ESTRUTURA DE AVALIAÇÃO
DE RISCOS A DIREITOS
E DE TRANSPARÊNCIA

USO DE INTELIGÊNCIA ARTIFICIAL PELO PODER PÚBLICO

TRANSPARÊNCIA BRASIL

REALIZAÇÃO

TRANSPARÊNCIA BRASIL

DIRETOR-EXECUTIVO
Manoel Galdino

DIRETORA DE OPERAÇÕES
Juliana Sakai

COORDENADORA DO PROJETO
Tamara Burg

PESQUISA
Jonas Coelho
Tamara Burg

TEXTO
Juliana Sakai
Manoel Galdino
Tamara Burg

DIAGRAMAÇÃO
Marina Atoji

www.transparencia.org.br

FINANCIAMENTO

ICNL
INTERNATIONAL CENTER FOR NOT-FOR-PROFIT LAW

PARCERIA

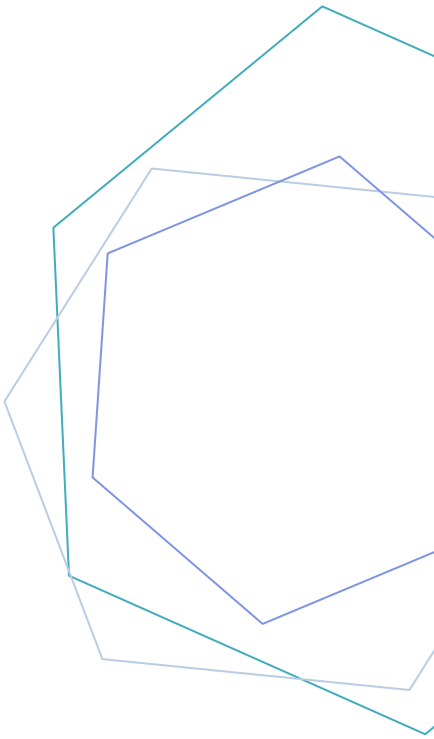
Northwestern University

AGRADECIMENTOS

Controladoria-Geral da União (CGU)
Ministério da Ciência, Tecnologia e Inovação (MCTI)
Centro de Estudos sobre Tecnologias Web (Ceweb.br)
Bruno Kunzler
Enrico Roberto (InternetLab)
Nathalie Fragoso (internetLab)
Nazareno Andrade

SUMÁRIO

INTRODUÇÃO



Este documento tem como objetivo apresentar uma proposta para avaliação de riscos no uso de algoritmos de inteligência artificial (IA) pelo poder público, numa busca de alinhamento entre promoção de inovação e tecnologia e responsabilidade pública e transparência.

Buscamos apresentar aqui uma metodologia simples que permita avaliar riscos envolvendo ameaças reais e potenciais a direitos e ao espaço cívico. A partir do resultado da avaliação seria possível entender a transparência necessária em processos de desenvolvimento, aquisição e implementação de IA pelo estado de forma a garantir o devido controle social democrático. Por isso, o framework divide-se em avaliação de riscos a direitos e avaliação de transparência.

A estrutura de avaliação de riscos está dividida da seguinte forma: i) riscos a direitos pela natureza da ferramenta; ii) riscos a direitos por discriminação algorítmica; iii) riscos ao direito à privacidade e; iv) potencial abuso autoritário do espaço cívico.

Em seguida, apresentamos a proposta de avaliação de transparência no uso de sistemas de inteligência artificial (item v). Não tratamos aqui transparência e *accountability* como direitos potencialmente ameaçados por IA, mas como ferramentas essenciais para garantir controle público e responsabilização. Assim, avalia-se o nível de transparência existente como instrumento necessário para acompanhar todo o

processo de utilização pública de sistemas de IA, tendo em vista os riscos a direitos elencados.

O objetivo deste documento é oferecer um guia na avaliação do uso de IA pelo estado com a finalidade de identificar potenciais ameaças a direitos e ao controle social. A proposta é que a partir da avaliação seja possível elaborar recomendações específicas quanto ao uso de determinadas ferramentas e sua governança, quanto às recomendações de diretrizes gerais para desenvolvimento, aquisição e implementação de IA.

Uma premissa importante no processo de avaliação de riscos a direitos é a inclusão de uma análise multissetorial que cubra diversos campos. Em função disso, esse framework deve ser aplicado com a contribuição de diversas organizações e movimentos da sociedade civil, atuantes na promoção de diferentes causas e direitos, de forma a enriquecer a análise.

ESTRUTURA DE AVALIAÇÃO DE RISCOS

1. Avaliação de riscos a direitos pela natureza da ferramenta

Busca-se aqui avaliar o potencial impacto a direitos que determinada ferramenta de IA pode causar em casos concretos com base no *output* ou em seus resultados, isto é, no que ela foi desenhada para entregar.

Algoritmos de IA podem ser utilizados para atingir diferentes finalidades dentro da esfera governamental. Há ferramentas desenhadas para acelerar procedimentos internos de gestão, como por exemplo algoritmos de processamento de linguagem natural que ajudam na triagem automatizada de ofícios. Há também algoritmos que podem ser usados para calcular a chance de reincidência criminal de um indivíduo, e impactam em sentenças condenatórias[1].

Um possível erro ou mal funcionamento do sistema no primeiro caso envolveria um reencaminhamento humano do ofício, prejudicando apenas a agilidade que o algoritmo oferecia. Sua natureza não envolve riscos substanciais de impactar acesso a direitos.

[1] Como divulgado pela ProPublica, o sistema COMPAS, usado nos Estados Unidos para avaliar o risco de reincidência criminal https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

No segundo caso, não é preciso discorrer muito que o impacto possível é grave porque envolve privação da liberdade de um indivíduo. Envolve o comprometimento do exercício à liberdade, da estigmatização de sujeitos e das próprias premissas da responsabilização penal. Este é um exemplo de uma ferramenta de IA que, por sua natureza, pode apresentar graves riscos a direitos.

Uma correta identificação inicial do risco que o algoritmo representa a diferentes direitos é necessária para que o desenho ou a implementação do algoritmo inclua formas de mitigar seus riscos previamente, e que haja um acompanhamento do seu funcionamento e resultados.

Para guiar a avaliação desses riscos, são necessárias, ao menos, as informações elencadas a seguir.

Quadros para avaliações de riscos a direitos pela natureza da ferramenta

QUADRO 1. DADOS GERAIS SOBRE A FERRAMENTA

Nº	INFORMAÇÃO A SER REGISTRADA
I	Órgão governamental responsável
II	Nome da ferramenta utilizada
III	Categoria de aplicação *Classificação de imagens (exceto reconhecimento facial), reconhecimento facial, sistema de recomendação, chatbot, estimativas de risco (incluindo detecção de fraudes), análise de sentimentos, outros.*
IV	Modelo estatístico envolvido *Regressão logística, regressão linear ou variações, métodos baseados em árvore de decisão (incluindo florestas aleatórias e XGBoost), redes neurais, processamento de linguagem natural, AutoML, outros.*
V	*Inputs* ou dados de entrada *Descrição das variáveis de entrada*
VI	*Output* ou resultado *Por exemplo, probabilidade de determinado caso apresentar fraude; ou concessão ou não de crédito.*
VII	Grau de apoio oferecido pela ferramenta *Faz diagnósticos e toma decisões; faz diagnósticos e sugere ações, faz diagnósticos, mas não sugere ações.*

QUADRO 2. AVALIAÇÕES A SEREM REALIZADAS

Nº	AVALIAÇÕES A SEREM REALIZADAS
I	No fluxo de utilização da ferramenta, há supervisão humana em todas as decisões sugeridas ou tomadas pelo algoritmo?
II	Em caso de erro do algoritmo corrigido por humano, essa informação é usada para aprimoramento do algoritmo?
III	A ferramenta, por sua natureza, pode impactar direitos fundamentais, seja por erro ou por design do seu algoritmo, seja direta ou indiretamente? Se sim, quais?
IV	Quais grupos ou populações serão afetadas por esse algoritmo? Esses segmentos foram considerados no processo de treinamento da ferramenta?
V	Existem órgãos ou pessoas dentro da entidade que utilizam o algoritmo que podem prestar informações sobre seu uso às autoridades competentes?
VI	O impacto negativo é criado ou acentuado a partir do algoritmo?
VII	Esse algoritmo é imprescindível para atingir o objetivo apontado? Se ele tem o potencial de afetar o exercício de direitos fundamentais ou de se colocar como intermediário para acesso a eles, existem formas alternativas para exercício de tal direito? Se sim, quais?
VIII	Existem evidências de que este algoritmo funcione no ambiente em que ele está sendo utilizado? As evidências são baseadas em experimentos científicos relevantes?
IX	Existe regulamentação específica sobre o uso deste algoritmo na área em que ele está sendo aplicado? Quais são? Se não, existe apoio de uma equipe jurídica especializada para garantir que haja respaldo jurídico?
X	Existe uma equipe técnica que acompanha e monitora a implementação deste algoritmo? Esta equipe contém funcionários do órgão capazes de analisar criticamente os caminhos tomados?

CONTINUA

CONTINUAÇÃO

Nº	AVALIAÇÕES A SEREM REALIZADAS
XI	A equipe responsável pelo desenvolvimento do algoritmo inclui especialistas da área na qual o algoritmo será aplicado?
XII	Um comitê de ética acompanha/acompanhou o desenvolvimento do algoritmo e os ritos envolvidos na coleta e uso de dados?
	AVALIAÇÃO: IMPACTO ALTO, MODERADO OU BAIXO

2. Avaliação de riscos a direitos por discriminação algorítmica

A discriminação algorítmica, tipicamente, surge a partir de bases de dados de treinamento insuficientemente representativas. Modelos são treinados a partir dos dados disponíveis, de forma que dados que não refletem diferentes grupos tendem a gerar modelos piores ou discriminatórios. Nesse sentido, a representatividade de uma base de dados irá impactar diretamente no resultado do sistema. Por exemplo, algoritmos de reconhecimento facial treinados com uma base de dados composta, em sua maioria, com rostos de pessoas brancas, terá menor precisão em reconhecer rostos fora de tal padrão, como rostos de pessoas negras[2].

Como o viés algorítmico pode refletir a falta de representatividade de determinados grupos, um algoritmo poderia acentuar diferenças sociais e a opressão a grupos marginalizados[3].

Para além da representatividade do dado, o viés pode surgir da validade deste dado ou no próprio desenho do algoritmo[4]. Os dados disponíveis para treinamento de um modelo podem não servir propriamente para o que foi desenhado, e seu resultado pode gerar discriminação entre grupos.

Um estudo publicado na revista Science[5] trouxe o exemplo do algoritmo usado em um hospital nos Estados Unidos,

[2] https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/
[3] Contribuições do InternetLab para a estratégia nacional de Inteligência Artificial: https://www.internetlab.org.br/pt/privacidade-e-vigilancia/as-contribuicoes-do-internetlab-para-a-estrategia-nacional-de-inteligencia-artificial/
[4] https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/
[5] https://science.sciencemag.org/content/366/6464/447

desenvolvido para orientar decisões médicas, ao classificar pacientes que necessitam de mais cuidados. O algoritmo privilegiava pacientes brancos e concluía falsamente que pacientes negros estariam mais saudáveis do que pacientes brancos. O viés ocorreu porque o algoritmo usava como dados pagamentos a planos de saúde como proxy para avaliar a condição médica do paciente, ignorando que brancos têm mais acesso a planos de saúde que negros.

Diversos ativistas, jornalistas, pesquisadores e funcionários de empresas de tecnologia vêm alertando sobre os perigos dos vieses em sistemas de IA por pelo menos uma década. Eles têm rigorosamente pesquisado, detectado e comprovado discriminação algorítmica em reconhecimento facial, direcionamento de anúncios em redes sociais, concessões de crédito, sistemas de previdência, e algoritmos usados em sentenças criminais[6].

Ademais, determinadas classes e regiões dominantes participam mais ativamente da comunidade tecnológica, influenciando o desenho dos modelos que compõem sistemas de classificação ou recomendação automatizados.

Para guiar a avaliação desses riscos, são necessárias as respostas às questões elencadas a seguir.

[6] AI Now Report, 2019. em https://ainowinstitute.org/discriminatingsystems.pdf

Quadro para avaliação de riscos a direitos por discriminação algorítmica

Nº	AVALIAÇÃO A SER REALIZADA
I	Foram considerados possíveis vieses no desempenho da ferramenta em seu desenvolvimento, aquisição e/ou implementação?
II	Se sim, os vieses foram corrigidos ou mitigados pelo código? De que maneira?
III	Foram feitos testes antes e durante a implementação para saber se taxas de erro são iguais ou menores em grupos minoritários?
IV	A amostra de treinamento é rica em quantidade e diversidade para um bom resultado da ferramenta com os diferentes grupos aos quais a ferramenta é aplicada?
V	Se é uma ferramenta que não foi desenvolvida internamente, ela foi desenhada especificamente para o público brasileiro ou ao público ao qual é aplicada? Se não, foi testada como sua acurácia difere para o público-alvo?
VI	Existe uma equipe que monitora a performance do algoritmo em relação a estes grupos periodicamente? Se sim, rotinas de retreinamento foram planejadas durante a implementação do algoritmo?
VII	No caso de ferramentas de interação com público externo, como chatbots, existe um responsável para receber reclamações de possíveis discriminações que a ferramenta esteja cometendo?
AVALIAÇÃO: IMPACTO ALTO, MODERADO OU BAIXO	

3. Avaliação de riscos ao direito à privacidade

Danos à privacidade podem ocorrer com a criação e/ ou disponibilização de bancos de dados massivos, já que o emprego de tecnologias de IA, em geral, demanda o processamento de grande quantidade de dados.

Órgãos governamentais vêm utilizando dispositivos tecnológicos capazes de coletar dados pessoais sobre os cidadãos capazes de monitoramentos massivos, como o monitoramento de localização por celular para acompanhamento de isolamento imposto pela COVID[7], e até ferramentas de identificação reconhecimento facial[8].

Algoritmos com a intenção de coletar informações estratégicas e melhorar serviços e políticas públicas também vêm sendo observados. Por exemplo, a cidade de San Diego instalou milhares de câmeras em postes de rua em um esforço para estudar as condições de trânsito e, embora os dados tenham se mostrado pouco úteis para melhorar o trânsito, a polícia utiliza filmagens sem supervisão ou responsabilidade[9].

A discussão das condições e dos limites para o uso de dados pessoais passa, primeiro, por uma compreensão a respeito do que é um dado pessoal. A Lei n. 13.709/2018, chamada de Lei Geral de Proteção de Dados no Brasil (LGPD) define dado pessoal como toda informação relacionada a pessoa natural identificada ou identificável. Isso inclui características pessoais, qualificação pessoal, dados genéticos etc.

[7] https://www.bbc.com/portuguese/brasil-52357879
[8] https://theintercept.com/2020/02/11/metro-sao-paulo-reconhecimento-facial/
[9] https://www.latimes.com/california/story/2019-08-05/san-diego-police-ramp-up-use-of-streetlamp-cameras-to-crack-cases-privacy-groups-raise-concerns

A LGPD prevê que o tratamento e compartilhamento de dados pela administração pública restringe-se aos dados necessários para a execução de determinada política pública (art. 7º, III). Ou seja, a coleta dos dados deve ser feita estritamente para prestação e melhoria do serviço proposto - com uma finalidade adequada, bem definida e com critérios.

Neste sentido, a administração pública precisa ser transparente com relação à coleta de dados de indivíduos identificáveis e identificados, assim como quanto ao compartilhamento e uso desses dados.

Para guiar a avaliação desses riscos, são necessárias as respostas às questões elencadas a seguir.

Quadro para avaliação de riscos ao direito à privacidade

Nº	AVALIAÇÃO A SER REALIZADA
I	Existe a necessidade da utilização de dados pessoais para treinar a ferramenta?
II	Se sim, as pessoas são informadas que seus dados pessoais estão sendo usados para alimentar a ferramenta?
III	Existe a necessidade da utilização de dados sensíveis ou sigilosos para treinar a ferramenta? Se sim, qual a base legal e quais camadas adicionais de segurança são aplicadas para proteger esses dados?
IV	As pessoas podem optar por retirar seus dados do treinamento da ferramenta ou exercer seus outros direitos sobre seus dados pessoais, como acesso, portabilidade etc.?
V	Os dados pessoais são coletados/produzidos pelo próprio órgão ou são compartilhados por outras fontes? As outras fontes têm base legal legítima para compartilhar os dados ao órgão em questão?
VI	Os dados pessoais são compartilhados com terceiros? Há base legal legítima para esse compartilhamento? Se sim, foram utilizadas técnicas de anonimização antes do compartilhamento?
VII	Dados pessoais/sigilosos são utilizados pelo algoritmo, foram empregadas técnicas de anonimização/despersonalização durante o pré-processamento destes dados?
VIII	Dados pessoais/sigilosos são utilizados para treinamento do algoritmo, foram empregadas técnicas de preservação de privacidade durante o treinamento dos modelos?

CONTINUA

CONTINUAÇÃO

Nº	AVALIAÇÃO A SER REALIZADA
IX	A ferramenta é desenvolvida por um terceiro, existe um documento que regula o compartilhamento e uso dos dados por este ente?
X	O cidadão pode escolher não ter seus dados analisados por um algoritmo?
XI	As pessoas deram consentimento para o uso de seus dados para treinamento deste algoritmo?
XII	Os dados utilizados para treino e os dados capturados por este algoritmo estão armazenados em um servidor na nuvem (no Brasil ou fora) ou em um servidor local? Existem protocolos de segurança desenvolvidos para acesso a estes dados?
AVALIAÇÃO: IMPACTO ALTO, MODERADO OU BAIXO	

4. Avaliação de potencial abuso autoritário e restrição do espaço cívico

Ferramentas que coletam e cruzam informações pessoais podem ser úteis a algumas políticas públicas – notadamente segurança pública –, mas podem também representar uma ameaça à sociedade civil e uma grande arma na mão de governos autoritários, que podem usar esses dados para implementar um estado de vigilância que persegue opositores e diminui o espaço cívico por consequência.

Dados sensíveis são informações sobre a esfera íntima do titular de dados, que têm maior potencial de abuso nesse contexto: podem ser utilizados de forma a perseguir minorias ou opositores políticos, por exemplo. Por sua natureza, dispõem de proteção especial pela lei. A LGPD define dado sensível como "dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural" (art. 5º, II). A essa categoria de dados pessoais, a Lei confere um maior grau de proteção e estabelece hipóteses mais restritas para o seu tratamento.

O objetivo desta seção é avaliar se a ferramenta oferece riscos ao espaço cívico por um potencial uso autoritário.

Para guiar a avaliação desses riscos, são necessárias as respostas às questões elencadas a seguir.

Quadro para avaliação de riscos ao espaço cívico

Nº	AVALIAÇÃO A SER REALIZADA
I	A ferramenta produz ou coleta informações que podem ser utilizadas para monitorar indivíduos ou grupos políticos, étnicos ou religiosos, bem como ativistas? Se sim, quais ferramentas são utilizadas para evitar esse uso excessivo dos dados?
II	O algoritmo usa dados sensíveis ou potencialmente discriminatórios? Se sim, quais camadas adicionais de segurança são aplicadas para proteger esses dados?
AVALIAÇÃO: IMPACTO ALTO, MODERADO OU BAIXO	

ESTRUTURA DE AVALIAÇÃO DE TRANSPARÊNCIA

Para viabilizar a avaliação de riscos e, logo, a defesa de direitos e do espaço cívico, é fundamental que haja transparência. Assim, o uso de sistemas de IA pelo estado deve considerar os princípios e regramentos de transparência pública, de forma a garantir o controle social do uso de IA e eventuais responsabilizações.

Além das informações já questionadas ao longo da avaliação de riscos, somam-se as seguintes questões para a avaliação de transparência.

Nº	QUESTÕES A SEREM RESPONDIDAS
I	Antes de a ferramenta ser colocada em uso, é possível ter acesso a relatórios de impacto prévios onde constam testes feitos para avaliar vieses e o que foi feito para contornar o comportamento discriminatório da ferramenta?
II	Antes de a ferramenta ser colocada em uso, é possível ter acesso a relatórios que apontem informações como o fato de o modelo estar sendo desenvolvido, qual seu propósito e quais as populações potencialmente afetadas, e quais os direitos fundamentais potencialmente afetados pelo sistema e quais mecanismos estão sendo usados para mitigar tais questões?
III	É possível ter acesso à informação sobre as variáveis de entrada ou inputs do sistema?

CONTINUA

CONTINUAÇÃO

Nº	QUESTÕES A SEREM RESPONDIDAS
IV	Há métricas para aferir a acurácia da ferramenta?
V	É possível ter acesso ao algoritmo desenvolvido/utilizado pela ferramenta?
VI	Depois de colocada em uso a ferramenta, há relatórios periódicos de impacto com atualização de testes de vieses e de acurácia, correções e melhorias da ferramenta, bem como prestação de contas quanto ao impacto nas pessoas e populações afetadas pela ferramenta?
VII	Há um responsável na implementação da ferramenta capaz de explicar a uma pessoa afetada por ela sobre os motivos do resultado da ferramenta?
VIII	Há um responsável por acompanhar a forma com a qual o algoritmo afeta uma decisão tomada por um humano? Existe um estudo, relatórios ou pesquisas que analisam o fenômeno da interação entre humano e máquina?
AVALIAÇÃO: TRANSPARÊNCIA ALTA, MODERADA OU BAIXA	

CONCLUSÃO

O uso de sistemas de inteligência artificial para auxiliar na realização ou prestar serviços públicos deve ser exercido com especial atenção aos seus potenciais efeitos nos direitos e liberdades individuais e coletivos, assim como nas regras aplicáveis à administração pública, em especial a exigência de transparência.

Documentos como Relatórios de Impacto – prévios e contínuos - devem ser exigidos sempre que um sistema de inteligência artificial for desenvolvido para finalidades públicas, devendo ser publicamente disponibilizado, por exemplo, no site do órgão que oferece ou faz uso da ferramenta.

Esta estrutura de avaliação de riscos a direitos e de transparência tem como objetivo apoiar o controle social no monitoramento de sistema de IA. Ela funciona como um guia para encontrar pontos críticos e a partir disso elaborar recomendações ao governo, exigir mais transparência pública, correções ou testes para garantir não-discriminação e erros algorítmicos, ou até a eventual descontinuidade de alguma ferramenta cujo risco seja incontornável.

TRANSPARÊNCIA
BRASIL

transparencia.org.br