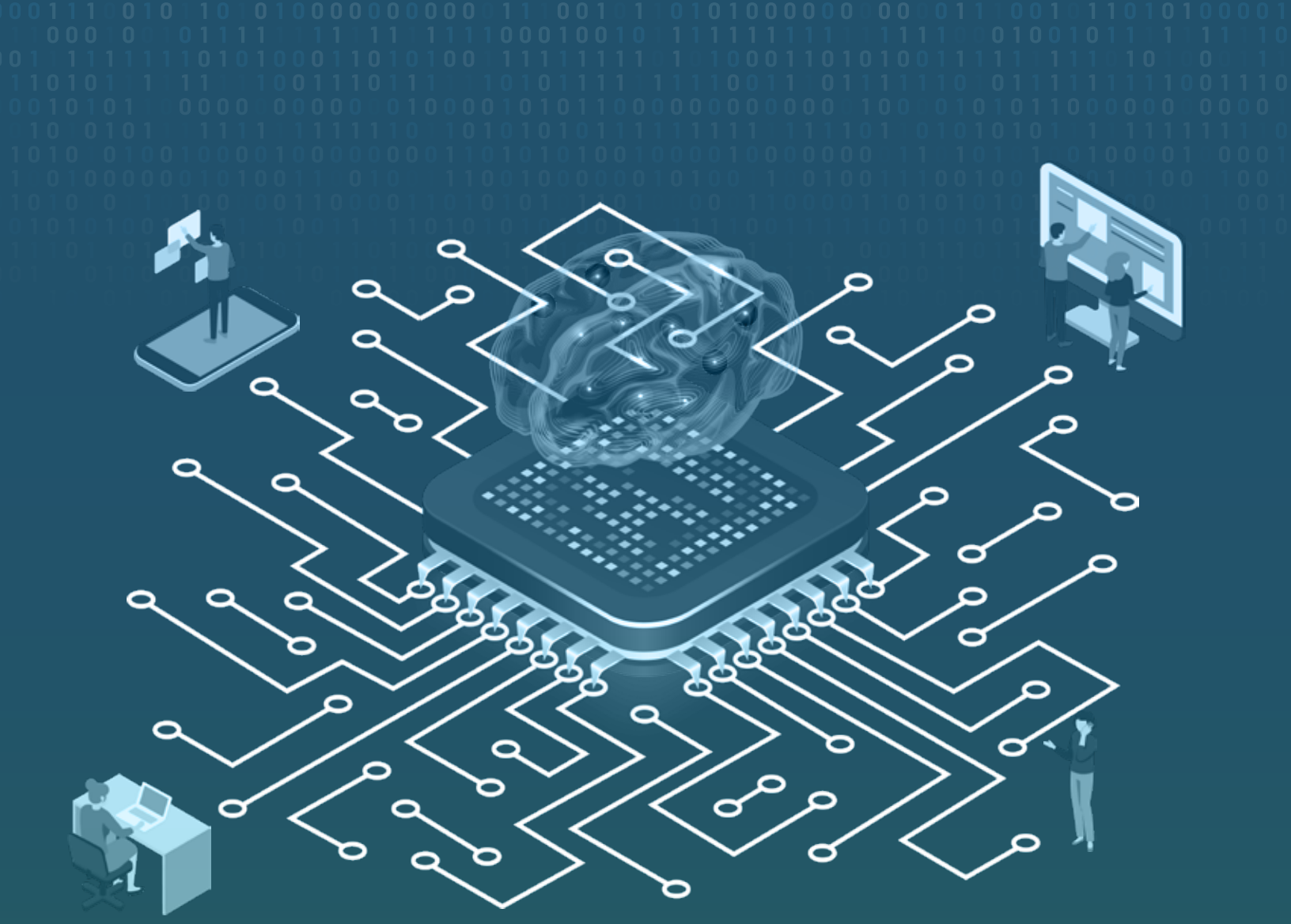


Auditoria algorítmica para sistemas de tomada de decisão ou suporte à decisão

Matías Aránguiz Villagrán

BID

Auditoria algorítmica para sistemas de tomada de decisão ou suporte à decisão

Matías Aránguiz Villagrán*
Professor e Subdiretor do Programa de Direito, Ciência e Tecnologia.
Pontifícia Universidade Católica do Chile

Março de 2022

Agradecimientos
A autora agradece todos os comentários de Cristina Pombo e a pesquisa de Sebastián Dueñas durante a redação deste documento. Da mesma forma, um agradecimento especial à AGESIC, Agência de Governo Eletrônico e a Sociedade da Informação e do Conhecimento do Uruguai, especialmente Maximiliano Maneiro pela revisão e comentários, e à Eticas Consulting pelas discussões iniciais para este documento.

https://www.iadb.org/

Copyright © 2022 Banco Interamericano de Desenvolvimento. Esta obra está licenciada sob uma licença Creative Commons IGO 3.0 Atribuição-NãoComercial-SemDerivações (CC BY-NC-ND 3.0 IGO) (http://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode) e pode ser reproduzida com atribuição ao BID e para qualquer finalidade não comercial. Nenhum trabalho derivado é permitido.

Qualquer controvérsia relativa à utilização de obras do BID que não possa ser resolvida amigavelmente será submetida à arbitragem em conformidade com as regras da UNCITRAL. O uso do nome do BID para qualquer outra finalidade que não a atribuição, bem como a utilização do logotipo do BID serão objetos de um contrato por escrito de licença separado entre o BID e o usuário e não está autorizado como parte desta licença CC-IGO.

Note-se que o link fornecido acima inclui termos e condições adicionais da licença.

As opiniões expressas nesta publicação são de responsabilidade dos autores e não refletem necessariamente a posição do Banco Interamericano de Desenvolvimento, de sua Diretoria Executiva, ou dos países que eles representam.

Tradução de Mariana Fagundez, otraspalabras.com.

* Gostaria de agradecer todos os comentários de Cristina Pombo e a pesquisa de Sebastián Dueñas durante a redação deste documento. Todos os erros são exclusivamente meus.

Índice

1. Introdução **4**

1.2 Para que serve uma auditoria? 5

1.3 O que é uma auditoria algorítmica? 5

1.4 A quem se destina este guia? 6

1.5 Como usar este guia 6

2. A auditoria algorítmica **7**

2.1 Por que realizar uma auditoria algorítmica? 7

2.2 Requisitos para a realização de uma auditoria algorítmica 9

2.3 Quando uma auditoria algorítmica deve ser realizada? 9

2.5 Que informações são compartilhadas e como são compartilhadas em um processo de auditoria 11

2.6 O que a documentação precisa conter? 12

2.7 Em que condições os dados devem ser entregues ao auditor? 13

2.8 Quem deve ter acesso aos resultados da auditoria? 13

2.9 Considerações para realizar uma auditoria algorítmica 14

2.10 Determinação de danos 14

2.11 Perfis e funções das pessoas que colaboram com a auditoria 16

3. Etapas de uma auditoria algorítmica **17**

4. Princípios orientadores para os ADSs **19**

5. Caso de uso: ADS no policiamento preditivo **22**

6. Comentários finais **24**

Referencias **25**

1. Introdução

A tomada de decisão é uma das principais habilidades do ser humano. Decidir entre mais de uma alternativa permite discernir e escolher formas melhores de fazer as coisas. A tomada de decisão é um processo por meio do qual uma pessoa pondera as informações disponíveis e incorpora sua experiência anterior para escolher a opção que, naquele momento, parece mais conveniente.

Daniel Kahneman, Prêmio Nobel de Economia em 2002, distingue no ser humano duas formas de pensar que operam na tomada de decisão: um primeiro sistema rápido, intuitivo e emocional, e um segundo sistema mais lento, deliberativo e lógico. O primeiro nem sempre é eficaz, enquanto o segundo, embora demorado, permite chegar a conclusões que incorporam um número maior de elementos de análise, um nível maior de reflexão e eficácia nas decisões.

Kahneman mostra que a tomada de decisão desses dois sistemas é complementar: a velocidade é essencial em algumas ocasiões, enquanto análises complexas e abrangentes são cruciais em outras.

Governos, empresas, instituições e uma ampla variedade de grupos tomam decisões que afetam a vida de outras pessoas (promoções no trabalho, benefícios sociais, condenações criminais etc.). É por isso que a tomada de decisão deve ser um processo cuidadoso e completo, ao qual todas as informações corretas, atualizadas e relevantes devem ser incorporadas, garantindo que tudo isso seja feito de forma eficiente.

Considerando o número de pessoas afetadas e o grau de impacto que têm na vida das pessoas afetadas[1], os processos de tomada de **decisão governamental** devem ser conduzidos com cuidado especial e incorporar critérios de participação democrática e prestação de contas.

1.1 O que é um sistema automatizado de tomada de decisão ou de suporte à decisão?

Um sistema automatizado de tomada de decisão ou suporte à decisão (ADS, na sigla em inglês de *Automated Decision Support*) é um sistema computacional que pode,

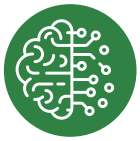
1 Por razões estritamente estilísticas, o gênero masculino é usado neste documento como genérico, independentemente do gênero das pessoas.

para um determinado conjunto de objetivos definidos pelo ser humano, fazer previsões e recomendações, ou tomar decisões que afetam ambientes reais ou virtuais. Esses sistemas são projetados para operar com vários graus de autonomia[2].

Nos últimos anos, os ADSs cresceram exponencialmente em número e esferas de aplicação. Hoje, estamos interagindo com cada vez mais ADSs, muitas vezes sem perceber. No entanto, a falta de conscientização sobre seu uso não reduz os riscos sociais, caso esses sistemas sejam mal formulados ou tenham sido criados sem os devidos cuidados.

Se os ADSs forem usados com grupos ou comunidades vulneráveis, como crianças, pessoas com deficiência e populações historicamente desfavorecidas ou em risco de exclusão[3], serão necessárias previsões ainda maiores no momento de sua implantação.

Neste guia, os ADSs são definidos em dois grupos de acordo com seu grau de autonomia:



ADSs nos quais as informações geradas pelos **modelos de aprendizado de máquina** são usadas como contribuições para a **tomada de decisão de uma pessoa**.

ADSs nos quais as decisões finais e suas ações resultantes são tomadas **sem intervenção humana direta**[4].

Este guia não pretende ser apenas uma ferramenta prática para a identificação e mitigação de riscos ou perigos que podem não ser óbvios à primeira vista. Também tem a intenção de servir como um instrumento para ajudar a aumentar a conscientização sobre as implicações e consequências da implantação de sistemas automatizados na tomada de decisão ou suporte à decisão que afeta a vida das pessoas.

1.2 Para que serve uma auditoria?

Em geral, qualquer sistema pode apresentar falhas ou riscos que não são detectados à primeira vista ou cuja relevância é negligenciada devido à frequência com que determinados processos são realizados. Quanto mais complexos os sistemas, maiores as chances de ocorrerem erros. Ao mesmo tempo, a complexidade dos sistemas muitas vezes permite uma maior adaptabilidade à realidade sobre a qual esses sistemas fazem previsões.

De acordo com o padrão ISO 19.011 sobre "Diretrizes para a auditoria de sistemas de gestão", uma auditoria deve ser um processo sistemático, independente e documentado, com o qual se busca coletar e avaliar evidências para determinar o grau em que determinados critérios previamente determinados são atendidos[5].

Uma auditoria deve incorporar os objetivos da entidade, a proteção dos interesses e necessidades dos beneficiários, colaboradores e demais possíveis interessados, bem como os requisitos de segurança e privacidade da informação[6]. Assim, existem auditorias de naturezas diversas: auditorias contábeis, jurídicas e de processos e informática, entre outras. A utilidade das auditorias reside no fato de permitirem uma avaliação objetiva dos possíveis riscos, sua quantificação e a priorização de sua mitigação.

Embora as auditorias tenham se tornado um componente fundamental no campo em expansão da governança algorítmica[7], por si só, elas não são suficientes para mitigar os impactos da implantação e execução de um sistema; em essência, elas constituem um

2 OECD, Recommendation of the Council on Artificial Intelligence. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

3 BID. Guía de aplicación Autoevaluación ética de IA para Actores del Ecosistema Emprendedor. Disponível em: https://publications.iadb.org/publications/spanish/document/Autoevaluacion-etica-de-IA-para-actores-del-ecosistema-emprendedor-Guia-de-aplicacion.pdf

4 BID: IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial. Disponível em: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

5 ISO 19011-2018, disponível em: https://www.iso.org/standard/70017.html

6 ISO 19011-2018, disponível em: https://www.iso.org/standard/70017.html

7 Ada Lovelace Institute. Examining the Black Box: Tools for assessing algorithmic systems. Disponível em: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

processo por meio do qual se determina o cumprimento de certos padrões. No entanto, as auditorias desempenham um papel indispensável na avaliação do impacto e da coleta e disponibilização de informações, tanto na esfera da própria entidade, como na dos órgãos reguladores, bem como para os potencialmente afetados e para a sociedade como um todo.

1.3 O que é uma auditoria algorítmica?

Uma auditoria algorítmica é um estudo que busca avaliar um ADS e seu processo de desenvolvimento, incluindo a formulação e os dados usados para treinar o sistema.
Da mesma forma, os impactos são avaliados em termos de precisão, justiça algorítmica[8], vieses, discriminação, privacidade e segurança, entre outros[9].

As auditorias algorítmicas podem ser realizadas como uma medição em relação a determinados padrões (auditorias de desempenho) ou como uma análise de cumprimento de padrões particulares (auditorias de cumprimento), com o objetivo de gerar recomendações sobre métricas específicas[10].

1.4 A quem se destina este guia?

Este guia é dirigido aos responsáveis pela formulação de políticas na América Latina e no Caribe, e/ou aos responsáveis pela condução de projetos de ADS encarregados de mitigar os impactos produzidos por seu uso. A ideia é que este documento sirva de guia para supervisionar esses desenvolvimentos a partir das etapas anteriores à formulação, passando pela implementação, até os possíveis ajustes e atualizações necessários para o bom uso do modelo de inteligência artificial. Trata-se de dar apoio ao leitor, orientando-o sobre a necessidade de realizar auditorias de sistemas de inteligência artificial e indicando os elementos a serem considerados durante sua realização.

1.5 Como usar este guia

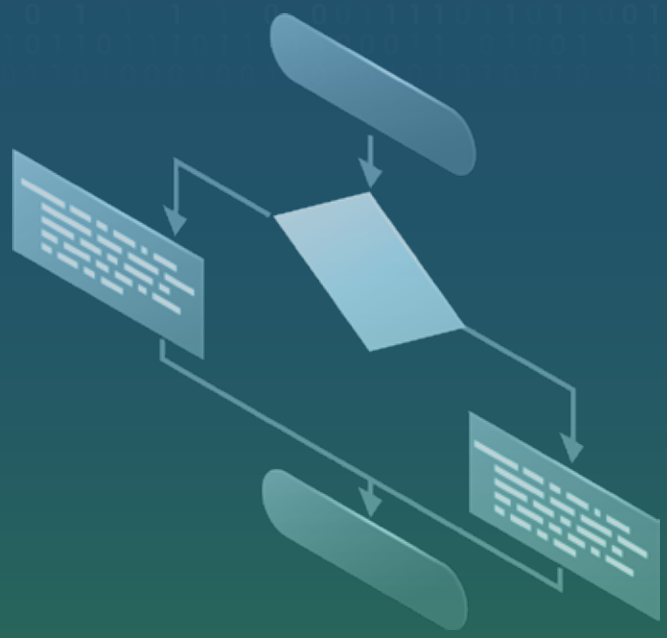
Este documento busca introduzir o leitor ao assunto por meio de perguntas estruturadas para tomar a decisão de realizar uma auditoria e o processo que ela envolve. O guia deve ser utilizado como acompanhamento durante o ciclo de vida do sistema[11]: de sua conceitualização e formulação, passando pelo uso, até a prestação de contas necessária. Também foram incluídas referências para os interessados em se aprofundar em determinados temas, o que permitirá destacar aqueles que são particularmente relevantes em função do tipo de entidade, da origem dos dados e/ou do modelo utilizado, entre outros fatores.

8 Nesse contexto, a justiça algorítmica é entendida como a característica de um algoritmo que, quando aplicado, não causa danos nem discrimina uma pessoa ou grupo.

9 Algorithmic Accountability Act of 2019

10 INTOSAI. Performance Audit Principles, disponible en https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

11 Uso responsable de IA para política pública: manual de formulación de proyectos, disponível em: https://publications.iadb.org/es/uso-responsable-de-ia-para-politica-publica-manual-de-formulacion-de-proyectos



2. A auditoria algorítmica

2.1 Por que realizar uma auditoria algorítmica?

Com a ampla adoção de ADSs nos setores público e privado, cada vez mais dimensões da vida das pessoas estão sob sua influência. Do tempo de espera dos usuários do transporte público até a correta atribuição dos serviços estatais, sempre se busca otimizar o benefício.

A implementação de sistemas automatizados traz desafios que muitas vezes não são abordados com profundidade suficiente. Entre eles estão, por exemplo, violações de direitos fundamentais devido ao uso de dados pessoais, discriminação indesejada por parte de entidades e até mesmo a tomada de decisões difíceis ou mesmo impossíveis de justificar.

Diante do exposto, é imperativo tomar medidas de controle e revisão internas e externas, principalmente no setor público. Aqui, as auditorias algorítmicas são de grande utilidade, pois são processos práticos e eficazes realizados por terceiros cuja finalidade é garantir que as decisões sejam tomadas corretamente à luz de considerações éticas e técnicas, sempre respeitando os direitos dos cidadãos.

Embora seja um assunto cuja regulamentação ainda está em discussão e desenvolvimento em vários países, as diferentes políticas nacionais relacionadas à inteligência artificial, bem como as diversas diretrizes internacionais, destacam a necessidade de contar com mecanismos de controle adequados.

Nos órgãos públicos, a realização dessas auditorias permite verificar o cumprimento das seguintes finalidades [12]:

Respeitar o direito dos cidadãos de saber com que sistemas estão interagindo em suas vidas, listando e descrevendo publicamente os ADSs que afetam significativamente as pessoas e comunidades.

Fortalecer a capacidade interna dos órgãos públicos de avaliar os sistemas que constroem ou adquirem, e ajudá-los a obterem mais experiência nessas tarefas. Tudo isso para que consigam antecipar problemas que possam surgir de situações indesejadas, incluindo a atribuição incorreta de benefícios ou violações do devido processo, para citar apenas dois.

Garantir uma maior responsabilidade no uso de ADSs, elaborando-se um método útil e contínuo, para que terceiros revisem e avaliem esses sistemas, de modo que os problemas possam ser identificados e resolvidos ou mitigados.

Garantir que os cidadãos tenham a oportunidade de responder e, se necessário, contestar a utilização de um determinado sistema ou as diretrizes utilizadas para seu desenvolvimento por um órgão público.

Devemos apontar, ainda, a existência de vários riscos inerentes ao desenvolvimento de ferramentas de inteligência artificial. Aqui encontramos, por exemplo, uma ênfase exagerada na otimização de métricas específicas de desempenho em detrimento das dimensões de transparência e equidade. Outro risco evidente decorre da falta de recursos necessários para desenvolver os modelos internamente nos órgãos que os requerem e que frequentemente optam por comprar ferramentas que, embora tenham sido projetadas por terceiros para usos diferentes, acabam sendo adaptadas para o fim específico do adquirente. Também se pode acrescentar o risco de os dados nos quais o sistema se baseia não serem igualmente representativos para todos os casos, criando-se um sistema onde a desigualdade é o correto. Isso pode gerar dificuldades na adequação do modelo, considerando-se seus requisitos operacionais e as exigências regulatórias para seu funcionamento.

Ainda assim, a realização de auditorias algorítmicas permite que as entidades cumpram as exigências que os órgãos públicos e privados

12 IA Now Institute. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. Disponível em: https://ainowinstitute.org/aiareport2018.pdf

precisam satisfazer em termos de eficiência e eficácia, seja pela regulamentação existente ou pelas exigências dos cidadãos em matéria de transparência e eficiência.

2.2 Requisitos para a realização de uma auditoria algorítmica

Os ADSs podem ser desenvolvidos internamente pelo serviço ou entidade governamental, ou por terceiros. Neste último caso, seu desenvolvimento é formalizado por meio de um contrato de aquisição de produtos ou serviços, licitação ou compra direta. No momento de tomar a decisão de adquirir um ADS, é fundamental que a instituição conte com um gerente de projeto que possa gerenciar o contrato e que tenha um certo nível de conhecimento técnico, para que não perca o controle do processo de desenvolvimento e implantação.

Ao adquirir o serviço de terceiros, será necessário incorporar ao contrato de licitação ou compra direta cláusulas que permitam a auditoria do sistema. É aconselhável não limitar o número de auditorias nem suas condições, a fim de não impedir sua realização quando necessário. Ao mesmo tempo, a licitação deve deixar claro que a auditoria pode ser feita diretamente ou por terceiros em nome do órgão público.

Algumas empresas podem mostrar-se mais relutantes em serem auditadas, principalmente se houver a possibilidade de as informações se tornarem públicas. Portanto, deve-se exigir acesso ao código-fonte dos ADSs, para que possam ser auditados *ex post*. É necessário apontar que as regras de cumprimento regulamentar dos fornecedores nunca poderão ser utilizadas como obstáculo à realização de auditorias.

Além disso, nas especificações da licitação, deve-se acrescentar como requisito para a prestação do serviço que os ADSs sejam acompanhados de toda a documentação técnica relativa a seu desenvolvimento. A documentação abrange de manuais do usuário e políticas a descrições técnicas do processo de treinamento, formulação e implementação. Tudo isso é essencial para poder manter um registro que permita que os auditores revisem os ADSs.

2.3 Quando uma auditoria algorítmica deve ser realizada?

Diferentemente de outros tipos de avaliação, as auditorias algorítmicas são realizadas após a implantação do sistema, quando ele já estiver em operação. Dessa forma, a formulação e o desenvolvimento de um ADS podem ser contrastados com os efeitos de sua implementação, principalmente porque já houve casos em que algum risco ou dano ocorreu.

Determinar o momento exato para realizar uma auditoria algorítmica não é um exercício trivial. Isso porque, em geral, os efeitos de um ADS na população tornam-se evidentes para o líder do projeto somente após os danos terem sido causados. Caso não tenham ocorrido, recomenda-se fortemente realizar as auditorias no fim do período-piloto do projeto, que ocorre quando é feita uma implementação controlada em uma amostra do universo total. Por exemplo, no caso da implementação de um ADS que ajude a classificar o risco socioeconômico das famílias para conceder-lhes benefícios sociais, o projeto-piloto deve começar em uma pequena cidade, não em uma região ou em todo o país.

Ao concluir o projeto-piloto, ou seja, uma vez realizados os testes de implantação no período determinado e em uma amostra definida, recomenda-se avaliá-lo. Casos anteriores à implementação do piloto podem ser considerados, como, por exemplo, em uma simulação que permita detectar possíveis erros com base em diferentes resultados. Se a avaliação preliminar interna da equipe que desenvolveu o ADS revelar a existência de complicações, recomenda-se realizar a auditoria.

A realização periódica dessas auditorias também é altamente recomendada, principalmente em contextos sociais em constante fluxo e/ou em sistemas cujo funcionamento pode evoluir em função de uma

quantidade e uma variedade maiores de dados utilizados, entre outros. A periodicidade dessas auditorias também deve ser determinada com base no risco de erro apresentado pelo sistema.

No caso de perceber erros ou efeitos negativos decorrentes de seu uso, a realização de uma auditoria periódica torna-se imprescindível.

Criticidade dos sistemas

A criticidade refere-se à importância e ao risco dos ADSs em sua formulação e implementação. Por importância entende-se a função que um ADS cumpre, quer dentro de uma cadeia de processos durante a qual outros sistemas são alimentados com as informações geradas, quer pelo papel que desempenha em uma tarefa específica. Exemplos de sistemas críticos incluem aqueles que concedem diretamente direitos, auxílios ou subsídios, ou aqueles que operam em áreas de natureza sigilosa, como defesa nacional, saúde ou sistema carcerário.

Entende-se por risco a possibilidade de, durante sua utilização, um ADS cometer erros que causem danos à população envolvida. Seria o caso dos ADSs que tomam decisões sobre a liberdade de um réu, sobre a alocação de recursos para fins sociais ou sobre a resposta a situações de conflito. Ao analisar o dano, pelo menos três elementos devem ser levados em consideração: (i) a **probabilidade de ocorrência do dano**, que geralmente é medida com base em uma determinação da precisão do sistema ao cumprir sua tarefa ou, no extremo oposto, estabelecendo-se quantas vezes ele se equivoca; (ii) a **profundidade do impact**o, ou seja, as consequências leves ou graves que o erro pode ter (os erros mais graves seriam aqueles que causam danos à vida, liberdade ou propriedade de um indivíduo ou grupo social, ou aqueles que os privem de um serviço ou ajuda pública essencial para sua sobrevivência); e (iii) a **distribuição do erro**, ou seja, quando o erro cometido pelo ADS afeta um subgrupo da população mais que outros, como tem acontecido com segmentos de baixa renda, imigrantes e minorias raciais. Um exemplo que ocorre constantemente são os sistemas de reconhecimento facial, que funcionam melhor com pessoas de pele mais clara e tendem a cometer mais erros com pessoas de pele mais escura. A discriminação contra esses grupos causa injustiças nas concessões ou classificações, e produz um desconforto considerável nos indivíduos do subgrupo discriminado.

Analisando-se as informações de ambos os grupos de elementos, ou seja, importância e risco, é possível determinar a criticidade de um sistema. Quanto mais crítico for, maiores serão os cuidados que devem ser tomados. Alguns sistemas podem gerar um risco tão grande, que sua implementação deixe de ser justificada, como acontece com aqueles cujo nível de precisão é muito baixo e que afetam diretamente o bem-estar de um grupo da população. Nesses casos, ao deliberar sobre a conveniência de sua implantação, deve-se considerar a opção de não o implementar.

Nos casos em que a implementação é essencial e o risco é suportável, medidas de mitigação e controle devem ser tomadas para reduzir suficientemente os erros e seus efeitos. As opções aqui são diversas: de revisão constante dos resultados do sistema até a intervenção humana nas decisões sobre um determinado subgrupo social ou sobre todos os envolvidos, passando pela transparência algorítmica, entre outras.

Vale apontar que também haverá casos em que as auditorias não cumprem seu propósito, pois são realizadas em momentos não ótimos, ou seja[13]:

» **Auditorias prematuras:** Nestes casos, a auditoria é realizada em um estágio muito inicial, antes que aspectos importantes do sistema sejam totalmente implementados ou antes que seja possível avaliar corretamente possíveis danos.

» **Auditorias tardias:** Por serem as auditorias um exercício ex post, este tipo é útil para remediar situações futuras, mas não para lidar com danos passados. Por esse motivo, muitas vezes são detectadas situações em que os danos são causados por um período considerável sem que sejam avaliados e/ou mitigados quando apropriado.

» **Auditorias esporádicas:** As auditorias são um mecanismo em constante desenvolvimento que evolui e amadurece de acordo com os avanços da tecnologia e da sociedade. Por isso, pode ocorrer que os impactos em potencial previstos em uma auditoria realizada nos primeiros meses após a implantação do sistema não sejam os mesmos que aparecem em etapas posteriores. Insiste-se, portanto, na recomendação da realização desses procedimentos periodicamente, de modo a manter o mecanismo de auditoria e as informações fornecidas pela entidade envolvida atualizados.

2.4 Quem deve realizar uma auditoria algorítmica?

» Para avaliar a eficácia e possíveis consequências do sistema, essas auditorias podem ser realizadas tanto externa quanto internamente. Dependendo de quem as executa, podem ser classificadas em três tipos[14]:

» **Auditorias realizadas por terceiros:** São aquelas realizadas por agentes externos à entidade auditada, que avaliam o comportamento de um sistema com base apenas em seus resultados.

» **Auditorias realizadas por segundos:** São aquelas realizadas por um fornecedor, cliente ou contratado da instituição auditada que tem acesso ao servidor (back-end) do sistema e avalia seu comportamento, considerando tanto a parte técnica quanto os resultados.

» **Auditorias internas:** São aquelas realizadas por um membro ou equipe da entidade envolvida, com o objetivo de avaliar os receios próprios dessa entidade. Normalmente, esses receios decorrem dos desafios comuns associados ao desenvolvimento responsável de sistemas de inteligência artificial, como transparência e equidade. Esse tipo de auditoria busca atingir metas relacionados ao próprio sistema, considerando seus próprios critérios de sucesso.

Vale apontar que, nas três categorias anteriores, e independentemente de o auditor ser interno ou externo, um requisito indispensável e comum é que o responsável pela auditoria não tenha estado envolvido no desenvolvimento do sistema.

É possível que seja realizada mais de uma auditoria, em mais de uma das modalidades indicadas. Nesses casos, será importante não duplicar esforços desnecessariamente e que as auditorias complementem umas às outras. Entre as competências que a equipe de auditoria deve possuir, estão: (i) conhecimento técnico; (ii) conhecimento da área específica na qual o ADS é implementado; e (iii) conhecimento sólido dos princípios éticos que devem ser incorporados aos procedimentos nos sistemas autônomos.

13 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, disponível em: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

14 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, disponível em: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

O fato de a equipe de auditoria possuir **conhecimento técnico** significa que ela domina as linguagens de programação e as metodologias específicas utilizadas nos ADSs, sendo capaz de validar a seleção e tratamento dos dados. Por exemplo, se alguém que se autodenomina auditor não for especialista na tecnologia de IA utilizada no ADS da entidade, terá pouco a dizer sobre a conformidade desse sistema com os critérios da auditoria.

O **conhecimento da área específica na qual o ADS** é implementado é essencial para que a equipe de auditoria possa avaliar e fazer recomendações sobre melhorias no sistema. Por exemplo, se for um sistema que avalia a periculosidade dos indivíduos envolvidos em processos criminais (veja o caso do COMPAS na seção 5) e, para isso, leva em consideração a cor da pele do réu e/ou o bairro onde mora (o que, por si só, vai de encontro aos direitos humanos), os auditores devem ser capazes de identificar o viés se quiserem cumprir o propósito de melhorar o sistema.

A equipe de auditoria deve ter um **conhecimento sólido dos princípios éticos a serem incorporados nos sistemas autônomos** para poder avaliá-los durante sua revisão. Hoje, existem diferentes quadros éticos que permitem isso, como veremos a seguir. Da mesma forma, é necessário adotar princípios relativos à proteção de dados pessoais e ferramentas que contribuam para evitar ou reduzir os casos de discriminação. A equipe de auditoria deve ser capaz de determinar se as regras e princípios éticos que devem reger o funcionamento dos ADSs para proteger a dignidade das pessoas estão sendo cumpridos ou não, sem se focar apenas na eficiência ou nas falhas do sistema. Do ponto de vista da estrutura da auditoria, é necessário estabelecer canais de comunicação entre a equipe responsável pela auditoria e a equipe da instituição auditada, dando-se atenção especial aos cargos e funções que tenham a ver com o desenvolvimento do sistema e que são descritos no Item 2.10.

2.5 Que informações são compartilhadas e como são compartilhadas em um processo de auditoria

Para auditar um sistema, deve haver primeiro um registro da documentação detalhada sobre os processos de treinamento, a realização e validação de testes, e sua implementação. Quanto mais detalhada for a documentação, mais fácil será o trabalho dos auditores. No entanto, o custo e os esforços investidos em documentar o ciclo de vida do sistema autônomo dependerão do nível de criticidade do processo. A documentação detalhada permite que os auditores revisem o histórico de desenvolvimento do algoritmo, incluindo os propósitos para os quais foi criado, a equipe que trabalhou nele, os testes realizados e as modificações pelas quais passou. Dessa forma, o sistema pode ser comparado em diferentes estágios de sua vida, o que é extremamente útil para determinar a instância exata em que uma anormalidade pode ocorrer.

2.6 O que a documentação precisa conter?

A realização de uma auditoria algorítmica envolve vários processos de avaliação. Estes vão do **modelo de governança da entidade** cujo sistema será auditado (organograma e funções da equipe envolvida em seu desenvolvimento, planos estratégicos para seu uso e implementação, partes interessadas e afetadas por isso, entre outros elementos), passando pelas **bases de dados** utilizadas (método de coleta de dados e sua qualidade, relevância, gerenciamento e manipulação, entre outros elementos), até chegar ao **modelo computacional** (algoritmos utilizados, sensibilidade e especificidade do sistema, impacto e distribuição de erros, entre outros elementos). Portanto, a documentação entregue pela empresa auditada deve permitir o entendimento do modelo de governança do sistema, bem como a construção de um perfil adequado dos dados[15] e do modelo algorítmico propriamente dito[16].

15 BID. IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial, p. 55. Disponível em: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

16 BID. IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial, p. 57. Disponível em: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

Para construir o perfil dos dados utilizados no modelo, será necessário contar com informações sobre sua origem, coleta, governança e estrutura, bem como uma avaliação de sua qualidade. Para elaborar o perfil do modelo, será necessário dispor de informações sobre sua conceitualização e formulação; sobre a origem e o tratamento dos dados; seu desenvolvimento, uso e monitoramento; e sobre a prestação de contas correspondente. O documento "Uso responsável de IA para políticas públicas: manual de ciência de dados"[17] oferece uma aproximação detalhada dos itens contidos em cada um deles.

2.7 Em que condições os dados devem ser entregues ao auditor?

Dependendo do nível de criticidade do sistema, será importante definir as condições de entrega da documentação. Para esse fim, os usos permitidos e proibidos dos algoritmos e dos dados devem ser definidos explicitamente. Aqui, será muito útil contar com um Acordo de Compartilhamento de Dados (*Data Sharing Agreement*). Esse documento permite estabelecer a responsabilidade e as funções de cada uma das partes; esclarecer a finalidade da transferência de dados; detalhar o que acontece com eles em cada etapa; e estabelecer padrões de uso, segurança e privacidade[18]. Dessa forma, tanto o auditor quanto o auditado terão um documento que esclarece as responsabilidades de cada parte em relação aos dados. O exposto acima é relevante principalmente, por exemplo, nos casos em que a confidencialidade é crucial, como os que envolvem o tratamento de dados pessoais, informações de segurança nacional ou de ordem pública, ou informações comercialmente sigilosas.

Os dados do treinamento podem ser entregues aos auditores, para que reproduzam o processo e avaliem se há uma forma melhor de trabalhar com essas informações. Para realizar esse procedimento, será imprescindível cumprir as normas de proteção de dados pessoais, se aplicável. Por exemplo, seria conveniente anonimizar o banco de dados para proteger totalmente essas informações. Da mesma forma, o acordo de compartilhamento deve especificar que esses dados são apenas para fins de auditoria e que não serão usados para outros fins.

2.8 Quem deve ter acesso aos resultados da auditoria?

A regra geral da administração pública é o princípio de transparência, segundo o qual os atos, resoluções, procedimentos e documentos da administração do Estado devem ser públicos. Esse princípio permite a prestação de contas do Estado perante a sociedade civil, representada diretamente por organizações sociais ou comunitárias, universidades e centros de estudos.

No caso dos resultados de uma auditoria algorítmica, é importante determinar quem são os terceiros que terão acesso às informações e à avaliação elaborada pelo(s) auditor(es). O relatório de auditoria conterá uma análise da eficácia dos algoritmos, mas também pode mostrar como eles funcionam, os tipos de dados usados e possíveis vulnerabilidades. Assim, é fundamental analisar e determinar o grau de divulgação dos relatórios.

Para isso, primeiro, deve-se identificar quais informações do processo automatizado são sigilosas e quais podem ser livremente conhecidas por terceiros. Em seguida, é necessário identificar quais dos elementos a serem analisados pelos auditores podem constituir um risco para a continuidade operacional do sistema e para a proteção dos beneficiários. Por último, deve-se lembrar de que o feedback da sociedade civil permite melhorar procedimentos em direções não necessariamente previstas e rever democraticamente os processos que afetam a vida das pessoas. Recomenda-se que o feedback dos cidadãos e da sociedade civil seja obtido através de meios diretos, onde exista pelo menos um e-mail por meio do qual possam ser recebidas reclamações ou sugestões. A transparência não é satisfeita apenas com a publicação de informações, mas também com mecanismos participativos, nos quais as pessoas podem expor suas preocupações diretamente à autoridade.

17 Ibid.

18 UK Information Commissioner's Office. Data sharing code of practice. Disponível em: https://ico.org.uk/media/for-organisations/guide-to-data-protection/ico-codes-of-practice/data-sharing-a-code-of-practice-1-0.pdf

O sigilo das informações contidas será determinado com base na possibilidade de seu conhecimento causar danos aos beneficiários, prejudicar a continuidade operacional do sistema ou afetar a eficácia do serviço. Pelo contrário, informações que contenham vieses em relação a determinados grupos ou evidências de discriminação não devem ser consideradas sigilosas nem confidenciais. Elas deverão ser publicadas, para que os beneficiários possam se defender e proteger seus interesses em caso de erros registrados.

Dependendo da necessidade de manter a confidencialidade, as auditorias podem ser classificadas de acordo com seus graus de transparência e divulgação. Um caso de **máxima confidencialidade** (menor divulgação) seria aquele em que os resultados da auditoria são conhecidos apenas pelo órgão que está implementando o sistema. Um caso de **confidencialidade média** seria aquele em que as mesmas informações podem ser compartilhadas com organizações hierarquicamente superiores ou avaliadores. Um caso de **menor confidencialidade** ocorre quando as informações utilizadas podem ser compartilhadas com órgãos públicos similares que as possam aproveitar para melhorar seus próprios processos.

O nível **mínimo de confidencialidade** é aquele em que as informações são compartilhadas com terceiros fora da administração pública, como instituições internacionais, universidades ou centros de estudos. Aqui, a integridade dos dados pode ser assegurada pela confirmação de que as informações compartilhadas são regidas por acordos de confidencialidade (também conhecidos como NDAs, sigla em inglês de *Non-Disclosure Agreements*). Isso garantiria que as informações não fossem divulgadas ou usadas para fins diferentes da auditoria.

Por fim, as informações serão **amplamente divulgadas** quando seu conhecimento não constituir perigo para a continuidade operacional do sistema, para as pessoas e/ou para a eficiência do serviço. Também é possível que a auditoria seja publicada quando os riscos identificados já tiverem sido neutralizados e as vulnerabilidades, corrigidas.

2.9 Considerações para realizar uma auditoria algorítmica

As suposições subjacentes à realização de uma auditoria algorítmica incluem as seguintes:

(i) Quem audita (seja uma entidade interna ou externa à instituição) é independente e alheio ao desenvolvimento e implementação do sistema.

(ii) Quem desenvolve e implementa o sistema deve ser capaz de fornecer informações adequadas sobre ele a quem estiver realizando a auditoria.

(iii) O auditor deve ser capaz de entender corretamente o sistema com base nas informações fornecidas, na documentação relevante e nos efeitos que podem ser percebidos sobre os impactos do sistema.

(iv) O comportamento do sistema durante seu uso e monitoramento é consistente com seu comportamento no momento da auditoria[19]. É extremamente importante levar isso em consideração, pois seu funcionamento pode variar de acordo com o contexto ou os dados com os quais é alimentado. Este último é o que justifica a necessidade de realizar essas auditorias periodicamente, para levar em conta possíveis mudanças de cenário, inclusão de novas funções ou supressão de outras.

(v) Sempre que possível, a auditoria deve ser conduzida em termos binários, o que significa que as avaliações devem ser emitidas em um formato que não permita nuances (por exemplo, em conformidade ou não). O motivo para isso é que uma graduação da avaliação pode levar a áreas cinzentas que prejudicam a clareza e a confiança exigidas na auditoria.

19 Ada Lovelace Institute. Algorithmic Accountability for the Public Sector. Disponível em: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

2.10 Determinação de danos

Durante a auditoria, os danos que um sistema causou devido a uma operação defeituosa, imperfeita ou abaixo do ideal serão revelados e poderão ser medidos.

Os danos são perdas sofridas tanto pelos beneficiários quanto por terceiros fora do sistema. Será sempre necessário definir corretamente cada grupo afetado, descrevendo-se claramente as características de cada um. Isso ajudará a reconhecer padrões presentes ou qualidades que podem estar sujeitas a mais ou menos escrutínio em relação ao nível considerado ideal no contexto em que o sistema é usado.

Exemplo de grupos afetados

Deve-se ter em mente que os grupos afetados pelos ADSs não são apenas aqueles que os utilizam ou que estão diretamente envolvidos nas ações ou recomendações provenientes desses sistemas. Muitas vezes, também existem terceiros não considerados, alheios a eles, que são afetados por seu uso.

Tomemos como exemplo um sistema que determina a frequência do transporte público e cujas decisões afetam de forma clara e direta seus usuários ao indicar se deve haver mais frequência em um período do dia, a fim de otimizar o uso dos recursos públicos e a satisfação dos usuários. No entanto, existem outros grupos que também são afetados, como usuários de veículos particulares, pedestres e ciclistas, pois seus tempos de deslocamento também seriam afetados por essa frequência.

Conforme indicado anteriormente no quadro sobre criticidade dos sistemas, os danos podem ter um impacto sério em dois tipos de fator-chave: (i) aqueles que influenciam diretamente a concessão ou restrição de direitos, auxílios ou subsídios; e (ii) aqueles que fazem parte de uma cadeia de processos e que, em caso de falha ou erro, podem afetar algum dos elementos que afetam a prestação de um serviço governamental. É nesta segunda instância que surgem as principais ameaças de ataques cibernéticos que têm afetado alguns dos serviços do Estado.

A União Europeia catalogou quatro níveis de risco para modelos de IA:

(i) Risco inaceitável: Aplicações prejudiciais à saúde e à integridade das pessoas, e que violam os direitos fundamentais. São proibidas.

(ii) **Alto risco:** Aplicações que tenham um impacto negativo na segurança das pessoas ou que sejam componentes de segurança de sistemas maiores. Devem ser avaliadas por terceiros e estar em conformidade com as regulamentações setoriais antes de entrar em operação.

(iii) **Risco limitado:** Aplicações que apresentam um baixo nível de risco, mas que devem cumprir os requisitos de transparência e informação para os cidadãos sujeitos a um tratamento automatizado.

(iv) **Risco mínimo:** Qualquer sistema cuja aplicação não envolva nenhum risco. Os desenvolvedores podem aderir aos códigos de conduta de forma voluntária.

Infelizmente, quando se trata de sistemas de ADS, nem sempre é possível obter uma explicação sobre o motivo e a causalidade do erro e seus consequentes danos. Isso é conhecido como um problema de “caixa preta”. Por isso, as auditorias técnicas limitam-se a avaliar se os cuidados necessários foram ou não tomados no desenvolvimento do sistema.

Caixa preta

A metáfora da "caixa preta" é usada para referir-se aos sistemas cujo funcionamento interno é desconhecido, seja porque é impossível compreendê-lo ou porque isso seria muito caro e, portanto, pouco razoável (por exemplo, tentar interpretar uma rede neural). Nesses casos, não é possível para um humano discernir a forma em que certas entradas (por exemplo, dados) levam o sistema a fornecer um resultado (por exemplo, uma determinada ação ou indicação)[a].

Embora a utilização desses modelos possa ser justificada pelo seu melhor desempenho, certamente vai de encontro à busca por transparência na implementação do ADS.

[a] Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponível em: https://www.auditingalgorithms.net/

2.11 Perfis e funções das pessoas que colaboram com a auditoria

Caso sejam necessários esclarecimentos ou mais informações sobre o histórico do sistema sujeitado ao processo de auditoria, deverão ser cadastrados os contatos relevantes, com suas respectivas funções no órgão que o implementa, pois serão eles os responsáveis pelo fornecimento das informações solicitadas. Os cargos e funções que podem existir em uma entidade média para esses sistemas são descritos a seguir[20]:

- **Diretor de Informações** (*CIO*)**:** Responsável pelos sistemas informáticos e tecnológicos de uma entidade. É quem decide e direciona desenvolvimentos tecnológicos para atingir os objetivos da instituição.
- **Diretor de privacidade** (*CPO*)**:** Pessoa encarregada de tomar decisões sobre a privacidade da instituição e garantir a proteção dos interesses dos beneficiários nessa área.
- **Diretor de Segurança da Informação** (*CISO*)**:** Pessoa encarregada de zelar pela segurança das informações que a entidade produz e possui.
- **Diretor Jurídico:** Pessoa responsável pelos assuntos jurídicos e pela garantia do cumprimento regulatório pela instituição.
- **Desenvolvedor de software:** Pessoa encarregada de programar o sistema e converter os requisitos institucionais em um software que atenda aos propósitos técnicos desejados.
- **Analista de dados:** Pessoa encarregada de analisar, ordenar e depurar os dados, para que contribuam para a tomada de decisão dentro da instituição.
- **Engenheiro de dados:** Pessoa encarregada de construir, manter e preparar os bancos de dados, para que possam ser utilizados posteriormente pelo analista de dados.
- **Gerente de projeto:** Responsável pelo projeto que está sendo realizado, mantendo sua coesão e distribuindo as tarefas entre os membros da equipe.
- **Proprietário do produto** (*Product Owner*)**:** Responsável pelas tarefas práticas do desenvolvimento do sistema relacionadas à estratégia, execução e lançamento.
- **Especialista na área:** Pessoa que tem conhecimento sobre o campo de atividade que corresponde à instituição e que pode contextualizar as necessidades dos usuários.

20 Haverá entidades de menor porte nas quais existe apenas um responsável, bem como entidades de maior porte que dispõem de equipes inteiras dedicadas ao cumprimento das tarefas de determinada função.

3. Etapas de uma auditoria algorítmica

Embora atualmente não exista um modelo unificado para conduzir auditorias algorítmicas, este guia utiliza o de Raji e Smart, et al. (2020), que compreende seis etapas: (i) definição do escopo da auditoria; (ii) mapeamento das partes interessadas; (iii) coleta da documentação; (iv) testes; (v) análise dos resultados; e (vi) pós-auditoria[21].

As tarefas a serem realizadas em cada uma das seis etapas mencionadas acima estão listadas abaixo:

» Definição do escopo da auditoria

- Compilação do documento de requisitos do produto (PRD, na sigla em inglês)
- Revisão dos princípios considerados na formulação do sistema
- Análise de casos de uso semelhantes
- Avaliação do impacto social com base na finalidade do ADS

» Mapeamento de partes interessadas

- Formulação de perguntas e entrevistas com a equipe
- Transcrição e sistematização das respostas

» Coleta da documentação

- Elaboração de listas de verificação com os pontos a serem auditados
- Criação do perfil de dados
- Elaboração do perfil do modelo

» Realização de testes

- Revisão da documentação
- Simulação de falhas e busca de vulnerabilidades
- Elaboração da matriz de risco correspondente ao uso do sistema

21 Raji et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Disponível em: https://arxiv.org/pdf/2001.00973.pdf

» Análise dos resultados

- Atualização e formalização da matriz de riscos
- Elaboração do plano de ação ou mitigação de riscos
- Compilação dos detalhes e evolução do desenvolvimento do sistema
- Relatório de auditoria

Observe que nem sempre essas etapas serão sequenciais, sendo possível que se descubra, logo no início da auditoria, que o sistema é inviável, tornando desnecessário prosseguir para etapas posteriores.
O anexo deste documento inclui perguntas-guia para realizar corretamente uma auditoria.
Até o momento, as várias regulamentações existentes na América Latina e no Caribe sobre inteligência artificial não fazem referência direta à responsabilidade algorítmica, ao contrário de países com jurisdições mais maduras no desenvolvimento dessas questões, como Canadá[22], Suécia[23] e Reino Unido[24].

Devido ao crescimento do número de iniciativas legais que influenciam o desenvolvimento de sistemas autônomos, bem como das políticas sobre o uso de IA nos países da região, a resposta será determinada pelo cumprimento tanto da legislação quanto das políticas que buscam garantir o uso adequado do sistema, bem como as próprias normas que regulamentam o setor em que é implementado. É necessário levar em conta que existem áreas relacionadas que terão um impacto na revisão dos padrões de conformidade, como, por exemplo, em matéria de proteção de dados, cibersegurança, leis antidiscriminação ou mesmo regulações setoriais.

No que diz respeito ao setor público, atualmente, não existe nenhuma prática padronizada para a realização de auditorias algorítmicas. No entanto, têm surgido algumas iniciativas que visam a consolidar experiências em vários casos, áreas e jurisdições. A título de exemplo, temos o documento intitulado "A White Paper for Public Auditors"[25], elaborado pelas autoridades de auditoria da Finlândia, Alemanha, Países Baixos, Noruega e Reino Unido, com base em sua experiência na matéria.

Após a auditoria e em virtude dos resultados obtidos, será necessário que a entidade determine se é possível continuar utilizando o sistema ou se deverá modificá-lo parcial ou totalmente, de acordo com as respostas obtidas (classificadas no anexo, de acordo com sua relevância, como de grande urgência, grande importância ou revisão recomendada). Se necessário, o plano de ação ou mitigação de riscos deve ser implementado, somado posteriormente a um acompanhamento contínuo de sua implementação.

22 Por exemplo, no Canadá, a diretriz de tomada de decisão automatizada de 2019 visa a reduzir o risco que esses sistemas apresentam e obter decisões administrativas mais eficientes, precisas, consistentes e interpretáveis, de acordo com a legislação canadense. Para tanto, amplia-se a realização de auditorias, facilita-se o acesso às informações e eleva-se o padrão de qualidade dos dados. Directive on Automated Decision-Making, disponível em: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

23 Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up, disponível em: https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

24 Guidance on the AI auditing framework, Draft guidance for consultation, disponível em: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

25 Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponível em: https://www.auditingalgorithms.net/

4. Princípios orientadores para os ADSs

Com o objetivo de promover a implementação e uso ético de sistemas baseados em inteligência artificial, diversas jurisdições e órgãos adotaram princípios com os quais buscam orientar sua implementação, tanto em assuntos específicos quanto como um todo. Existem, por exemplo, os princípios estipulados no Artigo 5.º do Regulamento Geral de Proteção de Dados europeu (RGPD)[26], que regulam especificamente o tratamento de dados pessoais. Para o sistema como um todo, existem os princípios estabelecidos no documento "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"[27] do Berkman Klein Center for Internet and Society da Universidade de Harvard.

Este guia levará em consideração a lista de princípios éticos propostos pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE), em seu documento "Recommendation of the Council on Artificial Intelligence"[28]. Trata-se do primeiro conjunto de padrões de políticas intergovernamentais sobre inteligência artificial integrado pelos princípios traduzidos no documento "Adoção Ética e Responsável de Inteligência Artificial na América Latina e no Caribe"[29], que estão resumidos abaixo.

Crescimento inclusivo, desenvolvimento sustentável e bem-estar. As partes interessadas precisarão participar ativamente do gerenciamento responsável de uma IA projetada para alcançar resultados que beneficiem as pessoas e o planeta. Com o uso adequado da IA, será possível promover o aumento das capacidades humanas e da criatividade, a inclusão de populações minoritárias, a redução das desigualdades econômicas e sociais, bem como a proteção dos ambientes naturais, estimulando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar.

26 https://gdpr-info.eu/art-5-gdpr/

27 Fjeld e Nagy. Principled Artificial Intelligence. Disponível em: https://cyber.harvard.edu/publication/2020/principled-ai

28 OECD. Recommendation of the Council on Artificial Intelligence. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

29 BID, fAIr LAC. Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe. Disponível em: https://publications.iadb.org/publications/spanish/document/fAIr_LAC_Adopción_ética_y_responsable_de_la_inteligencia_artificial_en_América_Latina_y_el_Caribe_es.pdf

Valores centrados no ser humano e na equidade. Os atores do ecossistema de IA devem respeitar o Estado de Direito, os direitos humanos e os valores democráticos durante todo o seu ciclo de vida. Dentre estes, destacam-se a liberdade, a dignidade e a autonomia; privacidade e proteção de dados; a não discriminação e igualdade; e a diversidade, equidade, justiça social e direitos trabalhistas reconhecidos internacionalmente. Para esse fim, os atores da IA devem implementar mecanismos e salvaguardas para proteger direitos como o de autodeterminação dos indivíduos. Estes devem se adequar ao contexto e ser consistentes com o estado da arte.

Transparência e explicabilidade. Os ADSs devem permitir que os atores do ecossistema entendam seu funcionamento e possíveis resultados. Assim, os sistemas implementados devem ser regidos pelos princípios da transparência e da divulgação responsável e justa das informações.

Informações relevantes devem ser fornecidas tanto a quem utiliza os sistemas quanto a quem é sujeito passivo da análise. As informações devem ser ajustadas ao contexto do receptor das informações, de forma que ele seja capaz de as compreender completa e corretamente.

Os objetivos são: (i) promover uma compreensão geral do funcionamento dos sistemas de IA; (ii) garantir que as partes interessadas estejam plenamente cientes de suas interações com esses sistemas; (iii) garantir que os beneficiários e sujeitos passivos entendam os possíveis resultados e riscos do uso dos ADSs; e (iv) permitir que as pessoas afetadas adversamente por um sistema de IA contestem seus resultados com base em informações claras e fáceis de entender sobre os fatores e a lógica que serviu de base para a previsão, recomendação ou decisão que se pretende refutar.

É fundamental que os tomadores de decisão também entendam o funcionamento e os riscos em potencial associados ao uso dos ADSs, a fim de incorporar sua própria análise nos pontos em que a máquina pode falhar ou apresentar riscos. Conforme descrito mais detalhadamente na Seção 5 deste documento sobre o caso de uso do COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*)[30], o objetivo era identificar o risco de as pessoas processadas voltarem a cometer crimes.

O uso do COMPAS gerou um grande rebuliço, pois mostrou um viés favorável para as pessoas de pele branca e desfavorável para aqueles de pele mais escura. Algo semelhante aconteceu no caso de crimes cometidos por homens e mulheres, sendo estas as mais punidas.

Como nenhum dos sujeitos passivos do sistema COMPAS sabia como ele funcionava por falta de transparência, os juízes que o utilizaram não questionaram suas recomendações. Isso só aconteceu mais tarde, com base em uma reportagem publicada na imprensa, em decorrência da qual o sistema deixou de ser utilizado.

Se o sistema fosse transparente, seria evidente que levava em conta elementos que não eram típicos de uma sanção judicial, como origem étnica, composição familiar e/ou escolaridade dos réus. Da mesma forma, estes teriam a possibilidade de se defender das penalidades que lhes fossem impostas com base no ADS, por serem obviamente contrárias ao devido processo legal.

A transparência dos sistemas, além de permitir que os sujeitos passivos ou destinatários de suas ações exerçam seus direitos, também ajuda os tomadores de decisão a ponderarem e analisarem a validade da recomendação, de modo que a compreendam plenamente e determinem se ela constitui um elemento que respeita a dignidade das pessoas, os direitos humanos e o Estado de Direito.

Robustez, segurança e proteção. Estes são três elementos essenciais em qualquer sistema de IA, pelos seguintes motivos:

» Os sistemas de IA devem ser robustos, seguros e protegidos durante todo o seu ciclo de vida, para que, sob uso normal, uso

30 Brennan, T. e Dieterich, W. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Disponível em: https://www.research-gate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.

previsível, uso indevido ou outras condições adversas, funcionem adequadamente e não representem um risco excessivo à segurança.

» Para isso, os atores da IA devem garantir a rastreabilidade permanente dos conjuntos de dados, processos e decisões tomadas durante o ciclo de vida do sistema de IA. Dessa forma, será possível analisar corretamente e de acordo com o estado da arte seus resultados e respostas às questões formuladas.

» Dependendo de suas tarefas, do contexto e de sua capacidade de atuação, os atores da IA devem aplicar continuamente uma abordagem sistemática à gestão de riscos em cada fase do ciclo de vida do sistema. O objetivo disso é abordá-los da melhor forma, incluindo os relacionados à privacidade, segurança digital e possíveis vieses.

Prestação de contas. Os atores da IA devem ser responsáveis pelo bom funcionamento desses sistemas e pelo respeito aos princípios acima mencionados, dependendo de suas funções, do contexto e do estado da arte da tecnologia.

5. Caso de uso: ADS no policiamento preditivo

Para destacar a relevância da realização de uma auditoria algorítmica, seu uso no contexto do **policiamento preditivo** será analisado a seguir.

Nos Estados Unidos, o uso de ADSs foi implementado na análise de risco para a comunidade formada por réus criminais em diferentes estados. O sistema utilizado, desenvolvido pela empresa Northpointe, foi chamado de COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Esse sistema fornece uma pontuação ao tribunal em questão com base nas respostas a um questionário composto por 137 perguntas com o seguinte teor: Seu pai/mãe já foi preso alguma vez? Com que frequência você participou de brigas na escola? Essas perguntas eram respondidas pelos réus ou obtidas de seus antecedentes criminais.

Em 2013, Eric Loomis foi preso por dirigir um veículo com pessoas que haviam se envolvido recentemente em um tiroteio. Em decorrência da recomendação do sistema COMPAS, que o classificou como sujeito de alta periculosidade para a comunidade, ele foi condenado a seis anos de reclusão e cinco anos de regime ampliado[31].

Nesse caso, surgem alertas em diferentes matérias: quanto à idoneidade da utilização do sistema no referido contexto, sua precisão e os vieses que os dados utilizados podem conter, entre outros. Assim, antes de proceder à utilização de um ADS em contextos tão complexos como o da administração da justiça, é sempre necessário responder às questões formuladas a seguir, que serão exemplificadas pelo caso do COMPAS.

Foi definido um propósito claro para o uso do sistema?

Como garantir que o sistema não seja utilizado para fins diferentes daqueles para os quais foi desenvolvido?

Em relação a essas perguntas, no caso do COMPAS, Tim Brennan, fundador da

31 Wisconsin Supreme Court. State v. Loomis. Disponível em: https://harvardlawreview.org/2017/03/state-v-loomis/

Northpointe, apontou que seu foco ao projetar o sistema era reduzir a criminalidade, não determinar penas. Conforme descrito anteriormente sobre esse caso de uso, o sistema degenerou-se e acabou sendo utilizado como base para apuração da culpa do réu, o que está longe do objetivo original no momento de seu desenvolvimento.

O sistema foi testado em diferentes grupos demográficos para mitigar os vieses existentes?

Foram tomadas medidas para mitigar vieses históricos nos bancos de dados usados?

O fato de o questionário do COMPAS ser composto por perguntas sobre infância, antecedentes ou bairro onde o réu morava deve ter alertado sobre os vieses que poderiam ser introduzidos nos dados. Nesse caso, não foram tomadas medidas apropriadas para mitigar os vieses históricos existentes nos dados usados, resultando na atribuição incorreta de pontuações de risco a indivíduos com registros de antecedentes díspares.

A definição da arquitetura e das técnicas utilizadas está de acordo com as necessidades de transparência e explicabilidade das decisões exigidas pelo setor de atividade em que o sistema está inserido?

Existem certos setores em que a explicabilidade das decisões é essencial para a aceitação adequada dos sistemas de IA pela sociedade. Em matéria de administração da justiça, a explicabilidade é um imperativo. No caso mencionado acima, as partes interessadas não tinham conhecimento de como a pontuação era atribuída, pois a Northpointe argumentava que essa informação era um segredo comercial confidencial.

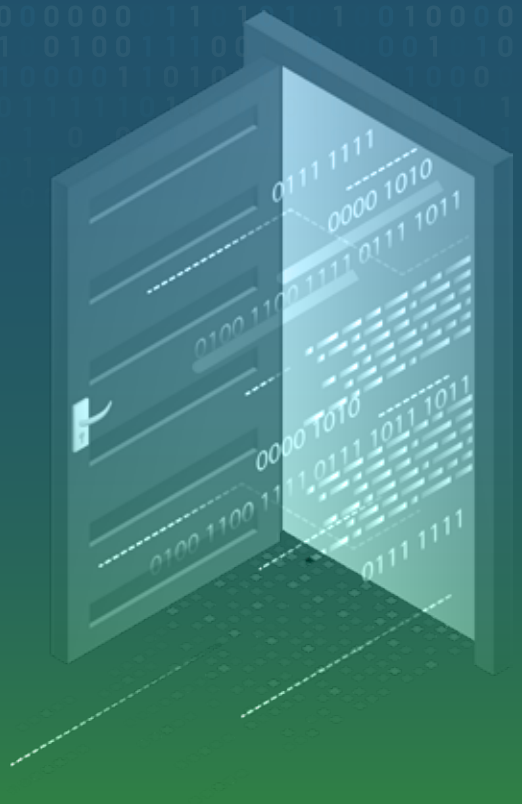
Considerando-se a instabilidade que caracteriza vários modelos de aprendizado de máquina, o modelo foi validado em várias ocasiões e casos para garantir que o sistema respondesse corretamente em diferentes contextos?

Como foram definidos os pontos ótimos de sensibilidade e especificidade na curva de ROC?32 São adequados para o setor no qual o sistema está sendo implementado?

Uma avaliação do sistema feita depois, na qual foram analisados 16 mil casos, revelou que sua precisão se aproximava de 71%[33]. Por tratar-se de um contexto de implementação altamente sensível como o da administração da justiça, é evidente que a precisão revelada está longe do que se poderia considerar ideal. Com base no exposto acima, pode-se concluir que foram necessários períodos de teste mais longos e parâmetros de validação mais elevados.

32 A curva de ROC (característica operacional do receptor, na sigla em inglês) é uma ferramenta estatística que permite avaliar a precisão das previsões de um modelo. Por exemplo, caso esteja tentando implementar um modelo que classifica as pessoas de acordo com seu risco de cometer crimes, a curva de ROC pode avaliar a precisão desse modelo.

33 Angwin et al. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



6. Comentários finais

A inteligência artificial desempenha um papel fundamental no nosso cotidiano e na nossa convivência como sociedade, a ponto de tornar-se cada vez mais difícil pensar em um caso no qual não interajamos com sistemas inteligentes no mundo atual. Dispositivos móveis, eletrodomésticos, meios de transporte, entre muitos outros, contam com esse tipo de sistema para facilitar as nossas atividades, tornando-as mais confortáveis e seguras.

Assim como a tecnologia de IA, o campo da auditoria algorítmica está avançando a um ritmo acelerado, aumentando a importância de seu uso. É um trabalho em contínua evolução, cujo escopo está em permanente fluxo. Assim, seu conteúdo deve ser atualizado regularmente, à medida que novas ferramentas tecnológicas e a regulamentação correspondente são desenvolvidas.

Dado o uso generalizado dos ADSs na sociedade, principalmente em áreas nas quais seu uso pode exigir precauções extraordinárias, é necessário realizar revisões constantes para orientar sua correta implementação. Já houve casos de sistemas que, por falhas na formulação ou no desenvolvimento, geram impactos negativos significativos nas nossas vidas e na sociedade: do aumento das tarifas de transporte público a sentenças injustas.

Entre os desafios que essa área representa para o setor público, está o movimento da realização de auditorias algorítmicas como um mecanismo voluntário rumo à sua inclusão como parte de uma política estruturada sobre o assunto ou como parte de uma ampla regulamentação sobre o assunto de responsabilidade algorítmica.

Este guia não pretende ser apenas uma ferramenta prática que contribua para monitorar áreas cruciais e mitigar riscos ou perigos que podem não ser óbvios à primeira vista. Espera-se, também, que sirva como um instrumento para ajudar a aumentar a conscientização sobre as implicações e consequências da implementação de um ADS. Esperamos que todas as equipes, tanto de entidades públicas como de promotores de ADSs, estejam cientes da relevância de seu trabalho. Se quisermos garantir um futuro mais justo e seguro, é necessário entender e esclarecer plenamente sua relevância.

Referências

Ada Lovelace Institute (2021). Algorithmic Accountability for the Public Sector. Disponível em: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

---------- (2020). Examining the Black Box: Tools for assessing algorithmic systems. Disponível em: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Angwin, J., Larsson, J., Mattu, S. e Krichner L. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

BID (2021) Uso responsable de IA para política pública: manual de formulación de proyectos. Disponível em: https://publications.iadb.org/es/uso-responsable-de-ia-para-politica-publica-manual-de-formulacion-de-proyectos

BID Lab (2021). Guía de aplicación. Autoevaluación ética de IA para actores del ecosistema emprendedor. Disponível em: https://publications.iadb.org/publications/spanish/document/Autoevaluacion-etica-de-IA-para-actores-del-ecosistema-emprendedor-Guia-de-aplicacion.pdf

BID/ fAIr LAC (2020). Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe. Disponível em: https://publications.iadb.org/publications/spanish/document/fAIr_LAC_Adopción_ética_y_responsable_de_la_inteligencia_artificial_en_América_Latina_y_el_Caribe_es.pdf

Brennan, T. e Dieterich, W. (2017). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Disponível em: https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS

Fjeld, J. e Nagy, A. (2020). Principled Artificial Intelligence. Disponível em: https://cyber.harvard.edu/publication/2020/principled-ai

Government of Canada (2021). Directive on Automated Decision-Making. Disponível em: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

IA Now Institute (2028). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. Disponível em: https://ainowinstitute.org/aiareport2018.pdf

Intersoft Consulting. S.f. Data Protection Regulation. Disponível em: https://gdpr-info.eu/art-5-gdpr/

INTOSAI (2019). Performance Audit Principles. Disponível em: https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

ISO 19011-2018 (2018). Disponível em: https://www.iso.org/standard/70017.html

Moss, E., Watkins, E. A., Singh, R., Elish, M. C. e Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Disponível em: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

OECD (2019). Recommendation of the Council on Artificial Intelligence. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD/BID (2020). IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial. Disponível em: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

Raji, I. D. et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Disponível em: https://arxiv.org/pdf/2001.00973.pdf

Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponível em: https://www.auditingalgorithms.net

Swedish National Audit Office (2020). Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up. Disponível em: https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

UK Information Commissioner's Office (2020). Guidance on the AI auditing framework, Draft guidance for consultation. Disponível em: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Congresso dos EUA (2019). Algorithmic Accountability Act of 2019.

BID

Melhorando vidas