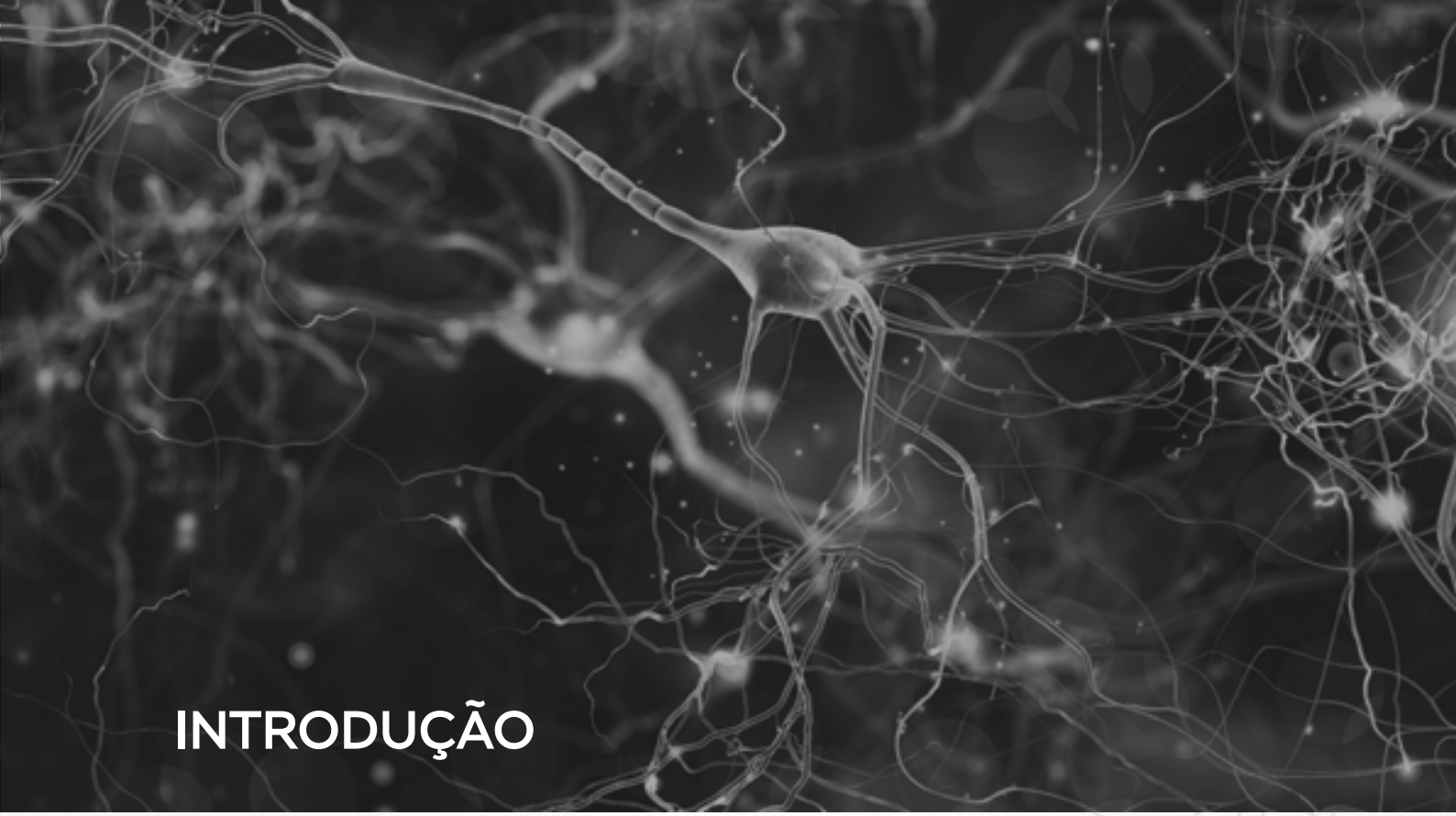
RELATÓRIO DE IMPACTO DE INTELIGÊNCIA ARTIFICIAL

ARTIFICIAL INTELLIGENCE IMPACT ASSESSMENT

Metodologia Daniel

DANIEL



INTRODUÇÃO

Cada vez mais aplicações de inteligência artificial são utilizadas para impulsionar negócios. Contudo, com essa utilização têm surgido também casos relevantes de vieses discriminatórios. São casos de algoritmos de seleção de profissionais para vagas que tendem a recrutar menos mulheres; modelagens para concessão de crédito que preterem pessoas negras; definição de perfil sobre identidade sexual ou de gênero a partir de reconhecimento biométrico ou por extração de conteúdo online, entre tantos outros.

Vieses podem ser inerentes aos dados utilizados e também emergentes dos algoritmos utilizados pela aplicação de inteligência artificial. Eles fazem parte das decisões tomadas por esses algoritmos. Porém, sempre que revelarem aspectos discriminatórios, nós os consideramos antiéticos, ilegais e, consequentemente, riscos patrimoniais e reputacionais são gerados às organizações.

Nos últimos anos, diversas iniciativas governamentais, legislativas e acadêmicas têm se dedicado a traçar diretrizes não apenas principiológicas, mas principalmente procedimentais para lidar com essa questão. Nesse contexto, há um consenso sobre a necessidade de uma governança direcionada ao desenvolvimento e utilização ética de inteligência artificial, que passa, necessariamente, pela análise de risco sobre o impacto que cada aplicação tem sobre os direitos e liberdades das pessoas.

Evidentemente, direitos e liberdades individuais já são previstos constitucionalmente e podem ser exigidos a qualquer tempo. Mas além deles, a legislação de proteção de dados pessoais, nos termos da União Europeia e do Brasil, exige que as organizações prestem contas (accountability) sobre as cautelas adotadas para que esses direitos e liberdades sejam preservados.

A Lei Geral de Proteção de Dados Pessoais brasileira (LGPD) é ainda mais específica ao tratar de possíveis aspectos discriminatórios de decisões automatizadas.

Não poderia ser diferente. O Brasil é um país incrivelmente diverso e, por isso, naturalmente suscetível a múltiplos vieses discriminatórios em aplicações de inteligência artificial. Aqui, a complexidade dos algoritmos soma-se à complexidade da diversidade e exige maior rigor na adoção de metodologias para análise de risco.

Como advogados e pesquisadores, desenvolvemos na área de Tecnologia, Privacidade e Proteção de Dados da Daniel uma metodologia para detecção e análise de risco de vieses discriminatórios nas aplicações de inteligência artificial que pode ser aplicada tanto pelas empresas desenvolvedoras dessas tecnologias, como pelas organizações que as adquirem. O resultado dessa análise é consolidado em um documento essencial para a prestação de contas: o Artificial Intelligence Impact Assessment (AIIA) / Relatório de Impacto de Inteligência Artificial (RIIA).

O AIIA / RIIA é um documento-chave para as aplicações de inteligência artificial em todo o mundo, mas, embora haja um consenso geral sobre a sua importância na prestação de contas inserida na governança de inteligência artificial, o desenvolvimento de uma metodologia para elaborá-lo é ainda muito incipiente em todo o mundo e inédito no Brasil.

Esta proposta de metodologia apresenta nossa contribuição para o tema: a metodologia Daniel de Artificial Intelligence Impact Assessment (AIIA) / Relatório de Impacto de Inteligência Artificial (RIIA), desenvolvida a partir de anos de pesquisa e experiência prática com compliance de Tecnologia e Proteção de Dados, especialmente para auxiliar as organizações a desenvolver ou utilizar aplicações de inteligência artificial nos negócios de forma ética.

Nuria López

Thamilla Talarico

SUMÁRIO

COMO UTILIZAR APLICAÇÕES DE IA DE FORMA ÉTICA E SEGURA?

AIIA - METODOLOGIA DANIEL 05

· Metodologia mais específica, voltada para uma aplicação concreta
· Mais foco em diversidade
· Análise integrada com as probabilidades de riscos no Brasil

PASSOS PARA O AIIA 06

· Pilares da Daniel Plural
· Banco de Questões Daniel
· Opacidade, nitidez e explicabilidade
· Análise humana
· Relatório Final

QUAL FOI A BASE PARA O DESENVOLVIMENTO DO AIIA - METODOLOGIA DANIEL?

CONCEITOS 10

· Inteligência Artificial
· Vieses
· Vieses Discriminatórios

DIRETRIZES PARA O DESENVOLVIMENTO E O USO ÉTICO DE IA 14

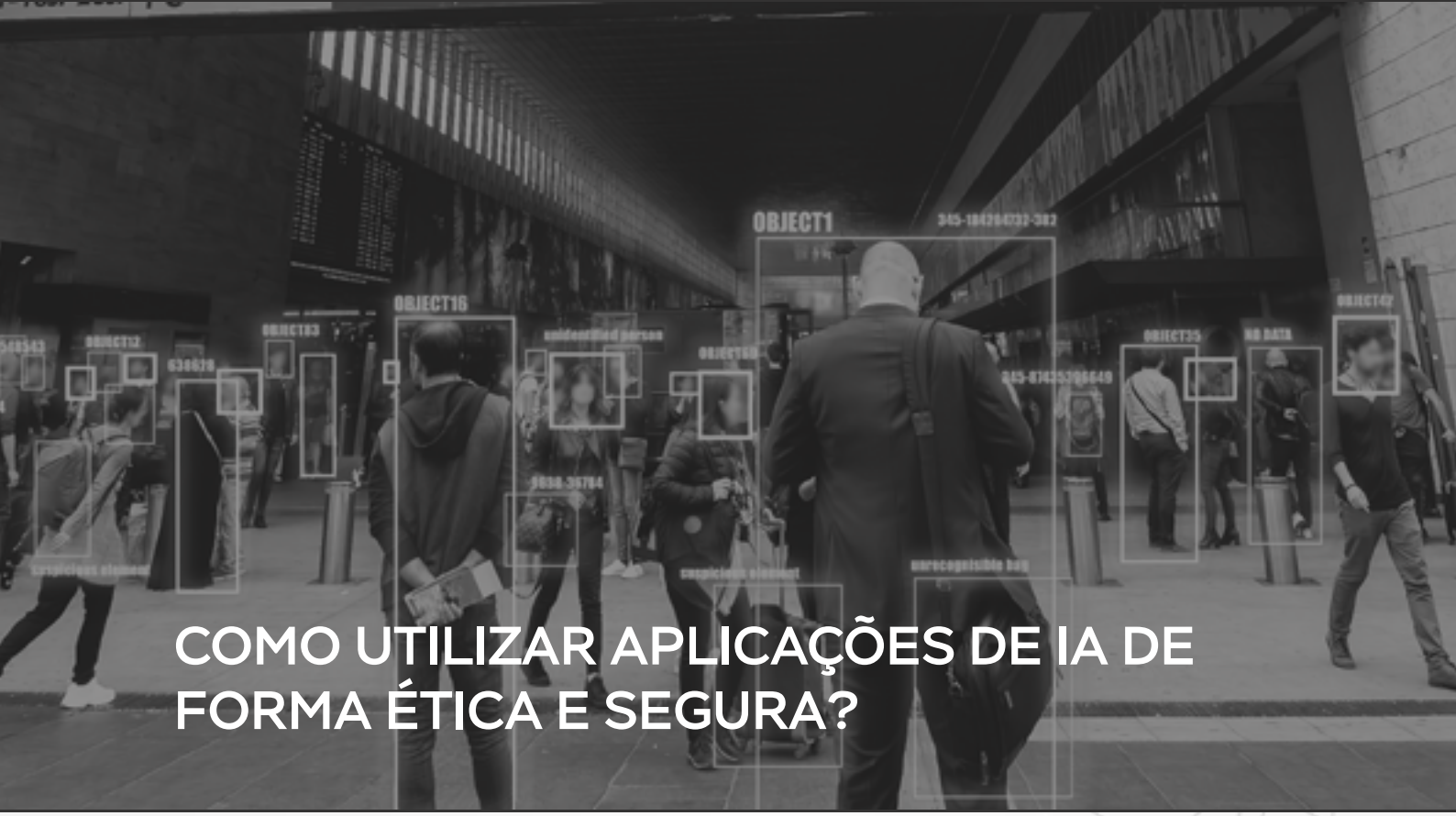
· Princípios

DOS PRINCÍPIOS À PRÁTICA 17

· Padrões técnicos
· Regulatório de IA

O INÍCIO DO REGULATÓRIO: A ENTRADA PELA LEGISLAÇÃO DE PROTEÇÃO DE DADOS PESSOAIS 19

· Europa
· Brasil



COMO UTILIZAR APLICAÇÕES DE IA DE FORMA ÉTICA E SEGURA?

AIIA - METODOLOGIA DANIEL

O que é o AIIA?

Como mencionado acima, a Avaliação (ou na tradução para o português brasileiro, Relatório, para manter a nomenclatura do Relatório de Impacto à Proteção de Dados Pessoais) específica para detecção de aspectos discriminatórios no tratamento de dados por aplicações de inteligência artificial é o Relatório de Impacto de Inteligência Artificial / *Artificial Intelligence Impact Assessment.*

Nesse sentido, existem algumas iniciativas em curso para criação de metodologias de avaliação de impacto de IA no mundo. Concretamente, podemos mencionar duas, mais acadêmicas, de 2018: um framework sobre Algorithm Impact Assessments para prestação de contas no setor público publicado pelo *AI Now Institute da New York University* e o *Artificial Intelligence Impact Assessment* da holandesa ECP – Plataform for the Information Society.

E outras duas, publicadas em 2020, por instituições governamentais: a *Assessment List for Trustworthy AI* (ALTAI) sugerida em 2020 pelo Grupo de Altos Especialistas em IA da Comissão Europeia para operacionalizar os sete pilares para IA confiável e o *Algorithm Impact Assessment* obrigatório para a maior parte dos órgãos federais no Canadá pela Diretiva sobre tomadas de decisão automatizadas (*Directive on automated decision-making*) em vigor desde 1º de abril de 2019 e aplicável a tecnologias a partir de 1º de abril de 2020.

Por que uma metodologia Daniel?

Acompanhando as discussões globais sobre o tema e o dia a dia da prática das aplicações de IA por organizações com atuação no Brasil, sentimos a necessidade

de desenvolver uma metodologia que atendesse especificamente algumas preocupações:

· **Metodologia mais específica, voltada para uma aplicação concreta**. Sem dúvidas, os riscos de aplicações de IA entrarão no radar da Governança Corporativa e são muito positivas as iniciativas mais abrangentes que exigem a criação de todo um contexto de Governança de IA. Contudo, é importante também que as organizações possam contar com uma metodologia mais específica, que verifique os riscos de uma aplicação em concreto.

· **Mais foco em diversidade**. Uma dificuldade que encontramos em outras metodologias foi a falta de diversidade contemplada quando falamos em não discriminação. De fato, o Brasil talvez seja o país do mundo com maior diversidade e a nossa legislação de proteção de dados pessoais (LGPD) foi bastante específica ao adotar o princípio da não discriminação e exigir informações a respeito de eventuais aspectos discriminatórios em decisões automatizadas, inclusive sob pena de auditoria algorítmica pela ANPD. Por isso, elaboramos uma metodologia que pudesse ser dinâmica a ponto de atender à nossa diversidade e auxiliar as organizações a fazer a prestação de contas especificamente sobre não discriminação já para os termos da LGPD.

· **Análise integrada com as probabilidades de riscos no Brasil**. Os riscos mapeados com a análise de IA não são abstratos. Eles têm probabilidade maior ou menor de materialização em um contexto específico, que deve ser considerado na mensuração do impacto. Isto significa que a análise deve levar em consideração pontos como o setor de atuação da organização, a atuação do Judiciário, da Autoridade Nacional de Proteção de Dados (ANPD) e de eventuais autoridades regulatórias sobre o tema.

Como funciona a Metodologia Daniel?

Aliamos nossas pesquisas à nossa experiência prática e elaboramos a Metodologia Daniel para o *Artificial Intelligence Impact Assessment* (AIIA), que passa pelos seguintes passos:

1. Pilares da Daniel Plural

A **Daniel Plural** é o Programa de Diversidade e Inclusão da Daniel Advogados (saiba mais em https://www.daniel-ip.com/plural-pt/) e nossa inspiração para estruturar os pilares de classificações de análise do AIIA.

Partimos dos grupos de afinidades **raça, gênero, LGBTQIA+ e parentalidade** para classificar as questões do Banco de Questões Daniel.

2. Banco de Questões Daniel

Uma das perguntas de ouro nos centros de pesquisa de Ética em IA é saber quais são as questões que devem ser feitas em uma avaliação de impacto de IA. Que perguntas devemos fazer para verificar eventuais vieses discriminatórios e que sirvam a tantas aplicações distintas abrangidas pelo amplo conceito de "inteligência artificial"?

Sabemos pelos estudos desenvolvidos, abaixo referenciados, que apenas questões mais genéricas são capazes de lidar com todo o espectro de aplicações de IA. Além disso, dada nossa preocupação em atender a diversidade brasileira precisamos também nos perguntar: que perguntas devemos fazer para endereçar tantas formas diversas possíveis de discriminação?

Não tem segredo: ao invés de elaborarmos algumas questões que sirvam para todas as aplicações de IA e para qualquer tipo de risco, **elaboramos um Banco de Questões a serem articuladas de forma customizável a cada caso concreto**. Assim, é possível realizar a avaliação dentro das especificidades de cada aplicação de IA e dos riscos mapeados para o contexto.

Além disso, o Banco de Questões é uma construção contínua, colaborativa e dinâmica: está sempre sendo alimentado com novas questões, por especialistas com representatividade diversa, que contribuem com novas perspectivas de diversidade e inclusão.

3. Opacidade, nitidez e explicabilidade

Outra questão difícil sobre as avaliações de IA é o que chamamos de opacidade algorítmica. Em algumas aplicações de IA os desenvolvedores têm pleno controle sobre a razão dos resultados obtidos pelos algoritmos.

É o caso de aplicações mais simples como as dos sistemas de conhecimento. Contudo, nos casos em que há redução do controle humano sobre os parâmetros de atuação da IA, como no caso de *machine learning* não supervisionados ou redes neurais (que são IA mais complexas), é mais difícil compreender a razão de seus resultados precisamente porque não conhecemos completamente os parâmetros, padrões ou pesos estabelecidos internamente. Por isso, dizemos que os algoritmos podem ser "opacos" para nós.

A opacidade é um problema relevante para o desenvolvimento e uso ético de IA porque nos impede de conhecer as razões de determinado resultado, que pode, por exemplo, ser discriminatório. A opacidade dificulta a explicabilidade sobre determinadas decisões algorítmicas. Como a explicação é importante, inclusive sob a perspectiva das atuais legislações de proteção de dados, uma grande tendência atual é o desenvolvimento de *explainable AI* (ou X-AI).

Entretanto, nem sempre a aplicação de IA sob análise será uma *explainable AI* (ou X-AI). Em regra, ela será mais ou menos opaca. Dessa forma, nossa metodologia precisa ser útil para todos os graus de opacidade algorítmica.

Na pior das hipóteses, podemos pensar nos algoritmos analisados como uma caixa totalmente fosca. Mesmo nessas condições adversas, podemos analisar os dados de origem (que entram como *input* para alimentar os algoritmos) e os dados dos resultados (*output*).

Grande parte dos vieses decorre dos dados de origem (as bases de dados utilizadas inicialmente) e, sobre esse ponto, é comum termos visibilidade. Mesmo que existam muitas bases convergentes em um *data lake*, por exemplo, é possível mapeá-las. Além disso, os resultados também são nítidos e devem ser avaliados sob a perspectiva da não discriminação.

A nitidez (ou transparência) no algoritmo ajuda a compreender como o resultado discriminatório foi obtido e, consequentemente, a encontrar possibilidades técnicas de correção do viés.

Contudo, mesmo que não saibamos como se deu o resultado discriminatório, é possível saber que ele ocorreu e endereçar soluções no nível corporativo em cada caso. Ao mesmo tempo, é

importante, documentar quando o resultado não é discriminatório, a fim de obter as informações necessárias para garantia do direito do artigo 20, LGPD e para prestação de contas para autoridades judiciais e administrativas, inclusive a ANPD.

4. Análise humana

Outro ponto relevante da metodologia Daniel é a análise humana dos vieses mapeados em cada caso. Colocar um "humano no loop" de análise de decisões automatizadas é algo que surgiu com a regulamentação específica sobre decisões automatizadas pela legislação de proteção de dados pessoais. O GDPR em sua Consideranda 71 e no artigo 22º, 3, estabelece a intervenção humana como uma garantia dada ao titular de dados.

Como perspectiva de futura regulamentação europeia, o estudo da Comissão Europeia sobre o tema declara que "a supervisão humana contribui para garantir que um sistema de IA não põe em causa a autonomia humana nem produz outros efeitos negativos. O objetivo de uma IA fiável, ética e centrada no ser humano só pode ser alcançado através da garantia de um envolvimento adequado dos seres humanos em aplicações de IA de alto risco."

A importância da análise humana em qualquer avaliação ética e jurídica de inteligência artificial tem uma razão clara. Os vieses nada mais são do que tendências, distribuições estatísticas não normais. Detectar vieses não significa realizar uma avaliação ética e jurídica de seus impactos. Vieses podem ter impactos diversos em contextos diversos. Podemos ter, por exemplo, vieses que reforçam políticas corporativas de Diversidade & Inclusão e são, portanto, desejáveis.

A compreensão de que determinado viés constitui aspecto discriminatório exige uma cognição mais complexa sobre os aspectos sociais, nos quais a aplicação de IA está inserida, e que seja capaz de produzir juízos morais.

Os juízos morais são decisões que tomamos sobre o que é certo ou errado, justo ou injusto. Embora essas decisões nos pareçam simples, elas são fruto de uma rede neuroquímica complexa e dinâmica, que se retroalimenta de forma reflexiva a partir de nossas interações sociais. É um processo sofisticado que demanda características humanas.

Trabalhar com inteligência artificial exige reconhecer quais atividades podem ser melhor executadas por habilidades específicas não-humanas e quais são melhor executadas por humanos. A análise voltada para questões éticas, inclusive como as de não discriminação, são melhor executadas pelo processo mais complexo – o da nossa cognição.

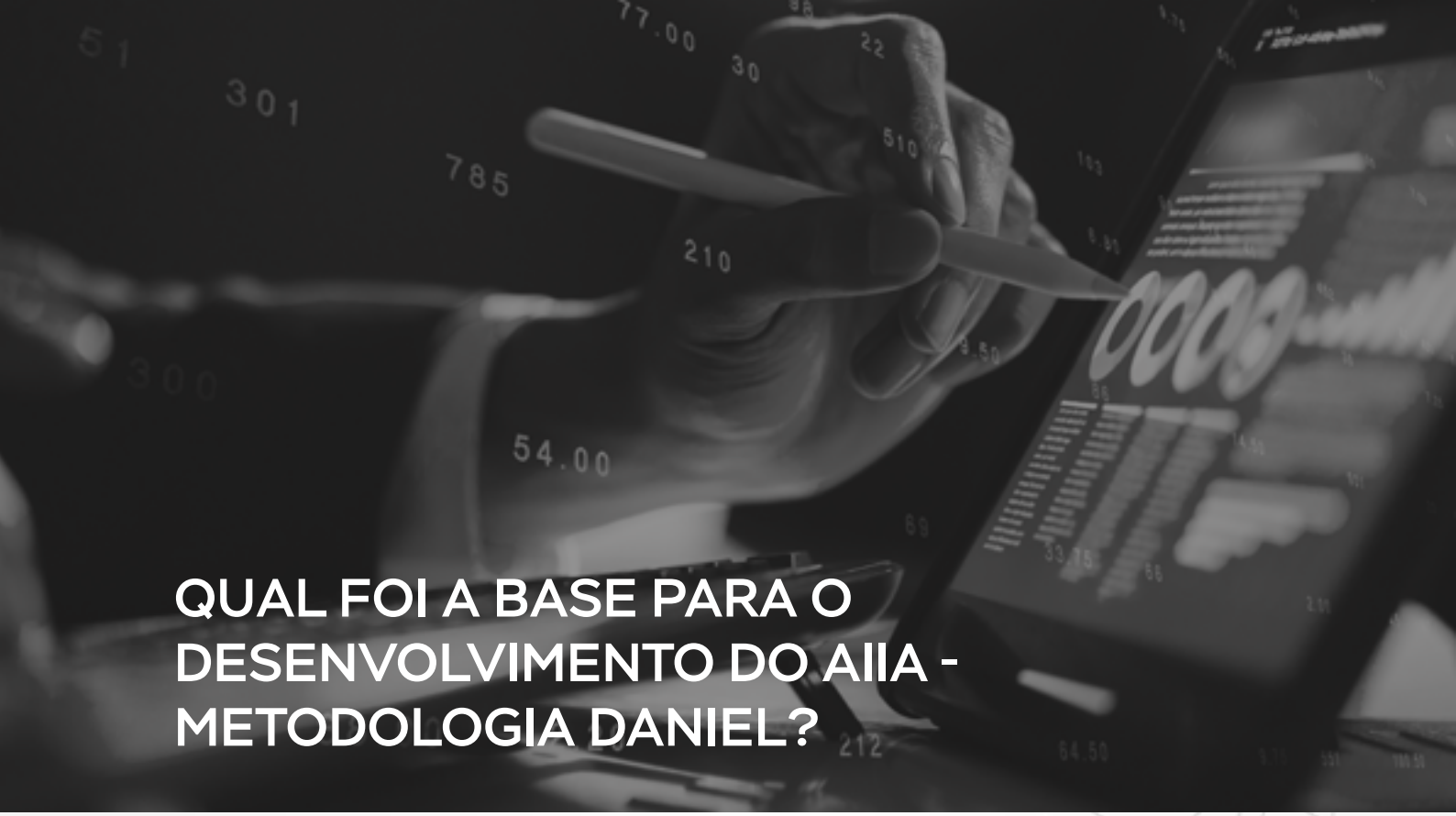
Incluir um "humano no loop" é uma prática que pode ser realizada a partir do AIIA.

Não significa que um humano revisará cada decisão tomada por uma aplicação de IA, o que na maior parte dos casos não seria possível ou mesmo desejável, se os algoritmos estiverem cumprindo seus papeis melhor do que nós faríamos. Mas, em algum momento, é importante inserir um "humano no loop" para que ele possa verificar se as decisões tomadas pela IA são éticas e não impactam negativamente os direitos e liberdades de outras pessoas em razão de raça, etnia, identidade de gênero, orientação sexual, dentre outros.

5. Relatório Final

O produto de um AIIA é seu Relatório Final. Ele apresenta a comprovação dos vieses enquanto distribuições estatísticas não normais, somada à demonstração da análise ética feita para determinar se aquela aplicação em concreto apresentou ou não aspectos discriminatórios. Quando houver possibilidade de viés discriminatório, o Relatório apresentará os riscos decorrentes e as recomendações práticas para endereçá-los.

Essas informações são imprescindíveis para o atendimento ao direito do artigo 20, da LGPD e, portanto, ao compliance com a legislação de proteção de dados pessoais. Mas não apenas isso. O Relatório Final do AIIA é sobretudo um instrumento para governança de riscos corporativos referentes à determinada aplicação de IA.



QUAL FOI A BASE PARA O DESENVOLVIMENTO DO AIIA - METODOLOGIA DANIEL?

CONCEITOS:

1. INTELIGÊNCIA ARTIFICIAL

Inteligência Artificial é um termo bastante amplo. Ele aponta para a programação de diversos tipos de algoritmos que buscam reproduzir habilidades humanas. Nos últimos setenta anos, o projeto de ensinar uma máquina a pensar saiu das universidades e mudou a realidade como a vivenciamos hoje.

Criamos bases de dados robustas. Imputamos estratégias de resoluções de problemas simples. Depois, de problemas complexos. Automatizamos funções. Desenvolvemos reconhecimento de imagens. Reconhecimento de discurso, em texto, em fala, processamento de linguagem natural. Até a habilidade mais importante: a de aprender, com supervisão, sem supervisão, tentando cada vez mais uma aproximação com a nossa rede neural.

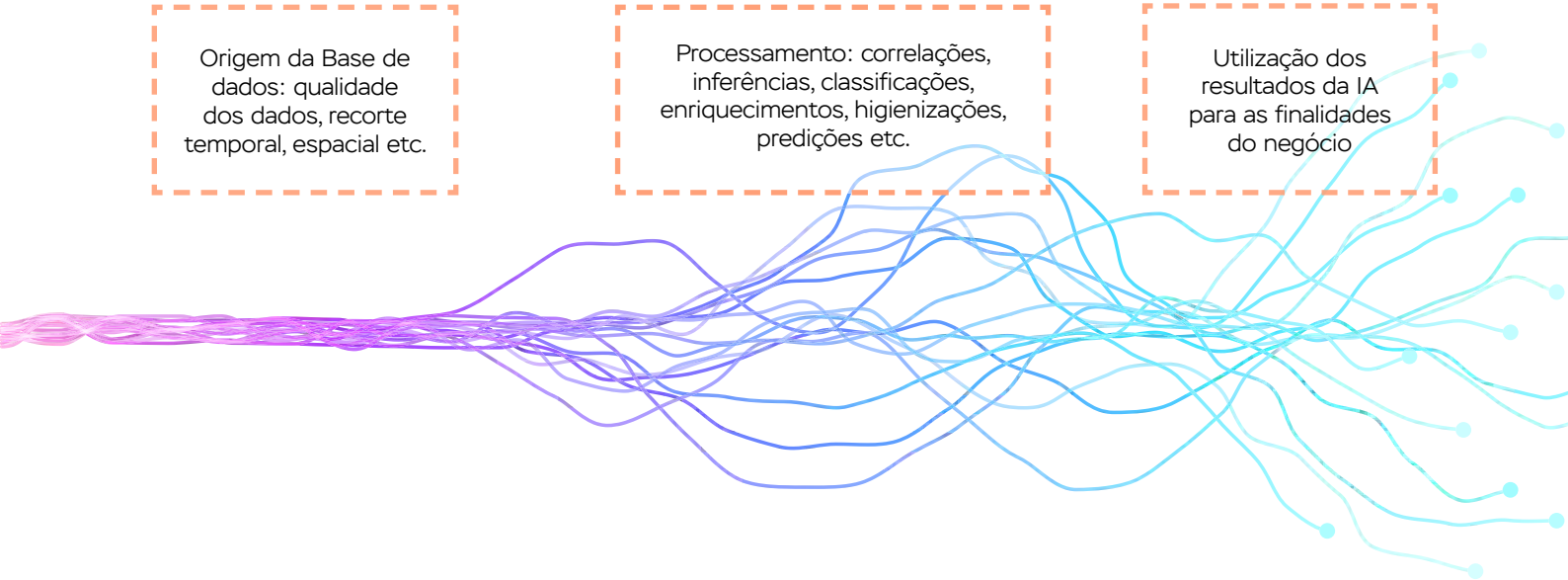
Aplicações de inteligência artificial estão presentes em todos os setores do mercado e ajudam a navegar as complexas tomadas de decisões exigidas nos dias de hoje. Quando nos referimos aos data driven businesses, melhor seria falar dos negócios guiados por aplicações de inteligências artificiais. São elas que têm capacidade de nos guiar pelos dados com uma eficiência maior do que nossa biologia permitiria. Por isso, gostamos da expressão inteligência aumentada. Os algoritmos, muitas vezes, nos complementam, tornam nossas decisões e nossos processos mais eficientes. Na maioria das vezes, teremos uma rede de decisões em que se mesclam decisões humanas e não-humanas, uma impactando a outra.

Estão em atividades do dia a dia: autenticam identidades, monitoram nossa saúde, reconhecem padrões, classificam clientes, traçam perfil de comportamento de consumo e crédito, selecionam profissionais para vagas em aberto, indicam lugares para morar, por qual caminho seguir, quais tratamentos têm mais chance de nos curar.

Os impactos das aplicações de inteligências artificiais muitas vezes passam despercebidos no cotidiano, mas têm contribuído em larga escala para o aumento da qualidade de vida das últimas décadas e impulsionado a transformação digital de negócios e o desenvolvimento de inovações.

A expressão inteligência artificial engloba sistemas baseados em conhecimento (*knowledge-based systems*), em que são imputadas e mantidas regras conhecidas de determinada área do conhecimento, e sistemas de aprendizagem (machine learning), em que os algoritmos são capazes de se retroalimentar de seus próprios resultados de forma a maximizar sua performance com base na experiência. O aprendizado de máquina pode ser (i) supervisionado, em que os dados iniciais são classificados e o desenvolvedor estabelece parâmetros específicos de atuação; (ii) não supervisionado, no qual os algoritmos detectam padrões e regras em uma base de dados inicial; ou (iii) semi-supervisionado, em que há uma abordagem intermediária ou por reforço, em que determinados resultados são recompensados dentro do mecanismo de retroalimentação.

Nesta proposta de metodologia nos referimos à inteligência artificial ou às aplicações de inteligência artificial de forma ampla, para abranger quaisquer produtos ou serviços que incorporem componentes de algoritmos que possuem essa qualidade mimética de alguma habilidade originariamente humana.



2. VIESES

Ao selecionar as bases de dados a serem utilizadas por uma aplicação de inteligência artificial, escolhe-se dentre as disponíveis, as desejáveis e as que serão mais úteis a um determinado projeto. As bases de dados escolhidas representarão, necessariamente, um recorte de dados coletados em um determinado tempo e espaço.

Os algoritmos utilizados, ao executarem comandos, também traduzem tendências que podem se relacionar à subjetividade de seus desenvolvedores

ou à extrapolação de tendências pré-existentes no banco de dados.

A recente ISO/IEC 24027 classifica vieses algorítmicos (que ela chama de "vieses de dados", porque compreende que eles sejam uma grande fonte de vieses) de forma não taxativa em:

- De estatísticas – decorrentes, por exemplo, do espaço amostral, da falta de respostas, da atuação de diversas variáveis, e da não-normalidade na distribuição
- De classificação – decorrentes de classificações enviesadas (inclusive por vieses humanos, de nossa própria subjetividade)
- De amostra não-representativa – decorrentes do uso de amostras que não representam a diversidade necessária para treinar os algoritmos
- Da falta de características ou de classificações – baixa ocorrência de determinadas características ou de classificações em uma base de dados
- De pré ou pós-processamento pelo algoritmo – correção de valores, novos inputs etc.
- Simpson's paradox – o viés aparece em diversos grupos de dados, mas desparece ou é corrigido quando esses grupos são combinados.
- Da agregação de dados que tem distribuições estatísticas diferentes
- Do Treinamento distribuído – o aprendizado é distribuído em diferentes treinamentos em que cada um deles pode gerar vieses em razão de diferentes distribuições de características

Assim, cada avanço no desenvolvimento e na aplicação de uma inteligência artificial implica possibilidade de novos vieses, revelam novas tendências. São como rastros dos dados, das metodologias aplicadas nos algoritmos e das pessoas que passaram, de alguma forma, por esse processamento.
Toda aplicação de inteligência artificial tem vieses, isto é, todas elas apresentam tendências. Sem contar os nossos próprios vieses, na utilização dos resultados obtidos pelos algoritmos.

3. VIESES DISCRIMINATÓRIOS

Vieses são parte do tratamento de dados por inteligências artificiais.

Eles não são em si mesmos bons ou ruins. É importante notar que o principal ponto de referência quando classificamos o impacto dos resultados de uma IA são os direitos e liberdades das pessoas que estão sujeitas a eles. São esses resultados que podem causar impactos positivos, indiferentes ou negativos aos direitos e liberdades das pessoas.

Podemos ter impactos positivos, por exemplo, com algoritmos cujos parâmetros sejam programados para reforçar metas estabelecidas em políticas de diversidade & inclusão. Nesse caso, teremos um viés desejado e propositadamente embarcado na IA para ampliar oportunidades de acesso ao trabalho. Podemos ter impactos indiferentes para a finalidade, por exemplo, com um algoritmo que classifique os currículos por ordem alfabética (considerando que todos serão, posteriormente, igualmente analisados). Contudo, podemos ter também impactos negativos, por exemplo, se tivermos um algoritmo que deverá aprender qual perfil ideal para um cargo sênior a partir de uma base de dados que historicamente terá mais homens / brancos. O algoritmo reproduzirá o padrão histórico com o qual é desejável romper para incluir mais diversidade nas organizações.

Assim, em razão de determinados vieses refletidos nos resultados dessas aplicações de inteligências artificiais muitas pessoas podem ter seus direitos e liberdades ameaçados ou violados. Por exemplo, menos acesso a vagas de trabalho, piores scores de crédito, preços por serviços e produtos mais caros, serem indicados para remunerações menores, indicações menos precisas para tratamentos médicos ou fraudes graves em suas identidades por reconhecimento biométrico.

Os vieses são apenas tendências, mas dentro de determinados contextos, podem reforçar discriminações sociais, por exemplo em razão de raça ou etnia, de identidade de gênero ou orientação sexual, como um ciclo vicioso.

É nesse âmbito que se revelam os aspectos discriminatórios em vieses algorítmicos.

Portanto, há possibilidade de riscos, primeiramente, de danos aos direitos e liberdades dos indivíduos sujeitos ao resultado que reflete o aspecto discriminatório e, posteriormente, à toda coletividade discriminada. Se materializado, esse risco se traduz na manutenção de um ciclo vicioso de discriminação de grupos sociais inteiros. Para qualquer negócio, há a perda de não contar com diversidade e o peso de ser um agente de discriminação. Juridicamente, soma-se a indenização devida aos afetados, individual e coletivamente, e o imenso dano reputacional decorrente do caso.

Por muito tempo, o impacto dos vieses discriminatórios foi pouco explorado. Foi com o aumento exponencial do uso de aplicações de inteligência artificial em todo o mundo que cresceram casos de denúncias e consequentes condenações contra o uso de algoritmos com vieses discriminatórios.

Conhecer esses casos foi o primeiro passo para mudança.

DIRETRIZES PARA O DESENVOLVIMENTO E O USO ÉTICO DE IA

Ao conhecer casos de discriminação, empresas, associações e governos de todo o mundo passaram a desenvolver possíveis soluções para o desenvolvimento e o uso ético de IA.

PRINCÍPIOS

Um amplo levantamento de princípios de IA em todo o mundo, incluindo muitos dos documentos apontados no mapa acima, encontrou os seguintes temas em comum:

- Privacidade
- Prestação de contas [*accountability*]
- Segurança
- Transparência e Explicabilidade
- Justiça [*Fairness*] e não discriminação
- Controle humano da tecnologia
- Responsabilidade profissional
- Promoção de valores humanos

E destacou que "com as preocupações sobre vieses de IA já impactando indivíduos globalmente, os princípios de justiça e não discriminação nos sistemas de IA a serem desenhados e utilizados para maximizar justiça e promover a inclusão. Justiça e não discriminação são princípios presentes em 100% dos documentos de nossa base de dados."

O relatório destaca que no tema justiça e não discriminação encontrado nos documentos, 89% referem-se a não discriminação e a prevenção de vieses; 36% a dados representativos e de qualidade; 56% a justiça; 47% a inclusão no design; 42% a inclusão em impacto; e 25% a equidade.

Clique no bullet para ler mais:

- Partnership on AI
- Preparing for the future of AI
- Asilomar AI Principles
- 10 principles for ethical AI
- Whitepaper on AI standardization
- For a meaninful AI
- AI in the UK
- AI in Europe
- Toronto Declaration
- European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment
- Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector
- AI Principles & Ethics
- Social Principles of Human-Centric AI
- Ethically Aligned Design
- European High Level Expert Group on AI
- OECD Principles on AI
- Beijing AI Principles
- G20 AI Principles
- Governance Principles for a New Generation of AI
- Artificial Intelligence Ethics Framework for the Intelligence Community
- Recommendation on the Ethics of Artificial Intelligence
- Australia's Artificial Intelligence Ethics Framework
- Estratégia Brasileira de Inteligência Artificial
- NATO Artificial Intelligence Strategy

NÃO DISCRIMINAÇÃO COMO DIRETRIZ ÉTICA PARA IA:

"Os sistemas de IA devem ser projetados e operados para serem compatíveis com ideais de dignidade humana, direitos, liberdades e diversidade cultural."

ASILOMAR PRINCIPLES

"Igualdade, não discriminação e solidariedade – "Num contexto de IA, a igualdade implica que as operações do sistema não podem gerar resultados injustamente tendenciosos (p. ex., os dados utilizados para treinar os sistemas de IA devem ser o mais inclusivos possível, representando diferentes grupos da população)."

EUROPEAN HIGH LEVEL EXPERT GROUP ON AI

"Justiça e valores centrados no humano – "Os atores de IA devem respeitar o Estado de Direito, os direitos humanos e os valores democráticos, ao longo do ciclo de vida do sistema de IA. Estes incluem liberdade, dignidade e autonomia, privacidade e proteção de dados, não discriminação e igualdade, diversidade, justiça, justiça social e direitos trabalhistas reconhecidos internacionalmente. Para isso, os atores de IA devem implementar mecanismos e salvaguardas, como a capacidade de determinação humana, adequadas ao contexto e coerentes com o estado da arte." representando diferentes grupos da população)."

OCDE

DOS PRINCÍPIOS À PRÁTICA

As diversas iniciativas sobre os princípios que deveriam guiar o desenvolvimento e a utilização de inteligência artificial deram lugar a abordagens mais práticas, em especial nos últimos dois anos.

Padrões técnicos:

É o caso dos padrões da IEEE 7010-2020 - IEEE *Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* e da ISO/IEC TR 24027:2021 - *Bias in AI systems and AI aided decision making*, com controles técnicos específicos para endereçar, de forma ampla, os impactos de inteligências artificiais nos indivíduos.

Além das organizações de padronização, duas autoridades nacionais de proteção de dados pessoais, a ICO (*Information Commissioner's Office*) e a AEPD (*Agencia Española de Protección de Datos Personales*) lançaram guias práticos para verificações de risco em aplicações de IA.

A ICO com seu *AI and Data Protection Risk Toolkit* e a AEPD com seu guia de *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial*, ambos voltados ao Compliance com o Regulamento Europeu sobre Proteção de Dados (RGPD/ GDPR); *e os Requisitos de auditorias de tratamientos que incluyan IA*, este com controles para endereçar de forma mais ampla a governança em IA. Também como um avanço, cabe mencionar que a Espanha autorizou a criação da *Agencia Española de Supervisión de Inteligencia Artificial*.

Na esteira de abordagens práticas para avaliar aplicações de IA, o Fórum Econômico Mundial recém lançou um "Toolkit" voltado para o C-level, o *Empowering AI Leadership*.

Regulatório de IA:

Em paralelo, começam as movimentações para a criação de um regulatório específico para inteligência artificial. Em

2020, a Comissão Europeia sugeriu em seu estudo "Sobre a inteligência artificial - Uma abordagem europeia virada para a excelência e a confiança" alguns requisitos legais para aplicações de inteligência artificial:

- Dados de treino
- Conservação de registros e de dados
- Prestação de informações
- Robustez e exatidão
- Supervisão humana
- Requisitos específicos para determinadas aplicações de IA (p.ex. biometria à distância).

Atualmente, encontra-se em discussão a Proposta de Regulamento europeu para a inteligência artificial: o Artificial Intelligence Act:

"a proposta completa o direito da União [Europeia] em vigor em matéria de não discriminação com requisitos específicos que visam minimizar o risco de discriminação algorítmica, em particular no que diz respeito à conceção e à qualidade dos conjuntos de dados utilizados no desenvolvimento de sistemas de IA, complementados com obrigações de testagem, gestão de riscos, documentação e supervisão humana ao longo do ciclo de vida dos sistemas de IA."

O texto proposto classifica as aplicações de IA em quatro níveis de risco:

	Inaceitável (com exceções)
	Alto
	Limitado
	Mínimo

Quanto maior o risco, mais exigências em termos assessments (ou avaliações) desde antes da aplicação ser colocada no mercado e por todo seu ciclo de vida. As IA de risco alto, por exemplo, terão como requisitos:

✓ sistema de gestão de risco
✓ governança de dados
✓ documentação técnica
✓ manutenção de registro
✓ transparência e prestação de informações aos usuários
✓ supervisão humana
✓ exatidão, solidez e segurança cibernética.

No Brasil temos o Projeto de Lei nº 21/20, que discute o Marco Legal da Inteligência Artificial. Enquanto fomentam os debates sobre um regulatório específico em inteligência artificial, a legislação de proteção de dados pessoais já abre espaço e estabelece as primeiras normas sobre o tema.

O INÍCIO DO REGULATÓRIO: A ENTRADA PELA LEGISLAÇÃO DE PROTEÇÃO DE DADOS PESSOAIS

EUROPA

A legislação de proteção de dados pessoais surge voltada para o potencial impacto do tratamento automatizado de dados pessoais. Se voltarmos às primeiras legislações sobre o tema, veremos que a Lei sueca de 1973 tinha como escopo apenas o tratamento automatizado de dados pessoais, a Lei alemã de 1977 tinha disciplina distinta, mais rigorosa, para esses casos, a Lei francesa de 1978 é chamada, literalmente de a "lei de informática e liberdade" e a Convenção Europeia 108, razão pela qual se comemora globalmente o "dia da proteção de dados", foi assinada em 1981 como a "Convenção para a proteção de indivíduos com relação ao tratamento automatizado de dados pessoais".

Nesse sentido, a Diretiva europeia sobre proteção de dados pessoais (Diretiva 95/46/EC), incluiu no direito de acesso "o conhecimento da lógica subjacente ao tratamento automatizado dos dados que lhe digam respeito, pelo menos no que se refere às decisões automatizadas" (artigo 12). Além disso, ela impôs aos Estados-membros da União Europeia que reconheçam a "qualquer pessoa o direito de não ficar sujeita a uma decisão que produza efeitos na sua esfera jurídica ou que a afete de modo significativo, tomada exclusivamente com base num tratamento automatizado de dados destinado a avaliar determinados aspectos da sua personalidade, como, por exemplo, a sua capacidade profissional, o seu crédito, confiança de que é merecedora, comportamento" (artigo 15).

Talvez em 1995, ano da Diretiva Europeia, pudéssemos pensar que seria viável exigir o direito de não estar sujeito a uma decisão automatizada que impacte seus direitos e liberdades. Hoje, no entanto, o amplo alcance das aplicações de IA torna esse direito menos efetivo. Na prática, isso implicaria uma renúncia a produtos e serviços básicos do nosso dia a dia.

Apesar disso, o Regulamento Europeu sobre Proteção de Dados (GDPR), de 2018, reproduziu essas normas – mas acrescentou em seus considerandos a recente preocupação com a possibilidade de discriminação nas decisões automatizadas, nas quais se inserem as inteligências artificiais:

"A fim de assegurar um tratamento equitativo e transparente no que diz respeito ao titular dos dados, tendo em conta a especificidade das circunstâncias e do contexto em que os dados pessoais são tratados, o responsável pelo tratamento deverá utilizar procedimentos matemáticos e estatísticos adequados à definição de perfis, aplicar medidas técnicas e organizativas que garantam designadamente que os fatores que introduzem imprecisões nos dados pessoais sejam corrigidos e que o risco de erros seja minimizado, e proteger os dados pessoais de modo a que sejam tidos em conta os potenciais riscos para os interesses e direitos do titular dos dados e de forma a prevenir, por exemplo, efeitos discriminatórios contra pessoas singulares em razão da sua origem racial ou étnica, opinião política, religião ou convicções, filiação sindical, estado genético ou de saúde ou orientação sexual, ou a impedir que as medidas venham a ter tais efeitos" (Consideranda 71).

BRASIL

A Lei Geral de Proteção de Dados Pessoais (LGPD) brasileira avançou ainda mais na exigência de caráter não discriminatório.

A não discriminação é um dos princípios expressos na lei, que a descreve como a "impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos" (artigo 6º, IX, LGPD).

Além disso, no artigo 20, a LGPD estabelece o direito de solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses. A perspectiva que surge com esse direito é a de que os indivíduos possam ter acesso não apenas aos seus dados pessoais tratados (direito de acesso), mas também possam compreender como seus interesses podem ser afetados, inclusive para verificar que não há discriminação ilícita ou abusiva.

Por isso, o controlador (a quem competem as decisões referentes ao tratamento de dados pessoais) "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" (artigo 20, §1º, LGPD). Se não houver o fornecimento dessas informações, a Autoridade Nacional de Proteção de Dados (ANPD) "poderá realizar auditoria para verificação de aspectos discriminatórios em tratamento automatizado de dados pessoais" (artigo 20, §2º, LGPD).

Isto é, a preocupação central da LGPD no que concerne as decisões automatizadas é precisamente evitar que elas produzam discriminação.

Para estar preparado para garantir o direito do artigo 20, LGPD é preciso:

✓ Mapear quais são as aplicações de IA utilizadas pela organização
✓ Analisar quais são os impactos nos direitos e liberdades dos titulares (Relatório de Impacto à Proteção de Dados Pessoais)
✓ Analisar se existem aspectos discriminatórios

✓ Reunir as informações a serem apresentadas aos titulares
✓ Documentar as análises, em especial a referente aos aspectos discriminatórios, para eventual fiscalização da Autoridade, a fim de evitar a possibilidade de auditoria algorítmica

Os primeiros passos já são dados corriqueiramente em qualquer projeto de adequação à LGPD. É necessário mapear as atividades de tratamento de dados pessoais e, consequentemente, as que são realizadas por aplicações de IA, para: levantar as informações para o obrigatório Registro de Operações de Tratamento de Dados Pessoais - ROPA (artigo 37, LGPD); assim como em casos de tratamento realizado no legítimo interesse (artigo 10, §3º, LGPD), em atividades que revelem dados sensíveis (artigo 38, LGPD) e, como boa prática, para atividades de alto risco, realizar o Relatório de Impacto à Proteção de Dados Pessoais – RIIPD/DPIA (artigo 5º, XVII, LGPD).

No entanto, a LGPD não trouxe uma análise específica para verificação de aspectos discriminatórios. Apenas impôs a possibilidade de auditoria algorítmica, caso o controlador não forneça as informações solicitadas.

Em alguns casos, o controlador poderá ser o próprio desenvolvedor da AI. Mas, na maioria das situações o controlador será apenas usuário da aplicação de IA. Assim, a LGPD lhe impõe uma obrigação para a qual ele não dispõe de muitos meios para cumprir. Caso não a cumpra, poderá sofrer a referida auditoria, bem como arcar com os danos patrimoniais e reputacionais decorrentes de eventual discriminação resultante de seu uso da aplicação de IA.

Para suprir essa lacuna, desenvolvemos o AIIA – Metodologia Daniel.

Para saber mais, contate nossa equipe de Tecnologia, Privacidade e Proteção de Dados

equipetecnologia@daniel-ip.com

DANIEL

CC

www.daniel-ip.com