# Consulta Pública – Estratégia Brasileira de Inteligência Artificial

(*http://participa.br/estrategia-brasileira-de-inteligencia-artificial/legislacao-regulacao-e-uso-etico*)

## Contribuições InternetLab

*[Somente copiadas as perguntas e parágrafos onde se sugerem comentários]*

**I. Segurança Pública**

(parágrafo) "Por outro lado, a utilização de tais tecnologias tem também sido problematizada, principalmente no que diz respeito aos problemas associados a viés e discriminação decorrentes, em muitos casos, de bases de dados de treinamento insuficientemente representativas. Embora sistemas de reconhecimento facial para segurança pública tenham sido adotados em inúmeros contextos, inclusive no Brasil, os índices alarmantes de falsas identificações positivas suscitam preocupações..."

A possibilidade de viés e discriminação não se limita à simples falta de representatividade em bases de dados, e a busca por maior representatividade não necessariamente será bem sucedida em afastar ou minimizar tais riscos. A falta ou enviesamento dos dados sobre grupos minoritários, por exemplo, frequentemente será estrutural: dados do PNAD de 2014 mostram que somente 38,5% das pessoas brancas não usam a internet no Brasil, contra 60,5% da população negra (https://www.nexojornal.com.br/grafico/2016/05/30/Quem-%C3%A9-a-popula%C3%A7%C3%A3o-sem-acesso-%C3%A0-internet-no-pa%C3%ADs). Isso, dentre diversos outros fatores, resulta em menos dados sobre essa população (por exemplo, em menor quantidade de fotos em redes sociais que possam ser usadas para treinamento de algoritmos de reconhecimento facial).

Outros exemplos nesse sentido, que explicamos em seguida, são o COMPAS, software utilizado nos Estados Unidos para oferecer notas de risco de reincidência em crimes para réus, que são então utilizadas pelos juízes do processo em questão no sopesamento de pena; um algoritmo utilizado num hospital não identificado, também nos Estados Unidos, que busca direcionar automaticamente o uso de recursos hospitalares com seus pacientes; e o próprio Google Tradutor.

No caso do COMPAS, comprovou-se que, controladas as outras variáveis, as notas de reincidência de negros eram invariavelmente mais altas que as de brancos (https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis), discriminação essa que decorre do fato de o software ter sido treinado com base nos dados reais do sistema carcerário norte-americano: se negros estão mais expostos ao controle penal; se policiais abordam mais negros nas ruas; se promotores acusam mais negros e juízes dão penas maiores; tal padrão repercutirá nos resultados do algoritmo.

Já no caso do algoritmo de saúde, notou-se que os recursos que este direcionava a pacientes brancos eram muito maiores do que os direcionados a pacientes negros (https://science.sciencemag.org/content/366/6464/447). Novamente, o viés ocorreu porque o algoritmo foi treinado não com base em dados diretamente associados à saúde do paciente, e sim com base em dados referentes a recursos de saúde efetivamente despendidos com pacientes passados. Se, por acessarem hospitais de menor qualidade, terem menos condições de ir com frequência ao médico, ou situações similares, menos dinheiro for gasto com pacientes negros, o algoritmo falsamente concluirá que negros são mais saudáveis do que pacientes brancos igualmente doentes.

Por fim, outro exemplo interessante diz respeito à ferramenta de tradução do Google. Na tradução de línguas com substantivos de gênero neutro (como o inglês, por exemplo em "the doctor") para línguas com diferenciação de gêneros (como o português, por exemplo em "a doutora" ou "o doutor"), a ferramenta geralmente traduzirá as palavras de gênero neutro para a sua variante mais comumente utilizada na língua onde há tal diferenciação. Assim, no mesmo exemplo utilizado acima, *the doctor* seria traduzido como "o doutor", já que a língua aprendida, o português, encontra-se por razões históricas enviesada para o uso de "doutor" como uma profissão masculina. (https://techcrunch.com/2018/12/07/google-translate-gets-rid-of-some-gender-biases/)

Os exemplos acima mostram que o problema do viés não decorre meramente da falta de representatividade nos dados de treinamento, mas, principalmente, da reprodução de padrões sociais pré-existentes nos resultados de algoritmos de aprendizado de máquina. A inteligência artificial, assim, não é um ator neutro capaz de mostrar resultados objetivos, e sim um reprodutor de situações sociais pré-existentes e da subjetividade humana historicamente situada – subjetividade essa que se impõe ao algoritmo tanto por seus dados quanto pela finalidade de sua utilização.

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

Se se diz que um sistema de IA é "seguro", conclui-se que não oferece ou minimiza riscos de violação a direitos individuais ou coletivos. Se se diz que são "confiáveis", trata-se de situação em que indivíduos e a coletividade poderão contar com a segurança do dispositivo. Para um sistema seguro e confiável, portanto, deve-se: (i) ter consciência, e estimular a pesquisa para o contínuo avanço dessa consciência, dos riscos da tecnologia; e (ii) criar mecanismos de controle individual e coletivo desse tipo de sistema, de forma a permitir que se averigue se seus riscos estão sendo minimizados.

Dentre seus riscos, ressaltamos a capacidade de a Inteligência Artificial reproduzir e cristalizar discriminações (o que não decorre somente da falta de representatividade em base de dados), assim como a sua pouca confiabilidade em diversos casos. Detalhamos esses riscos acima, em nossos comentários aos parágrafos introdutórios a esse item.

Além disso, constituem risco inafastável desse tipo de tecnologia os falsos positivos, que, associados à expectativa generalizada de objetividade que se deposita na inteligência artificial, podem representar sérios riscos às liberdades e direitos de pessoas inocentes. As promessas de altíssima acurácia feitas pelos produtores da tecnologia, muitas vezes baseadas em situações e dados ideais ou de laboratório, não afastam a realidade de que, na prática do uso da tecnologia pelas forças policiais, sua precisão ainda deixa muito a desejar: dados de 2018 da Polícia Metropolitana de Londres mostram que, de 2685 suspeitos identificados por um sistema de reconhecimento facial em um evento público, 2451 - ou seja, 91% das pessoas - foram alarmes falsos (https://www.bbc.com/news/technology-44089161). Como o próprio Ministério apontou, tal ausência de acurácia será ainda maior caso falte representatividade nos dados de treinamento do sistema, o que comprovadamente ocorre com pessoas negras. Desse risco, resultam importantes limites à confiança que forças policiais podem depositar nesse tipo de algoritmo, assim como regras de utilização para mitigação desses problemas, conforme detalhamos abaixo.

Assim, para que possam ser seguros e confiáveis, diversas salvaguardas devem ser consideradas, conforme apontamos nas questões abaixo dessa consulta.

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

Na introdução a essa questão, o Ministério mencionou o viés e discriminação decorrentes de bases de dados de treinamento insuficientemente representativas. De fato, trata-se de risco importante: a falta de representatividade em uma base de dados fará com que a acurácia de determinado sistema seja menor em relação ao grupo sub-representado. Não é à toa, por exemplo, que algoritmos de reconhecimento facial não funcionam bem com pessoas negras (https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/). Isso ocorre porque foram treinados majoritariamente com base em rostos de pessoas brancas, o que pode ser resultado da negligência dos programadores em buscar maior representatividade no momento de treinar o algoritmo, mas revela questão muito mais profunda do que simples falta de representatividade: a cristalização e intensificação de vieses sociais por meio de algoritmos de inteligência artificial, conforme apontamos no nosso comentário acima. Os exemplos que apresentamos, do software COMPAS, de outro utilizado por um hospital estadunidense, e do Google Tradutor, deixam clara essa realidade. Outros valiosos exemplos podem ser encontrados aqui (https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo/).

Os exemplos mostram que o problema do viés não decorre meramente da falta de representatividade nos dados de treinamento, mas, principalmente, da reprodução de padrões sociais pré-existentes nos resultados de algoritmos de aprendizado de máquina. O aumento de eficiência prometido por algoritmos de IA deve ser a todo momento sopesado e controlado de acordo com a finalidade para qual um tal algoritmo foi criado, e suas limitações devem ser de conhecimento dos que o desenvolvem (para diminuição dessas limitações) e dos que o utilizam.

Desses riscos e situações, resultam as seguintes recomendações práticas:

(i) O uso de sistemas de inteligência artificial para a segurança pública é atividade que gera riscos às liberdades civis e a direitos fundamentais. Com isso, documentações similares ao Relatório de Impacto à Proteção de Dados Pessoais devem ser exigidas sempre que um sistema de inteligência artificial for utilizado para essas finalidades. Em todas as situações, e em especial se utilizado pelo poder público e perante um grupo indeterminado de pessoas sem conhecimento de que a tecnologia está sendo utilizada (como câmeras de segurança com reconhecimento facial ou no direcionamento automatizado de forças policiais), tal relatório deve ser publicamente disponibilizado, por exemplo no site da empresa ou órgão que o oferece ou dele faz uso. Trata-se de medida de controle público, responsabilização (*accountability*) e transparência, que, concretamente, deve fornecer informações claras quanto:

- Ao fato de tal sistema estar sendo utilizado, incluindo informações sobre sua finalidade e locais de uso;
- quais os mecanismos utilizados para controle de vieses a partir das bases de dados e da seleção do modelo;
- qual o modelo algorítmico utilizado, se o sistema é atualizado, e se sim, como e com que frequência;
- qual a origem da base de dados utilizada para seu treinamento; e
- informações de contato para o exercício de direitos individuais.

(ii)

-

(iii) Existência de cotas raciais e de gênero nos times de desenvolvimento e manutenção dessas tecnologias;

(iv) Instituição de comitês públicos, por exemplo por meio de órgãos associados à Autoridade Nacional de Proteção de Dados, para exigência de tais relatórios, controle de seu conteúdo e auditorias regulares de tais sistemas;

(v) Criação de comitês para a elaboração de normas técnicas e boas práticas vinculantes para a diminuição de vieses no desenvolvimento e uso dessas ferramentas, como as sugeridas em https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/;

(vi) A obrigação de que quaisquer sistemas automatizados utilizados para fins de segurança pública sejam previamente testados de acordo com as normas técnicas mencionadas acima, antes de sua implementação e uso públicos;

(vii) Garantia de direitos individuais perante decisões tomadas ou informadas por sistemas de inteligência artificial, como o direito à revisão humana, em especial por se tratar de situação com claro risco às liberdades individuais;

(viii) e

(ix) O estímulo à pesquisa acadêmica sobre o tema, para atualização da consciência pública em relação aos riscos da tecnologia.

Além do apontado acima, mencionamos em nossos comentários também o risco de falsos positivos. Como dissemos, as promessas de altíssima acurácia feitas pelos produtores da

tecnologia, muitas vezes baseadas em situações e dados ideais ou de laboratório, não afastam a realidade de que, na prática do uso da tecnologia pelas forças policiais, sua precisão ainda deixa muito a desejar: há casos relatados de 91% de taxas de falsos positivos. Desse risco, resultam as seguintes recomendações práticas:

(x) A inteligência artificial para a segurança pública não deve substituir o julgamento de um ser humano, nem pode ser cegamente confiada. Por exemplo, o reconhecimento de um indivíduo suspeito por um sistema de de reconhecimento facial nunca deve ser considerado suficiente para identificá-lo indubitavelmente, devendo sempre haver outras formas de averiguar sua identidade.

(xi) Devem existir regras de conduta por parte das forças policiais para evitar a baixa acurácia de sistemas utilizados para segurança pública. Por exemplo, no caso do reconhecimento facial, não devem ser utilizadas fotos de baixa resolução ou com partes do rosto cortadas, não devem ser feitas edições às fotos utilizadas para identificação dos suspeitos, desenhos não devem substituir as fotos, dentre outros. Recomendações valiosas para o reconhecimento facial na segurança pública podem ser encontradas em https://www.flawedfacedata.com/#footnote6_c28hfft.

**II. Uso pelo poder público**

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?

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.

Em especial se a tomada de decisão pelo poder público impactar o exercício de direitos e liberdades fundamentais, deve haver a produção de um relatório de formato similar a um Relatório de Impacto à Proteção de Dados Pessoais. Tal relatório deve ser publicamente disponibilizado, por exemplo no site do órgão que faz uso da inteligência artificial. Trata-se de medida de controle público, responsabilização (accountability) e transparência, que, concretamente, deve fornecer informações claras quanto:

- Ao fato de tal sistema estar sendo utilizado, incluindo informações sobre sua finalidade e locais de uso;
- quais os mecanismos utilizados para controle de vieses a partir das bases de dados e da seleção do modelo;
- qual o modelo algorítmico utilizado, se o sistema é atualizado, e se sim, como e com que frequência;
- qual a origem da base de dados utilizada para seu treinamento; e

informações de contato para o exercício de direitos individuais.

Além disso, o uso da inteligência artificial pelo poder público, em especial caso os sistemas tenham contato direto com a população (por exemplo, caso a prestação perante o público de

determinados serviços registrais seja automatizada) deve ser *centrado no humano,* isto é, deve ser fácil, rápido e acessível, utilizável por qualquer pessoa sem a necessidade de ajuda técnica.

Ainda, deve sempre haver a possibilidade de prestação do serviço em questão por seres humanos, assim como de revisão humana dos serviços prestados.

De que maneira é possível implementar mecanismos de monitoramento dos sistemas de IA ao longo do seu ciclo de vida, de modo a assegurar que tais sistemas estejam atingindo os seus objetivos e que consequências não pretendidas sejam identificadas?

Em especial caso o sistema de IA desempenhe papel em decisões que tenham impacto sobre o exercício de direitos e liberdades fundamentais, por exemplo se utilizados para fins de segurança pública, algumas medidas de monitoramento de seu ciclo de vida são recomendadas:

(i) Relatório Prévio

(ii) Deve haver a produção de um relatório de formato similar a um Relatório de Impacto à Proteção de Dados Pessoais. Tal relatório deve ser publicamente disponibilizado, por exemplo no site do órgão ou empresa que faz uso da inteligência artificial. Trata-se de medida de controle público, responsabilização (accountability) e transparência, nos termos do nosso comentário anterior.

(iii) Deve haver a instituição de comitês públicos, por exemplo por meio de órgãos associados à Autoridade Nacional de Proteção de Dados, para exigência de tais relatórios, controle de seu conteúdo e auditorias regulares de tais sistemas;

(iv) Deve haver a criação de comitês para a elaboração de normas técnicas e boas práticas vinculantes para a diminuição de vieses no desenvolvimento e uso dessas ferramentas, como as sugeridas em https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/; e

(v) Deve haver a obrigação de que quaisquer sistemas automatizados utilizados para fins com riscos potenciais a direitos e liberdades fundamentais sejam previamente testados de acordo com as normas técnicas mencionadas acima, antes de sua implementação e uso públicos.

**III. Legislação, regulação e uso ético**

(parágrafo) Muitos dos documentos acima citados indicam que o desenvolvimento de inteligência artificial deve atentar à harmonização dos princípios que guiam a noção de estado de direito, de modo que a inteligência artificial beneficie a sociedade, impulsionando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar.

> O tema da propriedade intelectual associada a obras acessadas ou criadas por sistemas de inteligência artificial também deve ser tratado aqui.

Em primeiro lugar, fazemos referência à proteção via direitos autorais conferida a bases de dados, conforme estabelecido internacionalmente (Art. 2, item 2, Convenção de Berna; Parte II, Seção 1, Art. 10, TRIPS e Art. 5 do *WIPO Copyright Treaty*) e também na nossa Lei de Direitos Autorais, proteção essa que pode se apresentar como entrave ao livre desenvolvimento de sistemas de IA. O acesso a grandes base de dados é etapa imprescindível para a criação de um sistema de aprendizado de máquina, ferramenta capaz exatamente de encontrar padrões em tais bases e, por exemplo, realizar predições com base em tais informações. Caso o direito autoral sobre bases de dados seja utilizado como ferramenta para seu bloqueio e impossibilidade de acesso, seriam potencializadas condutas anticompetitivas ou monopolistas por quem detém tais bases de dados, dificultando-se a entrada de novas e menores empresas no mercado e o desenvolvimento de ferramentas de IA alternativas às desenvolvidas por aqueles que detêm as bases de dados.

Tendo-se isso em mente, recomenda-se que existam exceções à proteção autoral sobre bases de dados para o caso de seu acesso e mineração por algoritmos de Inteligência Artificial, de forma a inibir condutas monopolistas e estimular a inovação e o acesso ao conhecimento. Uma exceção como essa, ainda que limitada ao acesso por universidades e instituições de herança cultural e para fins acadêmicos, já está em vigor na União Europeia por meio do Art. 3º da Diretiva 2019/790 (https://eur-lex.europa.eu/eli/dir/2019/790/oj). Naturalmente, tal direito de acesso deve apresentar mecanismos de proteção à privacidade dos afetados e a outros direitos que o detentor da base de dados possa ter sobre ela.

Além disso, deve ser dada atenção à proteção conferida a obras, ou outros bens imateriais passíveis de proteção, produzidos por algoritmos de inteligência artificial. O tema merece debates mais aprofundados, mas cabe desde já pontuar que uma proteção irrestrita ou excessiva a tais criações via direito autoral pode ter efeitos prejudiciais à inovação, ao desenvolvimento de mercados baseados em dados e em inteligência artificial, e ao equilíbrio entre monopólio e acesso à que correspondem os objetivos dos sistemas de propriedade intelectual. Por exemplo, sistemas de IA podem criar inúmeras variações próximas de determinado software, música, narrativa ou similares, o que pode resultar em dificuldades de licenciamento e acesso a tais obras, ou ainda em disputas jurídicas em torno de materiais similares produzidos por sistemas automatizados distintos. Afinal, esses sistemas produzem com base em instruções, e é possível que não haja muitas formas diferentes de se criar um determinado material à partir de um algoritmo e um conjunto de dados. Regimes para proteção de obras criadas por softwares devem também sopesar o interesse público no acesso a tais criações, tanto do ponto de vista do estímulo à criação quanto do acesso à informação.

Qual papel pode ser desempenhado por códigos de conduta, regras de boas práticas corporativas e padrões voluntários?

Em um guia publicado pela União Europeia em abril de 2019, sobre ética e IA (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai), são apontadas algumas medidas para a mitigação de discriminações antiéticas executadas por sistemas de IA. As diretrizes envolvem a exigência de maior

diversidade, *design* inclusivo, não discriminação, entre outros, em todas as etapas existentes durante o desenvolvimento de um sistema de IA.

Com uma proposta mais pragmática, uma pesquisa realizada pelo Instituto Brookings (https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/) **apresentou um modelo de questões a serem aplicadas pela equipe de um projeto de IA, enquanto desenvolvem o sistema**. O objetivo é auxiliar na identificação de possíveis vieses durante o desenvolvimento de um modelo de IA. O modelo é composto por questões como:

- Quem é o público do algoritmo e quem será mais afetado por ele?
- Os dados de treinamento são suficientemente diversos e confiáveis?
- Com quais grupos estamos preocupados quando se trata de treinar erros de dados e tratamento diferenciado?

É importante que investimento em pesquisa seja realizado para que haja não só novas contribuições, mas também divulgação e aplicação de um guia de condutas éticas para profissionais que desenvolvem estes sistemas.

Seria necessário estabelecer salvaguardas para o uso de IA em determinados campos particularmente sensíveis (por exemplo, no campo da segurança pública, na educação, na guerra ou na saúde)?

O uso de sistemas de inteligência artificial em campos sensíveis deve ser exercido com especial atenção aos seus potenciais efeitos nos direitos e liberdades individuais e coletivos. Os campos mencionados (segurança pública, educação, guerra e saúde) são exemplos de situações em que há claros riscos às liberdades e direitos fundamentais. Assim, considerando-se os riscos da tecnologia (como sua capacidade de discriminação e falsos positivos, que apontamos em outros comentários a essa consulta), recomendam-se as seguintes medidas concretas:

(i) O uso de sistemas de inteligência artificial em campos sensíveis é atividade que gera riscos às liberdades civis e a direitos fundamentais. Com isso, documentações similares ao Relatório de Impacto à Proteção de Dados Pessoais devem ser exigidas sempre que um sistema de inteligência artificial for utilizado para essas finalidades. Em todas as situações, e em especial se utilizado pelo poder público e perante um grupo indeterminado de pessoas sem conhecimento de que a tecnologia está sendo utilizada (como câmeras de segurança com reconhecimento facial ou no direcionamento automatizado de forças policiais), tal relatório deve ser publicamente disponibilizado, por exemplo no site da empresa ou órgão que o oferece ou dele faz uso. Trata-se de medida de controle público, responsabilização (accountability) e transparência, que, concretamente, deve fornecer informações claras quanto:

- Ao fato de tal sistema estar sendo utilizado, incluindo informações sobre sua finalidade e locais de uso;
- quais os mecanismos utilizados para controle de vieses a partir das bases de dados e da seleção do modelo;
- qual o modelo algorítmico utilizado, se o sistema é atualizado, e se sim, como e com que frequência;
- qual a origem da base de dados utilizada para seu treinamento; e

- informações de contato para o exercício de direitos individuais.

(ii) Existência de cotas raciais e de gênero nos times de desenvolvimento e manutenção dessas tecnologias;

(iii) Instituição de comitês públicos, por exemplo por meio de órgãos associados à Autoridade Nacional de Proteção de Dados, para exigência de tais relatórios, controle de seu conteúdo e auditorias regulares de tais sistemas;

(iv) Criação de comitês para a elaboração de normas técnicas e boas práticas vinculantes para a diminuição de vieses no desenvolvimento e uso dessas ferramentas, como as sugeridas em https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/;

(v) A obrigação de que quaisquer sistemas automatizados utilizados para finadalidades sensíveis sejam previamente testados de acordo com as normas técnicas mencionadas acima, antes de sua implementação e uso públicos;

(vi) Estabelecimento de direitos individuais perante decisões tomadas ou informadas por sistemas de inteligência artificial, como o direito à revisão humana, em especial por se tratar de situação com claro risco às liberdades individuais;

(vii) e

(vii) O estímulo à pesquisa acadêmica sobre o tema, para atualização da consciência pública em relação aos riscos da tecnologia.

Além do apontado acima, mencionamos em nossos comentários também o risco de falsos positivos. Como dissemos, as promessas de altíssima acurácia feitas pelos produtores da tecnologia, muitas vezes baseadas em situações e dados ideais ou de laboratório, não afastam a realidade de que, na prática do uso da tecnologia pelas forças policiais, sua precisão ainda deixa muito a desejar: há casos relatados de 91% de taxas de falsos positivos. Desse risco, resultam as seguintes recomendações práticas:

(i) A inteligência artificial utilizada em campos sensíveis não deve substituir o julgamento de um ser humano, nem pode ser cegamente confiada. Por exemplo, o reconhecimento de um indivíduo suspeito por um sistema de de reconhecimento facial nunca deve ser considerado suficiente para identificá-lo indubitavelmente, devendo sempre haver outras formas de averiguar sua identidade.

(ii) Devem existir regras de conduta por parte dos usuários do sistema para evitar sua baixa acurácia. Por exemplo, no caso do reconhecimento facial, não devem ser utilizadas fotos de baixa resolução ou com partes do rosto cortadas, não devem ser feitas edições às fotos utilizadas para identificação dos suspeitos, desenhos não devem substituir as fotos, dentre outros. Recomendações valiosas para o reconhecimento facial na segurança pública podem ser encontradas em https://www.flawedfacedata.com/#footnote6_c28hfft.

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

São diversos os tipos de vieses que, por sua vez, podem possuir diferentes origens. Eles podem ocorrer durante as etapas de coleta, limpeza e tratamento dos dados e/ou teste do modelo gerado. É importante que a detecção de viés ocorra antes de o modelo ser colocado em produção. Logo, é necessário que se realize um conjunto completo de testes do modelo para diferentes casos possíveis, comparando os resultados obtidos.

Vieses resultantes da etapa de coleta ocorrem quando os dados obtidos para formar o conjunto de treinamento não representam precisamente a diversidade do ambiente em que o modelo será aplicado. Este problema pode ser solucionado ou amenizado por meio de ajustes no balanceamento do *dataset*, isto é, verificando-se se os dados representam proporcionalmente a diversidade do domínio no qual o modelo será aplicado. Caso um determinado grupo de casos não esteja devidamente representado, é necessário adicionar mais dados referentes a este grupo ao conjunto de treinamento.

Na etapa de limpeza e tratamento dos dados, o(s) viés(es) pode(m) ocorrer devido à exclusão indevida de atributos em razão de crenças pré-existentes dos desenvolvedores do sistema no sentido de que tais atributos não possuem relevância para o contexto. Para evitar esse tipo de viés, recomenda-se o uso de ferramentas que calculam a importância e a (in)dependência entre os atributos de um conjunto de dados (https://fairmlbook.org/).

O(s) viés(es) relacionados à etapa de teste do modelo estão diretamente relacionados às crenças das pessoas que o produziram. Ocorre quando a equipe rejeita um grupo de possibilidades para teste do modelo baseado em crenças pré-determinadas pela equipe, estereótipos, influências culturais, entre outros. Estes casos comumente resultam em modelos que reproduzem preconceitos existentes no mundo real, visto que não foram submetidos a todos os casos em que o modelo estará exposto. A principal solução para evitar este tipo de viés é ter diversidade na(s) equipe(s) do projeto.

Reconhecendo que sistemas de IA podem ser utilizados em variados contextos, com diferentes níveis de risco para a esfera de direitos dos indivíduos (e.g. traduções automatizadas versus aplicações na medicina), em quais circunstâncias e contextos deve ser preservada a determinação humana em decisões tomadas por sistemas de IA?

Nesta consulta, defendemos regras de transparência, responsabilização pública (*accountability)*, criação de comitês públicos de controle, elaboração de relatórios de impacto prévios e posteriores à implementação e estabelecimento de direitos individuais em decisões tomadas ou informadas por sistemas de IA (vide comentários acima)

Tais salvaguardas devem ser aplicadas em todas as situações em que um sistema de inteligência artificial puder afetar direitos e liberdades individuais, seja tal uso pelo poder público ou por entidades privadas. Por exemplo, no seu uso para prestação de

serviços públicos, segurança pública, seleção de candidatos para emprego, saúde privada ou pública, como suporte ao serviço jurisdicional etc.

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

Muito se fala da *black box* da inteligência artificial, essa "caixa preta", cujo interior não pode ser visualizado, onde ocorre o processamento do sistema. Em determinadas maneiras de aplicação do aprendizado de máquina*,* especialmente em *deep learning,* as informações externas que são alimentadas ao sistema – os *inputs* – são direcionadas a uma rede de "neurônios artificiais" que processam os dados e distribuem os comandos necessários – os *outputs* – para operar o sistema. O funcionamento interno desses sistemas é tão intrincado que até mesmo os desenvolvedores que os projetam não são tecnicamente capazes de apontar motivos específicos que os levem a tomar determinada decisão. E, da mesma forma, não há ainda nenhuma maneira óbvia de projetar tais sistemas para que passem a ser capazes de fornecer tal explicação, por mais que pesquisas nesse sentido tenham sido realizadas nos últimos tempos. Tendo-se isso em mente, a explicabilidade é um conceito que deve ser aplicado não de forma absoluta, mas como mais um dos elementos da transparência necessária para a operação desse tipo de sistema.

A explicabilidade deve referir-se a uma maneira de permitir que indivíduos exerçam seus direitos e que o público tenha acesso às principais questões associadas a determinado sistema de IA. Assim, por exemplo, se houver transparência quanto aos efeitos esperados de determinado sistema, as finalidades de seu uso, quais bases de dados foram utilizadas para seu treinamento, e outros elementos que apontamos em outros locais dessa consulta, estará aberto o espaço para controle público e auditoria dos sistemas, assim como para exercício de direitos potencialmente afetados por eles.

Em que medida a legislação brasileira requer atualização para endereçar as diferentes questões decorrentes da crescente adoção de sistemas autônomos em diferentes campos de aplicação?

Atualizações às leis de proteção de dados ou a promulgação de novas normas voltadas especificamente ao setor podem ser úteis ao estabelecer diversas salvaguardas necessárias para um uso não nocivo da inteligência artificial. A criação de comitês públicos, regras de transparência, direito à revisão humana, bem como a exigência de relatórios de impacto prévios e posteriores e de responsabilização pelo uso de sistemas com potencial de lesão a liberdades individuais, dentre outras sugestões que trouxemos nessa consulta, poderão ser endereçados por lei.

Em quais campos de aplicação de IA há necessidade mais premente de atualização das normas atualmente vigentes (por exemplo, no campo da segurança pública, no campo dos veículos autônomos, no campo da saúde, etc)?

Em todas as situações em que um sistema de inteligência artificial puder afetar direitos e liberdades individuais, seja tal uso pelo poder público ou por entidades privadas. Por exemplo, no seu uso para prestação de serviços públicos, segurança pública, seleção

de candidatos para emprego, saúde privada ou pública, como suporte ao serviço jurisdicional etc.

Como deve ser tratada a responsabilidade civil, penal e administrativa por danos causados com uso da IA?

A black box da inteligência artificial, assim como sua capacidade de tomar decisões autônomas, i.e., de certa forma independentes da programação inicial dos desenvolvedores do sistema e da esfera de controle de seus usuários, não devem ser argumentos para o afastamento da responsabilidade pela criação ou inserção de um sistema autônomo na sociedade, tratando-se da assunção de um risco.

Para possibilitar a responsabilização e *accountability* pela comercialização de sistemas autônomos, em especial nos casos em que sua utilização apresentar riscos às liberdades e direitos fundamentais, ou se colocar como requisito para exercício de um direito, recomendam-se as medidas sugeridas em nossos comentários anteriores nesta seção.

**IV. Governança de IA**

Seria conveniente estabelecer a obrigatoriedade de elaboração de relatórios prévios de avaliação de impacto quanto ao uso de IA em determinados setores?

O uso de sistemas de inteligência artificial em atividades que gerem riscos às liberdades civis e a direitos fundamentais deve ser acompanhado de regras claras de transparência. Relatórios de avaliação de impacto, com conteúdo e formato similares ao Relatório de Impacto à Proteção de Dados Pessoais, podem ser uma valiosa ferramenta para permitir controle público, responsabilização (*accountability*) e transparência pelo uso de sistemas de IA. O relatório deve ser publicamente disponibilizado, por exemplo no site da empresa ou órgão que o oferece ou dele faz uso, e deve fornecer informações claras quanto:

Ao fato de tal sistema estar sendo utilizado, incluindo informações sobre sua finalidade e locais de uso;

- quais os mecanismos utilizados para controle de vieses a partir das bases de dados e da seleção do modelo;
- qual o modelo algorítmico utilizado, se o sistema é atualizado, e se sim, como e com que frequência;
- qual a origem da base de dados utilizada para seu treinamento; e
- informações de contato para o exercício de direitos individuais.

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[1] e do Automated Decision Systems Task Force de Nova Iorque[2]?

Sim. A criação de estruturas institucionais capazes de elaborar diretrizes concretas e técnicas para o desenvolvimento de sistemas de IA menos discriminatórios e mais socialmente responsáveis, assim como para sua utilização de forma ética, exemplos do que provemos extensivamente nos outros comentários a essa consulta, seria recomendada.