há um risco substancial de que sistemas de inteligência artificial sejam desenvolvidos e utilizados de forma injusta. Os autores relatam que, em 2017, pesquisadores de Stanford buscaram desenvolver um sistema de "gaydar" com base em inteligência artificial, objetivando que o sistema estabelecesse a orientação sexual de uma pessoa a partir de um banco de 35.326 imagens faciais.

Nesse sentido, verifica-se que, para além das implicações éticas, o desenvolvimento de tais sistemas pode resultar em violações de direitos humanos para a comunidade LGBTQIAPN+. Tomasev *et al* (2021) ressaltam que tais sistemas poderiam ser utilizados em escala e de forma persecutória[8], especialmente em Estados onde a dissidência da cisheteronormatividade é criminalizada não apenas no sentido sociológico, mas também jurídico, gerando tipos penais com previsão de pena de privação de liberdade e em alguns casos até mesmo com pena de morte, além

[7] O trabalho opta pelo termo LGBTQIAPN+ por permitir reconhecimento das identidades dissidentes de sexualidade e de gênero (lésbicas, gays, bissexuais, transvestigêeras, queer, interesexo, assexuais, pansexuais e não-binárias) e, com o sufixo "+", indicar uma abertura epistêmica necessária, para que não seja sugerido caráter fixo ou estável dessa identidade coletiva, e para que se reduza a objetificação do grupo.

[8] Conferir Machado; Negri; Giovanini (2020) sobre impactos desiguais de uso de coletas de dados no contexto da pandemia para grupos politicamente vulnerabilizados (nem invisíveis, nem visados).

de exacerbar riscos relacionados à privacidade.

Mesmo no contexto de países que oferecem algum nível de proteção jurídica à população LGBTQIAPN+, é preciso acompanhar de forma atenta os impactos dos desenhos e usos concretos de sistemas de inteligência artificial, uma vez que recaem de forma mais negativa sobre grupos com maior vulnerabilidade e essa proteção jurídica está em constante disputa, por desafiar, ainda que de modo tímido, a cisheteronormatividade.

Tomasev *et al* (2021) demonstram, ainda, que estes sistemas podem perpetuar ideias e crenças sobre a população LGBTQIAPN+, por exemplo, no caso de sistemas que utilizam a informação genética como input primário e reforçam visões biológicas, eugenistas e concepções errôneas que correlacionam biologia e aparência à orientação sexual e identidade de gênero.

Evidentemente, o exercício de outros direitos humanos também pode ser impactado por sistemas de inteligência artificial, como, por exemplo, a liberdade de expressão. Tomasev et al (2021) apontam que a moderação de conteúdo automatizada - baseada em ferramentas de detecção de conteúdos digitais - pode ser abusivamente utilizada para censura de conteúdo LGBTQIAPN+, fazendo com que identidades e produções não conformes às atitudes culturais cisheteronormativas sejam apagadas da esfera digital.

No que diz respeito à discriminação em razão de gênero, o estudo de Buolamwini e Gebru (2018) demonstra que - no conjunto de tecnologia avaliadas pelas pesquisadoras - as precisões de classificação por gênero variam entre 87,9% e 93,7%. No entanto, a classificação é 8,1% - 20,6% pior em indivíduos designados como do sexo feminino do que em indivíduos designados como do sexo masculino e, quando o resultado é avaliado a partir de subgrupos de mulheres de pele escura, mulheres de pele clara, homens de pele escura e homens de pele clara, observa-se que o subgrupo de mulheres de pele escura possui as taxas de erro mais elevadas para todos os classificadores de gênero, variando entre 20,8% - 34,7%.

Nesse sentido, Buolamwini e Gebru (2018) entendem que é necessário contar com conjuntos de dados de referência inclusivos e relatórios de desempenho e precisão de subgrupos para aumentar a transparência - entendida como fornecimento de informação sobre a composição demográfica e fenotípica dos conjuntos de dados de formação e de referência - e a prestação de contas no desenvolvimento de sistemas de inteligência artificial.

Em relação às questões envolvendo raça, um dos primeiros debates com grande repercussão ocorreu em 2016, quando a rede de jornalismo investigativo sem fins lucrativos

ProPublica divulgou uma matéria denunciando o COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) e demonstrando que o sistema que calculava o nível de periculosidade de criminosos nos Estados Unidos para estabelecer a pena funcionava a partir de um sistema de pontuação que atribuía mais pontos para pessoas de minorias étnicas.

Nesse sentido, Silva (2022) define o racismo algorítmico como o modo pelo qual a disposição de tecnologias e imaginários sociotécnicos em um mundo moldado pela supremacia branca realiza a ordenação algorítmica racializada de classificação social, recursos e violência em detrimento de grupos minorizados.

Destaca-se que, no âmbito legislativo brasileiro, o Projeto de Lei nº 2.338/2023, apresentado no âmbito do Senado Federal, define a discriminação como qualquer distinção, exclusão, restrição ou preferência, em qualquer área da vida pública ou privada, cujo propósito ou efeito seja anular ou restringir o reconhecimento, gozo ou exercício, em condições de igualdade, de um ou mais direitos ou liberdades previstos no ordenamento jurídico, em razão de características pessoais como origem geográfica, raça, cor ou etnia, gênero, orientação sexual, classe socioeconômica, idade, deficiência, religião ou opiniões políticas.

Além disso, o texto apresenta uma conceituação para a discriminação indireta, que ocorre quando normativa, prática ou critério aparentemente neutro tem a capacidade de acarretar desvantagem para pessoas pertencentes a grupo específico, ou as coloque em desvantagem, a menos que essa normativa, prática ou critério tenha algum objetivo ou justificativa razoável e legítima à luz do direito à igualdade e dos demais direitos fundamentais (art. 4º, VI e VII do PL 2338/2023).

Nessa direção, o texto também prevê que as pessoas afetadas por decisões, previsões ou recomendações de sistemas de IA têm direito a tratamento justo e isonômico, sendo vedadas a implementação e o uso de sistemas de inteligência artificial que possam acarretar discriminação direta, indireta, ilegal ou abusiva, como (i) impactos decorrentes de uso de dados pessoais sensíveis ou em razão de características pessoais como origem geográfica, raça, cor ou etnia, gênero, orientação sexual, classe socioeconômica, idade, deficiência, religião ou opiniões políticas; e (ii) estabelecimento de desvantagens ou agravamento da situação de vulnerabilidade de pessoas pertencentes a um grupo específico, ainda que se utilizem critérios aparentemente neutros.

Existem diversos exemplos de aplicações de inteligência artificial que apresentam vieses algorítmicos e geram discriminações. Por exemplo, Corrêa (2021), ao analisar os sistemas de

pontuação de crédito, aponta que a população negra no Brasil sofre um contexto de exclusão em relação a rendimentos, condições de moradia, pobreza, falta/déficit de inserção no mercado de trabalho e de acesso à educação formal, de modo que os impactos do score de crédito são ainda maiores em suas vidas. Corrêa (2021) destaca, ainda, que uma série de legados concretos da escravidão são refletidos em critérios utilizados na composição do score, como a distribuição geográfica da população negra, que possui localização periférica em relação às regiões e setores hegemônicos.

Assim, é possível verificar que o desenvolvimento e a aplicação de sistemas de inteligência artificial sem a realização de avaliações específicas para proteção adequada de direitos de grupos marginalizados, além de perpetuar discriminações já existentes, poderá aprofundar o contexto de vulnerabilidade, marginalização e violência a que tais indivíduos estão submetidos.

Importante, nesse sentido, a contribuição de Mohamed, Png, Isaac (2020) na tentativa de construir uma taxonomia no cenário que denominam de colonialidade algorítmica, isto é, de colonialismo de dados no contexto das interações de algoritmos entre as sociedades, afetando alocação de recursos, comportamento político e sociocultural humanos e sistemas discriminatórios existentes. Segundo os autores, a colonialidade se apresenta nos sistemas de tomada de decisão algorítmica à medida que geram novos mercados de trabalho, impactam a dinâmica do poder geopolítico e disputam o discurso ético.

Assim, Mohamed, Png, Isaac (2020) apresentam uma taxonomia de previsão decolonial, na qual situam os casos de: opressão algorítmica institucionalizada; exploração algorítmica; e desapropriação algorítmica. Dentro de cada um desses tipos, os autores posicionam uma significativa variedade de casos que expressam desigualdades estruturais historicamente contextualizadas como continuidades coloniais. Sugerem que esses "sítios" de colonialidade estejam sob especial monitoramento, pois seriam mais propícios a um descolamento da observação empírica quanto às molduras teóricas de poder na IA, uma vez que majoritariamente concebidas de modo a-histórico.

Em decorrência de tais impactos, iniciam-se movimentos regulatórios no campo da inteligência artificial, sendo possível observar a atuação de (i) governos, por meio de estratégias nacionais e propostas legislativas; (ii) de organizações internacionais e entidades da sociedade civil, através de *policy papers* e *guidelines*; e (iii) do setor privado, por meio de *frameworks* e códigos de boas-práticas.

3. Estratégias regulatórias para sistemas de inteligência artificial

Diversas são as possibilidades regulatórias quando se trata de regulação de determinada atividade econômica ou do desenvolvimento de determinada tecnologia. Em termos de regulação relacionada ao desenvolvimento e implementação de sistemas de inteligência artificial, é possível identificar a adoção de duas grandes abordagens regulatórias: i) uma abordagem baseada em preceitos éticos; e ii) uma abordagem normativa, notadamente baseada no conceito de risco.

No que diz respeito a uma abordagem regulatória baseada em preceitos éticos, Boddington (2020) destaca que a intensificação no desenvolvimento de sistemas de inteligência artificial traz à tona diversas preocupações relacionadas à implementação ética de tais sistemas, como por exemplo as questões relativas à produção de resultados discriminatórios por sistemas de reconhecimento facial. Nesse contexto, uma das formas de oferecer uma resposta a tais problemas é o estabelecimento de padrões éticos e códigos de conduta para o desenvolvimento de sistemas de IA. Acerca dessa movimentação, Mittelstadt (2019) argumenta que nos últimos anos houve um aumento significativo de iniciativas, majoritariamente financiadas por grandes corporações (Nemitz 2018), que se debruçaram sobre a definição de princípios, valores e molduras para o desenvolvimento e implementação éticos de sistemas de inteligência artificial. A esta abordagem é atribuído o caráter autorregulatório, ou seja, desenha-se um contexto em que os próprios agentes econômicos seriam responsáveis por aplicarem tais princípios, valores e molduras ao longo do processo de desenvolvimento de tais sistemas.

Considerando essa característica, Yeung, Howes e Pogrebna (2020) apontam a possível ineficácia desse sistema, uma vez que normas autorregulatórias não possuem força vinculante, tendo em vista que são pautadas em instrumentos regulatórios baseados em *soft-law*, isto é, são instrumentos que, apesar de terem conteúdo normativo, não vinculam formalmente os agentes regulados (Trubek, Cottrell, Nance, 2005).

Estas iniciativas regulatórias resultaram, portanto, em princípios éticos tanto vagos quanto abstratos, além de produzirem orientações que falham com o seu objetivo central de guiar o desenvolvimento de sistemas de IA, por serem incapazes de endereçar questões que são fundamentais a tais sistemas, como a proteção da privacidade e dos dados pessoais (Mittelstadt, 2019; Boddington, 2020).

Nessa mesma esteira, Boddington (2020) acrescenta que uma segunda preocupação acerca dessa abordagem reside no fato de que normas éticas podem ser compreendidas como mero formalismo, de modo que são vistas como obstáculos ao desenvolvimento da tecnologia e não como diretrizes a serem seguidas e que devem constituir verdadeiramente os sistemas de IA.

Apesar de tais questões relativas à abordagem baseada em preceitos éticos, esta tem sido a principal escolha quando do desenvolvimento de Estratégias Nacionais de Inteligência Artificial, por diversos países, como Alemanha e Canadá. Esse, também, é o caso da Estratégia Brasileira de Inteligência Artificial (EBIA), instituída pela Portaria MCTI nº 4.617, de 6 de abril de 2021 e alterada pela portaria MCTI nº 4.979 de 13 de julho de 2021.

O desenvolvimento da EBIA está fundamentado em cinco princípios definidos pela Organização para a Cooperação e Desenvolvimento Econômico, quais sejam: i) crescimento inclusivo, o desenvolvimento sustentável e o bem-estar; ii) valores centrados no ser humano e na equidade; iii) transparência e explicabilidade; iv) robustez, segurança e proteção e; v) responsabilização e prestação de contas (BRASIL, s.d.). Ainda, esta Estratégia está associada ao objetivo de orientar as iniciativas do Estado brasileiro em direção ao estímulo à pesquisa, inovação e desenvolvimento de sistemas de Inteligência Artificial, além de incentivar o uso ético e consciente destes sistemas (BRASIL, s.d.).

Para alcançar os seus objetivos, a EBIA foi desenvolvida em consonância com o quadro de governança digital e com as políticas públicas implementadas, no país, sobre este tema. Assim, a Estratégia é dividida em nove eixos temáticos, dentre os quais destacam-se: i) Legislação, Regulação e Uso Ético; e ii) Governança de IA.

Como um desdobramento das disposições da EBIA, está em andamento a discussão legislativa acerca do estabelecimento de um Marco Legal para Inteligência Artificial no Brasil. Nesse sentido, destacam-se o Projeto de Lei (PL) nº 5.051, de 2019, que estabelece os princípios para o uso da Inteligência Artificial no Brasil; o PL nº 21, de 2020, que estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil e que foi aprovado pela Câmara dos Deputados; e o PL nº 872, de 2021, que dispõe sobre o uso da Inteligência Artificial.

Em 3 de fevereiro de 2022, esses três projetos passaram a tramitar conjuntamente no Senado Federal e, em sequência, em 17 de fevereiro do mesmo ano, foi instituída a Comissão de Juristas destinada a subsidiar a elaboração de minuta de substitutivo a eles. Após um período de

consultas públicas, referida Comissão elaborou uma minuta de texto substitutivo àquele apresentado nos Projetos de Lei nºs 5.051, de 2019, 21, de 2020, e 872, de 2021. Posteriormente, no âmbito do Senado Federal, foi proposto o Projeto de Lei nº 2.338/2023, que adota o texto produzido pela Comissão de Juristas.

Em primeiro lugar, destaca-se que a exposição de motivos do texto substitutivo apresentado pela Comissão revela a intenção de conciliar uma abordagem baseada em riscos com uma modelagem regulatória baseada em direitos. No que se refere a uma abordagem regulatória baseada em risco, Black e Murray (2019), afirmam ser possível traçar um paralelo entre as iniciativas regulatórias da Internet, nos anos 1990, e os debates acerca da regulação de sistemas de IA que existem atualmente. Tais semelhanças podem ser vistas a partir de duas características: i) o fato de tanto a Internet quando sistemas de IA e *machine learning* não apresentarem elementos de escassez; e ii) os riscos de implementação de ambas tecnologias serem tratados como individuais e não como riscos estruturais.

Nesse sentido, os autores apontam para o fato de que a abordagem autorregulatória, baseada em princípios éticos e códigos de conduta, está afastando modelos de regulação baseados em legislações ou até mesmo baseados em comando e controle ou co-regulação (Black, Murray, 2019). Desse modo, os autores sustentam ser, esta abordagem, insuficiente para endereçar riscos compreendidos como estruturais, uma vez que iniciativas autorregulatórias individualizam os riscos e atribuem o ônus ao consumidor, o qual deve realizar a melhor escolha dentre as opções disponíveis no mercado.

Considerando esse contexto, diversos países estão se movimentando em direção ao estabelecimento de um marco legal para inteligência artificial, apostando, desta vez, em iniciativas baseadas em *hard law*, isto é, iniciativas que se propõem mais normativas e que vinculam os agentes regulados com maior grau de coercibilidade. Dentre as iniciativas em andamento, destaca- se a "Proposta de Regulamento do Parlamento Europeu e do Conselho que estabelece regras harmonizadas em matéria de Inteligência Artificial (Regulamento Inteligência Artificial) e altera determinados atos legislativos da União", também denominado de *Artificial Intelligence Act* (AIA), datada de 2021.

Em 2020, foi publicado um *white paper* pela Comissão Europeia, que tratava sobre a necessidade do desenvolvimento de uma regulação para sistemas de Inteligência Artificial no âmbito da União Europeia, considerando o aumento significativo de implementação e impacto

destes sistemas. Como um desdobramento deste documento, foi proposto o *Artificial Intelligence Act* (*AI Act*). Atualmente, o AI Act encontra-se na fase de trílogo, marcada por negociações entre a Comissão Europeia, o Conselho da União Europeia e o Parlamento Europeu.

Trata-se de proposta ancorada em uma abordagem baseada em risco (*risk-based approach*), de modo que o risco se torna o elemento a partir do qual serão calibradas as obrigações e responsabilidades dos agentes regulados (Mahler, 2021). Para que isso seja possível, é estabelecida uma gradação de riscos, de modo que são estabelecidas, inclusive, aplicações de sistemas de IA que são proibidas, em razão da inaceitabilidade de seus riscos.

Mahler (2021) argumenta que a abordagem baseada em risco adotada pela Proposta inclui a consideração de riscos a direitos fundamentais, como a não discriminação, uma vez que eles podem ser violados em situações em que há o enviesamento de determinado algoritmo de *machine learning*, o que pode resultar em discriminações ilegais. Para o autor, o fato de o AIA tratar de riscos a direitos significa que a proposta não está circunscrita somente a uma perspectiva técnica de gerenciamento de risco, mas que integra avaliações de impacto e de risco.

Percebe-se, assim, que a gramática de avaliações de risco assume papel preponderante no processo regulação de Inteligência Artificial, assim como ocorreu ao longo do desenvolvimento das legislações mais recentes sobre proteção de dados pessoais, em que é possível identificar um processo de "risquificação" (Spina 2017) desta matéria.

Evidentemente, o desenvolvimento de sistemas de inteligência artificial exige que uma série de considerações sejam feitas (por exemplo, definir a abordagem mais adequada, avaliar o banco de dados utilizados, entender restrições etc.). Por tal razão, Bigonha (2018) aponta que a tecnologia em si é apenas um dos vários elementos atuantes. Nesse contexto, entende-se que a realização de avaliações de impacto em sistemas de inteligência artificial é essencial para sinalizar aspectos técnicos, jurídicos, éticos e sociais.

4. Sistemas de inteligência artificial e avaliações de impacto para direitos humanos

As avaliações de impacto estão inseridas em uma lógica precaucionária, uma vez que, pela aplicação do princípio da precaução, seria possível reconhecer as assimetrias de poder entre os sujeitos envolvidos no processo regulatório (Bioni, Luciano, 2019). Assim, a promoção do princípio da precaução poderia ser uma porta de entrada para o envolvimento dos cidadãos no

processo de tomada de decisão informada (Costa, 2012). Desse modo, tais avaliações são compreendidas como ferramentas utilizadas na identificação de possíveis consequências de uma determinada iniciativa sobre interesses socialmente relevantes (Kloza *et al*, 2020), especialmente se tais consequências expressarem externalidades negativas.

Nessa mesma esteira, Koshiyama e Engin (2019) apontam que os objetivos de uma avaliação de impacto estão relacionados ao estabelecimento de limites, usos e prazos do sistema de inteligência artificial, à construção de confiança entre as partes interessadas e ao registro de aspectos do funcionamento do sistema para fins de *accountability*, para além disso, as avaliações de impacto são ferramentas capazes de auxiliar no processo de tomada de decisão informada, inclusive sobre a pertinência de se iniciar ou não determinada atividade, de modo que as avaliações de impacto se traduzem em mecanismos de proteção de interesses sociais relacionados (Kloza *et al*, 2020).

Assim, Koshiyama e Engin (2019) compreendem que a avaliação de impacto deve considerar se o sistema é (i) robusto, isto é, seguro, de modo que não seja alterado ou comprometido; (ii) justo, evitando tratamentos discriminatórios e observando as consequências para determinados grupos; e (iii) explicável, ou seja, se o funcionamento do sistema pode ser compreendido por usuários e desenvolvedores.

Princípios procedimentais centralizam a dimensão normativa sobre regulação da IA, especialmente a cobrança por transparência, e, como adverte Zalnieriute (2021), podem ser apropriados e fetichizados pela agenda corporativa no sentido apenas retórico de "*transparency-washing*" (Zalnieriute, 2021).

É fundamental caminhar para um debate mais substancial da regulação. A agenda democrática, que discute os sujeitos impactados pelas tecnologias e como se inserem no processo de tomada de decisão, tem bastante a contribuir. Cuida-se, porém, de um passo ao mesmo tempo complexo e insuficiente. Complexo porque enfrenta, de acordo com Schramm (2022), dois contrapontos argumentativos. O primeiro seria o de que a participação democrática demanda alguma unidade social, cultural, política - "demos". Sob a perspectiva individual do usuário, torna-se difícil essa construção, a não ser no sentido mais fraco (embora possível) de subjugação compartilhada às regras.

O segundo argumento se refere à natureza privada das plataformas *online*. As plataformas exercem poder em esferas que afetam questões políticas, sociais, pessoais, econômicas, mas não

se confundem com o Estado. Esse poder precisa ser exercido e legitimado de forma apropriada ao seu significado político. Em que pese a discussão teórica sobre os poderes privados e a eficácia dos direitos fundamentais entre particulares, a aproximação prática entre constitucionalismo digital, participação democrática e governança das plataformas releva-se desafiadora (Schramm, 2022). Nesse sentido, é necessária a compreensão de que corporações multinacionais detêm o controle da infraestrutura básica de comunicação, desde as redes físicas ao *software* que permite e restringe a comunicação online, administrando-a a partir de um modelo de negócios baseado em publicidade de vigilância e em vastos lucros privados, um modelo pouco propício para o debate político construtivo e para a mídia independente.

Assim, Mantelero (2022), ao tratar do crescente uso de sistemas de inteligência artificial, ressalta que as consequências do tratamento de dados já não se limitam às questões de privacidade e proteção de dados, mas abrangem diversos direitos humanos. No entanto, como os dados estão no centro do funcionamento dos sistemas de inteligência artificial, algumas perspectivas iniciais são extraídas da regulação de proteção de dados, tais como parâmetros para qualidade, segurança e instrumentos de governança de dados.

Desse modo, Gellert (2015) e Quelle (2018) afirmam não haver incompatibilidades entre as abordagens baseadas em risco e em direito, uma vez que a abordagem baseada em risco serve para tutelar direitos dos indivíduos, na medida em que as obrigações e responsabilizações são moduladas de acordo com o nível dos riscos identificados em determinada atividade de tratamento de dados ou processo de implementação de uma nova tecnologia.

Ainda, Kloza (2014) argumenta ser possível aplicar os princípios relativos à justiça procedimental quando da condução de avaliações de impacto, com intuito de se alcançar o maior nível de justiça deste processo de tomada de decisão. O autor destaca, especificamente, o papel da aplicação do princípio da participação quando da condução de avaliações de impacto, o qual pode ser compreendido como uma das primeiras formas de se fortalecer a participação pública durante os processos de governança.

Nesse contexto, por vezes, o relatório de impacto à proteção de dados é utilizado para endereçar riscos derivados de sistemas de inteligência artificial. No ordenamento jurídico brasileiro, o art. 5º, XVII, da Lei nº 13.709/2018 (Lei Geral de Proteção de dados, abreviada por "LGPD"), estabelece que o relatório é a documentação do controlador que contém a descrição dos processos de tratamento de dados pessoais que podem gerar riscos às liberdades civis e aos direitos

fundamentais de titulares, bem como medidas, salvaguardas e mecanismos de mitigação de risco.

Acerca das avaliações de impacto para proteção de dados, Kloza (2014) destaca que a partir de uma metodologia de gestão de risco adequada, os riscos possíveis e outros impactos negativos sobre a privacidade podem ser identificados e idealmente mitigados, de modo que eventuais riscos residuais sejam justificados e registrados.

No entanto, especificamente sobre as avaliações de impacto de sistemas de inteligência artificial, Mantelero (2022) entende que nem os modelos tradicionais de avaliação de impacto da proteção de dados, nem os procedimentos mais amplos de avaliação de impacto social ou ético parecem dar uma resposta adequada aos desafios levantados pelo desenvolvimento e uso de tais sistemas.

Para o autor, o caminho estaria em uma avaliação centrada nos direitos humanos, abrangendo não só avaliações sob a perspectiva de proteção de dados, mas também uma análise sobre impactos para outros direitos e liberdades fundamentais, abordando ainda aspectos éticos e sociais (*Human Rights, Ethical and Social Impact Assessment* - HRESIA).

Mantelero (2022) destaca que as avaliações de impacto centradas no ser humano podem consolidar um ambiente mais seguro, de modo que o ônus da avaliação dos potenciais benefícios e riscos para os direitos e liberdades não deve recair sobre os ombros dos indivíduos ou grupos atingidos. Em síntese, o autor afirma que, assim como os consumidores não precisam verificar a segurança dos automóveis que compram, os usuários finais de sistemas de inteligência artificial também não deveriam ter de verificar se os seus direitos e liberdades são salvaguardados.

Nesse sentido, Mantelero (2022) aponta que é necessário definir se o modelo para a elaboração da avaliação de impacto em direitos humanos e aspectos éticos e sociais será geral ou específico/setorial. Para o autor, um modelo de avaliação centrado em uma tecnologia específica (por exemplo, uma avaliação de impacto de Internet das Coisas ou avaliação de impacto de *smart cities*) parece inadequado ou apenas parcialmente eficaz, pois independentemente dos distintos aspectos técnicos das tecnologias, o foco de uma abordagem centrada no ser humano está necessariamente na proteção de direitos. Assim, uma abordagem setorial deve concentrar sua atenção não somente nas tecnologias, mas no contexto e nos aspectos que assumem relevância naquele setor.

A avaliação a partir de elementos sociais e contextuais é essencial, pois, conforme buscou- se demonstrar anteriormente, determinadas aplicações de inteligência artificial possuem

o potencial de impactar - em maior medida - grupos e comunidades específicas. Nesse sentido, vale destacar que pensar em interesses coletivos durante o processo de avaliação de impacto de sistemas de inteligência artificial torna-se extremamente relevante, pois há inúmeros modelos de negócios que tratam de dados pessoais para perfilização (*profiling*), classificação (*scoring*) e monitoramento do comportamento de grupos, evidenciando a tensão existente entre pessoa e mercado.

Além disso, as avaliações de impacto são instrumentos importantes para *accountability*, isto é, para concretizar práticas que remetem à responsabilidade com ética, à obrigação, à busca por transparência e à prestação de contas (Gutierrez, 2019). Gutierrez (2019) aponta que a *accountability* está relacionada ao fato de que aqueles que desempenham funções relevantes na sociedade deveriam dar transparência ao que estão fazendo, por quais motivos e como estão fazendo. Assim, a *accountability* também remete à necessidade de uma estrutura de governança.

Desse modo, a partir da ideia de *accountability*, é possível compreender que - mais do que uma ferramenta para registro do funcionamento de um sistema de inteligência artificial e adoção de medidas mitigatórias - as avaliações de impacto podem ser enxergadas como instrumentos de transparência e controle social.

A abordagem regulatória baseada no gerenciamento de riscos decorrentes de sistemas de inteligência artificial é acompanhada de obrigações relacionadas à prestação de contas e demonstração de evidências de conformidade, por exemplo, por meio de documentações técnicas, rotinas e procedimentos internos que farão parte de uma estrutura de governança. Tal abordagem também deve ser pautada em práticas de transparência que permitam controle social e escrutínio público acerca do desenvolvimento e aplicação de sistemas de inteligência artificial. Para que a transparência se efetive de forma substancial, riscos considerados inaceitáveis, como aqueles associados ao racismo algorítmico ou à exploração econômica de crianças e adolescentes, devem ser imediatamente descartados em qualquer projeto.

A construção de um modelo efetivo e substancial de avaliação de impacto para direitos humanos não se esgota, assim, em rotinas e procedimentos, mas passa necessariamente pela proibição de práticas e atividades consideradas inadmissíveis. Caso contrário, as ferramentas que compõem a estrutura de governança de sistemas de inteligência artificial podem legitimar práticas que violam direitos humanos, autorizando a internalização de riscos considerados intoleráveis.

No âmbito das propostas legislativas que buscam regulamentar o uso de sistemas de IA,

é interessante notar que o PL 2.338/23 prevê a obrigatoriedade de realização de avaliação preliminar de todo sistema de IA (art. 13) e, especificamente para sistemas classificados como de alto risco, é necessário realizar avaliação de impacto algorítmico, elaborada por profissional ou equipe de profissionais com conhecimentos técnicos, científicos e jurídicos necessários para realização do relatório e com independência funcional, conforme se extrai dos art. 22 e 23.

A metodologia da avaliação de impacto conterá, ao menos, as seguintes etapas: (i) preparação; (ii) cognição do risco; (iii) mitigação dos riscos encontrados; e (iv) monitoramento. Caberá à autoridade competente a regulamentação da periodicidade de atualização das avaliações de impacto, considerando o ciclo de vida dos sistemas de inteligência artificial de alto risco e os campos de aplicação, podendo incorporar melhores práticas setoriais. Além disso, a autoridade competente poderá estabelecer outros critérios e elementos para a elaboração de avaliação de impacto, incluindo a participação dos diferentes segmentos sociais afetados, conforme risco e porte econômico da organização.

Além disso, em relação ao escrutínio público, o PL 2.338/23 estabelece que, nos limites dos segredos industrial e comercial, as conclusões da avaliação de impacto serão públicas e seu processo de atualização contará também com participação pública, a partir de procedimento de consulta a partes interessadas, ainda que de maneira simplificada (art. 26 e art 25, §2º, respectivamente). Destaca-se, inclusive, que cabe à autoridade competente criação e manutenção de base de dados de inteligência artificial de alto risco, acessível ao público, que contenha os documentos públicos das avaliações de impacto, respeitados os segredos comercial e industrial.

No âmbito da União Europeia, o texto do AI Act também estabelece requisitos para o gerenciamento de riscos de sistemas de IA classificados como de alto risco. Nesse sentido, a proposta prevê que o sistema de gerenciamento de riscos deve consistir em um processo iterativo contínuo executado durante todo o ciclo de vida do sistema, exigindo uma atualização sistemática regular e incluindo etapas como (i) identificação e análise dos riscos conhecidos e previsíveis; (ii) estimativa e avaliação dos riscos que podem surgir quando o sistema de IA de alto risco for usado de acordo com a finalidade pretendida e em condições de uso indevido razoavelmente previsível; e (iii) avaliação de outros riscos que possam surgir pós-comercialização.

Em junho de 2023, o Parlamento Europeu estabeleceu suas posições de negociação em relação ao texto do AI Act, dentre as quais se destaca a imposição de que aqueles que implementam sistemas de IA classificados como de alto risco realizem uma avaliação de impacto sobre direitos

fundamentais, incluindo consulta à autoridade competente e às partes interessadas relevantes (European Parliament, 2023).

Independentemente dos modelos de gerenciamento de riscos em debate na esfera legislativa, exemplos de avaliações de impacto algorítmico já começam a surgir. Por exemplo, o Governo do Canadá conta com uma ferramenta denominada *Algorithmic Impact Assessment Tool* (Government of Canada) e, na mesma direção, o Governo de Singapura conta com a ferramenta "AI Verify" - *AI Governance Testing Framework and Toolkit* (IMDA).

O Governo do Canadá disponibiliza, em acesso aberto, avaliações de impacto já concluídas. Um exemplo é a avaliação de impacto algorítmico em projeto envolvendo a automatização de revisão de solicitações não complexas para Vistos de Residente Temporário e Permissões de Trabalho feitas sob a Autorização Canadá-Ucrânia para Viagem de Emergência (CUAET). Trata- se de sistema que automatiza decisões positivas de elegibilidade, registro de determinações positivas de admissibilidade e aprovação. O sistema não recusa candidaturas ou recomenda a recusa de candidaturas, nem faz determinações negativas.

A partir da ferramenta de avaliação de impacto disponibilizada pelo Governo do Canadá, o projeto foi avaliado como de impacto "Nível 2", que corresponde a um impacto moderado (Canada Department of Citizenship and Immigration). Os níveis de impacto são diferenciados com base em critérios de reversibilidade e duração esperada: decisões automatizadas com pouco ou nenhum impacto são reversíveis e breves, enquanto aquelas com impacto muito alto são irreversíveis e perpétuas. Os níveis de impacto determinam as mitigações exigidas pela Diretiva sobre Tomada de Decisão Automatizada, que incluem requerimentos como revisão por pares, transparência, envolvimento humano direto (human-in-the-loop), explicação e treinamento.

No caso do projeto de automação de revisão de solicitações não complexas para Vistos de Residente Temporário e Permissões de Trabalho feitas sob a Autorização Canadá-Ucrânia para Viagem de Emergência (CUAET), verifica-se que o resultado da avaliação de impacto algorítmico exigiu a revisão por pares (por exemplo, por meio de experts de instituições governamentais, pesquisadores de instituições não governamentais ou consultores externos), bem como a disponibilização de informações em linguagem acessível em todos os canais de serviço utilizados e garantia de explicabilidade sobre o funcionamento do projeto.

É importante perceber que processos de gerenciamento de riscos já fazem parte da própria gramática empresarial, como se nota na ampla difusão de modelos de *compliance*. Acontece que,

no caso de tecnologias que impactam diretamente os direitos humanos, como se observa com o reconhecimento facial, esses processos devem seguir uma lógica própria, diferente dos mecanismos relacionados à gestão do risco empresarial.

Se nos modelos tradicionais, a preocupação se volta para a própria empresa e para os riscos do negócio, ainda que eventuais violações possam impactar a própria atividade, o foco, no caso dos direitos humanos, deve estar sempre nos detentores dos direitos que deverão ser respeitados. Da mesma forma, grupos historicamente submetidos a opressões e violências podem se mostrar mais expostos a um risco maior de discriminação e de outros danos ocasionados pelo uso de novas tecnologias.

Nesse sentido, a elaboração de um modelo de avaliação de impacto para direitos humanos pressupõe a percepção de que os riscos associados aos usos de novas tecnologias podem não se distribuir de forma linear entre pessoas e grupos.

Considerações finais

O presente trabalho teve por objetivo explorar o cenário de crescente desenvolvimento e aplicação de sistemas de inteligência artificial. No entanto, buscou-se demonstrar que, para além dos avanços em termos de eficiência e benefícios econômicos trazidos por tais tecnologias, é importante observar impactos para direitos humanos, na medida em que os sistemas de inteligência atravessam interações humanas e passam a estar presentes em processos decisórios.

A partir do cenário de impacto de sistemas de inteligência artificial para direitos humanos, o trabalho ressalta como resultado parcial da pesquisa realizada a conclusão de que tais tecnologias podem aprofundar práticas discriminatórias já existentes, exasperando violências e opressões.

Nesse contexto, as avaliações de impacto se apresentam como um dos instrumentos que podem fazer parte de uma estrutura de governança voltada para o gerenciamento de riscos decorrentes do desenvolvimento e utilização de sistemas de inteligência artificial. No entanto, a pesquisa conclui pela necessidade de que tais avaliações levem em consideração os impactos para direitos humanos e reconheçam que determinadas pessoas e grupos sociais podem ser especificamente afetados.

O trabalho busca, por fim, contribuir para a agenda de pesquisa em que se insere indicando que o desenvolvimento e a aplicação de sistemas de inteligência artificial venham

acompanhados de rotinas e instrumentos que sejam capazes de oferecer transparência e de aferir mais substancialmente os riscos, possibilitando o controle social sobre o funcionamento da tecnologia e a construção de alternativas mais legítimas.

Referências

ANGWIN, Julia, LARSON, Jeff, MATTU, Surya e KIRCHNER, Lauren. Machine bias: there's software used across the country to predict future criminals. And it's biased against black. *Propublica*. 23 maio 2016. Disponível em: https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing .

BIGONHA, Carolina. Inteligência Artificial em perspectiva. *Panorama setorial da Internet*. 2018, vol. 10, no. 2. Disponível em: https://cetic.br/media/docs/publicacoes/1/Panorama_outubro_2018_online.pdf

BIONI, Bruno Ricardo; LUCIANO, Maria. O princípio da precaução na regulação de inteligência artificial: seriam as leis de proteção de dados o seu portal de entrada? *Inteligência Artificial e Direito*: ética, regulação e responsabilidade. São Paulo, SP: Revistas dos Tribunais, 2009.

BLACK, Julia; MURRAY, Andrew. Regulating AI and Machine Learning: Setting the Regulatory Agenda. *European Journal of Law and Technology*, v. 10, n. 3, 2019. Disponível em: http://eprints.lse.ac.uk/102953/ .

BODDINGTON, Paula. Normative modes: codes and standards. Em: *The Oxford Handbook of Ethics of AI*. Nova Iorque: Oxford University Press, 2020.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: intersectional accuracy disparities in commercial gender classification. *Proceedings Of Machine Learning Research*: Conference on Fairness, Accountability, and Transparency, [s. l], v. 81, n. 1, p. 11, 2018. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a.html.

CANADA DEPARTMENT OF CITIZENSHIP AND IMMIGRATION. Algorithmic Impact

Assessment Results - Automate the review of non-complex applications for Temporary Resident Visas and Work Permits made under the Canada-Ukraine Authorization for Emergency Travel (CUAET). Disponível em: https://opencanada.blob.core.windows.net/opengovprod/resources/ec3f935c-2192-4ebd-b21b-4265a9d0e08a/f01253979-annex-a-aia-on-cuaet-automation-en.pdf?sr=b&sp=r&sig=WhsgHgAqisI14fKcBAdk0DxpkSAFqgjVoeXHLxVacg0%3D&sv=2015-07-08&se=2023-07-01T22%3A21%3A47Z

CARNEIRO, Aparecida Sueli. *A construção do outro como não-ser como fundamento do ser*. Tese (Doutorado). São Paulo, SP: Universidade de São Paulo, São Paulo, 2005. Disponível em: https://repositorio.usp.br/item/001465832

CERKA, Paulius; GRIGIENE, Jurgita; SIRBIKYTE, Gintare. Liability for damages caused by Artificial Intelligence. *Computer Law & Security Review*, Elsevier. 2015, vol. 31, no. 3, p. 376-389. Disponível em: https://doi.org/10.1016/j.clsr.2015.03.008

CORRÊA, Bianca Kremer Nogueira. *Direito e tecnologia em pespectiva amefricana*: autonomia, algoritmos e vieses raciais. Tese (Doutorado). Rio de Janeiro, RJ: Pontifícia Universidade Católica do Rio de Janeiro, 2021. Disponível em: https://www.maxwell.vrac.puc-rio.br/58993/58993.PDF .

CORTIZ, Diogo. Inteligência Artificial: equidade, justiça e consequências. *Panorama setorial da Internet*. 2020. vol. 12, no. 1. Disponível em: https://cetic.br/media/docs/publicacoes/6/20200626161010/panorama_setorial_ano-xii_n_1_inteligencia_artificial_equidade_justi%C3%A7a.pdf

COSTA, Luiz. Privacy and the precautionary principle. *Computer Law & Security Review*. 2012. v. 28, no. 1, p. 14–24, 2012. Disponível em: https://doi.org/10.1016/j.clsr.2011.11.004

DOUZINAS, Costas. *O fim dos direitos humanos*. Tradução de Luzia Araújo. São Leopoldo: Unisinos, 2009.

EUROPEAN PARLIAMENT. *Parliament's negotiating position on the artificial intelligence act*. Disponível em: https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/747926/EPRS_ATA(2023)747926_ EN.pdf

GELLERT, Raphël. Data protection: a risk regulation? Between the risk management of everything and the precautionary alternative. *International Data Privacy Law*. 2015, vol. 5, no. 1, p.3-19. Disponível em: https://doi.org/10.1093/idpl/ipu035

GOVERNMENT OF CANADA. *Algorithmic Impact Assessment tool*. Disponível em: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

GUTIERREZ, Andriei. É possível confiar em um sistema de Inteligência Artificial? Práticas em torno da melhoria da sua confiança, segurança e evidências de accountability. Em: *Inteligência Artificial e Direito*. São Paulo, SP: Revista dos Tribunais, 2019.

HERRERA FLORES, Joaquín. *Teoria Crítica dos Direitos Humanos*: os direitos humanos como produtos culturais. Rio de Janeiro, RJ: Lúmen Júris, 2009.

IMDA. AI Verify AI Governance Testing Framework and Toolkit. Disponível em: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press releases/2022/singapore-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency-invites-companies-to-pilot-and-contribute-to-international-standards-development

KLOZA, Dariusz, *et al*. VAN DIJK, Niels, GELLERT, Raphaël Maurice, BOROCZ, Istvan Mate, TANAS, Alessia, MANTOVANI, Eugenio e QUINN, Paul, 2017. Data protection impact assessments in the European Union: complementing the new legal framework towards a more robust protection of individuals. *POLICY BRIEF D.PIA.LAB*. [Acesso em 31 julho 2022]. Disponível em: https://doi.org/10.31228/osf.io/b68em

KOSHIYAMA, Adriano; ENGIN, Zeynep. Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making. *Data for Policy 2019: Digital Trust and Personal Data*. Londres: Data for Policy, 2019. Disponível em: https://zenodo.org/record/3361708#.YnCBUtrMKiM

MACHADO, Joana de Souza; NEGRI, Sergio M. C. Avila; GIOVANINI, Carolina Fiorini Ramos. Nem invisíveis, nem visados: inovação, direitos humanos e vulnerabilidade de grupos no contexto da COVID-19. *Liinc em Revista*. 2020, vol. 16, p. 1-21. Disponível em: http://revista.ibict.br/liinc/article/view/5367

MAHLER, Tobias. Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal. *Nordic Yearbook of Law and Informatics*. 2021 Disponível em: https://ssrn.com/abstract=4001444 .

MALDONALDO-TORRES, Nelson. Sobre la colonialidad del ser: contribuciones al desarrollo de un concepto. Em: *Reflexiones para una diversidad epistémica más allá del capitalismo global*. Bogotá: Siglo del Hombre Editores, Universidad Central, Instituto de Estudios Sociales Contemporáneos y Pontificia Universidad Javeriana, Instituto Pensar, 2007.

MANTELERO, Alessandro. *Beyond Data*: human rights, ethical and social impact assessment in AI. Berlim, Alemanha: Springer, 2022. Disponível em: https://link.springer.com/book/10.1007/978-94-6265-531-7

MCCARTHY, John, et al. A proposal for the dartmouth summer research project on artificial intelligence, 1965. Disponível em: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html .

MITTELSTADT, Brent. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*. 2019, vol.1, no. 11, p. 1-19. Disponível em: https://doi.org/10.1038/s42256-019-0114-4

MOHAMED, Shakir, PNG, Marie-Therese e ISAAC, William. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy and Technology*. 2020. vol. 33, no. 4, p.659-684. Disponível em: http://dx.doi.org/10.1007/s13347-020-00405-8

NEGRI, Sergio Marcos Carvalho Avila. Robôs como pessoas: a personalidade eletrônica na robótica e na inteligência artificial. *Revista Pensar*, 2020. vol. 25, no. 3. Disponível em: http://dx.doi.org/10.5020/2317-2150.2018.10178

NEMITZ, Paul. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions Of The Royal Society*: Mathematical, Physical and Engineering Sciences. 2018. vol. 376, no. 2133, 2018. Disponível em: http://dx.doi.org/10.1098/rsta.2018.0089

PIRES, Thula Rafaela de Oliveira. Direitos Humanos traduzidos em pretuguês. Em: *Seminário Internacional Fazendo Gênero 11 & 13th Women's Worlds Congress*. Florianópolis, SC. 2017. Disponível em: http://www.en.wwc2017.eventos.dype.com.br/resources/anais/1499473935_ARQUIVO_Texto_completo_MM_FG_ThulaPires.pdf

QUARESMA, Alexandre. *Inteligência artificial e bioevolução: Ensaio epistemológico sobre organismos e máquinas*. Dissertação (Mestrado). São Paulo, SP: Pontifícia Universidade Católica de São Paulo (PUC/SP), 2020.

QUELLE, Claudia. Does the risk-based approach to data protection conflict with the protection of fundamental rights on a conceptual level? *Tilburg Law School Research Paper*, 2015, p. 1-36. Disponível em: https://dx.doi.org/10.2139/ssrn.2726073

QUIJANO, Anibal. Colonialidade do poder, eurocentrismo e América Latina. Em: *A colonialidade do saber: eurocentrismo e ciências sociais – perspectivas latino-americanas*. Ciudad Autónoma de Buenos Aires, Argentina: Clacso, 2005.

RUSSELL, Stuart, NORVIG, Peter. *Inteligência artificial*. 3. ed. Rio de Janeiro: Elsevier, 2013.

SCHRAMM, Moritz. Where is Olive? Or: Lessons from Democratic Theory for Legitimate Platform Governance. *The Digital Constitutionalist*, 2022. Disponível em: https://digi-con.org/where-is-olive-or-lessons-from-democratic-theory-for-legitimate-platform-governance/

SPINA, Alessandro. A Regulatory Mariage de Figaro: risk regulation, data protection, and data ethics. *European Journal of Risk Regulation*. 2017. vol. 8, no.1, p. 88-94. Disponível em: https://doi.org/10.1017/err.2016.15

STEIBEL, Fabro; VICENTE, Victor Farias; JESUS, Diego Santos Vieira. Possibilidades e potenciais da utilização da Inteligência Artificial. Em: *Inteligência Artificial e Direito*. São Paulo: Revista dos Tribunais, 2019. p. 55.

SILVA, Tarcízio. *Racismo algorítmico: inteligência artificial e discriminação nas redes digitais*. [S.I]: Democracia Digital, 2022.

TOMASEV, Nenad et al. Fairness for Unobserved Characteristics: insights from technological impacts on queer communities. P*roceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021. Disponível em: http://dx.doi.org/10.1145/3461702.3462540.

TRUBEK, David M.; COTRELL, Patrick; NANCE, Mark, 2005. "Soft Law," "Hard Law," and European Integration: Toward a Theory of Hybridity. *Legal Studies Research Paper Series*. 2005. no. 1002, p. 1- 42. Disponível em: http://dx.doi.org/10.2139/ssrn.855447

UNIÃO EUROPEIA. *Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206)*. Bruxelas: Comissão Europeia, 2021.

YEUNG, Karen, HOWES, Andrew, POGREBNA, Ganna. AI Governance by Human Rights-Centered Design, Deliberation, and Oversight: an end to ethics washing. Em: *The Oxford*

Handbook of Ethics of AI. Nova Iorque: Oxford University Press, 2020.

ZALNIERIUTE, Monika. "Transparency-Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. Em: *Critical Analysis of Law*. UNSWLRS 33, 2021. Disponível em: https://cal.library.utoronto.ca/index.php/cal/article/view/36284.

Como citar este artigo:

NEGRI, Sergio M. C. Ávila; MACHADO, Joana de Souza; GIOVANININ, Carolina Fiorini Ramos; BATISTA, Nathan Pascoalini Ribeiro. Sistemas de inteligência artificial e avaliações de impacto para direitos humanos. **Revista Culturas Jurídicas**, V. 10, n. 26, p. 153-181, 2023. Disponível em: https://periodicos.uff.br/culturasjuridicas/index.

NEGRI, Sergio M. C. Ávila; MACHADO, Joana de Souza; GIOVANININ, Carolina Fiorini Ramos; BATISTA, Nathan Pascoalini Ribeiro. Sistemas de inteligência artificial e avaliações de impacto para direitos humanos. **Revista Culturas Jurídicas**, V. 10, n. 26, p. 153-181, 2023. Available for access: https://periodicos.uff.br/culturasjuridicas/index.

NEGRI, Sergio M. C. Ávila; MACHADO, Joana de Souza; GIOVANININ, Carolina Fiorini Ramos; BATISTA, Nathan Pascoalini Ribeiro. Sistemas de inteligência artificial e avaliações de impacto para direitos humanos. **Revista Culturas Jurídicas**, V. 10, n. 26, p. 153-181, 2023. Disponible en: https://periodicos.uff.br/culturasjuridicas/index.