Policy Paper

Regulação de Inteligência Artificial no Brasil

Contribuição do Centro de Tecnologia e Sociedade (CTS) – Fundação Getulio Vargas (FGV Direito Rio) à Consulta Pública do Ministério da Ciência Tecnologia Inovações e Comunicações – MCTIC sobre a Estratégia Brasileira de Inteligência Artificial

Índice

Sumário Executivo 3

Primeiro Eixo: Legislação, Regulação e Uso Ético 6
1. Quais deveriam ser os princípios éticos a serem observados no Brasil? 6
2. De que maneira princípios éticos podem ser incorporados na pesquisa e na utilização de IA? 9
2.1 Medidas Técnicas 9
2.1.1 Medidas Técnicas para a incorporação de princípios na arquitetura dos sistemas de IA 9
a) Abordagens de baixo para cima (*bottom-up approaches*) 10
b) Abordagens "de cima para baixo" (*top-down approaches*) 10
2.1.2 Medidas Técnicas para fiscalizar a adoção dos princípios nos sistemas de IA 11
a) Métodos de Explicação 11
b) Testes e Validação 11
c) Indicadores de qualidade de serviço 11
d) Desenvolvimento de uma IA guardiã 12
2.2 Medidas de Regulação 12
3. De que maneira é possível concretizar a ideia de explicabilidade em sistemas de IA? 14
3.1 A demanda por explicabilidade em sistemas de IA 14
3.2 Desafios em se conferir explicabilidade a sistemas de IA 14
3.2 Formas de se concretizar a explicabilidade em sistemas de IA 15
4. Como é possível endereçar questões relacionadas à discriminação e ao viés em decisões tomadas por sistemas autônomos? 19

Segundo Eixo: Governança de IA 21
1. Devem ser criadas estruturas institucionais voltadas ao desenvolvimento, aplicação e monitoramento de padrões éticos em IA, a exemplo do Centre for Data Ethics and Innovation do Reino Unido e do Automated Decision Systems Task Force de Nova Iorque? 21

Terceiro Eixo: Aplicação no Poder Público 23
1. Há necessidade de salvaguardas específicas nos processos de tomada de decisão no poder público envolvendo sistemas de IA? Em quais circunstâncias ou em quais áreas? 23

Quarto Eixo: Segurança Pública 25
1. Quais são os métodos e técnicas que podem ser usados para incentivar o desenvolvimento de sistemas de IA seguros e confiáveis? 25
2. De que maneira pode-se apoiar esforços para criar métricas para avaliar a segurança, a proteção e a confiabilidade das aplicações da ciência e tecnologia em relação à inteligência artificial? 28
3. Quais salvaguardas, critérios e cuidados devem ser adotados na utilização de IA no campo da segurança? 30

Bibliografia 33

Autores 37

Sumário Executivo

O Centro de Tecnologia e Sociedade (CTS) apresentou contribuições à Consulta Pública sobre a Estratégia Nacional brasileira de Inteligência Artificial organizada pelo Ministério da Ciência Tecnologia Inovações e Comunicações (MCTIC) na plataforma Participa.br respondendo a 9 (nove) perguntas no âmbito de quatro eixos específicos: **i)** Lei, Regulação e Uso Ético; **ii)** Governança de IA; **iii)** Aplicações no Poder Público e **iv)** Segurança Pública. Este *policy paper* unifica as respostas fornecidas, a fim de englobar as recomendações e ser um ponto de partida para futuras discussões sobre o tema.

1. **Primeiro eixo: Legislação, Regulação e Uso Ético**

Foram submetidas 4 contribuições sobre os seguintes questionamentos:

- **Quais deveriam ser os princípios éticos a serem observados no Brasil?**

  Foi indicado um rol de princípios, buscando captar os consensos existentes no cenário regulatório mundial e na comunidade científica.

- **De que maneira princípios éticos podem ser incorporados na pesquisa e na utilização de IA?**

  Indicamos duas formas principais de se implementar princípios éticos em aplicações de IA, através de: i) medidas técnicas e de ii) medidas regulatórias tradicionais e/ou medidas de auto-regulação.

- **De que maneira é possível concretizar a ideia de explicabilidade em sistemas de IA?**

  Dentre os diversos mecanismos disponíveis, são indicadas oito possibilidades, dentre elas: revisão humana da decisão automatizada, uso de medidas técnicas e procedimentais, atribuição de certificações e realização de auditorias por uma autoridade especializada e independente.

- **Como é possível endereçar questões relacionadas à discriminação e ao viés em decisões tomadas por sistemas autônomos?**

  Observada a predominância da reprodução de vieses em processos decisórios diversos, é recomendado que (1) seja criada uma certificação de boas práticas de desenvolvimento; (2) a formação das equipes de desenvolvimento de IAs seja plural; (3) sejam seguidas as diretrizes de transparência e explicabilidade.

2. Segundo Eixo: Governança de IA

Foi submetida contribuição sobre o seguinte questionamento:

- **Devem ser criadas estruturas institucionais voltadas ao desenvolvimento, aplicação e monitoramento de padrões éticos em IA, a exemplo do Centre for Data Ethics and Innovation do Reino Unido e do Automated Decision Systems Task Force de Nova Iorque?**

Apresentamos pontos positivos na criação de autoridades específicas para a regulação de sistemas de Inteligência Artificial (IA), como, por exemplo, (1) a possibilidade de atuação por experts com *know-how* na área; (2) eficiência e celeridade nas decisões com edição de normas e portarias. Também foram observadas iniciativas como a criação de organismos internacionais que sejam aptas para coordenar a atuação dos órgãos regionais e evitar discricionariedades e discrepâncias na aplicação das normativas.

3. Terceiro Eixo: Aplicação no Poder Público

Foi submetida contribuição sobre o seguinte questionamento:

- **Há necessidade de salvaguardas específicas nos processos de tomada de decisão no poder público envolvendo sistemas de IA? Em quais circunstâncias ou em quais áreas?**

Demonstramos que há necessidade para setores em que se apresentam riscos mais altos à vida social, bem como para casos em que há menor capacidade de supervisão por seres humanos, de mais testes de certificação e mais rigor na governança institucional sobre o sistema de IA operado.

4. Quarto Eixo: Segurança Pública

Foram submetidas três contribuições sobre os seguintes questionamentos:

- **Quais são os métodos e técnicas que podem ser usados para incentivar o desenvolvimento de sistemas de IA seguros e confiáveis?**

Exploramos a combinação de abordagens técnicas e não técnicas para estimular o desenvolvimento de estratégias de aprendizado de máquina mais justas, responsáveis e transparentes. Algumas das alternativas envolvem a adoção de métricas mais rigorosas para análise dos dados pelos sistemas, bem como auditorias e revisões constantes dos códigos e resultados. Destacamos, entre outros aspectos, a importância de existir cada vez mais transparência perante o público sobre quais decisões, e de que maneira, são mediadas por sistemas de IA.

- **De que maneira pode-se apoiar esforços para criar métricas para avaliar a segurança, a proteção e a confiabilidade das aplicações da ciência e tecnologia em relação à inteligência artificial?**

  É essencial o desenvolvimento de uma estratégia sólida para que os princípios norteadores da aplicação de IA sejam observados, garantindo o respeito a direitos fundamentais. Práticas preventivas de monitoramento contínuo e práticas repressivas para contenção de danos devem ser incorporadas em tal estratégia. Mecanismos de integração entre o Poder Público e os atores da sociedade civil, a criação de instituições intermediárias para verificação de eventuais violações e a atenção à privacidade individual são algumas das medidas adequadas e eficientes para atender os objetivos de implantação dos sistemas de IA de forma segura e ética.

- **Quais salvaguardas, critérios e cuidados devem ser adotados na utilização de IA no campo da segurança?**

  As diferentes aplicações de IA no campo de segurança pública possuem em comum o objetivo de alocação dos escassos recursos públicos a fim de otimizar resultados e reduzir custos. Tal motivação faz com que o emprego dessas tecnologias possa levar à violação de direitos fundamentais, seja ao produzir resultados discriminatórios, os quais reproduzem estereótipos de grupos historicamente marginalizados, seja por reforçar uma sociedade de vigilância sem a devida proteção da privacidade e dos dados pessoais. Desse modo, apresentamos algumas abordagens possíveis para três diferentes aplicações de IA em segurança pública, quais sejam: i) tecnologias de policiamento preditivo; ii) tecnologias de reconhecimento facial, e iii) análises de comunicações interpessoais.

Primeiro Eixo: Legislação, Regulação e Uso Ético

Neste eixo, foram respondidas as seguintes perguntas:

1. **Quais deveriam ser os princípios éticos a serem observados no Brasil?**
2. **De que maneira princípios éticos podem ser incorporados na pesquisa e na utilização de IA?**
3. **De que maneira é possível concretizar a ideia de explicabilidade em sistemas de IA?**
4. **Como é possível endereçar questões relacionadas à discriminação e ao viés em decisões tomadas por sistemas autônomos?**

1. Quais deveriam ser os princípios éticos a serem observados no Brasil?

Existe uma preocupação crescente em todo o mundo com a definição dos padrões e limites éticos no uso da Inteligência Artificial (IA). Nesse sentido, já foram mapeadas pelo menos 84 iniciativas público-privadas (MITTELSTADT et al., 2016) descrevendo princípios para orientar o desenvolvimento ético e a governança da IA.

Considerando que o Brasil aderiu aos princípios enunciados pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE)[1], recomenda-se, inicialmente, que a legislação brasileira faça referência a esses standards, elencados resumidamente a seguir:

i) os sistemas de IA devem promover o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar;

ii) os sistemas de IA devem ser projetados de maneira a respeitar o Estado de Direito, os direitos humanos, os valores democráticos e a diversidade, liberdade, dignidade, autonomia, privacidade e proteção de dados, não-discriminação e igualdade, diversidade, equidade, justiça social e direitos trabalhistas internacionalmente reconhecidos. Para isso, devem incluir salvaguardas apropriadas, como possibilitar a intervenção humana sempre que necessário;

iii) deve haver transparência e explicabilidade sobre os sistemas de IA. Por esse motivo, os atores da IA[2] devem fornecer informações significativas e apropriadas ao contexto, de modo a permitir que os potenciais indivíduos afetados tenham acesso aos critérios que serviram de base para a previsão, recomendação ou decisão algorítmica;

[1] OECD (2019a). *Recommendation of the Council on Artificial Intelligence*. OECD/ LEGAL/ 0449, maio 2019. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. (Último acesso em: 10 fev. 2020)

[2] Atores de IA são aqueles que desempenham um papel ativo no ciclo de vida de um sistema de IA, incluindo organizações e indivíduos que implementam ou operam esse sistema (OECD, 2019).

iv) os sistemas de IA devem ser robustos, seguros e protegidos durante todo o ciclo de vida. Os riscos potenciais devem ser continuamente avaliados e gerenciados. Para esse fim, os atores da IA devem garantir a rastreabilidade, inclusive em relação aos conjuntos de dados, processos e decisões tomadas durante o ciclo de vida do sistema de IA[3], possibilitando a análise dos resultados do sistema e, em caso de solicitação ou investigação, devem oferecer as respostas apropriadas ao contexto e consistentes com o estado da arte; e

v) as organizações e os indivíduos que desenvolvem, implantam ou operam sistemas de IA devem ser responsabilizados pelo seu apropriado funcionamento, de acordo com os princípios acima.

Além dos standards indicados pela OCDE, é recomendável que a legislação brasileira adote outros princípios. Conforme indicado anteriormente, várias iniciativas de entidades públicas, privadas e centros de pesquisa enunciam princípios para a regulação da IA. Considerando este cenário, optou-se pela análise de dois documentos gerais: o Relatório "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights based Approaches to Principles for AI" do Berkman Klein Center da Universidade de Harvard (FJELD et al., 2020)[4] e os "Princípios de Asilomar" (FUTURE OF LIFE INSTITUTE, 2017)[5], os quais buscaram captar consensos existentes no cenário regulatório mundial e na comunidade científica.

A pesquisa elaborada pelo Berkman Klein Center identificou oito princípios-chave comumente relatados nas principais iniciativas sobre o tema (FJELD et al., 2020)[6]. Desses princípios, a fim de complementar os indicados pela OCDE, devem ser levados em conta:

i) privacidade (presente em 97% dos documentos analisados na pesquisa) – conceito mais amplo que engloba o debate sobre: consentimento, controle sobre o uso de dados, capacidade de restringir o processamento de dados, direito à retificação, direito ao apagamento e privacidade por design (FJELD et al., 2020, p. 21-7);

ii) equidade e não-discriminação (princípio mencionado em todos os documentos analisados na pesquisa) - tem como finalidade a prevenção de vieses na operação algorítmica, especialmente com relação a populações marginalizadas. Para isso, mostra-se necessária a utilização de bases de dados representativas, atualizadas e de alta qualidade, além do tratamento equitativo e imparcial dos titulares dos dados pelos sistemas de IA (FJELD et al., p. 47-50);

---

[3] As fases do ciclo de vida do sistema de IA envolvem: i) 'design, dados e modelos'; ii) 'verificação e validação'; iii) 'implantação'; e iv) 'operação e monitoramento'. Essas fases geralmente ocorrem de maneira iterativa e não são necessariamente sequenciais. A decisão de tornar um sistema de IA inoperante pode ocorrer a qualquer momento durante a fase de 'operação e monitoramento' (OECD, 2019).

[4] O relatório partiu de uma análise de 36 documentos elaborados por governos, empresas, centros de pesquisa e organizações civis em diferentes localidades da América Latina, Leste e Sul da Ásia, Oriente Médio, América do Norte e Europa, para elencar os principais temas comumente abordados.

[5] Os princípios, assinados por mais de 1.500 pesquisadores especializados no tema de IA e Robótica, foram desenvolvidos a partir de conferências científicas internacionais coordenadas pelo Instituto "Future of Life" – organização beneficente voltada ao apoio e desenvolvimento de pesquisas focadas no desenvolvimento das novas tecnologias, especialmente a Inteligência Artificial, de forma benéfica para a humanidade. Dentre os membros, encontram-se cientistas renomados como Stephen Hawkings e Max Tegmark, bem como empreendedores na área de tecnologia, como Elon Musk. Para mais informações, ver: https://futureoflife.org/ai-principles/ (Último acesso em: 10 fev. 2020)

[6] Os oito temas indicados são: i) privacidade; ii) accountability; iii) confiabilidade e segurança; iv) transparência e explicabilidade; v) equidade e não discriminação; vi) controle humano sobre a tecnologia; vii) responsabilidade profissional; viii) promoção de valores humanos (FJELD et al., 2020).

iii) controle humano sobre a tecnologia (referido em 69% dos documentos analisados) - de modo geral, preocupa-se em preservar a autonomia para que os seres humanos escolham como e quais decisões delegar aos sistemas de IA. Está diretamente relacionado ao "direito de obter uma revisão humana da decisão automatizada" (presente em 33% dos documentos analisados na pesquisa) (FJELD et al., p. 53).

Do mesmo modo, recomenda-se que a legislação observe os 23 Princípios de Asilomar (FUTURE OF LIFE INSTITUTE, 2017)[7], os quais congregam aspectos básicos comumente aceitos por grande parte da comunidade científica altamente especializada no tema. Dentre os standards indicados, seguem, de forma adaptada, os 6 (seis) principais pontos que merecem consideração legislativa:

i) financiamento da pesquisa: os investimentos em IA devem ser acompanhados de financiamento para pesquisas que garantam seu uso benéfico, incluindo questões complexas em ciência da computação, economia, direito, ética e estudos sociais;

ii) diálogo entre Ciência e Política: deve haver um intercâmbio construtivo e saudável entre pesquisadores de IA e políticos. Nesse sentido, é imprescindível que o processo legislativo seja instruído por consultas e audiências públicas com participação de cientistas especializados no tema para que a futura legislação esteja obrigatoriamente fundamentada em evidências científicas;

iii) transparência judicial: qualquer envolvimento de um sistema autônomo na tomada de decisões judiciais deve fornecer uma explicação satisfatória auditável por uma autoridade humana competente;

iv) autonomia sobre os dados pessoais: a aplicação da IA aos dados pessoais não deve restringir injustificadamente a liberdade real ou percebida das pessoas. Sendo assim, os usuários devem ter o direito de acessar, gerenciar e controlar os dados que geram, tendo em vista o poder dos sistemas de IA em analisar e utilizar esses dados;

v) prosperidade compartilhada: a prosperidade econômica criada pela IA deve ser amplamente compartilhada em benefício de toda a humanidade;

vi) mitigação de riscos: os riscos colocados pelos sistemas de IA, especialmente os riscos catastróficos ou existenciais, devem estar sujeitos a esforços de planejamento e mitigação proporcionais ao impacto esperado.

Por fim, é importante ressaltar que os princípios citados devem ser compreendidos e aplicados conforme o contexto cultural, linguístico e geográfico do país (FJELD et al., 2020). Ademais, esses princípios são o ponto de partida para delinear uma política mais completa de governança, a qual deve abranger o desenvolvimento de estratégias e processos, muitas vezes específicos para cada um desses valores, a fim de colocá-los em prática nas rotinas profissionais de empresas, organizações e governo.

[7] A lista completa com os 23 princípios está disponível em: https://futureoflife.org/ai-principles. (Último acesso em: 10 fev. 2020)

2. De que maneira princípios éticos podem ser incorporados na pesquisa e na utilização de IA?

Além de estabelecer princípios éticos, é desejável que a legislação oriente sobre os atores e os modos pelos quais os princípios devem ou podem ser inseridos nos sistemas de IA. É importante esclarecer que os princípios são implementados pelos diversos atores participantes do ciclo de vida dos sistemas de IA, sejam eles os criadores, os implementadores ou até mesmo os usuários finais, bem como a sociedade em geral. Segundo o Grupo de Especialistas de Alto Nível sobre Inteligência Artificial (em inglês, High-Level Expert Group on Artificial Intelligence - AI HLEG) da Comissão Europeia, os criadores são aqueles que investigam, concebem e/ou desenvolvem os sistemas de IA; os implementadores são as organizações públicas ou privadas que utilizam esses sistemas de IA nos seus processos empresariais e para oferecer produtos e serviços a terceiros; os usuários finais são as pessoas que interagem de forma direta ou indireta com o sistema de IA; e, por fim, a sociedade em geral engloba todas as outras pessoas que são direta ou indiretamente afetadas pelos sistemas de IA (AI HLEG, 2019, p. 14).

Esses atores desempenham diferentes papéis para garantir que os requisitos éticos sejam cumpridos: i) criadores devem adotar e aplicar os requisitos aos processos de concepção e desenvolvimento; ii) implementadores devem assegurar que os sistemas que utilizam e os produtos e serviços que oferecem cumprem os requisitos; iii) usuários finais e a sociedade em geral devem ser informados acerca desses requisitos e exigir que os mesmos sejam respeitados.

No que diz respeito aos meios de implementação dos princípios, tanto o relatório do Fórum Econômico Mundial (WORLD ECONOMIC FORUM, 2019) quanto o relatório do Grupo de Especialistas de Alto Nível sobre Inteligência Artificial (AI HLEG, 2019) indicam que existem ao menos duas formas principais de se implementar princípios éticos em aplicações de IA: i) medidas técnicas e ii) medidas regulatórias tradicionais e/ou medidas de autorregulação. Algumas dessas medidas são abordadas a seguir, sendo importante ressaltar que não se trata de um rol exaustivo e que os métodos indicados podem ser considerados complementares ou alternativos entre si, uma vez que diferentes princípios podem demandar diferentes abordagens de aplicação.

2.1 Medidas Técnicas

No campo das soluções técnicas, existe a possibilidade de adotar a conformidade ética "por design", isto é, a incorporação dos princípios diretamente na arquitetura do sistema de IA. Além disso, também é possível adotar medidas técnicas de *accountability* para verificar se a aplicação das regras pelos sistemas está sendo realizada da forma adequada.

2.1.1 Medidas Técnicas para a incorporação de princípios na arquitetura dos sistemas de IA

No primeiro caso, os critérios éticos de tomada de decisão podem ser inseridos nos sistemas de IA tanto por uma abordagem "de baixo para cima" (*bottom-up approaches*) quanto por uma abordagem "de cima para baixo" (*top-down approaches*) (ETZIONI, 2017) (FÓRUM ECONÔMICO MUNDIAL, 2019).

a) Abordagens de baixo para cima (*bottom-up approaches*)

Esse primeiro tipo de abordagem requer que robôs e demais sistemas de IA observem o comportamento humano em situações específicas e aprendam como tomar decisões com base nos padrões de comportamento assimilados. Isso, no entanto, pode fazer com que os sistemas passem a adotar um comportamento "comum" que, como se sabe, não necessariamente reflete a conduta ética esperada. Essa questão pode ser exemplificada com o caso da robô Tay, chatbot lançado em 2016 pela Microsoft para interagir com os usuários do Twitter, que começou a adotar discursos políticos agressivos e racistas ensinados pelos próprios usuários (HUNT, 2016). Outro exemplo é o dos carros autônomos, os quais poderiam aprender a partir de decisões éticas de humanos com base em algum tipo de sistema que agregaria dados de milhões de motoristas (ETIZIONI, 2013). Assim como no primeiro exemplo, essa alternativa poderia levar os carros a adquirirem preferências não desejadas, como dirigir acima da velocidade máxima permitida, pela assimilação desse tipo de comportamento humano.

b) Abordagens "de cima para baixo" (*top-down approaches*)

Como alternativa às abordagens "de baixo para cima", alguns especialistas observam que as medidas "de cima para baixo" seriam melhores para implementação de padrões éticos nos sistemas de IA (FÓRUM ECONÔMICO MUNDIAL, 2019)[8]. Esse tipo de abordagem requer que os princípios e regras sejam diretamente programados no sistema de IA, ou seja, eles são traduzidos em procedimentos ou restrições a serem incorporados na própria arquitetura do sistema.

Isso pode ser feito pela criação de uma "lista branca" (*white list*) - conjunto de regras (comportamentos ou estados) que o sistema deve sempre seguir, e uma "lista preta" ("*black list*") - conjunto de restrições (a comportamentos ou estados) que o sistema nunca deve transgredir. As duas listas podem ser combinadas e outras garantias mais complexas referentes ao comportamento do sistema também podem ser adotadas (AI HLEG, 2019, p. 21).

Algoritmos de aprendizado de máquina (*machine learning*) são frequentemente analisados segundo a perspectiva teórica de um ciclo de "percepção, planejamento e ação" (AI HLEG, 2019, p. 21). Trata-se de sistemas não determinísticos, isto é, algoritmos mais complexos, que conseguem adaptar dinamicamente o seu comportamento, e, por isso, podem apresentar um comportamento inesperado. Sendo assim, sistemas de *machine learning* exigem que os princípios e regras sejam integrados nas três etapas do ciclo: i) na etapa de "percepção", o sistema deve ser desenvolvido de modo a reconhecer todos os elementos ambientais necessários para assegurar que os princípios sejam respeitados; ii) na etapa de "planejamento", o sistema apenas deve considerar os planos que cumprirem com os requisitos; iii) na etapa de "ação", as ações do sistema devem restringir-se aos comportamentos que cumprem os requisitos (AI HLEG, 2019, p. 21).

Além disso, cabe ressaltar que geralmente a incorporação de princípios e regras na arquitetura dos sistemas de IA requer a adoção de uma "abordagem casuística" (OECD, 2019), ou seja, a programação do sistema deve estar alinhada ao caso concreto de modo que a máquina reaja de forma específica a determinada situação. Também deve ser levada em conta a finalidade da atividade a ser desenvolvida pelo sistema de IA. Exemplo disso são os robôs de assistência médica, que podem ser programados para sempre considerarem a vontade do usuário.

[8] Conforme indicado pelo relatório do Fórum Econômico Mundial: "Therefore, from a technical perspective, it appears that a top-down approach is better suited to implement ethics into AI. In such an approach, ethical principles would be programmed directly into an AI system" (WORLD ECONOMIC FORUM, 2019).

Ainda com relação a esse tópico, vale observar que não existe uma visão única do que seria uma decisão ética. Os sistemas de IA podem ser programados de acordo com uma cultura específica, uma determinada escola de pensamento ético ou de acordo com valores culturais e/ou religiosos diversos. Por isso, continua sendo um desafio para os designers de sistemas de IA decidirem sobre qual será a filosofia aplicada aos sistemas algorítmicos de tomada de decisão (WORLD ECONOMIC FORUM, 2019). Assim sendo, a recomendação de órgãos internacionais, como o Fórum Econômico Mundial (2019) e o AI HLEG (2019), é de que os requisitos éticos para os sistemas computacionais sejam desenvolvidos colaborativamente e sejam revisados para obter consistência no processo de tomada de decisão. É necessária uma estreita cooperação entre pesquisadores, desenvolvedores e formuladores de políticas para desenvolver um entendimento comum dos princípios éticos e como eles serão implementados na prática, tendo em vista a variedade de setores em que a IA se aplica.

2.1.2 Medidas Técnicas para fiscalizar a adoção dos princípios nos sistemas de IA

Uma vez decidida e incorporada a regra ética na arquitetura do sistema de IA, é necessário o desenvolvimento de mecanismos de *accountability* capazes de verificar se o sistema realmente está seguindo os padrões éticos determinados. Isso pode ser feito por meio de: a) métodos de explicação; b) testes e validação; c) indicadores de qualidade do serviço; e d) desenvolvimento de uma IA guardiã.

a) Métodos de Explicação

A finalidade de métodos de explicação é compreender por que razão o sistema de IA se comportou de determinada forma e/ou produziu determinada interpretação. A IA explicável (*Explainable AI — XAI*) é um domínio de investigação totalmente dedicado a essa questão e busca obter uma melhor compreensão dos mecanismos subjacentes ao sistema e encontrar soluções. Atualmente, no entanto, este é um desafio em aberto no caso de sistemas de IA baseados em redes neurais (*deep learning*) (AI HLEG, 2019, p. 21).

b) Testes e Validação

Devido à natureza contextual e não determinística dos sistemas de IA, especialistas argumentam que testes tradicionais não são suficientes (AI HLEG, 2019, p. 22). Isso porque as falhas dos conceitos e representações utilizados pelo sistema podem se manifestar apenas quando um programa é aplicado a dados "suficientemente realistas". Diante disso, para verificação e validação do tratamento de dados, o modelo subjacente deve ser cuidadosamente monitorado, tanto durante a fase de treinamento quanto durante a fase de implementação, para que seja assegurada a sua estabilidade e operação dentro de limites bem compreendidos e previsíveis. Ademais, é recomendado que: i) os testes sejam concebidos e executados por um grupo de pessoas o mais diversificado possível; ii) sejam desenvolvidas múltiplas métricas para analisar as categorias testadas, segundo diferentes perspectivas; iii) seja considerada a realização de *adversarial testing* por *red teams* confiáveis e diversificadas, que tentem deliberadamente "penetrar" e "quebrar" o sistema para encontrar vulnerabilidades (AI HLEG, 2019, p. 22).

c) Indicadores de qualidade de serviço

Podem ser definidos indicadores adequados de qualidade de serviço para os sistemas de IA, a fim de assegurar que eles foram desenvolvidos e testados à luz de considerações de segurança e proteção. Esses indicadores podem incluir medidas para avaliar os testes e o treinamento dos algoritmos, bem como os parâmetros tradicionais de avaliação de software: funcionalidade; desempenho; usabilidade; confiabilidade; segurança; e manutenção (AI HLEG, 2019, p. 22).

d) Desenvolvimento de uma IA guardiã

Para além das técnicas acima mencionadas, e conforme indicado no relatório do Fórum Econômico Mundial (WORLD ECONOMIC FORUM, 2019), é possível o desenvolvimento de uma "IA guardiã" para fiscalizar o cumprimento das regras e dos princípios essenciais. Esse sistema de monitoramento seria projetado para garantir a conformidade dos processos, podendo interferir tecnicamente no sistema da IA básica e corrigir diretamente decisões ilegais ou antiéticas com base no conjunto de leis e regras previamente estabelecidos. Percebe-se que esse sistema "guardião" poderia ser programado para relatar a decisão ilícita ou antiética da IA básica a determinada autoridade ou agência executiva com competência para atuar no caso.

2.2 Medidas de Regulação

Apesar da importância das soluções técnicas, elas são insuficientes para resolver questões éticas sozinhas. Nesse sentido, até os líderes da indústria de tecnologia estão sugerindo a adoção de uma fiscalização regulatória por parte dos governos (SCHERER, 2016)[9]. Conforme indicado por Etzioni et al. (2017), o ideal é que os princípios éticos sejam implementados em sistemas de IA por métodos regulatórios. Os autores argumentam que duas principais maneiras de se implementar valores morais e sociais é por meio das escolhas pessoais dos indivíduos, que são promovidas e influenciadas socialmente, e por meio da imposição legal. Realmente, verifica-se que, de modo geral, os valores morais são coletivamente formulados e, muitas vezes, incorporados em leis e regulamentos.

Considerando esse argumento, as máquinas devem ser orientadas (no caso, programadas) para seguir as normas sociais especificamente indicadas para seu meio de atuação, deixando também um espaço para as escolhas pessoais de seus usuários. No exemplo de carros autônomos, eles seriam programados para seguir as leis de trânsito do país em que se encontram, enquanto que nas situações não prescritas pela lei, as decisões éticas podem ser tomadas pelas empresas que desenvolvem o software, ou de maneira mais personalizada, pelos usuários. Os carros autônomos produzidos pela empresa *Tesla*, por exemplo, permitem que o usuário defina a distância entre o seu carro e o carro da frente.

Uma questão importante trazida por Etzioni et al. (2017) é que o debate sobre a adequação de sistemas de IA a padrões éticos não trata propriamente de implantar um pensamento ético nos sistemas de IA ou em ensinar princípios éticos às máquinas. Os autores entendem que mesmo os sistemas de IA considerados “autônomos” são programados e limitados conforme as orientações estabelecidas por um ser humano. Sendo assim, muitas decisões éticas que as máquinas inteligentes devem tomar não precisam e não devem ser tomadas por elas livremente, já que as orientações devem estar consolidadas na lei. Isto é, as escolhas éticas não são feitas pelas máquinas, mas pela sociedade, utilizando os métodos tradicionais de adequação normativa, como legislações e tribunais (ETZIONI et al., 2017).

Nesse mesmo sentido, o Fórum Econômico Mundial (2019) recomenda a adoção de uma variedade de mecanismos de governança para a regulação de sistemas de IA, tais como: i) legislação; ii) resoluções e acordos internacionais; ii) tratados bilaterais de investimento e a criação de incentivos econômicos; iii) autorregulação; iv) certificação; v) normas contratuais e vi) *soft law*.

[9] Como exemplo, o próprio Elon Musk, engenheiro, empreendedor na área de tecnologia e fundador e CEO da Tesla (companhia norte-americana de energia e automotivos) (SCHERER, 2016).

É certo que a abordagem regulatória pode deixar descobertos casos específicos, e frequentemente imprevisíveis, uma vez que sempre existirão episódios não regulamentados ou sem definição pelos usuários e desenvolvedores dos softwares. Assim como ocorre, de maneira geral, com os fatos da vida, a repetição de casos imprevistos demandará a atualização e/ou melhor detalhamento das regras adotadas para o caso, o que, naturalmente, não elimina a necessidade de reparação para as pessoas afetadas por eventual dano causado por máquina.

Ademais, cumpre salientar que a definição de normas e princípios sobre o tema, por mais coerentes, desenvolvidas e pormenorizadas que sejam, não elimina a necessidade de constante reflexão ética, a qual deve permanecer sensível aos aspectos contextuais que nem sempre são captados em orientações genéricas. Mais do que a definição de um conjunto de regras, a garantia de uma IA de confiança depende da construção e da manutenção de uma cultura e mentalidade ética por meio do debate público, da educação e da aprendizagem prática (AI HLEG, 2019).

3. De que maneira é possível concretizar a ideia de explicabilidade em sistemas de IA?

Conforme levantado pela literatura especializada, as decisões tomadas por sistemas de IA não podem ser consideradas como automaticamente objetivas, justas ou imparciais (BECKER; FERRARI, 2018; BURRELL, 2016; DIAKOPOULOS, 2013; GANDY, 2010; PASQUALE, 2015). O julgamento e os vieses humanos encontram-se embutidos no design dos algoritmos, já que são os humanos que definem suas características, pré-classificam os dados de treinamento e ajustam seus limites e parâmetros (BURRELL, 2016).

3.1 A demanda por explicabilidade em sistemas de IA

Considerando esse cenário, já é amplamente aceita a ideia de que decisões tomadas por um sistema algorítmico devem ser, de alguma maneira, explicáveis para as pessoas afetadas por essas decisões. Conforme indicado em estudo recente de pesquisadores do Berkman Klein Center (FJELD et al., 2020), a explicabilidade e a transparência são dois princípios que devem ser fundamentalmente observados para a contenção de vieses em decisões automatizadas.

A ideia da explicabilidade é expor a lógica que fundamenta determinada decisão de uma maneira clara e compreensível. Isto é, apresentar uma descrição compreensível sobre como o tomador de decisão partiu de um conjunto específico de inputs e chegou a um output específico (DOSHI-VELEZ; KORTZ, 2017, p. 2). Para isso, é importante que um observador externo possa compreender em que medida certo fator exerceu influência ou foi determinante para o resultado. Dito isto, o propósito da explicação é o de verificar quais critérios foram considerados e se eles foram utilizados de forma adequada, de modo a prevenir discriminações e erros na decisão. Em outras palavras, uma decisão é explicada quando é possível responder, ao menos, uma das seguintes perguntas: (i) quais são os principais fatores que levaram à decisão? (ii) alterar algum dos fatores mudaria a decisão? (iii) por que casos semelhantes tiveram decisões diferentes e vice-versa? (DOSHI-VELEZ; KORTZ, 2017, p. 3).

Ocorre que, quando se trata de decisões tomadas por sistemas de IA, questões de opacidade inerentes ao algoritmo e ressalvas quanto ao segredo comercial levantam dúvidas sobre em que medida a explicabilidade pode ser demandada; que tipos de exigências podem ser feitas por eventual lei regulamentadora do tema e quais procedimentos devem ser adotados para colocar em prática essas exigências. Essas questões são exploradas com mais detalhes a seguir.

3.2 Desafios em se conferir explicabilidade a sistemas de IA

Jenna Burrell (2016) distingue três formas de opacidade dos algoritmos, compreendidas como barreiras para a transparência e explicabilidade do código: (1) o ocultamento intencional de seu funcionamento por parte das corporações e outras instituições, com a finalidade de proteger a propriedade intelectual, o segredo industrial ou impedir que as pessoas “burlem” o sistema; (2) a falta de entendimento técnico do público, de forma que o acesso ao código fonte não é suficiente para que a pessoa possa compreender o modo como o sistema opera e a decisão tomada; e (3) a opacidade intrínseca, derivada da própria característica de algoritmos de *machine learning*, que demandam uma alta complexidade matemática, difícil de ser interpretada pelo raciocínio humano. Essa é uma questão amplamente reconhecida no campo da programação, denominada “problema da

interpretabilidade" (*interpretability problem*). Conforme indicado pela autora, reconhecer as diferentes formas de opacidade seria um fator chave para determinar que tipo de soluções técnicas e jurídicas devem ser aplicadas ao caso (BURRELL, 2016, p. 3).

3.2 Formas de se concretizar a explicabilidade em sistemas de IA

O primeiro tipo de opacidade pode ser tratado estabelecendo-se a obrigatoriedade de que os desenvolvedores do software mantenham o código aberto. Isso, no entanto, não impede que sejam levantadas escusas, por parte de desenvolvedores e organizações que utilizam esses sistemas, em relação a questões de propriedade intelectual, segredo comercial ou possibilidade de usuários "burlarem"[10] o algoritmo. Diante desse desafio, uma medida recomendada por diversos pesquisadores da área é tornar o código disponível para avaliação por um órgão regulador autônomo capaz de mantê-lo em sigilo, sem violar eventuais direitos à propriedade intelectual ou segredo comercial, e, ao mesmo tempo, atender ao interesse público (BURRELL, 2016; DIAKOPOULOS, 2013; PASQUALE, 2015).

Também existem alternativas para se implementar a explicabilidade mesmo na ausência de acesso ao código. Dwork et al. (2011) explicam que determinado efeito discriminatório gerado por algoritmos de classificação pode ser detectado sem que seja necessário extrair o "como" e o "porquê" daquela decisão específica. Sandvig et al. (2014) detalham e comparam várias formas de auditoria algorítmica (realizada com ou sem cooperação da empresa que detém o código). Um dos métodos de auditoria externa é baseado na comparação do output com o comportamento equitativo esperado (SANDVIG et al., 2014).

O segundo tipo de opacidade diz respeito à falta de conhecimento técnico por parte do público em geral para ler e interpretar o código. No momento, uma alternativa viável e útil seria o estabelecimento de um dever de informação ao consumidor e aos cidadãos de forma geral acerca das decisões que estão sendo tomadas de maneira automatizada (OSOBA; WELSER IV, 2017), o que valeria tanto para empresas que se utilizam de IA como para o setor público. Além disso, acredita-se que a questão da falta de conhecimento técnico deve ser tratada, a longo prazo, com a inserção do aprendizado de programação no ensino público, por exemplo (DANAHER, 2016).

Ademais, parece essencial que as pessoas afetadas por decisões algorítmicas possam pleitear uma explicação dos principais critérios e fundamentos utilizados para a decisão automatizada. Esse é um dever já estabelecido na Lei Geral de Proteção de Dados brasileira, Lei nº 13.709/2018 (BRASIL, 2018), que garante a explicabilidade de qualquer decisão automatizada que utilize dados pessoais de indivíduos, independente do contexto, setor e mercado[11]. A LGPD prevê dois principais mecanismos para a concretização do direito à explicação: a revisão das decisões automatizadas

[10] O termo é geralmente referido em inglês como "game the system".

[11] O direito à explicação sobre decisões automatizadas se encontra atualmente regulamentado em duas normativas brasileiras: a Lei do Cadastro Positivo - Lei 12.414/2011 e a recente Lei Geral de Proteção de Dados (LGPD) - Lei nº 13.709/2018. Enquanto a primeira possui uma aplicação setorial, estabelecendo o dever de explicabilidade sobre as decisões automatizadas no âmbito da concessão de crédito, a segunda possui uma aplicação mais geral, ao garantir a explicabilidade de qualquer decisão automatizada que se utilize de dados pessoais de indivíduos, independente do contexto, setor e mercado.

(artigo 20, caput e §1º)[12] e a realização de auditorias pela ANPD (Autoridade Nacional de Proteção de Dados) (artigo 20, §2º)[13].

O terceiro tipo de opacidade se refere ao "problema da interpretabilidade" de sistemas de *machine learning*. O funcionamento de sistemas de aprendizado de máquina pode escapar totalmente ao entendimento e a interpretação humana, mesmo para cientistas da computação ou pessoas que possuem conhecimento especializado no tema (BURRELL, 2016).

Devido à opacidade inerente ao código, uma abordagem sugerida por certos especialistas é evitar o uso de determinados algoritmos de *machine learning* em áreas críticas (BURRELL, 2016, p. 9), restringindo sua aplicação aos casos que demandam respostas mais objetivas (ROBBINS, 2019). Alguns tipos de algoritmos aprendem sozinhos, a partir de dados e situações que não estavam previamente representados no conjunto de treinamento (*training set*), o que pode levar a consequências imprevisíveis e inexplicáveis de antemão. Tais consequências podem ser até mesmo fatais, se considerarmos como exemplo os carros autônomos (BOTH, 2014). Nesse sentido, pesquisadores argumentam que certos algoritmos de *machine learning* devem ser utilizados apenas para aquelas funções, decisões ou ações que não possuem a característica de "exigir uma explicação" (ROBBINS, 2019).

Uma possível solução para o "problema da interpretabilidade" seria a adoção de designs alternativos, mais interpretáveis, na construção dos sistemas de IA (OSOBA; WELSER IV, 2017). Ressalta-se que essa alternativa, embora viável, pode também representar uma perda de desempenho em termos de precisão e eficiência do sistema (DANAHER, 2016).

No intuito de facilitar a explicação do resultado obtido pelo algoritmo, existem maneiras de simplificar o aprendizado de máquina, adotando, por exemplo, a técnica de "extração de características" (*feature extraction*). Essa abordagem analisa quais características realmente importam para o resultado da classificação, removendo todas as outras características do modelo (BURRELL, 2016, p. 9).

Além disso, diversas pesquisas (ATHEY, 2015; BOTTOU et al., 2013; PEARL, 2009) investigam caminhos para construir algoritmos de *machine learning* que apliquem raciocínio causal ou contrafactual. Essa seria uma solução importante, já que sistemas automatizados de raciocínio causal seriam capazes de apresentar narrativas claras de causalidade para julgar a qualidade de um processo de decisão levado a cabo por um algoritmo. Certos pesquisadores reportam que justificativas causais precisas para decisões automatizadas são as formas de auditoria mais confiáveis para algoritmos e especialmente importantes para justificar estatisticamente resultados desproporcionais. Algoritmos que podem ser auditados por meio de fatores causais seriam, portanto, capazes de fornecer narrativas claras, isto é, justificativas para os resultados apresentados (OSOBA; WELSER IV, 2017).

---

[12] Art. 20. O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade. § 1º. O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial. (BRASIL, 2018).

[13] Artigo 20, § 2º. Em caso de não oferecimento de informações de que trata o § 1º deste artigo baseado na observância de segredo comercial e industrial, a autoridade nacional poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais. (BRASIL, 2018).

Se, por um lado, há pesquisas que focam nos elementos causais da decisão, outros estudos alegam que a explicabilidade não exigiria a compreensão do processo pelo qual as decisões são alcançadas, mas de uma análise do resultado das decisões. Assim, não seria necessário investigar os processos levados a cabo pelo código, muitas vezes indicados como uma "caixa preta", mas apenas analisar se a decisão em si é coerente. Sandvig et. al (2014) argumentam que certos tipos de auditoria podem ser realizadas ignorando o funcionamento interno dos algoritmos e julgando-os de acordo com a "justiça de seus resultados". Este tipo de avaliação seria similar à análise de uma decisão humana, por exemplo, em que são observadas as consequências da decisão e ação humana e não o conteúdo dos pensamentos do agente. Abordagens desse tipo advogam por uma ética consequencialista para sistemas de IA (OSOBA; WELSER, 2017).

Uma questão chave associada à opacidade é o fato de que, atualmente, não existem procedimentos de documentação padronizados destinados a divulgar as características de desempenho de sistemas de IA (MITCHELL et al., 2019). Tendo em vista essa lacuna, mostra-se necessária a adoção de procedimentos destinados a esclarecer pontos-chave do funcionamento do algoritmo, bem como os usos que se pretende dar a determinado modelo de *machine learning*. Uma das abordagens proposta por pesquisadores da área é a utilização de "cartões de modelo" (*model cards*) (MITCHELL et al., 2019), que são documentos curtos (uma ou duas páginas) e objetivos, com a finalidade de detalhar o modelo proposto. A recomendação é de que os modelos postos em produção sejam acompanhados dessa documentação detalhando o contexto no qual o algoritmo pretende atuar, suas características de desempenho, procedimentos de avaliação de desempenho, bem como outras informações relevantes, como métricas que capturam considerações sobre vieses, equidade e inclusão.

Os denominados "cartões de modelo" devem ser utilizados conjuntamente às já conhecidas "Fichas Técnicas para Bases de Dados" (*Datasheets for Datasets*), que revelam detalhes acerca do *conjunto de dados utilizado para treinar e testar os modelos de* machine learning. Enquanto as "fichas" focam nas características dos dados utilizados para alimentar o modelo, os "cartões" focam nas características de treinamento do modelo, como o tipo de modelo, os usos pretendidos, as informações sobre atributos para os quais o desempenho do modelo pode variar e as medidas de desempenho do modelo. O ideal é que as medidas de desempenho do modelo contenham resultados de avaliações quantitativas subdivididos conforme o grupo cultural, demográfico ou fenotípico e as condições de domínio relevantes, seguidas por uma análise interseccional combinando dois ou mais grupos e condições (MITCHELL et al., 2019, p. 221). Esses procedimentos têm a finalidade de aumentar a transparência sobre como determinado sistema de IA funciona e, assim, minimizar o uso de algoritmos em contextos ou finalidades para os quais aquele modelo não é adequado.

Conforme abordado anteriormente, não se pode esperar que todos consigam compreender plenamente o funcionamento e os efeitos dos sistemas de IA, sendo portanto recomendável que organizações especializadas possam atestar perante o público em geral que um sistema de IA é transparente, responsável e equitativo. Isso pode ser realizado a partir de procedimentos de certificação dos sistemas. Essas certificações aplicarão normas concebidas para diferentes domínios de aplicação e técnicas de IA, adequadamente alinhadas com as normas setoriais produzidas em diferentes contextos (AI HLEG, 2019).

Por fim, uma solução que parece atender aos três tipos de opacidade seria a criação de um órgão regulador especializado e independente, capaz de revisar e licenciar os sistemas de decisão algorítmica. Essa autoridade teria como função definir quais os tipos de auditorias podem ser realizadas; quais exigências técnicas e/ou jurídicas devem ser feitas para cada caso; determinar eventuais tipos de decisão ou contextos em que deve ser vedado o uso de algoritmos de *machine*

learning, devido à sua "opacidade intrínseca"; enunciar eventuais tipos de decisão ou contextos que demandam uma explicação mais apurada da decisão ou a possibilidade de revisão humana; definir as exigências técnicas a serem seguidas pelas organizações tanto no desenvolvimento quanto na utilização de sistemas de IA.

Como visto, tornar a operação de algoritmos explicável é uma tarefa que depende de uma série de ferramentas e processos, que podem ser combinados de maneiras diversas a fim de desenvolver uma regulação coerente e centrada no ser humano. São diversos os mecanismos que podem "dar forma" e aplicabilidade ao direito à explicação. Por isso, concluímos indicando, em resumo, algumas das possibilidades elencadas:

(1) dever de informar que a decisão é automatizada;

(2) possibilidade de contestar a decisão automatizada;

(3) revisão humana da decisão automatizada;

(4) adoção de designs alternativos na construção dos sistemas de IA, que privilegiem sistemas mais simples e interpretáveis;

(5) uso de medidas procedimentais pelos desenvolvedores para tornar algoritmos de *machine learning* explicáveis, tais como: "fichas técnicas " e "cartões de modelo";

(6) uso de medidas técnicas embutidas em algoritmos de *machine learning* que permitam, por exemplo, a "extração de características" ou a "apresentação de justificativas causais";

(7) atribuição de certificações que assegurem a adequação dos sistemas às exigências legais;

(8) auditorias dos sistemas de IA realizadas por uma autoridade especializada e independente.

4. Como é possível endereçar questões relacionadas à discriminação e ao viés em decisões tomadas por sistemas autônomos?

A automação de decisões tem produzido ganhos significativos em termos de velocidade e eficiência para procedimentos aplicados em diversos campos da vida cotidiana, como na prestação de serviços privados, de serviços públicos e infraestruturais. No entanto, a literatura especializada tem demonstrado preocupação com algumas questões inerentes aos sistemas de decisões automatizadas. O principal alerta gira em torno da confiança na "infalibilidade" de seus resultados, que é normalmente atribuída pela sociedade em geral. São diversos os estudos que apontam erros nesses sistemas, tanto em termos de precisão estatística – como no caso do Google's Flu Trends (LAZER et al., 2014) e da estimativa de riscos antes da crise imobiliária nos Estados Unidos em 2008 (SALMON, 2012), como em termos éticos, que dizem respeito aos vieses incorporados a estes sistemas (CITRON, 2019; SILVA, 2019; TUFEKCI, 2015).

Ao abordar a existência desses problemas em sistemas autônomos, é necessário destacar a reprodução de discriminações acirradas pela "aura de objetividade que a automação tende a reforçar". Como Osoba e Welser IV (2017, p. 2) descrevem, as decisões algorítmicas não são "automaticamente justas por serem produtos de processos complexos". Isto é, a consistência procedimental dos algoritmos e a sua objetividade não necessariamente equivalem a inexistência de problemas éticos. Decisões humanas tampouco estão livre de vieses éticos e de falta de transparência. Entretanto, como defendem Kleinberg et al. (2018), as decisões automatizadas podem ser objetos de investigações mais profundas, a partir da implementação de regulação apropriada com foco na transparência procedimental e na certificação ética dos algoritmos.

Considerável esforço vem sendo empregado para demonstrar os potenciais danos provocados pelo uso de algoritmos de Inteligência Artificial em processos decisórios. Friedman e Nissenbaum já examinavam em 1996 a persistência de vieses em sistemas automatizados, que tanto podem causar como reproduzir múltiplas discriminações sociais com potencial de ofensa a direitos fundamentais. Como definem Fjeld et al. (2020, p. 27), "vieses algorítmicos são a sistemática sobre ou subrepresentação de probabilidades de uma população específica" que podem ser inseridas nos sistemas de IA de múltiplas formas. Nesse sentido, um sistema pode ser treinado com dados enviesados, falhos, ou em desacordo com a representação da realidade.

Esses vieses podem estar presentes em sistemas computacionais voltados à contratação de funcionários, à definição de rotas de voos, à concessão de benefícios sociais, ao reconhecimento facial para segurança pública, entre outros. Problemas relativos à obscuridade de algoritmos foram analisados profundamente por Citron e Pasquale (2014). Kleinberg et al. (2018), por sua vez, descreveram a aplicação defeituosa de inteligência artifical em processos seletivos de universidades e empresas. Citron (2019) analisou diversos problemas na automação de sistemas de órgãos públicos nos Estados Unidos – como aqueles relativos à concessão de crédito ou à assistência social – que, devido a defeitos nas bases de dados, e violando o devido processo legal, permitiram a retirada automática de benefícios, gerando ônus irreversíveis a assistidos por políticas sociais.

A respeito das discriminações identitárias (como as de gênero e de raça), Tarcízio Silva (2019, p. 5) observa que a "visão computacional" – a qual subsiste na coleta, análise e síntese de dados visuais e minimamente representativos por máquinas – pode ser responsável pela codificação de vieses. Em sua análise, o autor verificou esse tipo de enviesamento em diversos casos, como nos mecanismos de reconhecimento facial do *Google* que identificaram pessoas negras como "gorilas" e

cabelos de pessoas negras com perucas; casos de embranquecimento de pessoas pelo aplicativo *Faceapp* como critério de embelezamento; a associação de expressões faciais de pessoas negras como negativas por APIs; o não reconhecimento de gênero e de idade de mulheres negras por APIs, e de forma ainda mais grave, a maior chance de que pessoas negras sejam atropeladas por carros autônomos (SILVA, 2019).

Em resumo, os estudos apontam para a incorporação de vieses em decisões automatizadas em duas camadas possíveis: (1) na estruturação das bases de dados – isto é, na escolha das variáveis que serão consideradas; e (2) no conteúdo selecionado para a composição dessas bases de dados.

Para lidar com a discriminação e os vieses em decisões tomadas por sistemas autônomos, é necessário observar a construção dos *datasets*. Como observa Isaac (2018), a suposição de objetividade dos dados é falha, tendo em vista que todo comportamento humano ou fenômeno social que os algoritmos de aprendizado de máquina tentam prever provêm de um processo de geração de dados (DGP) que, por sua vez, é composto por interações sociais complexas. Nesse sentido, um modelo estatístico também pode projetar previsões e conclusões imprecisas e tendenciosas.

Conforme ressaltado anteriormente, a transparência e a explicabilidade são princípios essenciais para a contenção de vieses (FJELD et al., 2020). A transparência deve envolver os dados, o sistema e todo o modelo de negócios (AI HLEG, 2019, p. 18). Nesse sentido, uma certificação de boas práticas de desenvolvimento estaria vinculada à total transparência técnica, com a garantia de acesso ao código-fonte do sistema pelas autoridades públicas e a possibilidade de abertura do processo de desenvolvimento do sistema de IA. Já em relação à explicabilidade, Fjeld et al. (2020, p. 25) descrevem esse princípio como sendo a tradução de conceitos técnicos e decisões de uma forma inteligível e em formato compreensível para o escrutínio público. Transparência e explicabilidade, dessa forma, se complementam para assegurar a confiança no funcionamento justo dos sistemas de IA.

Para Fjeld et al. (2020, p. 48) é preciso observar a pluralidade na formação das equipes de desenvolvedores a fim de mitigar os vieses nos sistemas de IA. Nessa esteira, especialistas da Comissão Europeia afirmam que é necessário que componham as equipes mais mulheres e pessoas com trajetórias distintas, incluindo pessoas com deficiências psicomotoras (AI HLEG, 2019). Para o grupo de especialistas que redigiu o documento, com mais pessoas com identidades diversas sendo envolvidas nos processos de desenvolvimento de IAs – começando pela educação e treinamento desses sistemas, até a sua aplicação – a chance de que discriminações sejam perpetuadas diminuiria.

Finalmente, a participação da sociedade civil no processo de implementação dos sistemas de IA deve ser considerada para evitar potenciais malefícios. Nesse sentido, transparência e explicabilidade se alinham ao pluralismo de desenvolvimento e aplicação dos sistemas de IA para assegurar maior controle social e autodeterminação das pessoas que serão diretamente impactadas pela implementação das novas tecnologias.

Segundo Eixo: Governança de IA

1. Devem ser criadas estruturas institucionais voltadas ao desenvolvimento, aplicação e monitoramento de padrões éticos em IA, a exemplo do Centre for Data Ethics and Innovation do Reino Unido e do Automated Decision Systems Task Force de Nova Iorque?

A fim de assegurar a ética nos sistemas de IA, a formação de órgãos específicos – com intenção de monitoramento e implementação de regulação – tem sido uma medida visualizada como fundamental para a garantia de boas práticas no desenvolvimento e aplicação das novas tecnologias. Nesse sentido, organismos compostos por especialistas, que são capazes de compreender o funcionamento dos sistemas de IA, mitigar possíveis danos e evitar violações a direitos fundamentais e aos princípios democráticos, passam a ser uma forma de exercício de governança necessária e adequada diante do crescente uso da Inteligência Artificial na mediação de serviços.

Scherer (2016) observou que poucas pesquisas se dedicaram ao desenho institucional para a regulação de Inteligência Artificial. Para o autor, os mecanismos tradicionais de regulação, tais como licenciamento de produtos, supervisão de pesquisa e desenvolvimento e responsabilidade civil, não seriam suficientemente capazes de lidar com riscos associados ao uso da IA (SCHERER, 2016, p. 356). Isso, porque sua pesquisa e desenvolvimento podem ser discretos (com pouca estrutura física, por exemplo), difusos (com dezenas de indivíduos em localizações geográficas distintas participando de um mesmo projeto) e opacos (observadores externos podem não ser capazes de detectar os potenciais danos) (SCHERER, 2016, p. 369-73).

Tendo em vista a dificuldade de compreensão sobre a operação dos sistemas de IA, a literatura especializada tem visto com bons olhos a possibilidade de agências administrativas realizarem a regulação da IA. O principal ponto positivo é, justamente, a possibilidade de que esses órgãos sejam compostos por diversos agentes que possuam expertise sobre a técnica. Menciona-se também o caráter de flexibilidade e de independência que essas agências podem possuir, dependendo de seu desenho institucional, o qual pode incluir mandatos com tempo determinado para evitar a vinculação política e a captura por interesses privados.

Em relação ao papel a ser desempenhado pelas agências reguladoras, e a fim, principalmente, de se evitar a captura por interesses privados, Scherer (2016) propõe a criação pelo Poder Legislativo de uma espécie de "Artificial Intelligence Development Act", o qual seria uma base principiológica para que as agências possam exercer a regulação de maneira incisiva, com a emissão de certificados de confiança para os desenvolvedores de IA que cumpram certos requisitos. Em casos de danos materiais após a aplicação da IA, por exemplo, caberia aos tribunais avaliar se o sistema em questão possuía ou não tal certificado.

Para Scherer (2016), em vez de ser um órgão estritamente legislativo, judicial ou executivo, as funções de uma agência podem abranger aspectos desses três poderes. Em sentido similar, Tutt (2017) visualiza a criação de uma agência executiva como positiva para a regulação da IA.

Observando o funcionamento da Food and Drugs Administration (FDA) norte-americana, o autor verificou que os elementos fundamentais para o seu bom funcionamento são (1) o *know-how* do corpo técnico, que é capaz de compreender os potenciais riscos e mitigar danos na administração da saúde pública e; (2) a eficiência para responder a problemas iminentes de forma mais célere do que os demais poderes. Tutt (2017) também salienta que as indústrias farmacêutica e alimentícia trabalham com conceitos extremamente técnicos, semelhantes ao que se convencionou chamar de "*black box*" em referência ao funcionamento dos sistemas de IA. Um órgão executivo composto por especialistas que possuam capacidade de compreender a fundo as problemáticas técnicas, nesse sentido, é encarada como essencial para a gestão de possíveis crises que podem emergir do desenvolvimento e implementação dos sistemas de IA.

A criação de autoridades específicas para lidar com os sistemas de IA pode ser positiva, na medida em que essas agências podem compreender (1) a capacidade judicial de lidar com reivindicações privadas de forma mais eficiente e menos morosa; (2) a capacidade do executivo de executar decisões; (3) a capacidade de editar e definir comportamentos a partir da elaboração de normas e portarias. Desse modo, as agências seriam autoridades fundamentais para influenciar os processos de desenvolvimento e aplicação de sistemas de IA, seja de maneira sutil – por meio da coleta e publicação de informações relevantes sobre os riscos à segurança de um setor específico –, seja pela aplicação de regras ou princípios legislados.

Entretanto, existem alguns problemas relacionados ao desenho institucional dessas agências que devem ser considerados. Esses órgãos podem ter uma alta capacidade de edição normativa, acesso a informações privilegiadas cujo risco de captura por interesses particulares não deve ser desconsiderado e também a capacidade judicial de resolver conflitos com discricionariedade. Nesse sentido, o controle da atuação dessas agências deve ser pensado com grau elevado de detalhamento, a fim de assegurar o bom funcionamento de mecanismos de freios e contrapesos.

Para além dos esforços nacionais em torno do desenvolvimento de políticas de IA, Erdélyi e Goldsmith (2018) propõem a criação de uma estrutura reguladora internacional para evitar os riscos decorrentes de possíveis interpretações imperfeitas de princípios pelas agências domésticas. Nesse contexto, uma nova organização intergovernamental — que poderia ser chamada de Organização Internacional de Inteligência Artificial (IAIO) — serviria como um fórum transnacional para a discussão e definição de padrões comportamentais a serem seguidos pelos desenvolvedores dos sistemas de IA e pelo poder público. Essa organização deve, na proposta dos autores, unir um grupo diversificado e interdisciplinar de pessoas interessadas do setor público, da indústria e da academia (ERDÉLYI; GOLDSMITH, 2018).

Terceiro Eixo: Aplicação no Poder Público

1. ***Há necessidade de salvaguardas específicas nos processos de tomada de decisão no poder público envolvendo sistemas de IA? Em quais circunstâncias ou em quais áreas?***

Sistemas de IA têm grande potencial de colaboração para o desenvolvimento de políticas públicas, bem como para assegurar o seu monitoramento. Entretanto, é necessário o estabelecimento de princípios gerais que orientem a transformação digital do governo. Esses princípios podem variar e ser ponderados de acordo com outros valores presentes nas áreas específicas de aplicação dos sistemas de IA. No Plano de Governo Digital uruguaio para o biênio de 2018-2020 (AGESIC, 2020), por exemplo, foram estabelecidos nove princípios gerais que serviriam como guias para a implementação de sistemas de IA no poder público:

1. **Finalidade**: a incorporação de IA ao poder público deve ter como referência a potencialização das capacidades do ser humano.
2. **Interesse Geral**: o poder público não pode visar a satisfação de interesses particulares, desconsiderando o interesse público. Nesse sentido, a incorporação de IA deve ser realizada de forma a garantir a inclusão e a equidade geral. Esse princípio também abarca o repúdio às práticas discriminatórias e à reprodução de vieses pelos sistemas de IA.
3. **Respeito aos direitos humanos**: sistemas de IA devem respeitar os direitos fundamentais, como liberdades individuais, igualdade e diversidade. Dessa forma, a potencial realização de práticas discriminatórias em sistemas de IA deve ser combatida pelo poder público.
4. **Transparência**: os sistemas de IA devem possuir transparência e assegurar a inteligibilidade de seu funcionamento (incluindo algoritmos, dados, provas e validações realizadas, além da lista de quais áreas utilizam tais tecnologias), de acordo com a norma vigente de acesso à informação.
5. **Responsabilidade**: é necessária a existência de um responsável, claramente identificável, que responda pelas ações provenientes dos problemas no sistema de IA.
6. **Ética**: quando a aplicação da IA ou seu desenvolvimento implicar em dilemas éticos, esses devem ser abordados e revisados por seres humanos.
7. **Valor agregado**: soluções respaldadas em sistemas de IA devem ser utilizadas apenas quando, de maneira comprovada, sejam capazes de trazer mais eficiência ao processo.
8. **Privacidade por design**: o design das soluções de IA deve contemplar preocupações com a proteção da intimidade das pessoas.
9. **Segurança**: o desenvolvimento da IA deve cumprir com os princípios básicos de segurança da informação.

A Comissão Europeia publicou relatório em abril de 2019, no qual elencou orientações éticas para uma IA de confiança (AI HLEG, 2019). Em suma, o documento traz princípios que já são considerados pelos ordenamentos jurídicos dos Estados-Membros da União Europeia tais como

dignidade humana, liberdade, democracia, igualdade e direitos dos cidadãos perante a administração pública. Diante do quadro jurisprudencial, o desenvolvimento de sistemas de IA deve respeitar a autonomia humana, a prevenção de danos, a equidade e observar a transparência e a explicabilidade.

O relatório da Comissão Europeia chama a atenção para o fato de que todos os governos em democracias constitucionais devem ser limitados pela lei – seguindo o princípio do Estado de Direito (*rule of law*). Nesse sentido, os sistemas de IA devem servir para manter ou expandir os processos democráticos, além de respeitar a pluralidade dos valores e das escolhas individuais. Como princípio que deve ser observado, o relatório aponta para a prevenção de danos pelos sistemas de IA. Isso significa que estes não podem causar ou exacerbar danos à sociedade civil. A proteção da dignidade humana e da integridade física, portanto, deve ser pressuposto fundamental para a incorporação de sistemas de IA ao poder público (AI HLEG, 2019).

Como mecanismos de controle, além da possibilidade de criação de estruturas institucionais específicas voltadas ao desenvolvimento, à aplicação e ao monitoramento de padrões éticos de IA, de acordo com os especialistas da Comissão Europeia, a supervisão de IA pode ser também realizada de três formas, com as abordagens *Human–In–The–Loop*, *Human-On-The-Loop* e *Human-In-Command*. A primeira diz respeito à capacidade de intervenção humana em todos os ciclos de decisão do sistema. Já a segunda diz respeito à capacidade de intervenção humana durante o ciclo de design do sistema e monitoramento da operação do sistema. Por fim, a terceira se refere à capacidade de supervisionar a atividade geral do sistema de IA (incluindo seu impacto econômico, social, jurídico e ético mais amplo) (AI HLEG, 2019).

Nesse sentido, é possível incluir a decisão de (1) não usar um sistema de IA em uma situação específica, (2) estabelecer níveis de intervenção humana durante o uso do sistema, ou (3) garantir a capacidade de substituir uma decisão tomada por um sistema. Além disso, o relatório aponta para a garantia de que os agentes públicos possam ter capacidade de exercer a supervisão de acordo com o mandato estabelecido. O documento aborda também as possíveis vulnerabilidades dos sistemas de IA que, como quaisquer sistemas de software, podem ser alvos de ataques cibernéticos – com práticas como data poisoning em machine learning, que podem alterar o comportamento do sistema de IA levando o sistema a tomar decisões "erradas" (AI HLEG, 2019).

A aplicação de sistemas de IA no Judiciário brasileiro já é uma realidade com a implementação do software VICTOR, por exemplo, o qual é treinado para ler recursos extraordinários no Supremo Tribunal Federal que já foram temas de repercussão geral. A aplicação desse tipo de software tem como intenção colaborar para a maior eficiência e celeridade do tribunal. Os ganhos de sistemas afins são múltiplos, mas cabe ressaltar que a cooperação de outras instituições — como, por exemplo, o Conselho Nacional de Justiça — se faz necessária para assegurar que o funcionamento deste tipo de sistema não contradiga direitos, nem que a objetividade da IA impeça a atualização normativa com a reprodução exacerbada de precedentes.

Falhas na segurança podem também ter como resultado danos físicos. Para casos de risco alto à vida social, bem como para os casos em que há menor capacidade de supervisão por seres humanos – por exemplo para sistemas que envolvem dados pessoais sensíveis –, mais testes de certificação são necessários e mais rigorosa deve ser a governança institucional sobre o sistema de IA operado. Nesse sentido, mecanismos de supervisão podem ser implementados em múltiplas instâncias com a intenção de apoiar outras medidas de segurança e de controle – também dependendo da área de aplicação do sistema de IA e do risco potencial apresentado.

Quarto Eixo: Segurança Pública

Neste eixo, foram respondidas as seguintes perguntas:

1. **Quais são os métodos e técnicas que podem ser usados para incentivar o desenvolvimento de sistemas de IA seguros e confiáveis?**

2. **De que maneira pode-se apoiar esforços para criar métricas para avaliar a segurança, a proteção e a confiabilidade das aplicações da ciência e tecnologia em relação à inteligência artificial?**

3. **Quais salvaguardas, critérios e cuidados devem ser adotados na utilização de IA no campo da segurança?**

1. Quais são os métodos e técnicas que podem ser usados para incentivar o desenvolvimento de sistemas de IA seguros e confiáveis?

De acordo com o Grupo de Especialistas de Alto Nível sobre Inteligência Artificial (AI HLEG, 2019, p. 36, tradução nossa), os sistemas de Inteligência Artificial (IA) são sistemas de software e/ou hardware concebidos por seres humanos para atingir um objetivo a partir da interpretação de dados estruturados ou não[14].

Os sistemas de IA utilizam técnicas de aprendizado automatizado que, por meio de padrões e da construção de inferências, aprendem a concluir uma determinada tarefa sem serem explicitamente programadas para isso (OECD, 2019b, p. 27; WEBFOUNDATION, 2017, p. 5). As técnicas por trás das aplicações de IA podem variar de acordo com o objetivo a ser alcançado e seu uso responsável deve considerar essa diferença, uma vez que a busca por maior precisão nos resultados pode representar um limite à transparência (MARGULIES, 2016, p. 1050), requisito essencial a ser observado nesses sistemas.

O paradoxo entre transparência e precisão torna-se mais evidente quando analisamos os métodos primários de aprendizado de máquina que podem ser utilizados nas aplicações de IA, por exemplo, as árvores de decisão, as redes neurais artificiais (RNAs; em inglês, Artificial Neural Network - ANN) e as máquinas de vetores de suporte (MVSs; em inglês, Support Vector Machines -

[14] Trecho original: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions." (AI HLEG, 2019, p. 36).

SVMs)[15]. Enquanto as árvores de precisão têm maior grau de transparência, elas são menos precisas em relação aos outros dois métodos (RNAs e MVSs) cuja complexidade do processo de inferência fornece resultados mais apurados, porém inteligíveis à compreensão humana (MARGULIES, 2016, p. 1064).

A capacidade de aprender autonomamente é um benefício das RNAs em comparação às árvores de decisão, por exemplo. As RNAs podem discernir autonomamente padrões em fotos e gravações, sendo bastante aplicadas nas tecnologias de reconhecimento facial. Por outro lado, sua estrutura altamente não linear - resultado da conexão de outputs e inputs em diferentes níveis, por meio de camadas ocultas[16] - dificulta, ou quase impossibilita, a interpretação e explicação humana sobre os pesos que são dados pela máquina para cada variável (MARGULIES, 2016, p. 1067-72).

Do mesmo modo, as camadas ocultas são um recurso das MVSs, utilizadas em processamento de linguagem natural e na identificação de padrões de encriptação em comunicações, por exemplo. A derivação das variáveis em mais dimensões também faz com que essa tecnologia não seja capaz de proporcionar uma justificativa verbal inteligível para seres humanos acerca de seus resultados (MARGULIES, 2016, p. 1067).

Estudos sobre o desenvolvimento de estratégias de aprendizado de máquina mais justas, responsáveis e transparentes, argumentam por uma combinação de abordagens técnicas e não técnicas para lidar com essa questão. Entre as *abordagens estatísticas e algorítmicas*, citadas por Osoba e Welser IV (2017, p. 21-2) por exemplo, estão o uso de métricas modificadas de distância ou similaridade para analisar dados de um determinado indivíduo, impondo padrões de justiça rigorosos na comparação com outros conjuntos de informações; desenvolvimento de métodos para auditoria dos algoritmos e comparação entre os resultados e o comportamento previamente esperado; modelos de aprendizado de máquina com independência estatística entre os resultados e as variáveis protegidas; e a introdução de métricas sociais de otimização para avaliação dos algoritmos e correção estatística das disparidades.

No mesmo sentido, Koene et al. (2019, p. 32-3) destacam a necessidade de revisão dos dados utilizados para alimentar o sistema e dos resultados obtidos a fim de verificar se os objetivos propostos são atendidos sem discriminações a indivíduos ou grupos sensíveis.

Avanços nas pesquisas de *algoritmos de raciocínio causal* podem ainda contribuir para tornar as auditorias dos sistemas mais confiáveis. Especialmente quando os resultados são estatisticamente desproporcionais, essa técnica ajuda na apresentação dos fatores que causaram essas decisões, melhorando seu processo de verificação/validação (OSOBA; WELSER IV, 2017, p. 22-3). As

---

[15] As árvores de decisão são modelos estatísticos úteis para realização de previsões e verificações de causalidade, fornecendo respostas simples, diretas e de mais fácil compreensão para os seres humanos (MARGULIES, 2016, p. 1064). Nas RNAs, que tentam imitar o funcionamento do cérebro humano, os dados são adicionados numa camada de entrada, em seguida são modificados na(s) camada(s) oculta(s) para alcançar determinados resultados nas camadas de saída, baseados nos pesos aplicados aos nós de processamento interconectados na rede (HURWITZ; KIRSCH, 2018, p. 17). Já As MVs permitem a identificação pelo sistema de variáveis análogas, que não estavam entre os dados utilizados para treinar o algoritmo, mapeando diversas características ao mesmo tempo (MARGULIES, 2016, p. 1067).

[16] O aprendizado profundo (deep learning), considerado uma subdivisão do aprendizado de máquina, é uma técnica que usa redes neurais hierárquicas para aprender com uma combinação de algoritmos não supervisionados e supervisionados e, embora seja muito semelhante a uma rede neural tradicional, tem muito mais camadas ocultas (HURWITZ; KIRSCH, 2018, p. 18). Nesse caso, as tarefas são divididas e distribuídas em algoritmos organizados em camadas consecutivas. Cada camada se acumula na saída da camada anterior e, juntas, constituem uma rede neural artificial que imita a abordagem distribuída da solução de problemas realizada pelos neurônios no cérebro humano (WEB FOUNDATION, 2017, p. 5).

revisões de código são igualmente métodos utilizados para aumentar a confiabilidade de um sistema em desenvolvimento e garantir que ele atenda a determinados requisitos (KOENE et al., 2019, p. 31; 34).

O aprimoramento dos sistemas de IA e o combate ao problema do viés algorítmico podem ser feitos também pela combinação entre *alfabetização algorítmica e transparência*. Deve haver cada vez mais uma transparência perante o público sobre quais decisões e ações são mediadas por agentes artificiais. Além disso, o *aspecto humano* envolvido na concepção dos sistemas de IA precisa ser aprimorado e diversificado. Equipes interdisciplinares trabalhando no desenvolvimento desses sistemas de IA ajudariam em um processo de concepção, e em resultados algorítmicos, mais éticos, compreensíveis, seguros e confiáveis (AI HLEG, 2019, p. 23; OSOBA; WELSER IV, 2017, p. 23-4).

Além de medidas semelhantes às expostas acima, para contribuir com a concretização de um sistema de IA confiável e seguro, o grupo de especialistas da Comissão Europeia (AI HLEG) propõe que esses requisitos sejam traduzidos e incorporados em restrições ou procedimentos na própria arquitetura do sistema e que eles sejam éticos "by design". Métodos de explicação, adoção de indicadores de qualidade e certificações externas das aplicações de IA também são recomendações dos especialistas (AI HLEG, 2019, p. 21-3).

Outras sugestões são a realização de testes antagônicos por *red teams* que tentariam deliberadamente invadir o sistema para encontrar suas vulnerabilidades e o oferecimento de *bug bounties* como incentivo a terceiros pela indicação de erros e falhas do sistema de forma ética e responsável (AI HLEG, 2019, p. 22).

Por fim, além dos parâmetros técnicos, ressalta-se a importância dos sistemas de IA observarem os princípios recomendados pela Organização para a Cooperação e Desenvolvimento Econômico (OECD, 2019a).

2. *De que maneira pode-se apoiar esforços para criar métricas para avaliar a segurança, a proteção e a confiabilidade das aplicações da ciência e tecnologia em relação à inteligência artificial?*

Em um contexto de desenvolvimento e adoção de tecnologias de IA é importante considerar medidas que sejam adequadas e eficientes para atender os objetivos de implantação dos sistemas, respeitando direitos e garantias fundamentais previstos na legislação nacional e internacional.

É essencial que políticas nacionais sejam implantadas de maneira apropriada e busquem alcançar resultados benéficos para os cidadãos por meio de investimentos públicos e privados em tecnologia e capacitação humana para lidar com a evolução da IA. O investimento em pesquisas públicas, por exemplo, é uma maneira de incentivar e guiar a inovação em IA, bem como a estruturação de um ecossistema de tecnologia e infraestrutura apropriados à criação, testes e adesão desses novos sistemas (OECD, 2019b, p. 100-1).

A elaboração de uma estratégia sólida é primordial para garantir a observância dos princípios norteadores das aplicações de IA e para que o processo seja seguro, transparente e profícuo. Algumas alternativas, citadas pela OECD (2019b, p. 30) são: suporte a melhores conjuntos de dados de treinamento dos sistemas de IA; financiamento para pesquisa acadêmica e ciências básicas; políticas de incentivo a integração interdisciplinar entre as habilidades em IA e outras competências; e educação em computação.

Margulies (2016, p. 1052-4) sugere que os resultados obtidos por decisões de sistemas de IA passem por técnicas de validação humana. Para o autor, empregar esforços na criação de um tribunal de revisão de decisões algorítmicas e em um órgão independente são formas de garantir legitimidade e que direitos fundamentais sejam observados nos resultados. Conselhos revisores são necessários para verificar ocorrências discriminatórias realizadas, por exemplo, pelos sistemas de reconhecimento facial em tempo real.

As soluções expostas por Margulies (2016, p. 1052-4) são medidas a serem adotadas após decisões problemáticas, ou seja, de caráter mais repressivo e de contenção dos danos. Instrumentos de verificação prévia e planos preventivos, entretanto, são ainda mais essenciais. Desse modo, a criação de conselhos avaliadores das aplicações e algoritmos antes de sua implementação/entrada no mercado ou a produção de relatórios de supervisão por organizações da sociedade civil, pesquisadores ou instituições públicas são algumas ideias na literatura sobre o tema (ISAAC, 2018, p. 557-8).

Além da revisão prévia, o monitoramento contínuo também é um desafio com o qual o Estado deverá lidar. Em caso de incapacidade do poder público de desenvolver e manter seus próprios sistemas de IA, a tendência é que ele adquira produtos de terceiros (KLEINBERG et al., 2018). Caso isso ocorra, não poderá medir esforços para realizar atualizações periódicas do sistema, uma vez que predições baseadas em condições sociais passadas, além de proverem soluções que podem até mesmo não se aplicar mais à realidade, podem reproduzir cenários discriminatórios para grupos historicamente marginalizados.

Sob uma perspectiva voltada para a segurança pública, para minimizar os impactos negativos do policiamento preditivo, por exemplo, a integração da polícia local com a comunidade, somada à observação de suas necessidades reais para elaboração da estratégia de segurança, são essenciais. O cuidado com os incentivos policiais contribui para que a própria atuação individual desses agentes

não seja um fator para reforçar o enviesamento dos dados: não considerar o número de prisões por cada agente para fins de promoção, mas sim a produção de relatórios sobre o contexto onde ele esteve atuando. Como na política canadense (GOVERNMENT OF CANADA, 2015), é possível fazê-lo por meio de intercâmbios de informações com demais instituições do poder público para que a polícia tenha potencial de ser um canal de acesso a outros serviços sociais para pessoas em situações de risco. A alteração dos incentivos é um remédio igualmente aplicável a alguns dos riscos que surgem com a aplicação de sistemas de reconhecimento facial em tempo real (ISAAC, 2018, p. 554).

Cumpre ressaltar ainda a questão do acesso e compartilhamento de dados pessoais, aspecto de especial importância que pode acelerar ou retardar o progresso da IA dependendo da abordagem adotada pelo governo em relação ao tema (OECD, 2019b, p. 101-3). No Brasil, especificamente para questões de segurança pública, a Lei nº 13.709/18 (Lei Geral de Proteção de Dados - LGPD) não é aplicável, conforme alínea a do inciso III, artigo 4º (BRASIL, 2018). Apesar disso, o parágrafo 1º desse mesmo artigo afirma que a legislação específica deverá prever medidas proporcionais e estritamente necessárias ao atendimento do interesse público, além de observar o devido processo legal, os princípios gerais de proteção de dados e os direitos do titular dispostos na própria LGPD.

Entre os princípios da LGPD estão o respeito à privacidade, à inviolabilidade da intimidade, da honra e da imagem e a liberdade de expressão (BRASIL, 2018). Isso merece especial atenção durante o debate sobre a adoção de sistemas de IA como os expostos acima, já que tais medidas podem representar ameaças a esses princípios da LGPD e outros direitos fundamentais dos indivíduos.

Sugere-se, para manter o mínimo controle sobre possíveis violações à privacidade a fim de evitá-las, o uso de tecnologias públicas, desenvolvidas em território nacional por instituição de ensino e pesquisa, em detrimento de sistemas proprietários de terceiros cujos interesses podem estar desalinhados aos das políticas públicas nacionais. Isaac (2018, p. 554-5) aventa a possibilidade do emprego de tecnologias de *blockchain*[17] para garantir maior respeito à privacidade, sem violar os dados pessoais que constam nos datasets.

[17] "A blockchain is a digital, secure, public record book of transactions (a ledger). "Block" describes the way this ledger organizes transactions into blocks of data, which are then organized in a "chain" that links to other blocks of data." (GALEN et al., 2018).

3. Quais salvaguardas, critérios e cuidados devem ser adotados na utilização de IA no campo da segurança?

Em diversos âmbitos da segurança pública a aplicação de tecnologias de Inteligência Artificial (IA) está ganhando espaço: na segurança nacional, por meio de controle de comunicações em caso de "iminente perigo" (MARGULIES, 2016); na segurança local, onde a IA é empregada para prever estatisticamente em quais locais crimes são mais prováveis (ISAAC, 2018) e o uso de tecnologias de reconhecimento facial em tempo real para identificar correspondências biométricas de pessoas que tenham cometido crimes (FUSSEY; MURRAY, 2019).

Existem ainda outras aplicações, de certo modo até mais sofisticadas, como os policiais robôs (JOHN, 2016), mas o foco será expor alguns riscos e cuidados diante dos três exemplos acima. Todos esses sistemas possuem um objetivo em comum, qual seja, a alocação dos escassos recursos do poder público a fim de otimizar resultados e reduzir custos. Entretanto, veremos a seguir que todos também têm um grande potencial violador de direitos e garantias fundamentais.

Tecnologias de policiamento preditivo são empregadas para identificar por meio de predições estatísticas possíveis regiões ou indivíduos que representam um maior risco à segurança pública. Seja para justificar uma intervenção policial intensiva, atos de intimação ou como instrumento de apoio para resolução de crimes, o emprego desses instrumentos abre margem para padrões discriminatórios. Ao realizar uma revisão de estudos sobre policiamento preditivo, Isaac (2018) evidencia a incorporação de vieses em sistemas dessa natureza, os quais refletem as diferenças de tratamento por parte da justiça criminal.

Um dos problemas dessas técnicas preditivas é o efeito catraca gerado (no que diz respeito à reprodução de discriminações), pois a atuação policial direcionada por previsões algorítmicas também será usada para alimentar o sistema de policiamento preditivo, aumentando o índice de dados enviesados. A partir do reforço do contingente policial (baseado nos resultados algorítmicos enviesados) em bairros cujo histórico era de policiamento mais ostensivo, o resultado dos próximos ciclos também vai reproduzir os mesmos resultados discriminatórios, já que os dados dizem respeito sempre às mesmas regiões (ISAAC, 2018).

Isaac (2018) pontua que a conjuntura política e fatores demográficos influenciam no enviesamento dos dados desses sistemas. O baixo índice de denúncias de crimes é um dos componentes dessa conjuntura: os registros de ocorrência refletem a relação entre a criminalidade, a estratégia policial e as relações entre comunidade e polícia. Quando esses aspectos não são levados em consideração, o autor destaca: "(...) instrumentos de predição (...) poderão simplesmente perpetuar discriminações históricas contra grupos sub-representados, e violar seus direitos humanos e civis"[18] (ISAAC, 2018, p. 546, tradução nossa).

O efeito catraca leva a um problema de efetividade, pois ele faz com que dificilmente os locais que precisam receber mais policiamento sejam alterados ao longo do tempo. Com isso, novos focos de criminalidade não aparecerão nas previsões dos sistemas, reduzindo a responsividade e eficiência da tecnologia em relação às mudanças reais das condições sociais (ISAAC, 2018, p. 543). Desse modo, antes de adotar a tecnologia em larga escala, estudos de casos devem ser realizados

[18] Trecho original: "Theses predictive tools (...) may simply perpetuate historical discrimination toward underrepresented groups and violate their civil and human rights." (ISAAC, 2018, p. 546).

para verificar a eficiência dessa política na redução dos índices de criminalidade, especialmente no Brasil, onde há grande desigualdade social e baixo índice de registros de determinados crimes de acordo com os valores sociais predominantes.

Além dos cuidados inerentes ao processo de coleta dados criminais, considerando que os registros devem ser relativizados como fenômenos político-sociais, Kleinberg et al. (2018) sugerem que é indispensável uma legislação clara acerca da discriminação algorítmica, para que os resultados enviesados sobre determinados grupos sejam devidamente combatidos.

Um estudo da Web Foundation (2018) debateu, entre outros casos, a implementação pelo governo do Uruguai do sistema preditivo PredPol. Esse sistema define e atualiza pontos cruciais em um mapa para realocação dos recursos policiais com base em previsões algorítmicas. Com relação a eficácia do modelo, nenhuma informação oficial foi encontrada sobre a taxa de erro do PredPol e a substituição por uma ferramenta estatística básica sugeriu que a sua eficácia é menor do que a esperada. Quanto a eficácia da implementação, detectou-se uma redução de crimes nas áreas em que foi implementado, mas não em termos gerais. A ferramenta simplesmente ocasionou um deslocamento das forças policiais para pontos críticos. Como os dados do sistema não são públicos e os resultados não são explicáveis, isso compromete a legitimidade do sistema. Sobre a validade dos resultados, verificou-se o risco de discriminação, o que pode ser explicado no âmbito dos sistemas preditivos pela replicação de vieses nos dados de treinamento e a dinâmica histórica de poder entre a aplicação da lei e as minorias ou populações desfavorecidas, que é usada para justificar a ação da polícia (WEB FOUNDATION, 2018, p. 26-30).

As *tecnologias de reconhecimento facial*, por sua vez, permitem o processamento biométrico para identificar indivíduos particulares criando uma assinatura digital com base nas informações dos rostos identificados, para posterior análise de correspondências com informações contidas em uma base de dados. Sistemas de reconhecimento facial podem ser utilizados em blitz móveis, locais fixos e até mesmo nos uniformes de policiais. Nesse caso, a primeira forma de violação de direitos civis é a vigilância demasiadamente intrusiva sobre a privacidade dos indivíduos (FUSSEY; MURRAY, 2019, p. 19-20).

Para evitar maiores problemas, Fussey e Murray (2019) sugerem que todas as hipóteses de treinamento dos algoritmos de reconhecimento facial em ambientes simulados sejam esgotadas antes de implementá-los nos espaços para o grande público. Quando o teste é concomitante à implementação em larga escala inexiste a figura do consentimento para participação, o que abre margem para confusões em relação à natureza da aplicação do sistema, ou seja, se de fato são testes das tecnologias ou se são operações policiais oficiais com o intuito de identificar potenciais suspeitos. Para os autores, é necessário que o poder público institucionalize uma metodologia clara acerca das técnicas que devem ser empregadas nos sistemas de reconhecimento facial (FUSSEY; MURRAY, 2019, p. 28).

No estudo realizado por Fussey e Murray (2019, p. 31) sobre a aplicação de tecnologias de reconhecimento em tempo real em Londres, a própria Comissão de Câmeras de Vigilância (em inglês, Surveillance Camera Commissioner), entrevistada pelos autores, apontou que é imprescindível a realização de análises de riscos operacionais, de impacto às comunidades e de violações à privacidade e outros direitos humanos antes da aplicação desses instrumentos (FUSSEY; MURRAY, 2019, p. 33).

Já existem trabalhos demonstrando que minorias raciais e mulheres são especialmente afetadas por erros em sistemas de reconhecimento facial. Taxas de falsos positivos de 40% para pessoas não-brancas e de 5% para pessoas brancas foram identificadas na ferramenta Amazon

Rekognition, por exemplo (WHITTAKER et al., 2018, p. 15-6). Outra pesquisa (GROTHER; NGAN; HANAOKA, 2019), realizada pelo National Institute of Standards and Technology (NIST), avaliou 189 algoritmos de 99 desenvolvedores de reconhecimento facial para medir as ocorrências de falsos positivos e falsos negativos. Um falso negativo significa que o sistema não identifica que duas fotos mostram a mesma pessoa; os falsos positivos, entretanto, ocorrem quando imagens de duas pessoas diferentes são consideradas como representativas do mesmo indivíduo, o que pode resultar em acusações falsas. Verificou-se nos algoritmos norteamericanos uma taxa mais alta de falsos positivos para rostos asiáticos, negros e indígenas em relação a pessoas brancas. Mulheres negras são o grupo mais atingido, segundo o estudo (GROTHER; NGAN; HANAOKA, 2019).

Percebe-se que no campo da segurança pública, especialmente, o enviesamento é agravado devido às relações históricas de desigualdade existentes, as quais comprometem a alimentação das bases de dados de treinamento dos sistemas e sujeita grupos socialmente vulneráveis às decisões automatizadas amplamente discriminatórias (SILVA, 2019).

Sobre as *análises de comunicações interpessoais*, para que determinadas buscas sejam realizadas, violando o direito à privacidade dos indivíduos, uma justificativa razoável para fazê-lo deve ser fornecida pelo Estado, bem como sobre os métodos utilizados para alcançar tal objetivo. A adoção de técnicas de IA com camadas ocultas inerentes ao seu funcionamento colocam em risco a transparência e legitimidade dos resultados atingidos, tendo em vista que a precisão tem como ônus a dificuldade dos seres humanos em explicar substancialmente como o raciocínio ocorreu em cada uma das camadas do sistema (MARGULIES, 2016, p. 1067-72).

Uma vez que pesquisas autônomas estejam condicionadas à validação, o operador do sistema pode oferecer uma explicação metodológica para isso, minimizando o problema. Nesse cenário, justificativas metodológicas (de validação) podem preencher o vácuo deixado pela impossibilidade de justificativas precisas acerca dos parâmetros (e seus respectivos pesos) identificados pela máquina (MARGULIES, 2016, p. 1069).

Bibliografia

AGESIC. *Principios generales sobre Inteligencia Artificial para un Gobierno Digital*. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento (AGESIC). Disponível em: https://bit.ly/2Siy4tO. (Último acesso em: 10 fev. 2020).

AI HLEG. *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on Artificial Intelligence set up by European Commission, 2019. Disponível em: https://bit.ly/2GenWeu. (Último acesso em: 10 fev. 2020).

BECKER, Daniel; FERRARI, Isabela. O DIREITO À EXPLICAÇÃO SOBRE DECISÕES AUTOMATIZADAS: UMA ANÁLISE COMPARATIVA ENTRE A UNIÃO EUROPEIA E O BRASIL. *Revista de Direito e as Novas Tecnologias*, vol. 1, out./dez. 2018.

BOTH, Göde. *What drives research in self-driving cars?* (Part 2: Surprisingly not machine learning). Platypus, The CASTAC Blog, abr. 2014. Disponível em: https://bit.ly/2Spsjt9. (Último acesso em: 10 fev. 2020).

BRASIL. *Lei nº 13.709 de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: https://bit.ly/2GrnTML. (Último acesso em: 10 fev. 2020).

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, [s.l.], v. 3, n. 1, p.1-12, 5 jan. 2016. SAGE Publications. Disponível em: https://bit.ly/2GQcvdr. (Último acesso em: 10 fev. 2020).

CITRON, Danielle Keats; PASQUALE, Frank. The scored society: Due process for automated predictions. *Wash. L. Rev.*, [s.l.], vol. 89, n. 1, 2014. Disponível em: https://bit.ly/36WOQ5. (Último acesso em: 10 fev. 2020).

CITRON, Danielle Keats. Technological Due Process. *Washington Law Review*, 2019. Disponível em: https://openscholarship.wustl.edu/cgi/viewcontent.cgi?article=1166&context=law_lawreview. (Último acesso em: 10 fev. 2020).

DANAHER, John. *Algorithmic Decision-Making and the Problem of Opacity*. Tech Law for Everyone, ago. 2016. Disponível em: https://bit.ly/2Omaeei. (Último acesso em: 10 fev. 2020).

DIAKOPOULOS, Nicholas. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Tow Center for Digital Journalism*, 2013. Disponível em: https://bit.ly/31liG2k. (Último acesso em: 10 fev. 2020).

DOSHI-VELEZ, Finale; KORTZ, Mason. *Accountability of AI under the law*: the role of explanation. Berkman Klein Center Working Group on Explanation and the Law, 2017. Disponível em: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584. (Último acesso em: 10 fev. 2020).

ETZIONI, Amitai; ETZIONI, Oren. Incorporating Ethics into Artificial Intelligence. *J Ethics*, [s.l.], vol. 21, n. 4, p. 403-418, 2017. Disponível em: https://bit.ly/2OlV1JT. (Último acesso em: 10 fev. 2020).

FJELD, Jessica; ACHTEN, Nele; HILLIGOSS, Hannah; NAGY, Adam; SRIKUMAR, Madhulika. *Principled Artificial Intelligence*: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society, 2020. Disponível em: https://dash.harvard.edu/handle/1/42160420. (Último acesso em: 10 fev. 2020).

FRIEDMAN, Batya; NISSENBAUM, Helen. Bias in Computer Systems. *ACM Transactions on Information Systems*, vol. 14, n. 3, jul. 1996, p. 330-347. Disponível em: https://vsdesign.org/publications/pdf/64_friedman.pdf. (Último acesso em: 10 fev. 2020).

FUSSEY, Pete; MURRAY, Daragh. *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology*. The Human Rights, Big Data and Technology Project - Human Rights Center, University of Essex, 2019. Disponível em: https://bit.ly/36vQ3AP. (Último acesso em: 10 fev. 2020).

FUTURE OF LIFE INSTITUTE. *Princípios de Asilomar*, 2017. Disponível em: https://futureoflife.org/ai-principles/. (Último acesso em: 10 fev. 2020).

GALEN, Doug et al. *Blockchain for Social Impact*: Moving Beyond the Hype. Stanford Business Center for Social Innovation e RippleWorks, 2018. Disponível em: https://stanford.io/3aOhRUn. (Último acesso em: 10 fev. 2020).

GOVERNMENT OF CANADA. *Economics of Policing and Community Safety*: Policy Makers' Dialogue on Privacy and Information Sharing, 2015. Disponível em: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/2015-plcy-mdps/index-en.aspx. (Último acesso em: 10 fev. 2020).

GROTHER, Patrick; NGAN, Mei; HANAOKA, Kayee. *Face recognition vendor test part 3*: Demographic Effects. National Institute of Standards and Technology, dez. 2019. Disponível em: https://bit.ly/2U1I0co. (Último acesso em: 10 fev. 2020).

HUNT, Elle. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, 24 mar. 2016. Disponível em: https://bit.ly/2S5FYFe. (Último acesso em: 10 fev. 2020).

HURWITZ, Judith; KIRSCH, Daniel. *Machine Learning for dummies* - IBM Limited Edition. Hoboken: John Wiley & Sons, Inc., 2018. Disponível em: https://www.ibm.com/downloads/cas/GB8ZMQZ3. (Último acesso em: 10 fev. 2020).

ISAAC, William S. Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice. *Ohio State Journal of Criminal Law*, vol. 15, n. 2, 2018, p. 543-558. Disponível em: https://kb.osu.edu/handle/1811/85814. (Último acesso em: 10 fev. 2020).

JOHN, Elizabeth E. Policing Police Robots. *UCLA L. Rev. Disc*, 5160, 2016. Disponível em: http://euro.ecom.cmu.edu/program/law/08-732/AI/Joh.pdf. (Último acesso em: 10 fev. 2020).

KLEINBERG, Jon; LUDWIG, Jens; MULLAINATHAN, Sendhil; SUSTEIN, Cass R. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, vol. 10, p. 113-174, Oxford, 2018. disponível em: https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086. (Último acesso em: 10 fev. 2020).

KOENE, Ansgar et al. *A governance framework for algorithmic accountability and transparency*. Panel for the Future or Science and Technology (STOA); European Parliamentary Research Service (EPRS), abril 2019. Disponível em: https://bit.ly/3aEZFfM. (Último acesso em: 10 fev. 2020).

LAZER, David; KENNEDY, Ryan; KING, Gary; VESPIGNANI, Alessandro. The parable of Google Flu: traps in big data analysis. *Science* 343, n˚. 6176, p. 1203-1205, 2014. Disponível em: https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf. (Último acesso em: 10 fev. 2020).

MARGULIES, Peter. Surveillance By Algorithm: The NSA, Computerized Intelligence Collection, and Human Rights. *Fla. L. Rev.*, vol. 68, nº 4, 2016. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2657619. (Último acesso em: 10 fev. 2020).

MITCHELL, Margaret et al. *Model Cards for Model Reporting*. In: FAT*'19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA. Disponível em: https://arxiv.org/abs/1810.03993. (Último acesso em: 10 fev. 2020).

MITTELSTADT, Brent et al. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, [s.l.], dez. 2016. Disponível em: http://journals.sagepub.com/doi/pdf/10.1177/2053951716679679. (Último acesso em: 10 fev. 2020).

MONTEIRO, Renato Leite. *Existe um direito à explicação na Lei Geral de Proteção de Dados do Brasil?* Instituto Igarapé, dez. 2018. Disponível em: https://bit.ly/2v2fyfJ. (Último acesso em: 10 fev. 2020).

OECD (2019a). *Recommendation of the Council on Artificial Intelligence*. OECD/ LEGAL/ 0449, maio 2019. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. (Último acesso em: 10 fev. 2020).

OECD (2019b). *Artificial Intelligence in Society*. OECD Publishing: Paris, jun. 2019. Disponível em: https://bit.ly/2Rfow1Q. (Último acesso em: 10 fev. 2020).

OSOBA, Osonde; WELSER IV, William. *An Intelligence in Our Image*: The Risks of Bias and Errors in Artificial Intelligence. Santa Mônica: RAND Corporation, 2017. Disponível em: https://bit.ly/2RDXVut. (Último acesso em: 10 fev. 2020).

PASQUALE, Frank. *The Black Box Society*: The Secret Algorithms that Control Money and Information. Cambridge, MA: Harvard University Press, 2015. Disponível em: https://raley.english.ucsb.edu/wp-content/Engl800/Pasquale-blackbox.pdf. (Último acesso em: 10 fev. 2020).

ROBBINS, Scott. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, [s.l.], vol. 29, n. 4, dez. 2019. Disponível em: https://bit.ly/2Sazvcs. (Último acesso em: 10 fev. 2020).

SALMON, Felix. The formula that killed Wall Street. *Significance* 9, no. 1˚, p. 16-20, 2012. Disponível em: https://www.researchgate.net/publication/227733068_The_Formula_that_Killed_Wall_Street. (Último acesso em: 10 fev. 2020).

SANDVIG, Christian et al. *Auditing algorithms*: Research methods for detecting discrimination on internet platforms. In: Annual Meeting of the International Communication Association, Seattle, maio 2014. Disponível em: https://bit.ly/2Oml3Ns. (Último acesso em: 10 fev. 2020).

SCHERER, Matthew U. Regulating Artificial Intelligence Systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, [s.l.], vol. 29, n. 2, 2016. Disponível em: http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf. (Último acesso em: 10 fev. 2020).

SILVA, Tarcízio Roberto da. *Visão computacional e vieses racializados*: branquitude como padrão no aprendizado de máquina. In: II COPENE Nordeste, 2019, João Pessoa, 13p. Disponível em: https://bit.ly/2t7suAF. (Último acesso em: 10 fev. 2020).

TUFEKCI, Zeynep. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colo. Tech. LJ*, [s.l.], v. 13, 2015. Disponível em: https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf. (Último acesso em: 10 fev. 2020).

WEBFOUNDATION. *Artificial Intelligence*: The Road Ahead in Low and Middle-Income Countries, 2017. Disponível em: https://bit.ly/2GeQHIe. (Último acesso em: 10 fev. 2020).

WHITTAKER, Meredith et al. *AI Now Report 2018*. Disponível em: https://ainowinstitute.org/AI_Now_2018_Report.pdf. (Último acesso em: 10 fev. 2020).

WORLD ECONOMIC FORUM. *AI Governance*: A Holistic Approach to Implement Ethics into AI. Genebra, jan. 2019. Disponível em: https://bit.ly/2S8E32S. (Último acesso em: 10 fev. 2020).

Autores

Coordenação

Ivar A. Hartmann: Coordenador do Centro de Tecnologia e Sociedade da FGV Direito Rio. Doutor em Direito Público pela UERJ. Mestre em Direito Público pela PUC-RS. Mestre em Direito (LL.M.) pela Harvard Law School. Professor e Pesquisador da Escola de Direito da Fundação Getúlio Vargas - RJ. Coordenador Executivo da Revista Direitos Fundamentais e Justiça (A2). Ex-bolsista da CAPES, do DAAD e da Harvard Law School. Áreas de interesse: Direito e Tecnologia, Pesquisa Empírica no Direito, Direito Constitucional.

Pesquisadores

Bruna Franqueira: Pesquisadora no Centro de Tecnologia e Sociedade da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Programadora do Observatório Carioca, projeto do Instituto RIO 21. Graduada em Direito pela Fundação Getulio Vargas, com período de intercâmbio Acadêmico na Universitat Pompeu Fabra e na Université Sorbonne Paris-IV.

Julia Iunes: Pesquisadora no Centro de Tecnologia e Sociedade da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Mestre em Direito Público pela Universidade do Estado do Rio de Janeiro (UERJ). Bacharel em Direito pela Universidade Federal do Rio de Janeiro (UFRJ). Advogada.

Lorena Abbas: Pesquisadora no Centro de Tecnologia e Sociedade da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Doutoranda em Políticas Públicas pela Universidade Federal do Rio de Janeiro (PPED/UFRJ). Mestrado e Graduação em Direito pela Universidade Federal de Juiz de Fora-MG (UFJF).

Yasmin Curzi: Pesquisadora no Centro de Tecnologia e Sociedade da Escola de Direito da Fundação Getulio Vargas (FGV Direito Rio). Doutoranda no Instituto de Estudos Sociais e Políticos da Universidade do Estado do Rio de Janeiro (IESP/UERJ), com bolsa CAPES. Mestre em Ciências Sociais, pela Pontifícia Universidade Católica do Rio de Janeiro (2017), realizado com bolsa CAPES. Graduada em Direito (2018) e em Ciências Sociais (2014), com período de Intercâmbio Acadêmico na Université Sorbonne Paris-IV, pela FGV-Rio. Advogada.

Bolsistas

Beatriz Villa: Bolsista de iniciação científica (PIBIC) no Centro de Tecnologia e Sociedade (CTS) da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Graduanda em Direito na Fundação Getúlio Vargas (FGV Direito Rio).

Fernanda Abreu: Bolsista de iniciação científica (PIBIC) no Centro de Tecnologia e Sociedade (CTS) da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Graduanda em Direito na Fundação Getúlio Vargas (FGV Direito Rio).

Renan Dias: Voluntário de iniciação científica no Centro de Tecnologia e Sociedade (CTS) da Escola de Direito da Fundação Getúlio Vargas (FGV Direito Rio). Graduando em Direito na Fundação Getúlio Vargas (FGV Direito Rio).