Novos Estudos Jurídicos

[NEJ], vol. 30, n. 1, 2025

O PAPEL DO RELATÓRIO DE AVALIAÇÃO DE IMPACTO ALGORITIMICO PARA O FOMENTO DE SISTEMAS DE IA ÉTICOS BY DESIGN

Maria Edelvacy Marinho
Universidade Presbiteriana Mackenzie

Camilo Onoda Caldas
Escola Paulista de Direito - EPD

Tatiana Aguiar
Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa - IDP

Contextualização: Apesar de crescente, o grau de conscientização da sociedade quanto aos riscos e benefícios do uso de sistemas de inteligência artificial ainda é reduzido. Consequentemente, a criação de um modelo regulatório capaz de promover o adequado equilíbrio entre incentivo à inovação e à proteção da sociedade, requer o uso de ferramentas de atualização que permitam o constante aperfeiçoamento do modelo. O tipo de dano, a velocidade de disseminado e o impacto social provocados pela IA quando seu uso gera resultados injustos em desfavor de grupos vulneráveis tornam urgente a discussão de medidas de gestão de riscos e prevenção conflitos.

Objetivos: Neste artigo, analisamos o papel do relatório de risco e impacto para fomentar a adoção dos valores da diversidade e da inclusão no design desta tecnologia e, assim, prevenir conflitos de natureza discriminatória.

Métodos: Aplicou-se o procedimento de pesquisa bibliográfico-documental para coleta de dados e utilizou-se do método hipotético-dedutivo para o teste da hipótese.

Resultados: Apesar de ser constatado um elevado percentual de documentos de soft law que direta ou indiretamente recomendam o uso de mecanismos que minimizem resultados discriminatórios no uso dos sistemas de IA, e, dentre eles, o relatório de riscos e impactos para o uso de tais sistemas, o atual estágio de desenvolvimento normativo permite apenas identificar fatores que poderiam favorecer o uso do relatório como indutor de um design antidiscriminatório dos sistemas de IA e quais elementos este relatório deveria conter de modo a desempenhar tal papel.

Palavras-chave: Discriminação; Ética; Inteligência Artificial; Relatório de avaliação de impacto algorítmico.

EL PAPEL DE LOS INFORMES DE EVALUACIÓN DE IMPACTO ALGORÍTMICO EN LA PROMOCIÓN DE SISTEMAS DE IA ÉTICOS POR DISEÑO

Contextualización: A pesar de crecer, el grado de concienciación de la sociedad sobre los riesgos y beneficios del uso de sistemas de inteligencia artificial es todavía bajo. En consecuencia, la creación de un modelo regulatorio capaz de promover el equilibrio adecuado entre el fomento de la innovación y la protección de la sociedad requiere el uso de herramientas de actualización que permitan la mejora constante del modelo. El tipo de daño, la velocidad de difusión y el impacto social de los sistemas de IA cuando su uso genera resultados injustos en perjuicio de grupos vulnerables hacen urgente discutir medidas de gestión de riesgos y prevención de controversias.

Objetivos: En este artículo analizamos el papel del reporte de gestión del riesgos y évaluación de impacto para fomentar la adopción de valores de diversidad e inclusión en el diseño de esta tecnología y así prevenir controversias de carácter discriminatorio.

Método: Para la recolección de datos se aplicó el procedimiento de investigación bibliográfico-documental y para la prueba de hipótesis se utilizó el método hipotético-deductivo.

Resultados: A pesar de encontrarse un alto porcentaje de documentos de soft law que directa o indirectamente recomiendan el uso de mecanismos que minimicen los resultados discriminatorios en el uso de sistemas de IA y, entre ellos, el reporte de gestión del riesgos y évaluación de impacto por el uso de dichos sistemas, el estado actual de desarrollo normativo solo permite identificar los factores que podrían favorecer el uso del informe como inductor de un diseño antidiscriminatorio de los sistemas de IA y qué elementos debe contener dicho informe para desempeñar dicho papel.

Palabras clave: Discriminación; Ética; Inteligencia Artificial; Informe de evaluación de impacto algorítmico.

THE ROLE OF ALGORITHMIC IMPACT ASSESSMENT REPORTING IN PROMOTING ETHICAL AI SYSTEMS BY DESIGN

Contextualization: Despite the growing awareness in society about the risks and benefits of using artificial intelligence systems, the level of understanding about it is still low. Therefore, the creation of a regulatory model capable of promoting the appropriate balance between encouraging innovation and protecting society requires update tools that allow constant improvement of the model. The type of harm, the speed of dissemination and the social impact caused by unfair AI systems against vulnerable groups make it urgent to discuss risk management and conflict prevention measures.

Objectives: In this article, we analyze the role of risk and impact assessment in fostering the adoption of diversity and inclusion values in the design of this technology and, thus, preventing conflicts of a discriminatory nature.

Method: In this article we applied the bibliographical-documentary research procedure for data collection, and we used the hypothetical-deductive method to test the hypothesis.

Results: Although we verified the existence of a high percentage of soft law documents that directly or indirectly recommend the use of mechanisms that minimize discriminatory results in the use of AI systems, and, among them, the risk and impact assessment report, the current stage of normative development only makes it possible to identify factors that may act in favor of the use of this report as an inducer of an anti-discriminatory design of AI systems and which kind of data this report should contain to play such a role.

Keywords: Discrimination; Ethics; Artificial intelligence; Algorithmic Impact Assessment Report.

INTRODUÇÃO

A quarta revolução industrial se distingue das revoluções pretéritas em razão de três fatores: velocidade, escopo e impacto dos sistemas.[1] A velocidade se refere à redução do intervalo de tempo entre revoluções; a mudança no escopo está relacionada à transversalidade de setores afetados diretamente pelas inovações e, o impacto, nesse caso, é distintivo em três campos: produção, gestão e governança.[2] Neste artigo buscamos analisar os impactos no campo da governança de riscos resultantes do emprego dos sistemas de inteligência artificial (IA).

Dentre os desafios jurídicos para a regulação dos sistemas de IA está o desenvolvimento de procedimentos e ferramentas capazes de lidar com os potenciais efeitos discriminatórios do emprego de tais sistemas. Isso porque, a IA baseada nos modelos de aprendizado por máquina veem sendo alimentados por base de dados que, apesar de verídicos, muitas vezes são incompletos e apenas refletem o histórico de discriminação da nossa sociedade. A depender da forma como tais dados forem utilizados, é possível a geração de resultados que prejudiquem injustamente determinados grupos.

A partir dessa situação problema, buscamos avaliar nesta pesquisa se a obrigatoriedade da produção de um relatório de risco e impacto poderia contribuir para adoção dos valores da diversidade e da inclusão desde o design de sistemas de IA. Ao longo do trabalho, testamos a hipótese de que esta ferramenta poderia ter efeitos benéficos quando esta integra um arcabouço legal e institucional capaz de avaliar e monitorar os dados gerados por tais relatórios.

A pesquisa foi conduzida tendo por objetivo testar a viabilidade e adaptabilidade dessa solução tendo como referências: o modelo de regulação responsiva, a avaliação do risco que os vieses algorítmicos podem representar para determinados grupos, o arcabouço jurídico internacional e os projetos de lei em tramitação no congresso nacional. A utilização dessas ferramentas são fundamentais para reforçar sistemas éticos e prevenir conflitos, medidas indispensáveis nos modelos de gestão contemporâneos.

[1] SCHWAB, Klaus. **A quarta revolução industrial**. São Paulo: Edipro, 2018.

[2] SCHWAB, Klaus. **A quarta revolução industrial**...

1. A GESTÃO DE RISCO POR MEIO DA REGULAÇÃO DE TECNOLOGIAS DISRUPTIVAS DA QUARTA REVOLUÇÃO INDUSTRIAL[3]

A gestão dos riscos resultantes da quarta revolução industrial[4] exige a adoção de novas estratégias de governança regulatória a fim de viabilizar a identificação de situações de risco ou perigo, ferramentas de minimização de tais riscos, desenvolvimento de protocolos de contenção de danos e sistemas de prevenção de conflitos.[5]

O conhecimento acumulado sobre as revoluções industriais anteriores permite inferir que o investimento em instrumentos de minimização dos riscos sociais e ambientais resultantes do uso de tecnologia disruptiva é tão importante quanto a tecnologia em si. A prevenção de conflitos e de eventos adversos tornaram-se portanto essenciais, pois não apenas danos ambientais podem ser irreversíveis, mas também os de imagem e reputação. As tecnologias digitais além de permitirem a difusão quase instantânea de informações por diversos meios, viabilizam a manutenção e acessibilidade desta por um período indeterminado.

O Estado, como espaço ainda dotado de confiança e legitimidade para gestão de tais riscos, tem se deparado com desafios de diferentes naturezas para cumprir tal papel.[6] Um destes desafios está em repensar a abordagem dicotômica da atuação do Estado diante de uma inovação disruptiva no que se refere a dois elementos do desenho regulatório: eleição de prioridades e a definição do quando intervir[7]. A decisão sobre quem priorizar induz a contraposição da proteção dos consumidores ao desenvolvimento tecnológico das empresas. Por esse posicionamento, o Estado teria apenas duas alternativas: ou opta por incentivar

[3] Neste trabalho será adotado o conceito de regulação conforme definido pela OCDE: "regulation refers to the diverse set of instruments by which governments set requirements on enterprises and citizens. Regulation includes all laws, formal and informal orders, subordinate rules, administrative formalities, and rules issued by non-governmental or self-regulatory bodies to whom governments have delegated regulatory power'. (OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).** Paris: ODCE, 2019).

[4] A quarta revolução industrial é caracterizada por Klaus Schwab por tecnologias que fundem os elementos físico, o digital e o biológico na construção de uma solução. (SCHWAB, Klaus. **A quarta revolução industrial**...).

[5] Esse processo pode ser analisado nas obras a seguir: UNCTAD **-** UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. **Technology and Innovation Report 2021:** Catching technological waves – Innovation with equity. Geneva: UNCTAD, 2021. Disponível em: https://unctad.org/system/files/official-document/tir2020_en.pdf. Acesso em: 10 jul. 2023; WEF - WORLD ECONOMIC FORUM. **Agile Governance:** Reimagining Policy-making in the Fourth Industrial Revolution. Geneva: World Economic Forum, 2018. Disponível em: https://encurtador.com.br/UZVug. Acesso em: 03 jul. 2023.

[6] GUNNINGHAM, N; SINCLAIR, D. **Designing smart regulation. OECD Global Forum on Sustainable Development**. 2004. Disponível em: https://www.oecd.org/env/outreach/33947759.pdf. Acesso em: 10 jul. 2023.

[7] Essa mudança no modelo, é destacada pela OCDE. "(...) Realizing the full potential of innovation in high-uncertainty contexts require a paradigm shift in regulatory policy and governance, from the traditional "regulate and forget" to "adapt and learn". (OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464).** Paris: ODCE, 2021).

determinada inovação para permitir um melhor posicionamento do país frente aos seus competidores via redução/não implementação de exigências regulatórias, ou pauta sua atuação na adoção de regras para minimizar potenciais riscos desta nova tecnologia aos seus consumidores, aumentando assim os custos e a responsabilidade das empresas inovadoras em uma fase nascente da tecnologia.

Tal abordagem estatal é ineficaz diante da natureza das inovações da economia digital e da necessidade de cooperação entre empresas e Estados para uma efetiva gestão dos riscos advindos do uso de tais inovações.[8]Isso significa que o papel do Estado não estaria limitado a proibir ou permitir a entrada no mercado de uma nova tecnologia, mas caberia a Este criar processos e instrumentos capazes de mensurar e monitorar o desenvolvimento das inovações. A construção de um modelo regulatório apropriado teria o potencial de ampliar os ganhos sociais e reduzir impactos negativos já mapeados. Integrar a incerteza dentro do ciclo de desenvolvimento de uma política destinada às tecnologias emergentes requer o mapeamento de quais dados devem ser coletados, a criação de instrumentos de coleta adaptados a finalidade, o processamento e incorporação desses achados dentro do sistema de criação/adaptação da nova política.

A definição do quando o Estado deve intervir tem se apresentado como escolha entre: a criação de regras ainda no estado nascente da tecnologia, ou seja, quando o Estado pode ter um maior controle sobre seus efeitos, mas também ser um obstáculo ao desenvolvimento de soluções benéficas à sociedade, ou quando a tecnologia se tornar madura, momento esse em que se conhece os seus reais impactos, mas no qual o controle pela via da regulação estatal poderia ser menos efetivo. A definição do quando regular, se antes ou depois, poderia ser substituída, no caso das tecnologias emergentes, por uma atuação estatal no durante, na qual o conteúdo regulatório interfere e sofre interferência do processo de desenvolvimento da inovação.

Não obstante o reconhecimento da importância de uma regulação adaptada à tecnologia e aos interesses da sociedade no uso desta, entendemos que a qualidade do conteúdo regulatório deriva principalmente de um processo focado na adaptação fruto do aprendizado sobre a tecnologia emergente que se busca regular.

Essa mudança de abordagem foi uma das premissas para a OCDE recomendar a adoção de um modelo de governança regulatória ágil para melhor aproveitamento da inovação. Neste instrumento, a OCDE reconhece que "[d]iante dos desafios regulatórios colocados pela inovação, será essencial realizar uma mudança nos processos de política regulatória, na qual a mentalidade tradicional de 'regular e esquecer' deve dar lugar a

[8] SCHWAB, Klaus. **A quarta revolução industrial**...

abordagens de 'adaptar e aprender'".[9]

Dentre as recomendações aos Estados, a OCDE cita: a adaptação de ferramentas de gerenciamento regulatório para permitir conteúdos e processos que sejam adaptados ao futuro, o desenvolvimento de bases institucionais em favor da cooperação intra e entre jurisdições, promoção de estruturas de governança que permitam a adoção de modelo ágil e "à prova de futuro", a adaptação das estratégias de compliance e enforcement para auxiliar os inovadores e proteger a sociedade inclusive entre diferentes jurisdições.[10] A promoção de ciclos regulatórios "mais adaptativos, interativos e flexíveis", deverão ser baseados em evidências (análise de cenários, monitoramento e avaliação dos resultados) em um sistema de aprendizado e adaptação contínua e que incluam diferentes stakeholders desde da fase inicial do processo regulatório.

No mesmo sentido, o Fórum Econômico Mundial (WEF) defende uma abordagem ágil para regulação das tecnologias representativas da quarta revolução industrial.[11] A governança ágil além ser adaptada à velocidade das inovações, seria também caracterizada pela sua natureza inclusiva, centrada no ser humano e capaz de assegurar sustentabilidade em logo prazo.[12] Para concretizar tal modelo, o WEF propôs em uma publicação posterior, um conjunto de ferramentas regulatórias que poderiam ser adotadas pelos Estados em função da natureza da inovação, dos objetivos, do grau de conhecimento que se tem sobre a nova tecnologia e do grau de interdependência com outros atores.

As soluções visam auxiliar o Estado em diferentes momentos do desenvolvimento tecnológico e foram agrupadas em sete grupos: regulação antecipatória, regulação focada em resultados, regulação experimental, regulação guiada por dados, auto e corregulação, regulação conjunta e cooperação internacional em favor da interoperabilidade e da ação conjunta na gestão de riscos. Cada uma das abordagens foi acompanhada de uma análise sobre seus objetivos, limitações e casos práticos de aplicação.

Tanto na proposta da OCDE quanto do WEF, o Estado tem um papel relevante de coordenação dos diferentes atores envolvidos. Esta atuação pressupõe o investimento em uma estrutura dotada de certo grau de flexibilidade no processo de elaboração normativa a

[9] OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464).**

[10] OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464).**

[11] WEF - WORLD ECONOMIC FORUM. **Agile Regulation for the Fourth Industrial Revolution:** a Toolkit for Regulators. Geneva: World Economic Forum, 2020. Disponível em: https://www.weforum.org/about/agile-regulation-for-the-fourth-industrial-revolution-a-toolkit-for-regulators/. Acesso em: 03 jul. 2023.

[12] WEF - WORLD ECONOMIC FORUM. **Agile Governance:** Reimagining Policy-making in the Fourth Industrial Revolution.

fim de que esta reflita o equilíbrio dinâmico entre incentivo à inovação e segurança do consumidor e do meio ambiente. O conteúdo regulatório seria, portanto, fruto de um processo contínuo de aprendizagem e adaptação pautado na cooperação entre os atores e nos dados produzidos durante as fases de experimentação. O que podemos concluir neste tópico é que futuras regulações de IA contam com um arcabouço teórico e com ferramentas que podem ser adaptadas em favor da construção de sistemas que sejam éticos *by design* (diversos e inclusivos).

Em um cenário de incertezas, como é o caso da IA, ferramentas de testagens têm uma dupla função: viabilizam um espaço de aprendizagem tanto para o regulador como para as empresas, e estimulam a colaboração destes em favor da criação de regras adaptadas às peculiaridades dos setores, das técnicas de IA empregadas e de suas aplicações. Essa modalidade de política regulatória baseada em evidências pode desempenhar um papel relevante na mitigação dos potenciais impactos de sistemas de IA que tenham o potencial de gerar resultados discriminatórios em desfavor de grupos minorizados.

Nos tópicos seguintes, iremos aprofundar essa análise sobre o uso de instrumentos regulatórios que adotem uma abordagem mais focada nos processos e na cooperação entre os agentes em favor da construção de sistemas de IA éticos *by design*.

2. RISCOS SOCIAIS ASSOCIADOS AOS VIESES ALGORITMOS EMBARCADOS EM SISTEMAS DE IA

A construção de um arcabouço normativo institucional para os sistemas de IA nos moldes descritos no tópico anterior depende, inicialmente, da identificação de dois elementos. O primeiro destes seria um cenário de incerteza quanto aos resultados de tais inovações para sociedade, ou seja, que dada a natureza e estágio de desenvolvimento da tecnologia, ainda não existem dados suficientes para respaldar a criação de normas detalhadas de modo eficiente. O segundo, seria a necessidade de cooperação entre diferentes atores, Estado, empresas e sociedade civil para garantir efetividade das normas.

O primeiro campo de incerteza se observa na própria conceituação do que seriam os sistemas de Inteligência Artificial.[13] Ainda que reconheçamos a necessidade de sentidos comuns para fundamentar consensos internacionais sobre IA, optamos aqui por nos focar na incerteza resultante da aplicação da tecnologia. Por essa razão, partiremos da definição da OCDE, para qual a inteligência artificial é um sistema baseado em máquina que pode, para um

[13] RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence:** a modern approach. 2 ed. New Jersey: Pearson Education, 2010. p. 1-3. Nesta obra os autores identificaram 4 abordagens para a definição de IA: "systems that think like humans, systems that think rationally, system that act like humans and systems that act rationally."

determinado conjunto de objetivos definidos por humanos, fazer previsões, recomendações ou decisões que influenciam ambientes reais ou virtuais[14].

Neste artigo, iremos nos ater a uma técnica do campo da IA, o aprendizado de máquina. Esta técnica "refere-se a um processo automatizado de descoberta de correlações (às vezes chamadas de relações ou padrões) entre variáveis em um conjunto de dados, geralmente para fazer precisões ou estimativas de algum resultado."[15] O aprendizado de máquina se apresenta mais comumente segundo três modalidades: supervisionado, aprendizado não supervisionado e por reforço.[16] Cada uma dessas modalidades é formada por técnicas que dão origem a outros subcampos de estudo do processo de aprendizado por máquina. Dentre eles, tem se destacado o aprendizado profundo, técnica que utiliza de redes neurais com várias camadas que permitem processar uma elevada quantidade de dados e definir o peso de cada conexão para o resultado.[17]

Apesar do objetivo de o aprendizado por máquina ser acessível a compreensão dos juristas, o mesmo não pode ser dito sobre o processo de execução. Para auxiliar juristas no estudo do tema, David Lehr e Paul Ohm identificaram oito etapas do processo de aprendizagem por máquina: 1. definição do problema, 2. coleta dos dados, 3. filtragem da base 4. revisão das características estatísticas dos dados (variabilidade, distribuição, categorias) 5. particionamento de dados 6. seleção do modelo, 7. treinamento do modelo e 8. aplicação do sistema em produção. Os autores identificam que grande parte dos juristas concentram as análises sobre risco social da IA nas três primeiras etapas e destacam a importância que as demais podem vir a desempenhar para uma gestão mais eficiente dos impactos dessa tecnologia.[18]

A incerteza sobre os impactos éticos da IA pode ser organizada em duas categorias de vulnerabilidades: humanas e do processo de aprendizado de máquina. O primeiro tipo refere-se à possibilidade de tais programas influenciarem o processo de tomada de decisão de humanos. O segundo tipo congrega os riscos éticos advindos dos vieses existentes na coleta e processamento de dados que alimentarão os sistemas de IA e do grau de eficiência da

[14] OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464).**

[15] LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning. **UCDL Review**, v. 51, p. 653-717, dec. 2017. Disponível em: https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf. Acesso em: 10 jul. 2023.

[16] DE SPIEGELEIRE, Stephan; MASS, Matthijs; SWEIJS, TIM. **Artificial Intelligence and the future of defense**. Haia: Hague Center for Strategic Studies, 2017. Disponível; https://www.jstor.org/stable/resrep12564.7?seq=3#metadata_info_tab_contents. Acesso em: 10 jul. 2023.

[17] LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning.

[18] LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning.

capacidade de tais sistemas de terem suas decisões explicáveis.[19]

Na etapa de definição do problema, o viés está embarcado na própria pergunta, ou objetivo que o sistema visa alcançar.[20] Na fase de coleta, os vieses podem ser decorrentes da diversidade dos dados escolhidos para treinar os sistemas de IA, pois tais amostras podem não representar inteiramente o universo que se pretende analisar, ou os dados expressam os preconceitos/vieses da sociedade onde estes foram coletados.

Sobre os impactos dos vieses na eficácia de sistemas de IA aplicados aos instrumentos de reconhecimento facial, estudos realizados por Joy Buolamwini e Timnit Gebru, evidenciaram diferentes taxas de erro em razão de gênero e raça. O estudo demonstrou índices de erro bem menores para homens e para pessoas brancas do que para mulheres e pessoas negras. Para identificação de gênero, a variação de erro foi entre 8,1% a 20,6% e para diferença de tons de pele, este índice variou entre 11,8% a 19,2%. Em uma análise interseccional entre raça e gênero, mulheres negras apresentaram a maior variação de erro, 20,8 a 34,7%.[21] A partir dos resultados, as pesquisadoras concluíram pela necessidade do uso de base de dados de referência que sejam inclusivas e a apresentação de relatórios de acurácia por subgrupo em favor da transparência e do desenvolvimento de uma IA responsável. Entendeu-se, nesse caso, que a sub representatividade de mulheres e pessoas negras, apesar de não ser o único erro identificado neste processo, foi um dos elementos relevantes para explicar a variação das taxas de erros. Além desses pontos, as pesquisadoras também levantaram questionamentos quanto à utilização de um critério binário para classificação de gênero e suas repercussões para a acurácia do uso de sistemas de reconhecimento facial baseados em sistemas de IA e, por consequência, seus potenciais impactos discriminatórios.

Na etapa de limpeza dos dados, o objetivo é identificar dados ausentes ou imprecisos. Trata-se de uma etapa importante para garantir a qualidade da base que irá alimentar o modelo de inteligência artificial. A detecção dessas ausências e imprecisões bem como a aplicação das correções podem introduzir vieses, e por essa razão, tais decisões devem ser indicadas na documentação para permitir maior transparência e futuro controle do modelo.[22]

---

[19] LIAO, S. Matthew. **Ethics of Artificial Intelligence**. New York: Oxford University Press, 2000.

[20] HAO, Karen. This is how AI bias really happens - and why it's so hard to fix. **MIT Technology Review**, feb. 2019. Disponível em: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. Acesso em: 10 jul. 2023.

[21] BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: FRIEDLER, Sorelle A.; WILSON, Christo (ed.). **Proceedings of the 1st Conference on Fairness, Accountability and Transparency.** NY, USA: PMLR, 2018. v. 81, p. 77–91. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.Acesso em: 10 jul. 2023.

[22] HAO, Karen. This is how AI bias really happens - and why it's so hard to fix.

Na etapa de seleção de características, correlações não óbvias que tem natureza discriminatória podem ser revistas pelo cientista de dados. Trata-se de mais uma oportunidade para verificar se a amostra a ser usada é capaz de garantir uma representação estatística que resulte em níveis de acurácia semelhantes para os diferentes perfis.

Imaginemos o cenário de um processo seletivo. Por meio da identificação de padrões, um modelo é criado para selecionar os melhores currículos para entrevista. Como resultado, aponta-se uma determinada característica, jogar tênis, como algo de alta relevância para seleção do perfil de um bom profissional, ainda que esta habilidade não tenha relação direta com as atividades da função objeto da vaga a ser preenchida. Ao se analisar o perfil de raça de quem detém essa característica, se identifica que 90% da amostra seria composta de pessoas brancas. Trata-se nesse caso de um efeito de discriminação indireta.[23] Caberia, nesse caso, ao cientista responsável pelo modelo se atentar para esse fato e adotar providências para não gerar um modelo que tenha efeitos discriminatórios.[24]

Na etapa de particionamento dos dados, o programador divide base a ser utilizada em duas, uma para treinamento e outra para validação do modelo a fim de evitar "*overfitting*", (disparidade nas previsões) permitindo que o modelo possa ser objeto de generalização.[25]

A forma como o processo de desenvolvimento dos sistemas leva a divisão de uma mesma base para treinamento e outra para validação faz com que os vieses embutidos em uma estejam também presentes na outra.[26] Tal fato reforça a necessidade dos cuidados indispensáveis nas fases anteriores minimizar resultados que levem a discriminação indireta de grupos vulneráveis.

Nas etapas de escolha do modelo é importante que as decisões sobre quais

[23] Por discriminação indireta, partimos do conceito trazido Convenção interamericana contra o Racismo, e que foi adotado pela minuta de projeto substitutivo do senado para regulação da IA no Brasil. Nestes textos, discriminação indireta "ocorre quando normativa, prática ou critério aparentemente neutro tem a capacidade de acarretar desvantagem para pessoas pertencentes ao grupo específico, ou as coloquem 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" – Artigo 1, item 2 da Convenção Interamericana contra o Racismo, a Discriminação Racial e Formas Correlatas de Intolerância , ratificada pelo Brasil pelo Decreto 10.932 de 10 de janeiro de 2022. (OEA - ORGANIZAÇÃO DOS ESTADOS AMERICANOS. **Convenção Interamericana contra Toda Forma de Discriminação e Intolerância.** Aprovada na segunda sessão plenária da Assembleia Geral, realizada em 5 de junho de 2013. Antígua, Guatemala: OEA, 2013. Disponível em: https://www.oas.org/pt/cidh/mandato/basicos/discriminacioneintolerancia.pdf. Acesso em: Acesso em: 10 jul. 2023).

[24] Na etapa de preparação dos dados, o foco está na eleição dos atributos que serão escolhidos para treinar o modelo.

[25] Sobre esse tema na generalização, o viés de avaliação deve ser um elemento a ser considerado. Ver mais sobre: CAPPRA INSTITUTE FOR DATA SCIENCE. **Ética na IA:** como desenvolver um sistema ético para Inteligência Artificial. Miami: Cappra Institute, 2021. disponível em: https://www.cappra.institute/s/Etica_IA_PT.pdf. Acesso em: 10 jul. 2023.

[26] HAO, Karen. This is how AI bias really happens - and why it's so hard to fix.

ferramentas e/ou métodos serão utilizadas possam ser justificadas.[27] Em determinadas situações para as quais o uso de sistemas de IA tragam questões de natureza sensível, como por exemplo a decisão sobre oferta de crédito, a indicação de determinadas ferramentas que permitam um maior grau de explicabilidade do processo decisório pode ser um elemento a ser ponderado no momento da definição de regras para o setor de crédito.

Nas etapas de seleção do modelo, diferentes algoritmos são testados de modo a se identificar quais parâmetros serão aplicados à base selecionada para extração dos resultados. A escolha deste modelo envolve a busca por maior acurácia que pode impactar no grau de explicabilidade dos resultados do modelo posteriormente.

Em seguida temos o processo de treinamento do modelo. Este inclui uma série de ciclos de avaliações, ajustes e seleções de características. O responsável pelo modelo analisará a performance, o grau de acurácia, estabilidade e tomará decisões de como aperfeiçoar o algoritmo para que este alcance seu objetivo.[28] Quando as bases de dados refletem os vieses da sociedade contra grupos minorizados, a fase de treinamento do algoritmo pode ser utilizada para minimizar tais vieses, evitando assim resultados injustos e que reforçariam e potencializariam os preconceitos já presentes na sociedade. Isso porque os resultados preconceituosos gerados pelo sistema de IA teriam uma "roupagem científica", baseados em uma pseudoneutralidade da técnica, que, por sua vez, fundamentaria erroneamente a validade de ações discriminatórias. A introdução de ferramentas antidiscriminatórias contra grupos minorizados pode ser um campo de estudo e investimento fomentado a partir de um desenho regulatório adaptado à tecnologia e ao setor no qual ela será aplicada.

Para inclusão de uma ideia de justiça nessa etapa de desenvolvimento do sistema, Selbst et al propõem uma mudança do processo que eles denominam de "algorithmic frame" para "data frame". Os autores defendem que o primeiro processo visa "produzir um modelo que melhor capture a relação entre representações e rótulos", mas tal abordagem seria insuficiente para investigar a ideia de aprendizado de máquina justo. Para tanto seria necessário ampliarmos a análise para além do algoritmo, incluindo as entradas e saídas do algoritmo."[29]

Karen Hao reconhece também a dificuldade da inclusão de "justiça" na construção de sistemas de IA e lista três razões pelas quais os vieses nesses casos são difíceis de serem consertados: a dificuldade de identificar os vieses decorrentes de correlações não usuais, a

[27] LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning.

[28] LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning.

[29] SELBST, Andrew D. Et Al. Fairness and Abstraction in Sociotechnical Systems. *In:* **FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency**. New York: Association for Computing Machinery, 2019. p. 59–68. Disponível em: https://dl.acm.org/doi/10.1145/3287560.3287598. Acesso em: 10 jul. 2023. p.3

imperfeição do processo e a falta de contextos sociais para compreensão dos dados, as múltiplas definições de justiça e sua tradução em um modelo algorítmico. Os vieses não seriam tão fáceis de serem identificados no início do desenvolvimento do programa. Em razão da capacidade de relacionar dados do sistema, correlações que possam expressar vieses de gênero, raça ou outros podem advir de um padrão que ainda não é óbvio para os desenvolvedores do modelo.[30]

Um sistema que foi desenvolvido inicialmente para um objetivo em um determinado contexto ao ser utilizado para outras aplicações pode produzir resultados enviesados pois o contexto social pode apresentar elementos distintos. Selbst et al identificam cinco armadilhas que estão relacionadas a inclusão de "justiça" nos sistemas de aprendizado por máquina: enquadramento, portabilidade, formalismo, efeito em cascata e "solucionismo". Para os autores, tais armadilhas estão relacionadas à lógica e objetivos que fundamentam atualmente o desenvolvimento de sistemas de IA que não incluem a ideia de justiça (fairness), principalmente àquelas relacionadas às falhas de compreensão entre dois sistemas: sociais e técnicos. Na base da solução dessas armadilhas estaria a mudança de abordagem no desenvolvimento dos sistemas: de uma abordagem orientada para solução para uma orientada por um processo apto a incluir atores sociais, instituições e suas interações.[31]

3. CENÁRIO PARA A REGULAÇÃO DOS SISTEMAS DE IA

A construção de diretrizes em favor de um modelo regulatório adaptado às tecnologias da quarta revolução industrial vem sendo debatida em diferentes instituições internacionais. Dentre as características desse futuro modelo, se apontam a necessidade de um modelo ágil, fundado nas ideias de uma regulação responsiva, conduzido a partir de evidências, instrumentos de experimentação, e desenvolvido de maneira colaborativa entre os stakeholders sobre a liderança estatal.[32]

[30] HAO, Karen. This is how AI bias really happens - and why it's so hard to fix.

[31] SELBST, Andrew D. Et Al. Fairness and Abstraction in Sociotechnical Systems.

[32] A União Europeia adotou "better regulation agenda" baseada em três objetivos: políticas baseadas em evidências, tornar normas mais simples e melhores, o envolvimento dos cidadãos, empresas e diferentes stakeholders. EUROPEAN COMMISSION. **Better regulation.** Brussels: European Commission, 2023. Disponível em: https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en. Acesso em: 10 jul. 2023; OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464);** UNCTAD **-** UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. **Technology and Innovation Report 2021:** Catching technological waves – Innovation with equity; WEF - WORLD ECONOMIC FORUM. **Agile Regulation for the Fourth Industrial Revolution:** a Toolkit for Regulators...

No caso específico dos sistemas de IA, tem se observado dois movimentos normativos distintos: (1) criação/adoção de princípios para guiar o uso da IA por instituições internacionais Estados e empresas, e (2) a publicação de estratégias para o incentivo ao desenvolvimento tecnológico nessa área, criadas principalmente no nível nacional e regional. A primeira categoria estabelece princípios e objetivos éticos para guiar o desenvolvimento da tecnologia e a segunda recomenda diretrizes para adoção de políticas de incentivo ao setor tecnológico e a competitividade do Estado nesse campo.

O mapeamento desse arcabouço normativo tem sido objeto de diferentes estudos. Os números variam de acordo com a metodologia utilizada para definição de amostra e indicam como essa discussão tem sido pulverizada entre diferentes organizações internacionais, grupos técnicos e empresas de tecnologia.

Em estudo elaborado pela Algorithm Watch foram inventariados 167 *Guidelines* voltados à ética e IA.[33] Ao se utilizar como categoria instrumentos de *soft law*, esse número aumenta significativamente. Em estudo recente, a OCDE identificou 634 documentos dessa natureza tendo a IA como objeto.[34] Destes, 94% foram publicados no intervalo de 2015 a 2019 e 35,8% tiveram origem governamental, 18% foram produzidos por empresas com foco em parâmetros internos para o desenvolvimento de sistemas de IA.[35] Tais instrumentos foram elaborados majoritariamente em países desenvolvidos, nos quais 54% nos EUA, Organizações internacionais ou países europeus. No referido estudo, esses textos foram classificados a partir de sete categorias: recomendações/estratégias (54,6%), princípios (24,92%), standards (9,46%), guia de condutas para profissionais ou códigos de conduta (3,63%), parcerias (3,31%), certificações ou programas voluntários (2,52%), moratória voluntária ou banimento (1,89%)[36].

Essa quantidade de documentos leva a pergunta sobre possíveis consensos, se teríamos em construção uma base principiológica comum que pudesse servir de referência para futuras regulações nacionais. Em busca dessa convergência, Berkman klein Center desenvolveu uma pesquisa de profundidade a partir de uma amostra de trinta e seis documentos sobre princípios para IA. Desta análise comparativa, foram identificados quarenta e sete princípios, categorizados em oito temas: privacidade, prestação de contas,

[33] ALGORITHMWATCH. **AI Ethics Guidelines Global Inventory.** Berlin: AlgorithmWatch, 2019. Disponível em: https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/. Acesso em: 10 jul. 2023.

[34] GUTIERREZ, Carlos Ignacio; MARCHANT, Gary. **Soft law 2.0:** Incorporating incentives and implementation mechanisms into the governance of artificial intelligence. OECD.AI, 2019. Disponível em: https://oecd.ai/en/wonk/soft-law-2-0. Acesso em: 10 jul. 2023.

[35] Exemplos: PICHAI, Sundar. **AI at Google:** our principles. Google, jun. 2018. Disponível em: https://www.blog.google/technology/ai/ai-principles/. Acesso em: 10 jul. 2023; DOBRIN, Seth; MONTGOMERY, Christina. **Principles and Practices for Building More Trustworthy AI.** IBM Newsroom, 2021. Disponível em: https://newsroom.ibm.com/Principles-and-Practices-for-Building-More-Trustworthy-AI. Acesso em: 10 jul. 2023.

[36] GUTIERREZ, Carlos Ignacio; MARCHANT, Gary. **Soft law 2.0:** Incorporating incentives and implementation mechanisms into the governance of artificial intelligence.

proteção e segurança, transparência e explicabilidade, justiça e não discriminação, controle humano da tecnologia, responsabilidade profissional e promoção de valores humanos. [37]

Em razão do objetivo deste artigo, iremos concentrar nossa análise dos achados da pesquisa relacionados à categoria de princípios da justiça e não discriminação. No estudo, essa categoria ainda é subdividida em outras seis, que foram assim representadas: 89% dos documentos analisados citavam não discriminação e prevenção de vieses, 36% se referiam a representatividade e elevada qualidade dos dados, 56% citam justiça, 25% equidade, 42% inclusão no que concerne ao impacto da tecnologia e 47% se referem a um design inclusivo. Apesar de nem todos os documentos utilizarem de uma linguagem que permita a classificação acima, se concluiu que 100% destes abordam de alguma forma o princípio da justiça e da não discriminação em seus conteúdos.[38]

Os pesquisadores concluíram que os documentos se concentram mais na etapa de construção da base de dados e no receio de que preconceitos historicamente enraizados na sociedade possam ser reproduzidos e escalados com o uso de sistemas de IA. A ideia de qualidade dos dados aparece relacionada a acurácia, consistência e validade. Apesar das especificidades encontradas quando se define justiça, a pesquisa alcança um conceito comum entre os documentos analisados, segundo a qual, justiça seria o "tratamento equitativo e imparcial dos titulares de dados" [39]. O princípio da equidade expressaria que "pessoas, em situações semelhantes ou não, merecem as mesmas oportunidades e proteções"[40]. Na categoria design inclusivo, o princípio foi representado por duas abordagens: necessidade de diversidade na composição das equipes de desenvolvedores e em um processo inclusivo para a definição dos objetivos da IA.

Diretrizes antidiscriminatórias aparecem de modo transversal nas demais categorias de princípios. Intersecções foram identificadas nos princípios: da privacidade, no que se refere ao direito à retificação, (privacidade); da prestação de contas, no que tange à verificabilidade e replicabilidade; da proteção e segurança, no tocante à previsibilidade; da transparência em questões envolvendo a explicabilidade das decisões de sistemas de IA; do controle humano da tecnologia ao indicar a necessidade da IA ser focada no benefício da sociedade.

---

[37] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Cambridge, MA: Berkman Klein Center Research Publication, 2020. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482. Acesso em: 10 jul. 2023.

[38] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

[39] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

[40] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

Dentre os objetivos comuns destes instrumentos, está o desenvolvimento de um quadro normativo baseado em princípios em favor de uma IA que seja confiável, centrada nas necessidades humanas e em valores de justiça, não discriminação e equidade, capaz de garantir transparência e explicabilidade. Como tais comandos serão traduzidos em regras, processos e metodologias de desenvolvimento de sistemas de IA ainda resta a ser determinado.

Floridi identifica que esse período principiológico, da adoção de códigos de condutas de natureza voluntária sobre a governança da IA estaria terminando frente ao desenvolvimento de marcos regulatórios nacionais, baseados em regras e instrumentos próprios de monitoramento.[41] Podemos identificar duas razões para tanto: a necessidade de soluções específicas para a gestão do risco da IA para garantir maior segurança para investidores e usuários e, o baixo grau de efetividade da responsabilização dos agentes pela via do *soft law*.

Observa-se, portanto, um espaço a ser preenchido pelo legislador local. Dentre os temas se destaca para os fins desse artigo: os requisitos e instrumentos em favor do mapeamento dos riscos, o conteúdo do relatório de impacto, a obrigatoriedade de identificação dos potenciais usos indevidos, os procedimentos mínimos para gestão de dados, parâmetros que permitissem a transparência e prestação de contas relacionadas às decisões tomadas por sistemas de IA.

Ainda não há consensos internacionais sobre a natureza e quais os instrumentos devam compor a governança dos sistemas de IA nem qual o modelo de regulação desta tecnologia. Enquanto os EUA adotam uma modelo principiológico geral e delega às agências reguladoras o detalhamento de obrigações nos seus respectivos setores, a UE tem adotado uma abordagem híbrida, contendo obrigações substantivas e/ou procedimentais a depender do grau de risco que for estabelecido para uma determinada aplicação de IA[42].

O Brasil é um dos países que têm buscado desenvolver um marco legal sobre IA. Em 2021, foi publicada a Estratégia Brasileira de IA[43], e, atualmente, tramitam no congresso os PLs 21/20, 5051/19 e 872/21 nessa temática. Os PLs são recentes e demandam a reflexão dos legisladores e a contribuição da sociedade civil para seu aperfeiçoamento. No momento,

[41] FLORIDI, Luciano. Establishing the rules for building trustworthy AI, 2019. **Nature Machine Intelligence**, v. 1, p. 261-262, Jun. 2019. Disponível em: https://philarchive.org/rec/FLOETR. Acesso em: 10 jul. 2023.

[42] EUROPEAN COMMISSION. **Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.** COM(2021) 206 final. Brussels, 21 Apr. 2021. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Acesso em: 10 jul. 2023.

[43] BRASIL. Ministério da Ciência, Tecnologia e Inovação. **Estratégia Brasileira de Inteligência Artificial**. Portaria MCTIn º 4.617 de 6 de abril de 2021 alterada pela portaria MCTI nº 4.979 de 13 de julho de 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial. Acesso em: 10 jul. 2023.

os PLs seguem uma linha principiológica e ainda não avançaram na construção de obrigações substantivas e/ou procedimentais relacionadas aos sistemas de IA.

Deste tópico, podemos concluir que já há um arcabouço normativo de natureza principiológico que inclui a não discriminação como um elemento relevante para futuras regulações de sistemas de IA. No tópico a seguir, iremos analisar o papel de um instrumento em específico que poderia auxiliar na gestão dos riscos relacionados aos efeitos discriminatórios dos sistemas de IA.

4. O INSTRUMENTO DE IMPACTO DE SISTEMAS DE I.A. COMO UM DOS ELEMENTOS RELEVANTES PARA CRIAÇÃO DE SISTEMAS DE I.A. ÉTICOS BY DESIGN

O desenvolvimento de sistemas de IA centrados nas necessidades e valores humanos tem sido um ponto de consenso entre os instrumentos internacionais de *soft law* publicados pela OCDE[44] e Unesco e na proposta de regulação da IA pela UE. Os valores da diversidade e objetivos de natureza antidiscriminatória estão presentes em grande parte dos instrumentos normativos analisados pelo estudo já citado do Berkman Center.[45] Por reconhecermos que a nossa sociedade atual é atravessada e fundada em preconceitos e desigualdades, estamos cientes dos riscos advindos do uso dos dados coletados neste ambiente para alimentar sistemas de IA. Contudo, ainda não temos a exata noção dos impactos discriminatórios resultantes da correlação de dados, que isoladamente não teriam esse potencial. Diante dessa zona de incerteza, a adoção de instrumentos de regulação que permitam a coleta e monitoramento dos efeitos da aplicação da nova tecnologia pode ser uma solução visando equilibrar o incentivo à inovação e a proteção da sociedade.

Uma vez que alcançamos minimamente consensos sobre os princípios que deveriam guiar o desenvolvimento e aplicação dessa tecnologia, passamos para a fase de operacionalização de ferramentas que concretizem tais princípios. Neste tópico, iremos investigar se um instrumento já conhecido, relatório de impacto, poderia ser útil e minimamente eficaz para gestão de risco discriminatório de sistemas de IA desde seu design.

A obrigação da elaboração de estudos prévios sobre os potenciais impactos de novo projeto/tecnologia na sociedade é um dever já conhecido pelo sistema jurídico brasileiro. Inicialmente utilizado como uma solução normativa para gestão de riscos e

[44] OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).**

[45] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

impactos no meio ambiente,[46] essa solução expressa o resultado de uma construção normativa internacional e nacional que consagrou os princípios da precaução e da prevenção como resposta jurídica para impactos negativos da atuação dos seres humanos no meio ambiente.

Apesar do relatório de impacto ser uma solução já conhecida na área ambiental, esta é ainda pouco utilizada para gestão de impactos sociais das tecnologias. Um exemplo do uso dessa ferramenta, está consubstanciado na Lei Geral de Proteção de Dados (LGPD). Esta lei introduziu dois instrumentos para viabilizar a coleta, identificação, avaliação e monitoramento de riscos associados aos dados pessoais: a avaliação do legítimo interesse e o relatório de impacto à proteção de dados pessoais (RIPDP).[47]

O primeiro instrumento funciona como um teste para o equilíbrio mínimo entre os interesses dos titulares de dados e daqueles que farão uso destes. O RIPDP é definido pela LGPD como "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, bem como medidas, salvaguardas e mecanismos de mitigação de risco."[48]

Tal tema ainda será objeto de regulamentação da ANDP, segundo sua agenda regulatória. É esperado que esse futuro instrumento venha minimizar as dúvidas concernentes à sua obrigatoriedade, indique qual abordagem será adotada, quais elementos integrarão o Relatório e como poderá ser conduzido o monitoramento dos riscos identificados no instrumento.

Ainda que não tenha sido publicado tal regulamento[49], esse exemplo é de fundamental importância para se pensar na gestão de risco envolvendo o uso de sistemas de

---

[46] Art. 8º, II, BRASIL. **Lei nº 6.938, de 31 de agosto de 1981.** Dispõe sobre a Política Nacional do Meio Ambiente, seus fins e mecanismos de formulação e aplicação, e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, 2 set. 1981. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l6938.htm. Acesso em: 10 jul. 2023.; e, Art. 9º, CONSELHO NACIONAL DO MEIO AMBIENTE (CONAMA). **Resolução CONAMA nº 001, de 23 de janeiro de 1986.** Dispõe sobre critérios básicos e diretrizes gerais para a avaliação de impacto ambiental. Diário Oficial da União: seção 1, Brasília, DF, 17 fev. 1986. Disponível em: https://www.ibama.gov.br/sophia/cnia/legislacao/MMA/RE0001-230186.PDF. Acesso em: 10 jul. 2023, elevado a matéria constitucional pelo art. 225, IV.

[47] Art. 105, XVII. BRASIL. **Lei nº 13.709, de 14 de agosto de 2018.** Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Diário Oficial da União: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 10 jul. 2023.

[48] Art. 105, XVII. BRASIL. **Lei nº 13.709, de 14 de agosto de 2018.**

[49] A ANPD procedeu a consultas para ouvir a sociedade civil sobre o Relatório de Impacto e espera-se que dessa colaboração possa ser desenvolvida uma futura regulamentação. ANPD - AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS. **ANPD divulga cronograma completo de reuniões técnicas sobre relatório de impacto à proteção dos dados pessoais.** Brasília, DF: ANPD, 17 jun. 2021. Disponível em: https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-divulga-cronograma-completo-de-reunioes-tecnicas-sobre-relatorio-de-impacto-a-protecao-dos-dados-pessoais. Acesso em: 10 jul. 2023.

IA, pois, uma parte relevante do funcionamento de tais sistemas está na forma como os dados serão coletados e tratados nos ciclos de criação da IA.

O relatório de impacto é uma das ferramentas que viabiliza a adoção de uma regulação baseada no design, vez que inclui o mapeamento dos riscos e ferramentas para sua gestão desde uma fase inicial do projeto. Trata-se de um instrumento fundamental para prevenir futuros litígios, pois, torna-se possível a implantação de medidas voltadas a minimizar ou erradicar riscos existentes, evitando assim que haja conflito entre as partes envolvidas.

A regulação baseada no design visa promover resultados sociais considerados como desejáveis pela sociedade.[50] Essa abordagem pressupõe que as instituições competentes possam indicar quais são os objetivos sociais que se pretende atingir com determinada inovação, quais os valores a serem contemplados e quais são os usos considerados não desejados pela sociedade. Nesse sentido, a OCDE identifica o relatório de avaliação de risco como um instrumento de concretização da lógica da regulação fundada no design."[51]

Nos exemplos internacionais sobre regulação dos sistemas de IA estudados pelo projeto do Berkman Center, o relatório de impacto aparece como elemento integrante do princípio da prestação de contas. Este princípio foi citado em 97% dos textos dos textos analisados, e, em mais de metade (53%), os relatórios de avaliação de impacto são citados.[52] Apesar de ser citado, pouco se explora em tais documentos, se o relatório de impacto deverá ser produzido por uma organização externa ou pelo próprio desenvolvedor da IA, se o foco deve ser nos riscos ou nos danos, e se seriam adotados diferentes modelos de responsabilização em razão do risco verificado. Tais fatos revelam que o detalhamento sobre o conteúdo, sua obrigatoriedade e autoria ficariam sob a responsabilidade da regulação local.

Na Recomendação da OCDE sobre IA, o relatório de avaliação de impacto não aparece de maneira expressa, mas seus objetivos são contemplados na descrição dos princípios da segurança, proteção e robustez."[53]

Na Recomendação da UNESCO, também se identifica este cuidado. Ao descrever

---

[50] YEUNG, Karen. 'Hypernudge': Big Data as a mode of regulation by design. **Information Communication and Society**, **v. 20**, **p. 118-136**, **May. 2016.** Disponível em: https://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1186713. Acesso em: 03 jul. 2023. p.120.

[51] OECD - ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD Regulatory Policy Outlook 2021.** Paris: OECD Publishing, 2021. Disponível em: https://www.oecd.org/en/publications/oecd-regulatory-policy-outlook-2021_38b0fdb1-en.html. Acesso em: 30 maio 2025.

[52] FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

[53] OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).**

o sentido do princípio da proporcionalidade e de não causar danos, o texto evoca a implementação de procedimentos para avaliação de risco e a adoção de medidas preventivas que minimizem os impactos dos sistemas de IA aos direitos humanos. O texto avança em relação a outros instrumentos pois recomenda quais elementos podem compor futuras políticas nacionais para avaliação de impacto ético, dentre elas: medidas para gestão do risco, a indicação dos benefícios do sistema em análise para a sociedade e dos "efeitos sociológicos e psicológicos das recomendações baseadas em IA na autonomia de decisão dos seres humanos" , coleta de dados relativos ao do processo de acompanhamento e monitoramento de todo o ciclo de vida do sistema de IA, adoção de mecanismos de supervisão que permitam a auditabilidade, rastreabilidade e explicabilidade. A revisão externa dos sistemas também é expressamente citada como parte integrante do marco legal que é recomendado aos Estados para adoção dos relatórios de impacto ético. Por fim, a UNESCO também sugere que tais avaliações devam ser transparentes, abertas ao público quando possível e que tenham natureza multidisciplinar, multiparceira, multicultural, pluralista e inclusiva[54].

No Brasil, o Projeto de lei 21/20 é uma das propostas para regulação da IA em tramitação no congresso nacional. O relatório de impacto de inteligência artificial é definido como

> documentação dos agentes de inteligência artificial que contém a descrição do ciclo de vida do sistema de inteligência artificial, bem como medidas, salvaguardas e mecanismos de gerenciamento e mitigação dos riscos relacionados a cada fase do sistema, incluindo segurança e privacidade. [55]

A definição deste relatório, contudo, não é acompanhada de regras detalhadas sobre sua composição, requisitos setoriais específicos, nem como as informações ali apresentadas serão processadas e monitoradas de maneira clara, não há também critérios bem definidos sobre a prestação de contas e transparência desses sistemas.

Em 2022, uma comissão de juristas foi constituída com o objetivo de auxiliar o Senado Federal na redação de um projeto substitutivo às três propostas que tramitam no congresso – os PLs 5.051/19, 21/20 e 872/21. Em dezembro do mesmo ano, foi publicado um relatório com as propostas do grupo.

---

[54] UNESCO - ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_eng. Acesso em: 10 jul. 2023.

[55] BRASIL. **Projeto de Lei nº 21, de 2020.** Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências. Autor: Deputado Eduardo Bismarck (PDT/CE). Apresentado em 4 fev. 2020. Arquivado em 10 dez. 2024. Disponível em:https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340. Acesso em: 10 jul. 2023.

Na proposta de substitutivo, a não discriminação aparece como fundamento[56] para o desenvolvimento e aplicação de sistemas de IA, como princípio[57] e como direito[58]. O texto adota conceitos de discriminação direta e indireta da Convenção Interamericana contra o racismo, reconhecendo a possibilidade de efeitos discriminatórios não evidentes no processo de desenvolvimento do sistema de IA. Destacamos aqui duas modalidades de medidas antidiscriminatórios presentes na proposta: governança preventiva e rol de direitos dos consumidores, como o direito à correção de vieses discriminatórios, diretos, indiretos ilegais e abusivos.

O projeto adota princípios da "regulação ágil", em consonância com o que vem sendo prescrito pela OCDE e Fórum Econômico Mundial sobre o tema. Como exemplos, citamos a inclusão de regras sobre espaços de experimentação regulatória, a conciliação de instrumentos de diferentes naturezas para induzir o cumprimento dos seus dispositivos para além das sanções, o uso de ferramentas de monitoramento e gestão de risco durante todo o ciclo de vida da IA, mecanismos de coleta de dados. A depender da forma como tais regras forem regulamentadas, seria possível minimizar a assimetria de informação entre os atores e promover atualizações do arcabouço normativo em resposta aos novos riscos ou impactos não previstos na publicação da lei.

A governança preventiva proposta na minuta é fundada na regulação baseada em risco. Para os riscos considerados excessivos, o art. 14 apresenta um rol de vedações para implementação e uso de sistemas de IA deixando para a autoridade a ser designada regulamentar essa modalidade de risco futuramente. A definição de alto risco é determinada segundo as finalidades dos sistemas descritas em quatorze incisos do art. 17. A lista do que se considera tanto como risco excessivo como alto risco pode ser atualizada pela autoridade, desde que esta siga critérios mínimos definidos no art. 18.

Em particular, o papel do relatório de riscos e impactos de sistemas de IA como ferramenta indutora da adoção de um design antidiscriminatório pode ser observado nessa proposta em dois momentos: na avaliação preliminar e na avaliação de impacto algorítmico. O conteúdo, grau de monitoramento e obrigações decorrentes diferem entre as modalidades de avaliação.

A modalidade preliminar deve ser realizada para todos os sistemas de IA pelo

[56] BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.** Brasília, DF: Senado Federal, 2022. Disponível em: https://www.stj.jus.br/sites/portalp/SiteAssets/documentos/noticias/Relato%CC%81rio%20final%20CJSUBIA.pdf. Acesso em: 10 jul. 2023.

[57] Art. 3º, IV. BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.**

[58] Art. 5º,V. BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.**

fornecedor antes de sua oferta ao mercado com o objetivo de identificar o grau de risco do sistema. Esse documento será objeto de registro e poderá ser utilizado como referência para responsabilização e prestação de contas. A minuta não inclui quais os critérios que devem guiar a elaboração da avaliação preliminar, contudo obriga os agentes de IA a adotarem estruturas de governança que garantam a segurança, os direitos e o atendimento das pessoas afetadas pelos sistemas criados. Dentre os elementos que devem conter essa estrutura, a minuta cita "medidas de gestão de dados adequadas para a mitigação e prevenção de potenciais vieses discriminatórios;" e a necessidade das medidas de governança serem aplicáveis "ao longo de todo o ciclo de vida, desde a concepção inicial até o encerramento das atividades e descontinuação do sistema de IA".[59]

Por essa proposta, a classificação de risco é feita pelo próprio fornecedor, podendo essa ser alterada pela avaliação da autoridade competente. Caso o sistema seja classificado como de alto risco, se torna obrigatória a realização da avaliação de impacto algorítmico. Essa avaliação deverá ser resultado de um "processo iterativo contínuo, executado ao longo de todo o ciclo de vida dos sistemas de inteligência artificial de alto risco, requeridas atualizações periódicas"[60]. A referida minuta indica quatro etapas mínimas da avaliação de impacto algorítmico: preparação, cognição dos riscos, mitigação dos riscos encontrados e monitoramento. Quanto ao conteúdo, o fornecedor deve indicar obrigatoriamente qual a finalidade do sistema e quais medidas de mitigação de risco foram aplicadas. A descrição da participação de diferentes setores afetados, deve integrar o relatório apenas nos casos requeridos especificamente pela autoridade.

Da breve análise do conteúdo da minuta de projeto substitutivo para regulação da IA no Senado, observamos a preocupação de seus autores com possíveis riscos discriminatórios da aplicação dessa tecnologia. O relatório de impacto foi, nesse caso, uma das ferramentas propostas para a gestão de tais riscos. Neste modelo, o relatório preliminar foi desenhado para permitir ao próprio desenvolvedor indicar em qual nível de risco o sistema se enquadra a partir de uma lista de finalidades trazidas pelo próprio normativo. O relatório de impacto algorítmico seria reservado apenas para tecnologias autoavaliadas como de alto-risco ou que posteriormente fossem avaliadas pela autoridade como tal.

O potencial destes relatórios induzirem o desenvolvimento de um sistema de IA "antidiscriminatório" desde sua concepção ainda resta a ser determinado. Alguns elementos da minuta auxiliam nesse propósito, como por exemplo, a adoção da não discriminação como princípio/fundamento da futura lei, o reconhecimento do direito da correção de vieses às

[59] BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.**

[60] Art. 25. BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.**

pessoas afetadas por tais sistemas e da vedação do uso de sistemas que explorem vulnerabilidades de grupos específicos. Para a concreta execução de tais normas, se presume que o desenvolvedor realize o mapeamento de riscos e adote medidas para sua mitigação, sob pena de ser responsabilizado futuramente pelos conflitos emergentes e consequentes danos que vier a causar que poderiam ser evitados por uma avaliação de risco atenta.

O ponto chave para que a avaliação de risco se torne um instrumento capaz de influenciar o design do sistema de IA antidiscriminatório está no engajamento dos atores. A qualidade com a qual a avaliação preliminar será executada pelo fornecedor depende do grau de compromisso dos fornecedores de IA em construir e adotar um sistema de governança adaptado aos riscos e impactos que sua tecnologia pode causar. Como esse espaço de autorregulação será construído dependerá de mudanças tanto nos processos como tais sistemas são desenvolvidos, como no papel que os instrumentos de compliance das empresas irão desempenhar. A velocidade e adesão às mudanças depende também da qualidade do monitoramento a ser realizado pela autoridade competente e da capacidade da sociedade acessar e compreender as informações sobre o sistema de IA disponibilizadas publicamente. Acreditamos que nessas condições, o relatório de riscos e impactos tem o potencial de auxiliar a introjeção dos valores da diversidade e inclusão no processo de desenvolvimento de sistemas de IA, o que contribui decisivamente para tenhamos a prevenção de potenciais conflitos.

CONSIDERAÇÕES FINAIS

Diante do elevado grau de complexidade das mudanças trazidas pelas tecnologias da quarta revolução industrial, se torna indispensável o desenvolvimento de mecanismos em favor da cooperação entre diferentes atores por meio de espaços de experimentação e de instrumentos que dosem a responsabilidade dos agentes inovadores em relação ao nível de incerteza e ao potencial de dano e possíveis benefícios. Neste artigo se analisou a possibilidade do relatório de impacto ser um instrumento a ser incorporado nas legislações locais para induzir o desenvolvimento de sistemas de IA antidiscriminatórios desde a sua concepção. Para tanto, foram analisados se os fundamentos para a construção de um modelo regulatório adaptado a tais desafios encontrariam respaldo na literatura. As teorias de regulação responsiva e suas derivações justificam objetivos, instrumentos, atores e estratégias que podem servir de guia para as autoridades competentes construírem modelos adaptados às tecnologias emergentes, como é o caso da IA. Pelas fontes que conseguimos analisar neste artigo, o relatório de impacto teria o potencial de se tornar um destes modelos de regulação responsiva.

Em seguida, o arcabouço normativo internacional sobre o tema foi analisado para

verificar a presença dos valores da diversidade e da não discriminação nesses documentos e se, sendo este o caso, se o grau de relevância destes justificaria a adoção de ferramentas capazes de incluir tais objetivos de maneira concreta no próprio design na inovação. Constatamos que apesar do elevado número de documentos, estão presentes valores e objetivos da diversidade e inclusão entre os dispositivos dos documentos mais relevantes.

Por último, buscamos analisar como tal discussão está sendo construída no Brasil e se o relatório de risco e impacto para sistemas de IA integram tal discussão e que papel tem sido designado a esse instrumento. Concluímos que os projetos de lei em tramitação no congresso, em especial o PL 21/20 e o substitutivo elaborado pela comissão de juristas do Senado trazem esse instrumento como parte da estratégia de governança dos sistemas de IA. Entretanto, em razão do atual estágio de discussão da matéria, neste momento, os dados coletados permitem apenas apontar elementos que poderiam favorecer o uso do relatório como indutor de um design antidiscriminatório dos sistemas de IA e identificar alguns requisitos que este relatório deve conter para aprimorar a sua eficácia.

REFERÊNCIAS DAS FONTES CITADAS

ALGORITHMWATCH. **AI Ethics Guidelines Global Inventory.** Berlin: AlgorithmWatch, 2019. Disponível em: https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/. Acesso em: 10 jul. 2023.

ANPD - AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS. **ANPD divulga cronograma completo de reuniões técnicas sobre relatório de impacto à proteção dos dados pessoais.** Brasília, DF: ANPD, 17 jun. 2021. Disponível em: https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-divulga-cronograma-completo-de-reunioes-tecnicas-sobre-relatorio-de-impacto-a-protecao-dos-dados-pessoais. Acesso em: 10 jul. 2023.

BRASIL. **Lei nº 6.938, de 31 de agosto de 1981.** Dispõe sobre a Política Nacional do Meio Ambiente, seus fins e mecanismos de formulação e aplicação, e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, 2 set. 1981. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l6938.htm. Acesso em: 10 jul. 2023.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018.** Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Diário Oficial da União: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 10 jul. 2023.

BRASIL. **Projeto de Lei nº 21, de 2020.** Estabelece fundamentos, princípios e diretrizes para o desenvolvimento e a aplicação da inteligência artificial no Brasil; e dá outras providências. Autor: Deputado Eduardo Bismarck (PDT/CE). Apresentado em 4 fev. 2020. Arquivado em 10 dez. 2024. Disponível em:

https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340. Acesso em: 10 jul. 2023.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. **Estratégia Brasileira de Inteligência Artificial**. Portaria MCTIn º 4.617 de 6 de abril de 2021 alterada pela portaria MCTI nº 4.979 de 13 de julho de 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial. Acesso em: 10 jul. 2023.

BRASIL. Senado Federal. **Relatório final da Comissão de Juristas responsável por subsidiar a elaboração de substitutivo sobre inteligência artificial no Brasil.** Brasília, DF: Senado Federal, 2022. Disponível em: https://www.stj.jus.br/sites/portalp/SiteAssets/documentos/noticias/Relato%CC%81rio%20final%20CJSUBIA.pdf. Acesso em: 10 jul. 2023.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: FRIEDLER, Sorelle A.; WILSON, Christo (ed.). **Proceedings of the 1st Conference on Fairness, Accountability and Transparency.** NY, USA: PMLR, 2018. v. 81, p. 77–91. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Acesso em: 10 jul. 2023.

CAPPRA INSTITUTE FOR DATA SCIENCE. **Ética na IA:** como desenvolver um sistema ético para Inteligência Artificial. Miami: Cappra Institute, 2021. disponível em: https://www.cappra.institute/s/Etica_IA_PT.pdf. Acesso em: 10 jul. 2023.

CONSELHO NACIONAL DO MEIO AMBIENTE (CONAMA). **Resolução CONAMA nº 001, de 23 de janeiro de 1986.** Dispõe sobre critérios básicos e diretrizes gerais para a avaliação de impacto ambiental. Diário Oficial da União: seção 1, Brasília, DF, 17 fev. 1986. Disponível em: https://www.ibama.gov.br/sophia/cnia/legislacao/MMA/RE0001-230186.PDF. Acesso em: 10 jul. 2023,

DE SPIEGELEIRE, Stephan; MASS, Matthijs; SWEIJS, TIM. **Artificial Intelligence and the future of defense**. Haia: Hague Center for Strategic Studies, 2017. Disponível; https://www.jstor.org/stable/resrep12564.7?seq=3#metadata_info_tab_contents. Acesso em: 10 jul. 2023.

DOBRIN, Seth; MONTGOMERY, Christina. **Principles and Practices for Building More Trustworthy AI.** IBM Newsroom, 2021. Disponível em: https://newsroom.ibm.com/Principles-and-Practices-for-Building-More-Trustworthy-AI. Acesso em: 10 jul. 2023.

EUROPEAN COMMISSION. **Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.** COM(2021) 206 final. Brussels, 21 Apr. 2021. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Acesso em: 10 jul. 2023.

EUROPEAN COMMISSION. **Better regulation.** Brussels: European Commission, 2023. Disponível em: https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en. Acesso em: 10 jul. 2023.

FJELD, J.; Achten, *et al.* **Principled Artificial Intelligence:** Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Cambridge, MA: Berkman Klein Center Research Publication, 2020. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482. Acesso em: 10 jul. 2023.

FLORIDI, Luciano. Establishing the rules for building trustworthy AI, 2019. **Nature Machine Intelligence,** v. 1, p. 261-262, Jun. 2019. Disponível em: https://philarchive.org/rec/FLOETR. Acesso em: 10 jul. 2023.

GUNNINGHAM, N; SINCLAIR, D. Designing smart regulation. **OECD Global Forum on Sustainable Development**, 2004. Disponível em: https://www.oecd.org/env/outreach/33947759.pdf. Acesso em: 10 jul. 2023.

HAO, Karen. This is how AI bias really happens - and why it's so hard to fix. **MIT Technology Review**, feb. 2019. Disponível em: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. Acesso em: 10 jul. 2023.

LEHR, David; OHLM, Paul. Playing with data: What legal scholars should learn about machine learning. **UCDL Review**, v. 51, p. 653-717, dec. 2017. Disponível em: https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf. Acesso em: 10 jul. 2023.

LIAO, S. Matthew. **Ethics of Artificial Intelligence**. New York: Oxford University Press, 2000.

GUTIERREZ, Carlos Ignacio; MARCHANT, Gary. **Soft law 2.0:** Incorporating incentives and implementation mechanisms into the governance of artificial intelligence. OECD.AI, 2019. Disponível em: https://oecd.ai/en/wonk/soft-law-2-0. Acesso em: 10 jul. 2023.

OEA - ORGANIZAÇÃO DOS ESTADOS AMERICANOS. **Convenção Interamericana contra Toda Forma de Discriminação e Intolerância.** Aprovada na segunda sessão plenária da Assembleia Geral, realizada em 5 de junho de 2013. Antígua, Guatemala: OEA, 2013. Disponível em: https://www.oas.org/pt/cidh/mandato/basicos/discriminacioneintolerancia.pdf. Acesso em: 10 jul. 2023.

OECD - ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD Regulatory Policy Outlook 2021.** Paris: OECD Publishing, 2021. Disponível em: https://www.oecd.org/en/publications/oecd-regulatory-policy-outlook-2021_38b0fdb1-en.html. Acesso em: 30 maio 2025.

OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council for Agile Regulatory Governance to Harness Innovation (OECD/LEGAL/0464).** Paris: ODCE, 2021.

OCDE - ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).** Paris: ODCE, 2019.

PICHAI, Sundar. **AI at Google:** our principles. Google, jun. 2018. Disponível em: https://www.blog.google/technology/ai/ai-principles/. Acesso em: 10 jul. 2023.

RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence:** a modern approach. 2 ed. New Jersey: Pearson Education, 2010.

SCHWAB, Klaus. **A quarta revolução industrial**. São Paulo: Edipro, 2018.

SELBST, Andrew D. Et Al. Fairness and Abstraction in Sociotechnical Systems. *In:* **FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency**. New York: Association for Computing Machinery, 2019. p. 59–68. Disponível em: https://dl.acm.org/doi/10.1145/3287560.3287598. Acesso em: 10 jul. 2023.

UNCTAD - UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. **Technology and Innovation Report 2021:** Catching technological waves – Innovation with equity. Geneva: UNCTAD, 2021. Disponível em: https://unctad.org/system/files/official-document/tir2020_en.pdf. Acesso em: 10 jul. 2023.

UNESCO - ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_eng. Acesso em: 10 jul. 2023.

WEF - WORLD ECONOMIC FORUM. **Agile Regulation for the Fourth Industrial Revolution:** a Toolkit for Regulators. Geneva: World Economic Forum, 2020. Disponível em: https://www.weforum.org/about/agile-regulation-for-the-fourth-industrial-revolution-a-toolkit-for-regulators/. Acesso em: 03 jul. 2023.

WEF - WORLD ECONOMIC FORUM. **Agile Governance:** Reimagining Policy-making in the Fourth Industrial Revolution. Geneva: World Economic Forum, 2018. Disponível em: https://encurtador.com.br/UZVug. Acesso em: 03 jul. 2023.

YEUNG, Karen. 'Hypernudge': Big Data as a mode of regulation by design. **Information Communication and Society, v. 20, p. 118-136, May. 2016.** Disponível em: https://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1186713. Acesso em: 03 jul. 2023.

INFORMAÇÕES DOS AUTORES

Maria Edelvacy Marinho

Professora da Graduação do Curso de Direito da Universidade Presbiteriana Mackenzie do curso de Direito. Pesquisadora do Observatório Jurídico da Inovação do OIC/IEA-USP. Doutora em Direito pela Universidade Paris 1- Panthéon Sorbonne. ORCID: https://orcid.org/0000-0002-6957-3099. Endereço eletrônico: mariaedelvacy@gmail.com.

Camilo Onoda Caldas

Professor do Programa de Mestrado da Escola Paulista de Direito (EPD). Pesquisador na Faculdade Lumina de Direito e da Universidade Sao Judas Tadeu. Pós-doutor pela Universidade de Coimbra em Democracia e Direitos Humanos. Doutor em Filosofia e Teoria Geral do Direito pela Faculdade de Direito da Universidade de São Paulo. ORCID: https://orcid.org/0000-0002-6957-3099. Endereço eletrônico: camilo.onoda@gmail.com.

Tatiana Aguiar

Professora do Mestrado do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). Doutora em Direito do Estado pela PUC/SP. ORCID: https://orcid.org/0000-0002-8353-0167. Endereço eletrônico: taticris05@gmail.com.

COMO CITAR

MARINHO, Maria Edelvacy; CALDAS, Camilo Onoda; AGUIAR, Tatiana. O papel do relatório de avaliação de impacto algoritimico para o fomento de sistemas de IA éticos *by* design. **Novos Estudos Jurídicos**, Itajaí (SC), v. 30, n. 1, p. 103-128, 2025. DOI: 10.14210/nej.v30n1.p.103-129.

Recebido em: 16 de ago. de 2023.
Aprovado em: 29 de abr. de 2025.