Capítulo 5

Responsible AI: Princípios para o Projeto, Desenvolvimento e Implantação Responsável de Soluções Baseadas em Inteligência Artificial

Marcelo S. Locatelli*§, Mateus Zaparoli*§, Victor Thomé*, Marcelo M. R. Araújo*, Matheus Prado*, Thaís Ferreira*, Igor Joaquim Costa*, Tomas Lacerda*, Leonardo Augusto Ferreira◇, Marisa Vasconcelos*, Julio C. S. Reis†, Jussara M. Almeida*, Wagner Meira Jr.*, Virgílio Almeida‡*

*Departamento de Ciência da Computação – Instituto de Ciências Exatas
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

◇Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

†Departamento de Informática – Centro de Ciências Exatas
Universidade Federal de Viçosa (UFV) – Viçosa, MG – Brasil

‡Berkman Klein Center For Internet & Society
Harvard University – Boston, MA – Estados Unidos

{locatellimarcelo, mateuszaparoli, victorthome, marceloaraujo,
matheus.prado, thaisferreira, igor.joaquim, tomas.muniz,
marisavasconcelos, jussara, meira, virgilio}@dcc.ufmg.br,
leauferreira@cpdee.ufmg.br, jreis@ufv.br

Abstract

The use of Artificial Intelligence (AI) is increasingly common in various sectors of society, such as public safety, health, and economy, bringing benefits such as cost reduction and increased accuracy in decision-making. However, it is important to question what the associated risks are, especially in relation to possible biases that can be introduced into the results and the implications for fairness and transparency in the decisions made. In light of such concerns, the term "Responsible AI (RIA)" has been coined to refer to the practice of developing and using AI systems in a way that benefits society while minimizing the risk of negative consequences.

§Ambos autores contribuíram igualmente para o trabalho.

The objective of this chapter is to discuss the current scenario in the context of RIA, offering an introduction to researchers who intend to work on this topic, as well as an overview of the fundamental principles of RIA for the design, implementation, and deployment of AI-based systems. To this end, definitions are presented and metrics related to each of the pillars of RIA, such as fairness, transparency, privacy and security, and accountability, are discussed. After that, we discuss the concept of governance, which includes management principles and practices on laws and regulations, quality assurance (which also considers risk assessment, and continuous auditing), as something indispensable for the implementation of RIA principles. Finally, we provide a critical overview of the field, highlighting challenges and research opportunities in the Brazilian context with special attention to local practices and policies that may influence the development and application of responsible AI in Brazil.

Resumo

O uso da Inteligência Artificial (IA) é cada vez mais comum em diversos setores da sociedade, como segurança pública, saúde e economia, trazendo benefícios como redução de custos e aumento da precisão na tomada de decisões. No entanto, é importante questionar quais são os riscos associados, principalmente em relação a possíveis vieses que podem ser introduzidos nos resultados e as implicações para a justiça e transparência nas decisões tomadas. À luz dessas preocupações, o termo "IA Responsável (IAR)" foi cunhado para se referir à prática de desenvolver e usar sistemas de IA de uma forma que beneficie a sociedade, minimizando o risco de consequências negativas.

O objetivo deste capítulo é discutir o cenário atual no contexto da IAR, oferecendo uma introdução aos pesquisadores que pretendem trabalhar neste tópico, bem como uma visão geral dos princípios fundamentais da IAR para o design, implementação e implantação de sistemas baseados em IA. Para tanto, são apresentadas definições e discutidas métricas relacionadas a cada um dos pilares da IAR, como justiça, transparência, privacidade e segurança e responsabilização. Depois disso, discutimos o conceito de governança, que inclui princípios e práticas de gestão sobre leis e regulamentos, garantia de qualidade (que também considera avaliação de risco e auditoria contínua), como algo indispensável para a implementação dos princípios da IAR. Por fim, fornecemos uma visão geral crítica do campo, destacando desafios e oportunidades de pesquisa no contexto brasileiro, com atenção especial às práticas e políticas locais que podem influenciar o desenvolvimento e a aplicação da IA responsável no Brasil.

5.1. Introdução

O uso crescente de algoritmos baseados em Inteligência Artificial (IA) está se tornando comum em uma variedade de setores, como segurança pública, saúde e economia [Shah et al., 2024; Li et al., 2023; Zota et al., 2023]. Isso tem proporcionado benefícios significativos, como a redução de custos e uma tomada de decisões cada vez mais precisa. No entanto, é crucial examinar os potenciais riscos associados, especialmente os vieses que podem ser introduzidos nos resultados produzidos pelas soluções baseadas em IA, além das implicações para a justiça e transparência nas decisões. Para isso é necessário projetar, implementar e implantar soluções baseadas em IA que sejam transparentes, justas

(livres de vieses), explicáveis e centradas no ser humano.

Esse tema tem gerado debates acalorados em várias comunidades científicas, levando à formação de comitês e conselhos de pesquisadores globalmente focados na promoção do que recentemente tem sido referenciado como IA responsável (IAR). Exemplos de comitês internacionais com foco na discussão de temas relacionados à IAR são: *Global AI Ethics Consortium*[1], *Global Partnership on AI*[2], *IADB's fAIr LAC Initiative*[3] (América Latina e Caribe) e *ACM's US Technology Policy Committee*[4]. Neste contexto, é importante enfatizar que esse tema é extremamente novo e motiva uma reflexão científica que contemple questões específicas do cenário brasileiro, considerando aspectos culturais e vieses sociais (e.g., cor de pele, etnia, classe e condição econômica). Assim, acreditamos que é essencial fornecer à comunidade uma fundamentação conceitual de IAR, contribuindo assim para minimização dos impactos ocasionados por essas aplicações em nossa sociedade. Ademais, o desenvolvimento de habilidades, capacitação e compartilhamento de conhecimentos a respeito dos riscos associados aos sistemas que envolvem, em algum nível, soluções baseadas em IA é um fator que fomenta a discussão aqui apresentada.

Assim, o objetivo deste capítulo é proporcionar aos leitores uma compreensão abrangente dos conceitos relacionados à IAR. Inicialmente, apresentamos o estado da arte nesta área, incluindo definições e uma descrição de cada um dos pilares associados, incluindo *Justiça*, *Transparência*, *Responsabilização*, *Privacidade e Segurança* e *Governança*. Além disso, exploramos as principais métricas e/ou métodos atualmente empregados para atender ou mitigar os riscos associados a cada princípio. Em seguida, apresentamos considerações finais, incluindo desafios e oportunidades de pesquisa na área, com um enfoque especial no contexto brasileiro, abarcado por suas especificidades.

Por fim, esperamos que os leitores adquiram uma compreensão aprofundada da IAR, capacitando-os para projetar, desenvolver e implantar soluções baseadas em IA que promovam um ambiente digital mais saudável e confiável. Neste sentido, este capítulo visa não só uma exposição teórica dos conceitos e definições relacionadas à temática, mas também apresentar práticas introdutórias, as quais podem ser consideradas passos fundamentais no estudo e desenvolvimento de soluções responsáveis no contexto de IA, contribuindo com a formação de recursos humanos, capacitação e compartilhamento de conhecimentos a respeito dos impactos relacionados à área.

5.2. Inteligência Artificial Responsável (*Responsible* AI)

O termo "Inteligência Artificial (IA) Responsável" (IAR) denota a tentativa de encontrar maneiras práticas de lidar com as várias questões éticas, sociais, dentre outras, associadas ao uso de soluções de IA para resolver problemas da sociedade moderna. Ele se baseia em uma longa discussão sobre o conceito de responsabilidade no âmbito do direito, ciências sociais e filosofia moral [Stahl, 2023]. Embora a IA seja uma tecnologia estabelecida que se beneficia da vasta disponibilidade de dados e do avanço do hardware para processamento, a amplitude das aplicações de IA ainda cresce, trazendo desafios po-

[1] https://www.ieai.sot.tum.de/global-ai-ethics-consortium/
[2] https://gpai.ai/
[3] https://fairlac.iadb.org/pt
[4] https://www.acm.org/public-policy/ustpc

tenciais derivados de seu uso inadequado e decisões que podem resultar em injustiças ou vieses. Portanto, há uma crescente demanda por regulamentações que promovam o uso responsável dessa tecnologia.

Baseado no exame de sistemas e ecossistemas de IA, pode-se identificar os principais requisitos para a implementação eficaz de políticas de IA responsáveis. Neste contexto, um ecossistema equipado para lidar com questões éticas e sociais exigiria uma definição clara em termos de tempo, tecnologia e geografia para estabelecer uma estrutura precisa de (meta-)responsabilidade. Dentre os conceitos cruciais para o desenvolvimento responsável da IA, questões como justiça, transparência, entre outros, emergem como pontos centrais na Governança de IA [Dignum, 2020]. No entanto, a imensa quantidade de dados disponíveis, torna inviável a análise manual dessas informações que alimentam os conjuntos de treinamento dos sistemas de IA. Essa dificuldade tem contribuído para problemas emergentes, tanto nacional quanto internacionalmente, relacionados a vieses e violações de direitos humanos [Doneda et al., 2018]. Diante desse cenário, torna-se imperativo adotar princípios fundamentais e/ou medidas responsáveis no projeto, desenvolvimento e implantação de soluções baseadas em IA [Citron, 2007; Colombelli, 2024].

Em resumo, os princípios fundamentais da IA responsável buscam garantir que os sistemas de IA sejam seguros, aceitos e confiáveis. Para isso, o sistema deve ser projetado considerando princípios éticos e as consequências morais de suas ações e decisões, de forma responsável e transparente [Dignum, 2017]. Várias organizações internacionais desenvolveram conjuntos de princípios fundamentais para a adoção da IAR[5]. Esses princípios visam mitigar riscos específicos associados ao uso dessas tecnologias, como desigualdade, preconceito, manipulação, discriminação, violação de privacidade e falta de responsabilização.

No Brasil, visando guiar o desenvolvimento da IA no país, o governo brasileiro introduziu, em 2021, a Estratégia Brasileira de Inteligência Artificial (EBIA), e mais recentemente, o Plano Brasileiro de Inteligência Artificial (PBIA)[6]. Ambas iniciativas estão alinhadas com os pilares internacionais delineados pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [Masseno, 2020], que incluem princípios fundamentais sobre IA como desenvolvimento sustentável, bem-estar, valores centrados no ser humano, equidade, transparência, explicabilidade, segurança, e responsabilização. Neste capítulo, abordamos em mais detalhes cada um desses pilares de IA Responsável (IAR), conforme resumido na Figura 5.1.

Neste contexto, é importante mencionar que a literatura no contexto de IAR embora recente, é vasta – existem trabalhos relacionados que elencam pilares adicionais, no entanto, este capítulo foca nestes pilares cobrindo conceitos relacionados de forma distribuída nas seções subsequentes.

[5]https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

[6]https://www.gov.br/lncc/pt-br/assuntos/noticias/ultimas-noticias-1/plano-brasileiro-de-inteligencia-artificial-pbia-2024-2028

Figura 5.1. Pilares da IA Responsável (IAR).

5.3. Justiça (*Fairness*)

No âmbito da IAR, o princípio relacionado à justiça (ou *fairness*), envolve garantir que sistemas de IA sejam justos e tratem de forma equitativa todos os indivíduos afetados por decisões automatizadas ou assistidas por IA, independentemente de características como raça, gênero, idade, origem étnica, entre outros [Google AI]. Quando essas características são exploradas direta ou indiretamente, esses grupos podem ser tratados de forma desigual, o que pode resultar em oportunidades desproporcionalmente favorecidas ou prejudicadas para determinados grupos. Isso pode acarretar consequências adversas para indivíduos e para a sociedade como um todo, como observado em certas tarefas para predição de reincidência criminal [Publica, 2024] e sistemas de recrutamento de pessoal [Eun and Hwang, 2020].

No contexto de reconhecimento facial, por exemplo, muitos sistemas tendem a ter melhor desempenho ao classificar alguns gêneros ou etnias enquanto têm desempenho inferior ao considerar outros. Alguns estudos de caso, como o apresentado em [Buo-

lamwini and Gebru, 2018], já evidenciaram essas disparidades em sistemas correntes. Na obra especificada, os autores analisaram três sistemas comerciais a partir de um conjunto de dados existente, e demonstraram que as taxas de erro variavam significativamente, com um desempenho superior na categorização de homens com tom de pele claro em comparação com mulheres com tom de pele escuro.

Outro exemplo de discriminação algorítmica é o sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), utilizado para prever a probabilidade de um indivíduo reincidir criminalmente[7]. Avaliações mostraram que o COMPAS tem um viés racial, com uma taxa de falso positivo de 45% para réus afro-americanos, comparada a 23% para réus brancos [Angwin et al., 2016]. Isso significa que o sistema é mais propenso a prever incorretamente a reincidência para réus afro-americanos. Na área da saúde, um algoritmo usado para identificar pacientes que necessitam de cuidados adicionais apresentou viés. Ao usar os custos anteriores de saúde como critério, o algoritmo subestimou as condições de saúde dos pacientes negros, perpetuando desigualdades existentes [Obermeyer et al., 2019].

Dessa forma, conhecendo as extensas implicações sociais possíveis, é imperativa a definição do conceito de justiça (*fairness*), bem como quais as métricas devem ser utilizadas para sua medição, além da proposição de medidas corretivas para mitigação. Na próxima seção serão apresentados definições de justiça encontradas na literatura e das métricas propostas em diferentes cenários.

5.3.1. Definições

A literatura sobre justiça no contexto de IAR apresenta várias propostas de definições e abordagens. Ferrara [2023] explora as complexidades associadas às definições de justiça, destacando como o viés sistêmico pode ser introduzido em diversas fases do desenvolvimento e uso de IA. Nesse estudo, o autor explora as complexidades associadas às definições de justiça, destacando como o viés sistêmico pode ser introduzido em várias fases do desenvolvimento e uso de IA.

O termo justiça (do inglês *fairness*) é frequentemente definido a partir de diferentes perspectivas. Uma delas é a paridade demográfica, que busca igualdade de tratamento entre grupos [Yan et al., 2020]. Outra é a justiça individual, que se concentra no tratamento igualitário de indivíduos, independentemente de seu grupo [Zafar et al., 2017]. Além disso, a justiça contrafactual é discutida como um esforço para garantir decisões justas em condições hipotéticas, evidenciando a complexidade de desenvolver IAs que operem de maneira justa em cenários diversos e dinâmicos [Zafar et al., 2017]. Assim, não há definição padronizada de justiça na literatura; é um conceito fluido, cuja interpretação pode variar dependendo do contexto cultural, econômico, social e da aplicação em análise [Google AI].

5.3.2. Métricas

Diversas métricas foram propostas para avaliar justiça em aplicações de IA considerando diferentes contextos e objetivos [Mehrabi et al., 2021]. Entre essas métricas, destacam-se:

[7] `https://www.bbc.com/portuguese/brasil-37677421`

- *Equal Opportunity*: garante que todos os grupos tenham a mesma taxa de verdadeiros positivos (TTP) quando pertencem à classe positiva [Hardt et al., 2016a]. Por exemplo, em um banco que usa um modelo para prever a capacidade de pagamento de empréstimos, os candidatos são divididos em dois grupos: não-minoritário e minoritário. O modelo deve garantir que ambos os grupos tenham a mesma TTP, ou seja, a mesma taxa de aprovação para candidatos qualificados. Se o modelo aprova 90% dos candidatos não-minoritários que pagarão o empréstimo, mas apenas 70% dos candidatos minoritários, ele viola esse critério. Para corrigir, o modelo deve ser ajustado para que a TTP seja igual em ambos os grupos, garantindo que tanto não-minoritários quanto minoritários tenham uma taxa de aprovação de 90% entre aqueles que pagarão o empréstimo. Isso garante acesso justo e equalitário aos empréstimos para todos os candidatos qualificados.

- *Demographic Parity* ou *Statistical Parity*: visa garantir que a probabilidade de uma decisão positiva seja a igual para todos os grupos, independentemente de suas características [Feldman et al., 2015]. Por exemplo, em uma ferramenta de contratação, se 70% dos candidatos masculinos são selecionados para entrevistas, mas apenas 30% das candidatas femininas, isso indica um problema de justiça. Para cumprir a paridade demográfica, a empresa deve garantir que ambos os gêneros sejam selecionados em taxas semelhantes. Se 50% dos candidatos masculinos forem escolhidos, 50% das candidatas femininas também devem ser selecionadas, independentemente das qualificações. O objetivo é equilibrar as taxas de seleção, assegurando tratamento justo em relação ao gênero. No entanto, essa abordagem pode desconsiderar fatores como mérito, criando um dilema entre equidade e utilidade, conforme discutido no artigo.

- *Equalized Odds*: avalia se um modelo apresenta taxas semelhantes de falsos positivos (TFP) e falsos negativos (TFN) para diferentes grupos demográficos [Hardt et al., 2016a];

- *Disparate Impact*: mede a diferença nas taxas de aceitação ou rejeição entre grupos distintos, sinalizando possíveis discriminações [Zafar et al., 2017];

- *Disparate Mistreatment*: avalia se os erros de um modelo (TFP e TFN) são distribuídos de forma desigual entre diferentes grupos, identificando se algum grupo é sistematicamente mais prejudicado [Zafar et al., 2017].

No entanto, algumas dessas métricas enfrentam desafios devido à variabilidade nas taxas de TFP e TTP, especialmente em amostras de dados com distribuições diferentes entre grupos. A métrica *Conditional Equality of Opportunity* aborda essa limitação ao condicionar a análise a características específicas e avaliar a diferença de oportunidade entre grupos privilegiados e não privilegiados em contextos relevantes [Beutel et al., 2019]. O termo "condicional" significa que a métrica considera igualdade de oportunidades sob certas condições ou contextos adicionais relevantes para o problema específico, como subgrupos dentro dos dados ou a contextos específicos da aplicação do modelo. Dessa forma, ela garante que o modelo seja justo não apenas de forma geral, mas também em situações específicas importantes para o problema em questão.

Para cenários específicos, como reconhecimento facial, outras métricas incluem:

- *Fairness Discrepancy Rate* (FDR): mede a diferença máxima entre as taxas de falsas correspondências (FMR) e falsas não correspondências (FNMR) entre grupos demográficos, variando de 0 a 1, onde 1 indica equidade máxima [Pereira and Marcel, 2022];
- *Functional Fairness Measure Criteria* (FFMC): avalia a interpretabilidade das medidas de equidade, focando na contribuição das taxas de FMR e FNMR [Howard et al., 2022];
- *Inequity Rate* (IR): calcula a razão entre as taxas de erro máximas e mínimas entre grupos, com limitações quando a taxa de erro mínima é zero [Grother, 2021];
- *Pareto Curve Optimization with Overall Effectiveness*: avalia algoritmos com base no equilíbrio entre eficácia e equidade, usando o princípio da eficiência de Pareto para comparar visualmente as compensações entre diferentes algoritmos [Wei and Niethammer, 2022].

Outras métricas de justiça podem ser encontradas em [Mehrabi et al., 2021; Anahideh et al., 2021; Pagano et al., 2023].

5.3.3. Ações Mitigatórias

Na mitigação de viés em inteligência artificial, as técnicas são geralmente organizadas em três categorias principais: pré-processamento, no processamento e no pós-processamento [Mehrabi et al., 2021].

As técnicas de pré-processamento corrigem vieses nos dados antes do treinamento do modelo. Elas ajustam as distribuições amostrais das variáveis protegidas ou aplicam transformações específicas nos dados para eliminar discriminações [Kamiran and Calders, 2012; Calmon et al., 2017]. Essa abordagem é flexível, pois não depende da técnica de modelagem aplicada posteriormente.

Os métodos no processamento adaptam os algoritmos de aprendizado de máquina para torná-los mais resistentes a vieses. A regularização [Tian and Zhang, 2022], por exemplo, adiciona penalizações às funções objetivo do modelo para evitar *overfitting*. Ela também ajuda a reduzir a complexidade do modelo e melhora sua capacidade de generalização. No contexto de justiça, o *Fairness-Aware Learning* [Kamishima et al., 2011] introduz um termo de regularização, o *Prejudice Remover Regularizer*, para reduzir a dependência entre a saída do modelo e atributos sensíveis. O aprendizado adversarial [Zhang et al., 2018] ajusta o comportamento do modelo introduzindo distorções adversariais nos dados, tornando-o mais robusto e consistente entre grupos. Embora essas abordagens possam aumentar a justiça dos modelos, é importante notar que a acurácia pode ser afetada, exigindo outros métodos e ajustes para equilibrar justiça e desempenho.

As técnicas de pós-processamento ajustam as previsões do modelo após o treinamento para melhorar a justiça, sem modificar os dados de entrada ou o algoritmo. A

técnica *Equalized Odds*, por exemplo, garante taxas iguais de verdadeiros positivos entre grupos, dado que pertencem à classe positiva [Hardt et al., 2016b]. A *Reject Option Classification* ajusta previsões para indivíduos próximos à fronteira de decisão, oferecendo resultados mais favoráveis a grupos desfavorecidos [Kamiran et al., 2012]. Outra abordagem é o "detector de viés individual", proposto por Lohia et al. [2019], que usa um classificador treinado para identificar e ajustar amostras com possível viés individual. Determinar o nível de tolerância para desigualdade é um desafio para essas abordagens, e os ajustes necessários podem ser calculados manualmente ou com métodos estatísticos.

Para melhor compreender algumas das técnicas, considere um cenário de recrutamento de novos funcionários em uma empresa de tecnologia. O objetivo é ter um modelo automatizado para prever se os candidatos serão contratados ou não, com base em variáveis como experiência, educação e habilidades técnicas.

Em um dos possíveis panoramas, os dados históricos mostram um viés significativo a favor de homens, que são contratados com muito mais frequência que mulheres, mesmo quando ambos os grupos apresentam qualificações semelhantes. Suponha que o conjunto de dados original tenha 1.000 homens e 500 mulheres, dos quais 700 homens e 100 mulheres foram contratados. Para diminuir esse desbalanceamento, uma das abordagens de pré-processamento seria a aplicação de técnicas de reamostragem. No caso de *undersampling*, o número de homens contratados no conjunto de dados seria reduzido, removendo aleatoriamente algumas instâncias até que o número de contratações de homens e mulheres esteja mais equilibrado. De forma alternativa, há também o *oversampling*, que atuaria replicando instâncias de mulheres contratadas para atingir valores próximos àqueles de homens.

Outro possível panorama, na categoria de processamento, envolve o uso de técnicas de regularização durante o treinamento do modelo. Tratando-se do *Prejudice Remover Regularization*, um termo penalizador seria inserido na função de custo quando ocorresse uma correlação indevida entre a variável sensível e as previsões. Nesse contexto, isso é feito para reduzir a dependência entre a decisão de contratação e o gênero do candidato. Dessa forma, as contratações receberiam mais influências das outras variáveis, como experiência e habilidades técnicas, promovendo decisões mais justas sem prejuízo significativo a acurácia.

Em contextos sensíveis como reconhecimento facial, o estudo recente [Melzi et al., 2023] explorou como reduzir o viés demográfico usando dados sintéticos. A ideia foi criar conjuntos de dados ajustáveis, onde características demográficas pudessem ser controladas, permitindo que os modelos fossem treinados de maneira mais equilibrada. Os resultados mostraram que essa técnica pode ajudar a reduzir as disparidades entre grupos e tornar sistemas de reconhecimento facial mais justos.

5.3.4. Ferramentas

Existem diversas ferramentas (ou *frameworks*) relacionados à justiça auxiliam na auditoria de modelos de aprendizado de máquina e na capacitação sobre esse pilar da IA responsável. Abaixo, estão listadas algumas ferramentas *open-source* relevantes que podem contribuir para o entendimento e a aplicação dessa temática.

Tabela 5.1. Ferramentas de justiça em IA.

Ferramentas	Descrição
Aequitas	Realiza auditoria de preconceito e justiça em modelos de IA.
IBM's AI Fairness 360	Auxilia na avaliação, relato e mitigação de discriminação e preconceito em soluções de IA.
Google's What-If Tool	Permite explorar diferentes definições de justiça.
Microsoft's Fairlean	Avalia a imparcialidade e mitiga problemas de injustiça, oferecendo algoritmos de mitigação e métricas de avaliação do modelo.

- O *Aequitas*[8] - termo em Latim para equidade – é uma ferramenta de auditoria de justiça desenvolvida pelo *Center for Data Science and Public Policy* (DSaPP) da Universidade de Chicago. Ele identifica vieses e disparidades em modelos de aprendizado de máquina, facilitando a análise das decisões do modelo em relação a características sensíveis, como raça e gênero. A ferramenta gera relatórios que indicam discrepâncias entre grupos e fornece várias métricas de justiça[9], como *False Omission Rate*, *Prevalence* ou *False Positive Rate* (FPR), para avaliar a equidade dos modelos. Uma demonstração utilizando o COMPAS pode ser acessada em: `http://aequitas.dssg.io/example.html`.

- O *What-If Tool*[10] (WIT) é uma ferramenta interativa desenvolvida pelo Google para explorar diferentes definições de justiça e entender o comportamento dos modelos de aprendizado de máquina. Ela permite simular mudanças nos dados e parâmetros do modelo, visualizando seu impacto no desempenho. A ferramenta facilita a comparação do desempenho entre subgrupos, como gênero ou raça, e permite ajustes para melhorar a justiça do modelo. O WIT é integrado ao *TensorBoard* e ao *TensorFlow*.

Devido à pluralidade conceitual em torno da definição de justiça na literatura, diversas empresas desenvolvem ferramentas proprietárias para mitigar os efeitos negativos da falta de justiça em modelos de aprendizado de máquina, como o *Microsoft Fairlearn*[11] e o *IBM's AI Fairness 360*[12]. A Tabela 5.1 resume as ferramentas. Para fins de exemplificação, exploramos o repositório de dados *Loan Eligible*[13], que consiste em um repositório dados de aplicações para empréstimos imobiliários, apresentando o valor do empréstimo, número de membros da família, assim como suas rendas, histórico de crédito e caso o empréstimo foi aprovado ou não, o que permite comparar a decisão humana com a algorítmica.

[8] `https://dssg.github.io/aequitas`
[9] `https://dssg.github.io/aequitas/metrics.html`
[10] `https://pair-code.github.io/what-if-tool/`
[11] `https://github.com/fairlearn/fairlearn`
[12] `https://aif360.res.ibm.com/`
[13] `https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset`

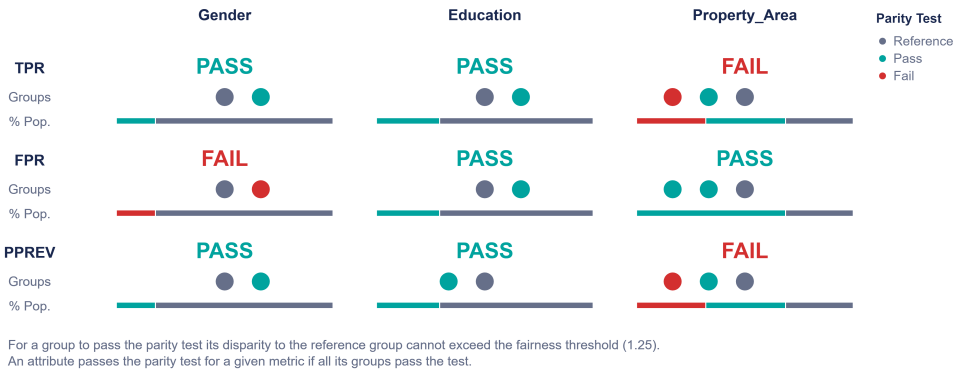

Figura 5.2. Resultado da execução do *Aequitas* para um classificador treinado na base de dados *Loan Eligible*. Os grupos de referência (cinza) são, respectivamente, homens, pessoas graduadas e propriedades urbanas. Em vermelho, pode-se ver casos em que a performance do modelo na métrica não ficou dentro do limiar de justiça definido pelo usuário (1.25). Nessa figura, esses casos são: Mulher (FPR 1.51 vezes maior que o da referência), Propriedade Rural (TPR 1.26 vezes menor que o da referência e PPREV 1.30 vezes menor que o da referência).

Nessa auditoria, consideramos as métricas *True Positive Rate* (TPR), *False Positive Rate* (FPR) e *Predicted Prevalence* (PPREV) e os atributos sensíveis *Gender* (gênero), *Education* (educação) e *Property_Area* (área da propriedade), buscando compreender se o modelo estaria favorecendo algum grupo. Para essa análise, utilizamos as configurações padrão da ferramenta, mas o limiar de justiça pode ser alterado pelo usuário para atender suas necessidades.

A Figura 5.2 mostra o resultado fornecido pela ferramenta. Pode-se observar que o classificador auditado é justo em relação à essas métricas para os diferentes níveis de educação da base. Porém, nota-se que o gênero feminino é favorecido com relação à FPR, isto é, estão sendo concedidos mais empréstimos para mulheres que não deveriam ter recebido empréstimos do que para homens. Por fim, nota-se que empréstimos para a compra de propriedades rurais são consistentemente desfavorecidos pelo nosso classificador, mostrando um possível viés.

5.4. Transparência (*Transparency*)

A transparência é um conceito fundamental no campo da IA, que se refere à capacidade de entender o processo de tomada de decisão dos modelos de IA com base nas informações que eles fornecem. A definição de transparência pode ser dada como a capacidade de um ser humano compreender o funcionamento de um modelo, mesmo que seu comportamento seja inesperado. É uma característica passiva do modelo, onde o entendimento se dá a partir do conhecimento do algoritmo, sem a necessidade de acesso aos dados ou ao modelo aprendido [Barredo Arrieta et al., 2020; Molnar, 2019; Vilone and Longo, 2020; Abdollahi and Nasraoui, 2018]. Logo, compreender como uma solução baseada em IA toma suas decisões é essencial para garantir que ela opere de maneira justa e não discriminatória.

Todavia, o aumento impactante da quantidade de publicações sobre o tema, con-

forme reportado por Jacovi [2023], trouxe à tona algumas questões que requerem um exame ainda mais minucioso. Um exemplo disso é a dificuldade em padronizar uma nomenclatura neste campo. Alguns autores como Rudin [2019] estabelecem uma diferença clara entre "explicabilidade" e "interpretabilidade"; "Interpretabilidade" refere-se a modelos inerentemente inteligíveis, enquanto a "explicabilidade" trata de fornecer esclarecimentos sobre modelos complexos e não transparentes. Lipton [2018] diferencia ainda mais esses conceitos, observando que interpretabilidade responde à pergunta "Como esse modelo funciona?", enquanto explicabilidade busca responder "O que mais esse modelo pode me dizer?". Entretanto, é possível encontrar autores como Molnar [2019], que utilizam os conceitos de interpretabilidade e explicabilidade como sinônimos.

A natureza da transparência varia entre diferentes modelos de IA, dependendo do contexto e do usuário [Rudin, 2019; Ribeiro et al., 2016]. No entanto, há uma noção comum que conecta os estudos sobre o tema: (i) a capacidade de soluções baseadas em IA de apresentarem informações de maneira compreensível para os humanos [Marcinkevics and Vogt, 2020; Doshi-Velez and Kim, 2017]. Essa noção, é definida por Vilone and Longo [2020] como entendibilidade (do inglês, *understandability*).

Portanto, para evitar ambiguidade, neste capítulo, abordamos a definição de que a transparência aborda como os resultados de uma solução baseada em IA são produzidos, a explicabilidade se refere à capacidade de fornecer esclarecimentos *post-hoc* sobre os modelos que são essencialmente opacos (comumente referenciados como caixas-pretas), e a interpretabilidade lida com o grau que um ser humano pode entender a causa de uma decisão.

5.4.1. Inteligência Artificial Explicável (XAI)

A Inteligência Artificial Explicável (XAI, do inglês *eXplainable AI*) visa tornar as soluções baseadas em IA mais claras e transparentes, o que amplia sua utilidade em setores onde é fundamental ter alta confiança nos resultados. XAI, de acordo com [Barredo Arrieta et al., 2020], pode ser definido como: "Dado uma audiência, explicabilidade se refere aos detalhes e razões que um modelo fornece para tornar seu funcionamento claro ou fácil de entender".

Algumas áreas da ciência, aplicam técnicas de XAI para diminuir a opacidade de métodos complexos com o intuito de atingir objetivos específicos. Muitos desses objetivos da utilização de XAI, catalogados em trabalhos anteriores [Vilone and Longo, 2020; Marcinkevics and Vogt, 2020; Barredo Arrieta et al., 2020] e resumidos na Figura 5.3, são elencados a seguir. Note, de forma geral, que eles são bastante relacionados com os princípios gerais de IAR.

- **Confiança**: A aceitação de soluções (i.e., modelos) de IA pela sociedade depende da confiança dos usuários. Em contextos de alto risco, como sistema carcerário/prisional, é essencial que os resultados gerados sejam compreendidos claramente;

- **Robustez / Estabilidade**: Explicações sobre soluções baseadas em IA garantem que os resultados não sejam vulneráveis a pequenas modificações nos dados de

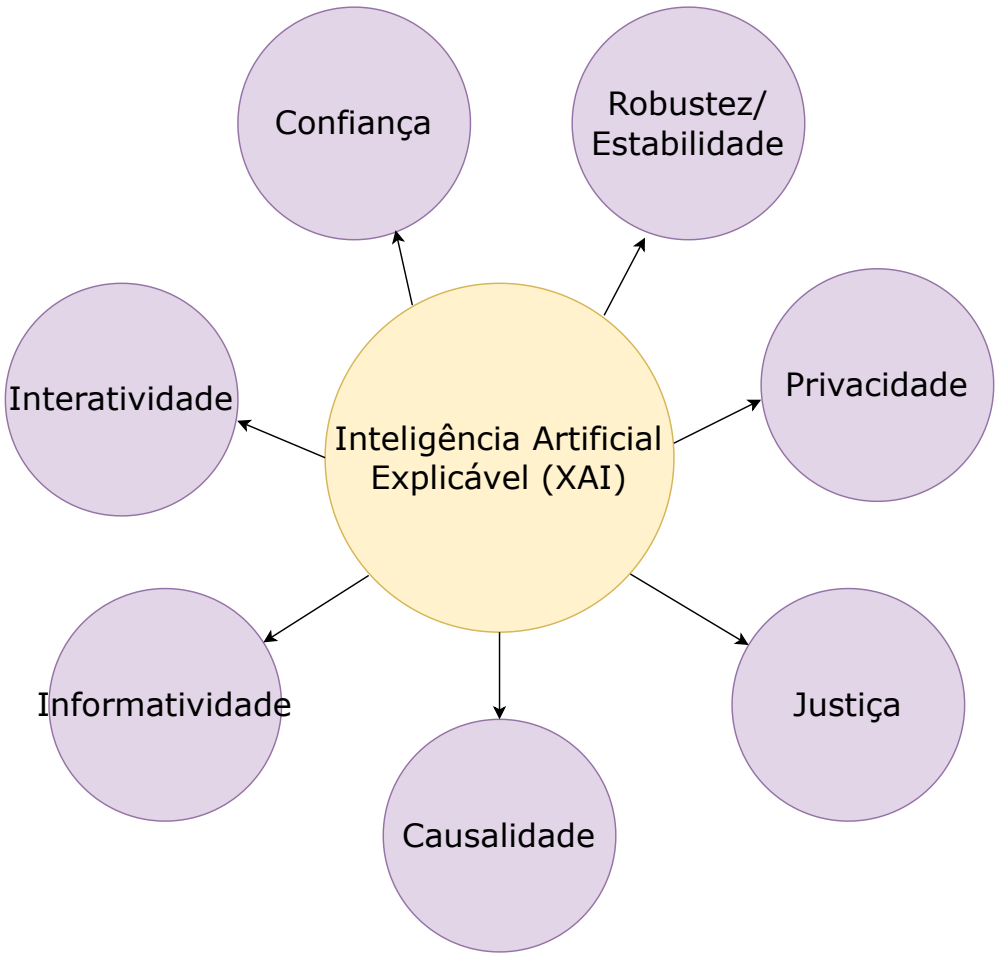

Figura 5.3. Principais objetivos da utilização do XAI.

entrada [Lipton, 2018; Vilone and Longo, 2020], assegurando estabilidade e previsibilidade;

- **Privacidade**: Conhecimento sobre o funcionamento interno das soluções/modelos pode ajudar a verificar se eles manipulam dados sensíveis, permitindo ajustes para a conformidades com normas previamente estabelecidas (e.g., LGPD [Doshi-Velez and Kim, 2017; Brasil, 2018] no contexto brasileiro);

- **Justiça (ou *Fairness*)**: Analisar a existência de discrepâncias no desempenho de soluções baseadas em IA, referentes à tarefa alvo, entre grupos ou instâncias com características distintas, a fim de identificar possíveis vieses;

- **Causalidade**: A capacidade de um método XAI esclarecer a relação entre as variáveis de entrada e saída;

- **Informatividade**: A capacidade de um método de explicabilidade para fornecer informações úteis aos usuários finais;

- **Interatividade**: Possibilitar que uma solução XAI de recapitular enunciados anteriores tanto para interpretar quanto para responder às perguntas de acompanhamento dos usuários.

5.4.2. Métodos

As estratégias de XAI podem ser divididas em (i) explicações globais e (ii) locais. As (i) explicações globais abrangem todas as instâncias, fornecendo uma visão geral sobre a relação entre as características e as saídas do modelo. Embora esse tipo de análise cumpra seu objetivo, é importante notar que algumas tendências podem ser mascaradas. Já as (ii) explicações locais analisam um (sub)espaço do problema, próximo a uma instância específica usada para gerar uma predição do modelo opaco. Essa abordagem pode revelar tendências não observadas na análise global, pois estas podem ter sido diluídas ao considerar todo o domínio do problema.

Dentre os diversos métodos existentes na literatura para a avaliar a transparência dos modelos de IA [Molnar, 2019], podemos citar PDP [Apley and Zhu, 2019; Molnar, 2019], LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017] e Captum [Kokhlikyan et al., 2020]. Detalhes acerca de cada um deles são apresentados nas seções subsequentes. Estes métodos abordam as categorias (i) e (ii). Neste cenário é importante mencionar que os exemplos relacionados nesta seção foram elaborados explorando o mesmo repositório de dados mencionado na Seção 5.3.4, a saber: *Loan Eligible*[14] – de aplicações para empréstimos imobiliários.

5.4.2.1. *Partial Dependence Plot* (PDP)

O *Partial Dependence Plot* (PDP) é uma estratégia global, que apresenta a contribuição marginal de uma determinada característica sobre a saída do modelo de IA, especificamente, aprendizado de máquina [Apley and Zhu, 2019; Molnar, 2019]. A análise do PDP gera informações sobre como é o comportamento desse relacionamento entre atributos (*feature*) e label (*target*) do modelo. A linearidade é um exemplo de tendência que pode ser identificada na análise do PDP. A Figura 5.4 apresenta uma análise marginal do atributo Valor do Empréstimo (*LoanAmount*) comparada com a probabilidade de conceder o empréstimo. Ambos oriundos do repositório de dados de empréstimos. Pode-se observar que pessoas que pedem empréstimos maiores tem menos chance de terem seu pedido aceito, especialmente quando comparado àquelas que pedem empréstimos nem tão grandes nem pequenos.

5.4.2.2. LIME (*Local Interpretable Model-Agnostic Explanations*)

O método LIME (*Local Interpretable Model-Agnostic Explanations*) [Ribeiro et al., 2016] gera explicações para qualquer modelo de classificação ao analisar localmente a vizinhança de uma instância específica e treinar um modelo de explicativo baseado nas distâncias entre as perturbações dessa instância e a original. Este modelo explicativo é projetado para ser de fácil interpretação e é especialmente útil em tarefas como reconhecimento

[14] `https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset`

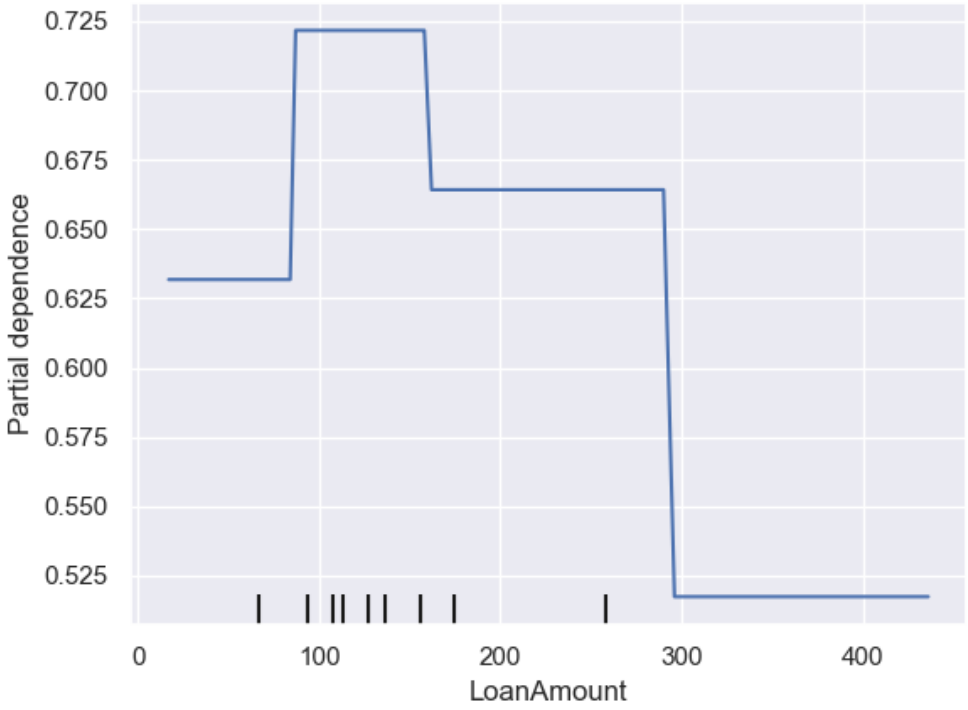

Figura 5.4. Explicação Global do PDP para o atributo Valor do Empréstimo (*LoanAmount*) presente no repositório de dados de empréstimos. Note que o pico na chance do empréstimo ser aceito coincide com valores intermediários.

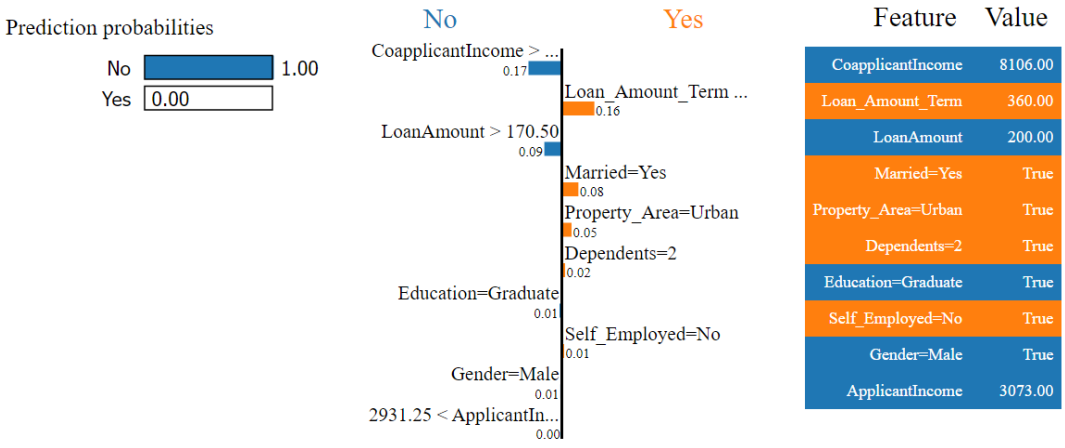

Figura 5.5. Exemplo de explicação do LIME para uma instância de um modelo de decisão sobre empréstimos. Note como a quantidade do empréstimo e o salário do co-aplicante aparecem como fatores para negação, enquanto estado civil e prazo surgem como fatores que favorecem a concessão do empréstimo.

de imagens. A Figura 5.5 apresenta um exemplo de explicação gerada pelo método para uma instância de um modelo de decisão para concessão de empréstimos. Nota-se que, para cada atributo, o LIME fornece o peso estimado que contribui para a predição ser positiva ou negativa.

Figura 5.6. Exemplo de explicação do SHAP para o conjunto de dados de teste de uma aplicação de empréstimos. Cada ponto representa uma instância de um atributo. Fatores como estado civil, histórico de crédito e a localização da propriedade têm grande influência nas decisões do modelo.

5.4.2.3. SHAP (*Shapley Additive Explanations*)

O método SHAP (*Shapley Additive Explanations*) utiliza valores de Shapley [Shapley, 1953], baseados na "Teoria dos Jogos Cooperativos" para determinar quanto cada fator (ou atributo) contribui para a decisão final do modelo [Lundberg and Lee, 2017]. Diferente de outros métodos, como o LIME, o SHAP garante uma explicação mais detalhada e matemática, sendo útil também para tarefas como reconhecimento de imagens [Molnar, 2019]. Uma vantagem do SHAP é que ele consegue resumir as explicações de várias previsões em uma única visualização. A Figura 5.6 ilustra isso: nela pode-se ver como cada atributo afeta a decisão final do modelo, onde pontos à direita indicam uma influência positiva. As cores ajudam a a identificar o valor de cada atributo, com o vermelho mostrando valores mais altos e o azul, valores mais baixos.

5.4.2.4. Captum

Por fim, o Captum [Kokhlikyan et al., 2020] é uma biblioteca *open source* desenvolvida sobre a plataforma *PyTorch*[15], voltada para a interpretabilidade de modelos de aprendizado de máquina. Ela oferece uma série de algoritmos avançados para atribuição de importância de características. A principal função do Captum é auxiliar pesquisadores e desenvolvedores a entender quais características influenciam as previsões de modelos complexos, aumentando a transparência de modelos que, por sua natureza, são difíceis de

[15] `https://pytorch.org/`

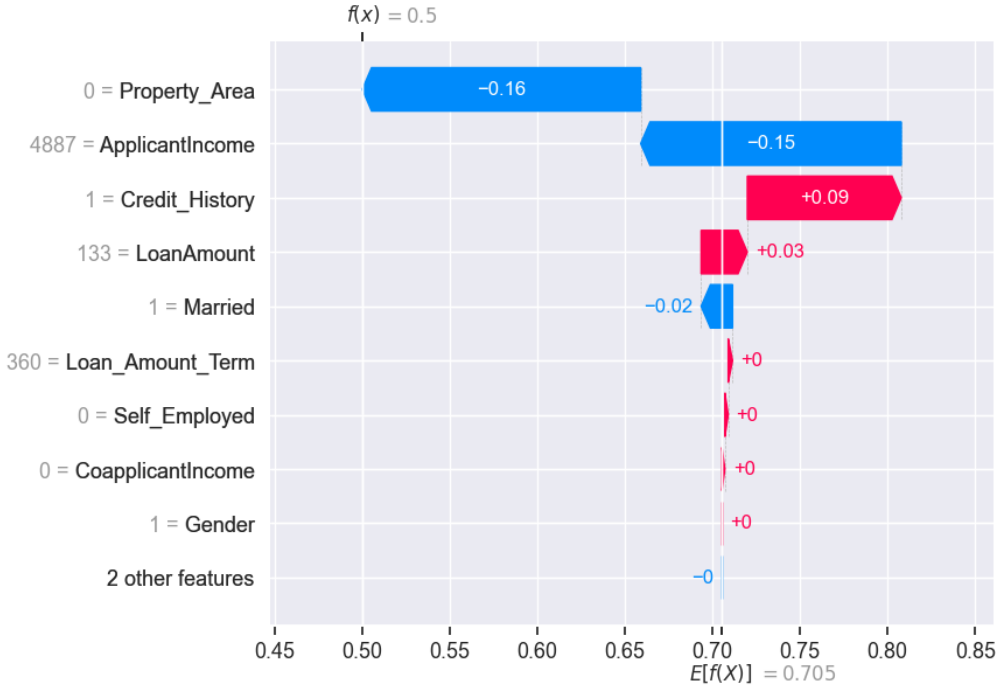

Figura 5.7. Previsão de empréstimos com SHAP: o valor esperado da chance do empréstimo ser concedido $E(f(X)) = 0,705$ não foi superado pela predição de $f(x) = 0,5$. A figura mostra como cada variável influenciou esse desvio. Os atributos `Property_Area` (area da propriedade) e `ApplicantIncome` (renda do aplicante) tiveram os maiores impactos, reduzindo o valor predito.

interpretar. Além disso, o Captum facilita a comparação de novos algoritmos com os já disponíveis e é útil tanto para depuração quanto para otimização de modelos, permitindo ajustes com base na identificação de características mais relevantes para as previsões.

5.4.2.5. Comparação dos Métodos de XAI

Os métodos PDP, LIME e SHAP podem ser aplicados a dados estruturados, oferecendo ao usuário explicações gráficas, textuais e simbólicas que ajudam a entender como um método opaco tomou uma decisão. É comum que essas explicações destaquem quais características são consideradas mais importantes pelo modelo. Para ilustrar isso, é apresentada uma comparação entre as explicações do SHAP e do LIME para uma mesma instância do repositório de dados *Loan Eligible*.

Na Figura 5.7 observa-se que o valor esperado da chance do empréstimo ser concedido é $E(f(X)) = 0,705$, enquanto a predição real é de $f(x) = 0,5$, um valor inferior. O objetivo da Figura 5.7 é mostrar como cada atributo contribuiu para o desvio do valor esperado. Os atributos *Property_Area* (area da propriedade) e *ApplicantIncome* (renda do aplicante) tiveram os maiores impactos, reduzindo o valor predito. É importante ressaltar que a explicação do PDP na Figura 5.4 corrobora a do SHAP, indicando que valores mais baixos de empréstimo impactam positivamente a chance de se conseguir um.

A Figura 5.8 apresenta uma explicação do LIME para a mesma instância anali-

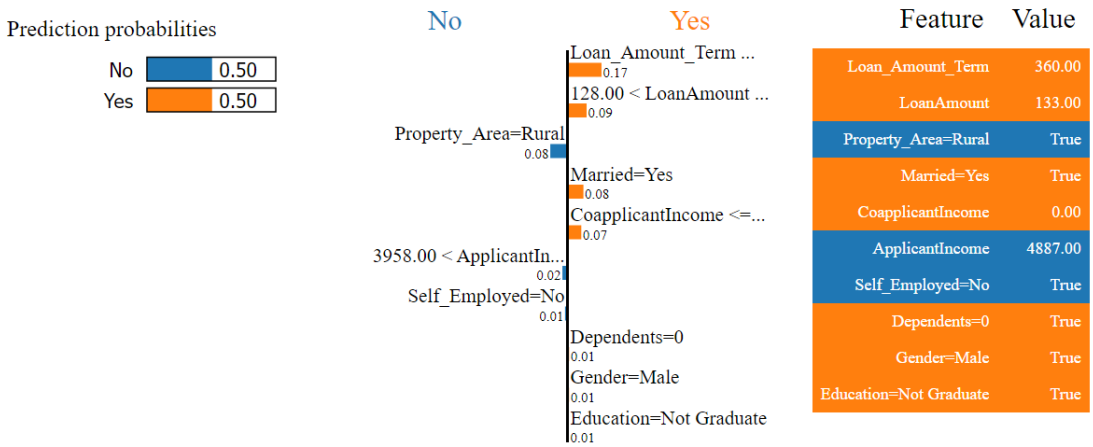

Figura 5.8. Explicação do LIME para previsão de aluguéis de bicicletas: a figura analisa a mesma instância previamente examinada pelo SHAP, mostrando uma discordância nos resultados. Enquanto o SHAP indica que o atributo `Married` (estado civil) tem impacto negativo, o LIME atribui uma contribuição positiva ao valor predito.

sada anteriormente pelo SHAP, na Figura 5.7. Nota-se uma certa concordância entre a importância atribuída pelos métodos: enquanto o LIME indica que o atributo *Married* contribui positivamente para o resultado, o SHAP apresenta uma perspectiva diferente. Autores como Molnar [2019]; Ferreira et al. [2020] apresentam evidências em seus trabalhos que a utilização de uma aproximação linear, mesmo que localmente, pode resultar em imprecisões nos resultados.

Os métodos LIME, SHAP e Captum podem ser aplicados a dados não estruturados, como imagens. Nesse contexto, é comum que o explicador destaque as regiões da imagem que mais contribuíram para a classificação, utilizando cores distintas para chamar a atenção do usuário. A Figura 5.9 apresenta uma explicação gerada para a classificação de raças de cães, destacando as áreas que ajudaram e as que prejudicaram a classificação.

Finalmente, é importante garantir que a transparência não comprometa a eficácia do algoritmo, expondo seu funcionamento ou vulnerabilidades não corrigidas.

5.5. Responsabilização (*Accountability*)

O conceito de responsabilização (do inglês, *Accountability*), é um princípio fundamental para o desenvolvimento e aplicação da IA responsável. Ele se refere à responsabilidade de indivíduos em proteger e controlar equipamentos, materiais de codificação e informações, assumindo a autoria por qualquer perda ou uso indevido desses recursos [Xia et al., 2024].

5.5.1. Definições

Apesar de o termo responsabilização ser bem definido no âmbito jurídico, significando a admissão de responsabilidade por ações, decisões e suas consequências [Thomson Reuters Practical Law, 2024], ainda existem desafios na sua definição no contexto específico de IA. De modo geral, responsabilização envolve agir com responsabilidade ética, trans-

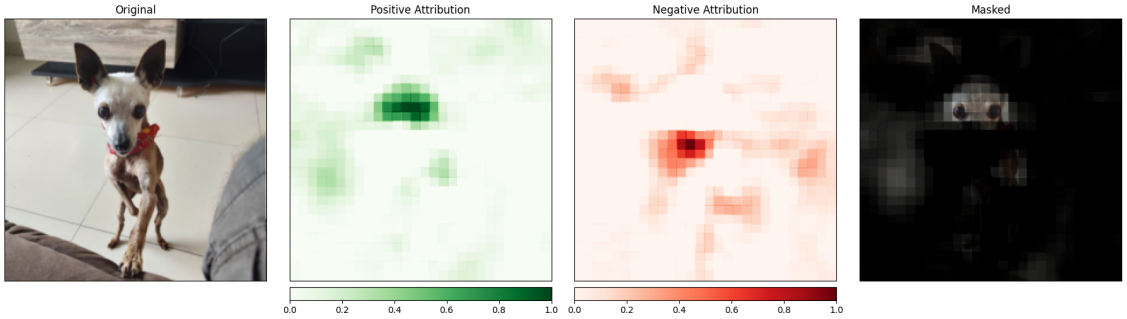

Figura 5.9. Explicação de modelos em dados não estruturados com Captum: a figura mostra como a ferramenta gera explicações para a classificação de raças de cães. As regiões da imagem que mais contribuíram para a decisão são destacadas, com pixels exibidos em cores distintas para facilitar a interpretação, ressaltando as áreas que ajudaram quanto as que prejudicaram a classificação.

parência nas ações, e prestar contas desses atos. Isso é essencial para assegurar que sistemas baseados em IA operem de forma justa, sem violar os valores éticos das pessoas impactadas. O termo responsabilização é frequentemente traduzido para expressões que abrangem suas diferentes dimensões. Um exemplo é a Lei Geral de Proteção de Dados (LGPD) no Brasil [Brasil, 2018], que utiliza os termos "Responsabilização e Prestação de Contas". Essa tradução destaca a importância da responsabilização conforme os requisitos legais.

Na literatura, o termo responsabilização é composto por três componentes principais que podem ser observados na Figura 5.10, a saber: (i) responsabilidade, (ii) auditoria e (iii) retificação [Xia et al., 2024], que são descritos a seguir:

- **Responsabilidade** refere-se à atribuição de responsabilidade por ações a indivíduos ou entidades, estabelecendo claramente quem é responsável. Uma vez estabelecido o responsável, é preciso entender sobre o que exatamente essa responsabilidade abrange.
- A **auditoria** é o processo de rastrear e documentar essas ações, permitindo que sejam revisadas e justificadas.
- Por fim, **retificação** ou **reparação** envolve corrigir ou compensar qualquer dano ou injustiça causada, garantindo que a responsabilidade seja efetivamente aplicada.

Compreender a relação de responsabilidade na responsabilização exige considerar os seguintes aspectos elencados em [Novelli et al., 2023]:

- o contexto (*para quê?*);
- o alcance das ações (*sobre o quê?*);

Figura 5.10. Três facetas da responsabilização.

- o agente responsável (*quem é o responsável?*);
- o fórum (*ao qual se deve prestar contas*);
- os padrões de avaliação (*de acordo com quais critérios?*);
- o processo de prestação de contas (*como é realizado?*), e;
- implicações das ações.

Para fins de ilustração, considere o seguinte cenário: uma solução baseada em IA para reconhecimento facial é utilizada por uma empresa no contexto de vigilância. A partir deste cenário, a Tabela 5.2 apresenta exemplo de cada um destes aspectos aplicados.

5.5.2. Métodos

A responsabilização de um algoritmo pode ser avaliada através da análise da relação de responsabilidade e da realização de auditorias que vão além do código-fonte, focando em aspectos como registros de treinamento e calibragem dos sistemas [Pinto, 2019]. Essa

Tabela 5.2. Exemplo de responsabilização no uso de IA para reconhecimento facial.

Contexto	Utilização de IA para reconhecimento facial por uma empresa
Alcance das ações	Processamento de dados biométricos dos funcionários ou clientes
Agente responsável	Empresa que implementa e gerencia a IA
Fórum	Autoridades de proteção de dados ou reguladores de privacidade
Padrões de avaliação	Regulamentos de privacidade e proteção de dados, como LGPD ou GDPR
Processo de prestação de contas	Relatórios periódicos, auditorias e revisões de conformidade
Implicações das ações	Possíveis violações de privacidade, discriminação ou erros no reconhecimento facial

Tabela 5.3. Catálogo de métricas em nível de sistema para responsabilização de IA.

Categoria	Subcategoria	Descrição
Responsabilidade	Supervisão de Sistemas de IA	Garantir que os sistemas sejam devidamente supervisionados.
	Sistemas de IA Competentes	Assegurar que o sistema apresente as funcionalidades propostas, sem ignorar princípios de IA responsável.
Auditoria	Supervisão Sistemática	Manter um sistema de registro e logs abrangentes.
	Compliance Checking	Realizar auditorias periódicas nos registros.
Retificação	Reparação por Design	Incorporar mecanismos no sistema para detecção e gerenciamento de problemas e erros.

abordagem permite uma compreensão mais profunda das questões levantadas no desenvolvimento de sistemas de IA, considerando, princípios como justiça, transparência, limitação de finalidade, minimização de danos, limitação de armazenamento, precisão e confidencialidade [Xia et al., 2024].

Os métodos de responsabilização incluem perguntas-chave, como: (i) "quem é responsável se o sistema A tomar a decisão B no contexto C?". A resposta a essas perguntas permite rastrear as ações de um sistema de IA, assegurando que processos adequados estejam em vigor para responder a qualquer falha ou uso indevido. A operacionalização da responsabilização envolve a manutenção de registros detalhados, documentação extensiva, análises de risco, além da implementação de ferramentas de vigilância e mecanismos de prestação de contas. Componentes novos do sistema devem ser validados para garantir conformidade com as regras de responsabilidade estabelecidas [Kroll, 2020].

Diante deste cenário, pesquisadores têm buscado formas de operacionalizar o conceito de responsabilização usando diferentes métricas como o número de auditorias realizadas, a rastreabilidade dos processos decisórios, o nível de conformidade com padrões regulatórios, a transparência dos dados e modelos utilizados, e a eficácia das medidas corretivas adotadas após falhas detectadas [Novelli et al., 2023]. Inicialmente, muitas abordagens adotam uma perspectiva binária, onde o sistema atende ou não a certos requisitos [Xia et al., 2024]. Para supervisionar sistemas de IA, podem-se definir papéis e responsabilidades claras na organização, documentando essas funções em relatórios de avaliação e manuais de processos. Dessa forma, é possível quantificar o conceito nos sistemas ao definir requisitos não funcionais que devem ser atendidos, tais como os relacionados na Tabela 5.3.

Para cada requisito não funcional, é possível gerar métricas de processos e produtos que garantam sua implementação adequada [OECD, 2023]. Por exemplo, na supervisão de soluções baseadas em IA, podem existir documentos internos às organizações que definem as pessoas responsáveis pela supervisão e seus papéis. Isso pode ser cobrado por meio de relatórios de avaliação, apresentações do comitê de supervisão, manuais de

processos e assegurado por ações governamentais. A operacionalização do conceito, portanto, se dá pela existência ou não de mecanismos que garantam a implementação e seguimento dos princípios estabelecidos [Xia et al., 2024].

No entanto, apesar das metodologias promissoras, ainda existem lacunas significativas na pesquisa. Faltam diretrizes específicas para Relatórios de Impacto de Proteção de Dados (RIPDs) em diversas tecnologias (e.g., reconhecimento facial). É crucial identificar claramente os agentes responsáveis pela responsabilização em sistemas baseados em IA, especialmente em aplicações complexas, como uma empresa que implementa reconhecimento facial para segurança em eventos. Deve-se definir se a responsabilidade pela proteção de dados cabe ao desenvolvedor da IA, à empresa que a implementa ou ao operador que lida diretamente com o sistema. Outro desafio é garantir que a responsabilização seja mantida ao longo de todo o ciclo de vida do sistema, desde o desenvolvimento até a implantação e manutenção. A falta de métricas robustas para avaliar a responsabilização também representa um obstáculo considerável.

Finalmente, a responsabilização deve ser considerada não apenas sob uma perspectiva técnica, mas também em termos de governança organizacional e legal. Integrar a responsabilização em um arcabouço mais amplo de governança de IA é essencial para garantir que os princípios de responsabilização e prestação de contas sejam efetivamente implementados.

5.6. Segurança e Privacidade (*Safety and Privacy*)

No estudo sobre IAR, a privacidade e a segurança também emergem como pilares fundamentais. Com o avanço do uso de IA em sistemas de mídias digitais, foram criadas diversas leis para proteger os direitos e as liberdades dos cidadãos [Banciu and Cîrnu, 2022]. No entanto, há desafios na definição, medição e gestão da privacidade no contexto dos algoritmos, uma vez que ela pode ser violada em diversos momentos da aplicação, desde sua concepção até sua utilização.

5.6.1. Definições

O conceito de privacidade é amplamente discutido, especialmente no Direito [de Miranda and Cavalcanti, 2013; Amaral, 2018]. No entanto, no contexto de Ciência da Computação, especificamente IA, esse conceito ainda não está plenamente consolidado, tanto em termos de definição quanto de medição. Existem diferentes definições geradas por grandes instituições, órgãos governamentais e empresas. A seguir, são apresentadas algumas dessas definições, adotadas no contexto deste capítulo.

- **Privacidade**: liberdade de intrusão na vida ou nos assuntos privados de um indivíduo quando essa intrusão resulta da coleta e uso indevido ou ilegal de dados sobre esse indivíduo[16] [Garfinkel et al., 2023];

- **Privacidade ética**: um conjunto contextual de valores relacionados à privacidade e à satisfação de uma estrutura de expectativas, incluindo a preservação da autonomia,

[16]Tradução livre de: *"Freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual"*.

autodeterminação e comunidades, autosselecionadas[17] [Patricia and Ali, 2022];

Para o conceito de segurança em IA, destacamos a definição do *National Institute of Standards and Technology* (NIST) [Ross and Pillitteri, 2024]:

- **Segurança**: condição resultante do estabelecimento e manutenção de medidas de proteção que permitem a uma organização desempenhar suas funções críticas, apesar dos riscos representados por ameaças aos seus sistemas. Essas medidas podem envolver uma combinação de dissuasão, prevenção, prevenção, detecção, recuperação e correção, integradas na abordagem de gestão de riscos da organização[18].

Os conceitos de segurança formulados pela definição providenciada pelo NIST, como, por exemplo, a ideia de que segurança é uma condição de funcionamento de empresas e sistemas sob medidas de proteção, são essenciais para traçar estratégias e para que se atente à importância da formulação de estratégias para garantir que os sistemas de IA desempenhe suas funções ancorado em estratégias que mitiguem danos, envolvendo medidas preventivas ou reativas para garantir a integridade, confidencialidade e disponibilidade deses sistemas mesmo com os ricos associados à operação deles.

5.6.2. Métodos

Medir e determinar se um algoritmo preserva a privacidade dos dados aos quais foi exposto é uma tarefa de grande importância atualmente. Isso se deve à ampla presença de algoritmos de IA no cotidiano e aos possíveis ataques que podem expor dados sensíveis de seu treinamento [Smith et al., 2024]. Com foco nos grandes modelos, descreveremos a seguir um método [Abadi et al., 2016] para identificar a robustez do modelo em relação à proteção da privacidade.

O método em questão, que visa prevenir a exposição dos dados de treinamento por parte do modelo, é o Gradiente Descendente Estocástico com Privacidade Diferencial (DP-SGD, em inglês). O DP-SGD modifica o algoritmo padrão de Gradiente Descendente Estocástico (GDE) para incorporar a Privacidade Diferencial(PD), reduzindo o risco de exposição de dados sensíveis durante o treinamento [Privacy, 2021; Abadi et al., 2016]. A PD introduz parâmetros que limitam a probabilidade de exposição de dados e mudanças no comportamento do modelo [Abadi et al., 2016]. Maiores detalhes sobre o cálculo desses valores podem ser encontrados em [Mironov et al., 2019]. A seguir, apresentamos uma explicação do GDE com o uso da PD, incluindo uma breve descrição do algoritmo, a fim de elucidar seu entendimento e aplicação neste contexto.

Para começar, descrevemos o GDE. Este é um algoritmo de otimização utilizado no aprendizado de máquina para encontrar os parâmetros ótimos de um modelo. O GDE

[17]Tradução livre de: *"A contextual set of values pertaining to privacy and the satisfaction of a framework of expectations (preservation of autonomy, self-determination, and self-selected communities/locum and intimacies)"*.

[18]Tradução livre de: *"A condition that results from the establishment and maintenance of protective measures that enable an organization to perform its mission or critical functions despite risks posed by threats to its use of systems. Protective measures may involve a combination of deterrence, avoidance, prevention, detection, recovery, and correction that should form part of the organization's risk management approach"*.

Algorithm 1 GDE com Privacidade Diferencial (DP-SGD)

Entrada: Exemplos $\{x_1, \ldots, x_N\}$, função de perda $\mathscr{L}(\theta) = \frac{1}{N}\sum_i \mathscr{L}(\theta, x_i)$
Parâmetros: Taxa de aprendizado η_t, escala de ruído σ, tamanho do grupo L, limite da norma do gradiente C
Inicializar θ_0 aleatoriamente
for $t = 1$ até T **do**
    Selecionar uma amostra aleatória L_t com probabilidade de amostragem L/N
    Calcular gradiente:
    for cada $i \in L_t$ **do**
        Calcular $g_t(x_i) \leftarrow \nabla_\theta \mathscr{L}(\theta_t, x_i)$
    end for
    Limitar gradiente:
    for cada $i \in L_t$ **do**
        $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$
    end for
    Adicionar ruído:
    $\tilde{g}_t \leftarrow \frac{1}{L}(\sum_i \bar{g}_t(x_i)) + \mathscr{N}(0, \sigma^2 C^2 \mathbb{I})$
    Descida(Atualização):
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}_t$
end for
Retornar θ_T e calcular o custo total de privacidade (ε, δ) utilizando um método de contabilidade de privacidade

atualiza os parâmetros do modelo iterativamente com base em amostras dos dados de treino [Abadi et al., 2016]. Em seguida, abordamos a PD mais detalhadamente. A PD adiciona ruído aos resultados do modelo, sejam eles finais ou intermediários, a fim de que os dados de treinamento não possam ser inferidos com grande precisão [Dwork, 2008].

No algoritmo que estamos tratando o ruído é adicionado ao gradiente durante o treinamento a fim de garantir a privacidade. Isso faz com que o modelo treinado usando este algoritmo seja diferencialmente privado. Assim, mesmo que se tenha acesso ao modelo final, não será possível identificar informações específicas dos dados usados no treinamento.

O Algoritmo 1 [Abadi et al., 2016] apresenta os dados de entrada, a função de perda e alguns parâmetros importantes, como a taxa de aprendizado η_t, a escala de ruído σ, o tamanho do lote L (número de amostras usadas a cada iteração) e o limite da norma do gradiente C. O algoritmo começa inicializando os parâmetros θ_0 aleatoriamente. Em cada iteração, é selecionado um subconjunto dos dados com probabilidade L/N. Para cada dado do lote, calcula-se o gradiente da função de perda e ajusta-se o gradiente para que sua norma não exceda C. Em seguida, adiciona-se ruído ao gradiente médio, garantindo a privacidade diferencial. Por fim, atualizam-se os parâmetros do modelo. Ao final, o algoritmo retorna os parâmetros finais θ_T e calcula o custo total de privacidade (ε, δ), que quantifica o nível de privacidade diferencial garantido.

Assim, mostramos que o algoritmo DP-SGD que é uma importante ferramenta na prevenção da privacidade de modelos de IA. As fontes citadas detalham o uso com-

pleto do algoritmo para garantir a privacidade, bem como os cálculos e parâmetros aqui apresentados aqui brevemente, a fim de favorecer a didática.

Outra abordagem reconhecida e divulgada que visa aumentar o nível de privacidade de modelos de IA é chamada de Sanitização dos dados de treinamento[Smith et al., 2024][19]. Esse método visa remover informações sensíveis dos dados de treinamento, sendo muito utilizado no contexto da saúde para eliminar dados de identificação pessoal. No contexto de grandes modelos, porém, a identificação desses dados sensíveis pode ser difícil e trabalhosa. Assim, a sanitização passa a ser feita majoritariamente de maneira automática, através de padrões nos dados que para isso precisam seguir certo contexto. Embora seja uma maneira efetiva para evitar o vazamento de informações pessoais, esse método pode não ser aplicável em todos os contextos devido à diversidade e complexidade dos dados de treinamento.

Concluímos que os mecanismos de segurança dos dados que citamos, como o DP-SGD e a sanitização dos dados de treinamento, têm a finalidade de garantir que o algoritmo mantenha a privacidade dos indivíduos cujos dados foram usados no treino. Isso contribui para a prevenção de danos que podem ser causados pelo sistema de IA, em relação à violação de privacidade.

Além disso, os mecanismos de prevenção de danos já incluídos nas práticas de segurança, como a detecção e correção de erros, são cruciais. A detecção rápida de falsos positivos permite ações corretivas imediatas, minimizando os danos potenciais. A correção desses erros garante que o sistema se adapte e melhore continuamente, prevenindo futuros incidentes similares. Erros em sistemas de reconhecimento facial, por exemplo, podem levar a casos de prisões indevidas[20][21] evidenciando a necessidade de medidas robustas de segurança.

Em relação a isso, o Benchmarking Test Empresa Gryfo [Gryfo, 2021] propõe métricas relevantes para avaliar a eficácia da segurança e privacidade em sistemas de reconhecimento facial:

- **Confiança**: taxa de confiança da biometria facial, indicando a probabilidade de que o reconhecimento esteja correto;
- **Cobertura**: métrica que indica o alcance do reconhecimento facial, relevando a quantidade de rostos reconhecidos versus não reconhecidos;
- **Erro**: taxa de erro no reconhecimento facial;
- **Precisão**: proporção entre os verdadeiros positivos e o total das faces classificadas como positivas para reconhecimento.

Assim, vale notar que a avaliação de tais métricas e a apresentação e adequação delas para níveis aceitáveis dentro cada aplicação é imprescindível para um uso social

[19]Tradução livre de "*Sanitizing training data*".

[20]https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig

[21]https://arstechnica.com/tech-policy/2023/01/facial-recognition-error-led-to-wrongful-arrest-of-black-man-report-says/

seguro desses sistemas de IA tornando os mais seguros para um uso real em importantes tarefas no ciclo social.

No contexto brasileiro, a Lei Geral de Proteção de Dados Pessoais (LGPD), sancionada pela Lei Nº 13.853 de 8 de julho de 2019, e a Autoridade Nacional de Proteção de Dados [Brasil, 2018, 2019] têm sido fundamentais para a proteção dos dados pessoais. Essas normas estabelecem critérios claros para a coleta e manipulação de dados pessoais, garantindo respeito à privacidade e a definição clara das responsabilidades.

5.7. Governança (*Governance*)

Os pilares para IA responsável apresentados nas seções anteriores são um ponto de partida, mas eles só fornecerão o que as instituições precisam se forem combinados com práticas de governança que efetivamente ajudem a orientar o desenvolvimento e o uso de aplicativos de IA. Em resumo, a governança refere-se aos processos e estruturas utilizados para dirigir e governar organizações, sistemas sociais e tecnológicos. Em computação, a governança pode ser aplicada de várias formas, como a governança de algoritmos [Doneda and Almeida, 2016] e a governança de IA [Dafoe, 2018]. Especificamente no contexto de IA, a governança abrange a regulamentação e a supervisão dos algoritmos e das tecnologias associadas para assegurar que sua implementação seja ética e conforme as normas estabelecidas [Gasser and Almeida, 2017; Mäntymäki et al., 2022].

5.7.1. Definições e Principais Aspectos

Com o aumento da complexidade e ubiquidade dos algoritmos, especialmente em IA, as decisões geradas por esses sistemas podem ter impactos significativos e amplos. Neste contexto, a opacidade desses algoritmos representa um desafio considerável, dificultando a explicação de suas decisões e levantando preocupações sobre direitos autorais, justiça, privacidade e segurança [Gillespie, 2014; Katzenbach and Ulbricht, 2019].

A governança em IA visa coordenar diferentes *stakeholders* para mitigar as consequências negativas do uso das tecnologias emergentes, ao mesmo tempo, em que mantém sua eficácia. Assim, é essencial para garantir os princípios da IA responsável, como justiça, responsabilização, transparência, segurança e privacidade, sejam respeitados e que os sistemas de IA estejam alinhados com os interesses da sociedade [Gasser and Almeida, 2017; Mäntymäki et al., 2022].

Os principais aspectos da governança incluem: (i) Governança de Algoritmos, e (ii) Governança de IA, conforme apresentado na Figura 5.11.

- **Governança de Algoritmos**: envolve a regulamentação e controle dos algoritmos utilizados em diversos sistemas, para assegurar que operem de forma justa e transparente. Isso inclui a gestão de questões como viés, privacidade e segurança, além da manutenção da explicabilidade dos modelos [Doneda and Almeida, 2016];

- **Governança de IA**: abrange a supervisão e regulação das tecnologias de IA em geral, desde a formulação de políticas até a implementação de técnicas e *frameworks* que garantam a conformidade com os princípios de justiça, responsabilidade e segurança [Dafoe, 2018].

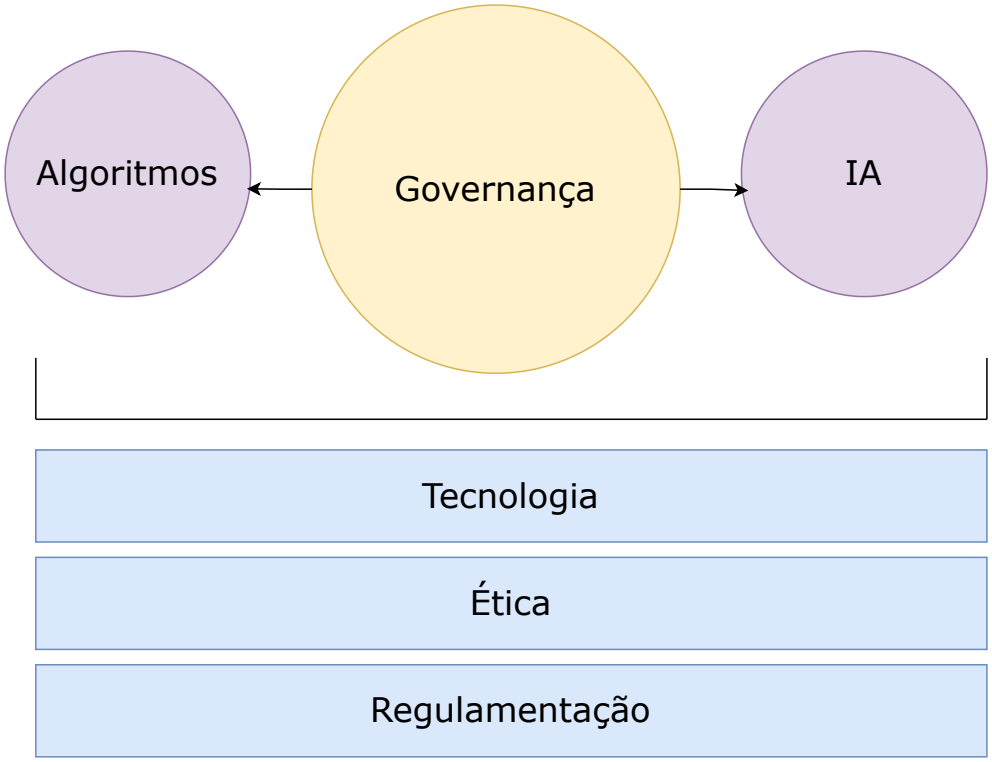

Figura 5.11. Principais aspectos da governança.

5.7.2. Métodos

Os métodos de governança de algoritmos e IA incluem, aspectos também representados na Figura 5.11, tais como:

- **Tecnologia**: envolve o desenvolvimento de algoritmos com considerações éticas integradas desde o design, incluindo a conformidade com padrões estabelecidos e medidas de segurança para mitigar vieses e desigualdades fundamentais para garantir que os algoritmos atendam aos critérios desejados. Também envolve a pesquisa de ferramentas de auditoria algorítmica que escalam com o tamanho e complexidade dos serviços baseados em IA [Almeida et al., 2024];

- **Ética**: orienta a aplicação de IA com base em princípios éticos e direitos humanos, frequentemente adotados por meio de autorregulação, como as diretrizes da IEEE para sistemas autônomos e inteligentes [Chatila and Havens, 2019];

- **Regulamentação**: refere-se à criação e aplicação de leis e normas que regulam o uso de IA. Exemplos incluem a Lei Geral de Proteção de Dados (LGPD) [Brasil, 2018], o *AI Act* proposto pela União Europeia [Edwards, 2021] e o *Model Artificial Intelligence Governance Framework* de Singapura [Personal Data Protection Commission et al., 2020].

Além desses métodos, as auditorias de algoritmos desempenham um papel crucial na governança de IA. Elas avaliam se os algoritmos estão em conformidade com normas e padrões estabelecidos. Existem diferentes tipos de auditoria, como a auditoria *black-box*, que analisa apenas entradas e saídas geradas por uma solução baseada em IA sem acesso ao código-fonte ou ao modelo em si. Por exemplo, uma auditoria pode avaliar se um algoritmo atende ao princípio de minimização de dados da GDPR [European Union, 2016] usando apenas consultas [Netherlands Court of Audit, 2022].

Ademais, auditorias podem verificar se os algoritmos utilizam apenas os dados necessários e identificar possíveis vieses nos dados de treinamento e nas saídas dos modelos [Clavell et al., 2020; Lunter, 2020]. É crucial interpretar os resultados das auditorias com cautela para evitar um falso senso de segurança e progresso, que pode não refletir a realidade [Raji et al., 2020a].

5.8. Iniciativas de Governança de IA e Proteção de Dados no Brasil e no Mundo

Recentes iniciativas no Brasil têm fortalecido o desenvolvimento responsável de IA, alinhadas às diretrizes internacionais e à crescente demanda por tecnologias éticas e inclusivas. A seguir, são apresentadas algumas iniciativas importantes:

- **LGPD (Lei Geral de Proteção dos Dados)**: essa lei visa proteger a privacidade e a liberdade individual no tratamento de dados pessoais, tanto em ambientes físicos, quanto digitais. Ela regula a coleta, o armazenamento e o uso dessas informações exigindo o consentimento do titular. A lei define dados sensíveis como aqueles que exigem proteção adicional, incluindo origem racial ou étnica, religião e dados de saúde. Além disso, assegura direitos como acesso, correção e exclusão de dados, garantindo transparência no uso dessas informações [Brasil, 2018];

- **EBIA (Estratégia Brasileira de Inteligência Artificial)**: a EBIA visa orientar o uso responsável da IA no Brasil. Ela visa definir diretrizes para governança de IA, e focar na eliminação de vieses, a curadoria dos dados utilizados, responsabilização (*accountability*) para garantir uma IA confiável [Brasil, 2021];

- **Projeto de Lei 2338/2023 ou Marco Legal da IA**: visa regulamentar o uso e desenvolvimento de tecnologias de IA no Brasil. O PL estabelece normas para garantir a ética, transparência e segurança no uso da IA, alinhadas à LGPD. Exige sistemas de IA transparentes e explicáveis, com foco na minimização de vieses e discriminação, e promove responsabilidade e governança para um desenvolvimento sustentável e equitativo [Brasil, 2023];

- **PBIA (Plano Brasileiro de Inteligência Artificial)**: Esse plano prevê investimentos em diferentes áreas que envolvem Inteligência Artificial que totalizem R$ 23,03 bilhões. Os principais alvos são infraestrutura e desenvolvimento de IA, IA para inovação empresarial e para melhorias nos serviços públicos. Com isso o governo brasileiro procura promover o desenvolvimento e o uso de IA nos principais setores econômicos e sociais do país [Governo Federal et al., 2024].

As principais iniciativas internacionais sobre governança de IA incluem:

- **Declaração Ministerial do G20 (2024)**: destaca a importância da cooperação internacional, capacitação e compartilhamento de tecnologia para promover práticas de IA transparentes. A declaração se baseia nas recomendações da UNESCO sobre Ética da IA [G20, 2024].
- **Recomendação da UNESCO sobre a Ética da Inteligência Artificial**: estabelece valores para promover a responsabilidade na IA, enfatizando direitos humanos, dignidade e sustentabilidade ambiental. Destaca princípios de transparência, responsabilidade, aborda governança de dados e a igualdade de gênero, e inclui diretrizes para monitoramento e avaliação para garantir a implementação efetiva e o impacto ético das práticas de IA [UNESCO, 2021].
- **AI Act**: regula o uso da IA na União Europeia com uma abordagem baseada em risco, estabelecendo regras diferenciadas conforme o nível de risco. Proíbe práticas de IA inaceitáveis e impõe requisitos de transparência e gerenciamento de riscos para outras. Define obrigações para modelos de IA de uso geral e prevê penalidades substanciais por não conformidade, visando um desenvolvimento responsável da IA que garanta segurança, proteção de direitos e a confiança na tecnologia [European Union, 2024].
- **GPAI (Global Partnership on Artificial Intelligence)**: é uma iniciativa que promove a colaboração global para o desenvolvimento responsável da IA. A GPAI reúne governos, empresas, academia e sociedade civil para criar diretrizes e melhores práticas em IA, focando em ética, transparência e inclusão. Apoia projetos, fomenta pesquisa e estabelece grupos de trabalho em áreas-chave [GPAI, Acesso em: 19/09/2024].
- **Declaração de Toronto**: estabelece princípios para assegurar que a IA respeite os direitos à igualdade e não discriminação. Ela amplia normas internacionais de direitos humanos no desenvolvimento e uso de IA, buscando proteger direitos fundamentais e prevenir preconceitos. A declaração destaca transparência, responsabilidade e a participação de diversas partes interessadas visando influenciar políticas e práticas para promover justiça social e valores éticos fundamentais na IA [Toronto Declaration, 2018].

As leis atuais de regulamentação de dados, como a GDPR (*General Data Protection Regulation*) [European Union, 2016] e a LGPD [Brasil, 2018], enfrentam desafios ao lidar com o contexto das IAs generativas, como o ChatGPT[22]. Essas leis foram criadas antes do surgimento dessas tecnologias e têm dificuldades para gerenciar com o grande volume de dados necessário para treinar esses modelos com a proteção da privacidade dos usuários [Zhang et al., 2024]. A crescente popularidade do ChatGPT levou a revisões no rascunho do *AI Act* da União Européia, evidenciando a dificuldade dos reguladores em criar normas que abordem adequadamente os riscos associados a tecnologias emergentes.

[22]`https://chatgpt.com/`

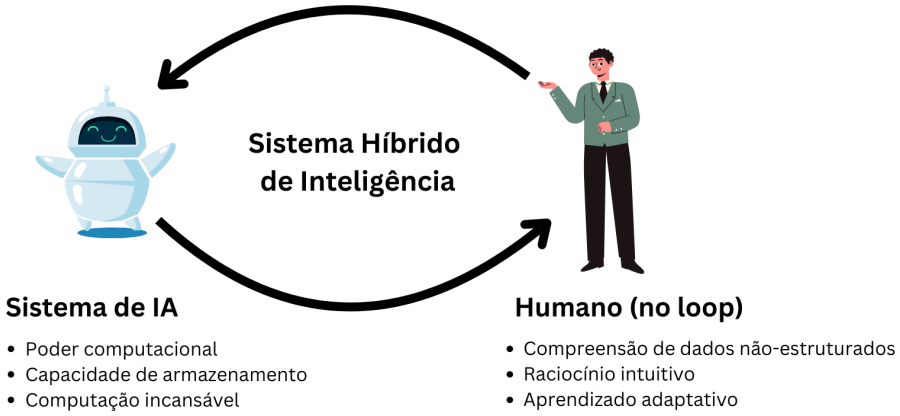

Figura 5.12. Sistema de inteligência híbrido onde a inteligência humana e a artificial se complementam.

Embora o Parlamento Europeu não tenha classificado o ChatGPT como um sistema de alto risco, os reguladores ainda enfrentam o desafio de ajustar suas abordagens para lidar as diversas implicações dessas tecnologias. Isso levanta questões cruciais sobre como as experiências anteriores com a GDPR podem auxiliar na formulação de políticas para novos riscos tecnológicos e como as instituições regulatórias precisam evoluir para enfrentar o impacto transformador dessas inovações [Wolff et al., 2024].

5.9. A Importância de Humanos no Processo (*Human-in-the-Loop*)

A participação humana é essencial em todas as etapas do ciclo de vida de sistemas de IA, desde o projeto até a implantação. Embora, as tecnologias de IA tenham avançado, elas ainda não conseguem interpretar de completamente os contextos sociais, culturais e emocionais em que operam. Além disso, como os modelos são treinados em dados históricos, podem herdar e até amplificar os vieses presentes nesses dados.

Por isso, a intervenção humana (*human-in-the-loop*) é é fundamental, especialmente em decisões de alto impacto, como questões de equidade, onde diferentes subgrupos podem exigir tratamentos diferenciados. Por exemplo, quando os valores de corte de equidade variam entre grupos, os humanos podem ajustar esses parâmetros, calibrando o sistema para minimizar disparidades injustas. Sem essa supervisão, há um risco maior de vieses e discriminação, que podem ser intensificados por algoritmos incapazes de compreender as complexidades da condição humana.

Neste contexto, estudos anteriores de pesquisadores e profissionais destacam que soluções baseadas em IA podem se tornar mais inclusivas com a criação de um "Sistema Híbrido de Inteligência" [Rai et al., 2019], conforme ilustrado na Figura 5.12. Além disso, é essencial desenvolver sistemas de *IA Ambient Intelligent* (AmI) para ampliar a colaboração entre humanos e IA [Gams et al., 2019]. Nesses ambientes, a IA interage com os humanos, recebendo informações e aprendendo tanto com eles quanto com o ambiente, em um processo contínuo de *feedback* colaborativo [Ramos et al., 2008]. Esse avanço transforma as tradicionais "caixas pretas" da IA em "caixas de vidro" [Rai, 2020], criando aplicações que incorporam características explicáveis (XAI), uma transparência

essencial para reduzir preconceitos e promover a inclusão.

Em cenários onde a IA pode impactar a vida das pessoas — como na saúde, na segurança pública e na justiça — soluções de IA supervisionadas por humanos ajudam a garantir que haja um controle efetivo sobre os resultados. Isso não apenas promove confiança nas tecnologias, mas também permite que as organizações aprendam com os erros e façam ajustes necessários, minimizando danos potenciais. Portanto, a colaboração entre humanos (i.e., inteligência humana) e máquinas (i.e., IA) é essencial para criar um futuro em que a IA beneficie a sociedade de maneira justa.

5.10. Desafios e Oportunidades de Pesquisa

Desenvolver e operacionalizar sistemas de IA de forma responsável é um desafio complexo, uma vez que os conceitos, como justiça, transparência e privacidade possuem dimensões subjetivas. Operacionalizá-los e quantificá-los de maneira eficaz, considerando diferentes contextos e legislações, é um desafio adicional, especialmente em aplicações globais. A seguir, estão destacados alguns dos principais desafios e oportunidades de pesquisa nessa área:

- **Criação de Métricas**: a criação e implementação de métricas eficazes para a governança de IA é um desafio multifacetado. Um dos principais obstáculos é garantir a disponibilidade e qualidade dos dados, pois dados imprecisos ou incompletos podem comprometer a eficácia do programa de governança e levar a resultados não confiáveis. Além disso, as métricas devem equilibrar avaliações quantitativas e qualitativas, que podem ser influenciadas por viés subjetivo, ou experiências pessoais. Para minimizar isso, é essencial usar procedimentos padronizados e garantir a representação diversificada dos *stakeholders* [Selbst et al., 2019; Raji et al., 2020b].

- **Integração e Monitoramento Contínuo**: a integração das métricas de governança com os arcabouços (i.e., *frameworks*) e processos operacionais é crucial. Isso envolve alinhar as métricas com as políticas de governança, definir papéis e responsabilidades claros, criar processos padronizados para coleta e validação de dados, e promover a colaboração entre diferentes equipes. A revisão e adaptação contínua das métricas são essenciais para garantir sua eficácia e relevância ao longo do tempo [Minkkinen et al., 2022; Mäntymäki et al., 2022].

- **Plataformas de Auditoria**: desenvolver plataformas de auditoria que considerem diferentes contextos e regras de governança de diversos países também é um desafio significativo. Essas plataformas precisam ser capazes de verificar e adaptar métricas conforme as nuances regulatórias e culturais de cada contexto. A criação de ferramentas e tecnologias que suportem essas necessidades pode oferecer um suporte valioso para a governança global de IA [Mökander, 2023].

- **Projeto da Aplicação de IA**: incorporar práticas de IA responsável desde a concepção do sistema é um desafio fundamental. Isso envolve considerar aspectos éticos e governança durante a engenharia de software e o desenvolvimento do sistema. A integração de princípios de responsabilidade desde o início pode ajudar a evitar

problemas e garantir que as soluções de IA sejam projetadas com a devida consideração de suas implicações sociais e éticas [McLennan et al., 2022; Kijewski et al., 2024].

Por fim, enfrentar esses desafios é essencial para estabelecer uma governança robusta e confiável de IA. A criação de métricas bem definidas e a contínua revisão e adaptação das práticas de governança permitem assegurar a conformidade com requisitos internos e externos e fomentar a confiança nas aplicações de IA.

5.11. Considerações Éticas

Assim como explorado ao longo desse capítulo, o uso ético de sistemas de IA é permeado por diversos fatores e tem como base princípios. Estes princípios foram delimitados aqui como: justiça, transparência, responsabilização, segurança e privacidade e governança.

Nesse contexto, para que se atinja um uso ético de um sistema de IA é necessário que ele seja justo, sendo imprescindível que se tenha transparência no processo de desenvolvimento e utilização de sistemas baseados em IA. Tal transparência se traduz, na maior parte dos casos, em um alto grau de explicabilidade em relação ao sistema [Rudin, 2019], uma vez que, grande parte dos modelos que a população tem contato são opacos aos usuários.

Ainda, mesmo que a justiça e a transparência sejam, em grande parte, atingidas, deve se pensar, em conjunto a elas, o princípio da responsabilização. Isso se dá pois, caso ocorram falhas e as decisões tomadas contenham viés, ou ainda, ocorram violações de privacidade e de segurança do sistema, a responsabilização por eventuais danos é uma questão de grande relevância, principalmente no âmbito jurídico e social.

Por fim, conforme mencionado na Seção 5.7, é necessário combinar os pilares ou princípios mencionados com técnicas de governança que guiem sua criação e utilização, para que tudo o que foi citado acima seja aplicado na prática, e para que os sistemas de IA sejam integrados na sociedade de maneira efetiva e ética, suprindo o que as instituições precisam. Visando assim, com a integração de todos esses princípios, atingir um sistema que possa atender de maneira satisfatória princípios de IAR.

5.12. Conclusão

Nesse capítulo foram abordados os principais conceitos e práticas fundamentais da Inteligência Artificial Responsável (IAR). Espera-se que os leitores tenham adquirido uma visão completa das ferramentas e estratégias necessárias para criar soluções de IA que sejam justas, transparentes e eticamente sólidas. A implementação de IA responsável demanda, além de competência técnica, um compromisso com princípios éticos que consideram o impacto social dessas tecnologias. Ressalta-se, ainda, a relevância de incorporar esses valores desde as fases iniciais do desenvolvimento de sistemas, assegurando que as soluções sejam construídas com uma base ética e de governança robusta, prevenindo potenciais desafios futuros e promovendo uma IA mais inclusiva e confiável.

5.12.1. Material Adicional

O material complementar deste capítulo, incluindo códigos de exemplo bem como repositório de dados, está disponível em um repositório no GitHub, permitindo a aplicação prática dos conceitos apresentados: `https://github.com/marceloslo/responsible_ai_tutorial`.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Ministério Público de Minas Gerais (MPMG), projeto Capacidades Analíticas, pelo projeto CIIA-Saúde, pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In Conference on Computer and Communications Security, 2016.

B. Abdollahi and O. Nasraoui. Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems, pages 21–35. Springer International Publishing, Cham, 2018.

V. Almeida, J. M. Almeida, and W. Meira. The role of computer science in responsible ai governance. IEEE Internet Computing, 28(3):55–58, 2024.

F. Amaral. Direito civil - 10ª edição. Saraiva Jur, 2018.

H. Anahideh, N. Nezami, and A. Asudeh. On the Choice of Fairness: Finding Representative Fairness Metrics for a given context. arXiv preprint arXiv:2109.05697, pages 1–25, 2021.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016.

D. Apley and J. Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv preprint arXiv:1612.08468, 2019.

D. Banciu and C. Cîrnu. AI Ethics and Data Privacy Compliance. In Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2022.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58:82 – 115, 2020.

A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In Conference on AI, Ethics, and Society (AIES), 2019.

Brasil. Lei nº 13.709, de 14 de agosto de 2018. Diário Oficial [da] República Federativa do Brasil, 2018.

Brasil. Lei nº 13.853, de 8 de julho de 2019. Diário Oficial [da] República Federativa do Brasil, 2019.

Brasil. Estratégia Brasileira de Inteligência Artificial - EBIA. Ministério da Ciência, Tecnologia e Inovações Secretaria de Empreendedorismo e Inovação, 2021.

Brasil. Projeto de lei nº 2338, de 2023. https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1726246471835&disposition=inline, 2023. Acesso em: 19/09/2024.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency, 2018.

F. Calmon, D. Wei, B. Vinzamuri, K. Ramamurthy, and K. Varshney. Optimized pre-processing for discrimination prevention. In Conference on Neural Information Processing Systems, 2017.

R. Chatila and J. Havens. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Robotics and well-being, 95:11–16, 2019.

D. Citron. Technological due process. Washington University Law Review, 85:1249, 2007.

G. Clavell, M. Zamorano, C. Castillo, O. Smith, and A. Matic. Auditing algorithms: On lessons learned and the risks of data minimization. In Conference on AI, Ethics, and Society (AIES), pages 265–271, 2020.

W. Colombelli. Regulamentação da IA (Inteligência Artificial) na administração pública brasileira: análise do Projeto de Lei n° 21 de 2020 e Projeto de Lei n° 2338 de 2023. B.S. thesis, 2024.

A. Dafoe. AI Governance: a Research Agenda. Governance of AI Program, Future of Humanity Institute, 1442:1443, 2018.

P. de Miranda and F. Cavalcanti. Tratado de Direito Privado. Editora Revista dos Tribunais Ltda, 2013.

V. Dignum. Responsible Autonomy. In Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2017.

V. Dignum. Responsibility and artificial intelligence. The oxford handbook of ethics of AI, 4698:215, 2020.

D. Doneda and V. A. Almeida. What is algorithm governance? IEEE Internet Computing, 20(4):60–63, 2016.

D. Doneda, V. A. Almeida, and F. Bruno. O que é a governança de algoritmos. Tecnopolíticas da vigilância: Perspectivas da margem, pages 141–148, 2018.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

C. Dwork. Differential Privacy: A Survey of Results. In Theory and Applications of Models of Computation, 2008.

L. Edwards. The EU AI Act: a Summary of its Significance and Scope. Artificial Intelligence (the EU AI Act), 1, 2021.

J.-H. Eun and S.-S. Hwang. An exploratory study on policy decision making with artificial intelligence: Applying problem structuring typology on success and failure cases. Informatization Policy, 27(4):47–66, 2020.

European Union. General data protection regulation. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679`, 2016. Acesso em: 19/09/2024.

European Union. Artificial intelligence act (eu ai act). `https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32024R1689`, 2024. Acesso em: 19/09/2024.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and Removing Disparate Impact. In Conference on Knowledge Discovery and Data Mining (SIGKDD), 2015.

E. Ferrara. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, 6(1):3, dec 2023.

L. Ferreira, F. Guimarães, and R. Silva. Applying Genetic Programming to Improve Interpretability in Machine Learning Models. In Congress on Evolutionary Computation, 2020.

G20. G20 Ministerial Declaration. `https://www.gov.uk/government/publications/g20-ministerial-declaration-maceio-13-september-2024/g20-ministerial-declaration-13-september-2024`, 2024. Acesso em: 19/09/2024.

M. Gams, I. Y.-H. Gu, A. Härmä, A. Muñoz, and V. Tam. Artificial Intelligence and Ambient Intelligence. Journal of Ambient Intelligence and Smart Environments, 11 (1):71–86, 2019.

S. Garfinkel, J. Near, A. Dajani, P. Singer, and B. Guttman. De-Identifying Government Datasets: Techniques and Governance. In US Department of Commerce, National Institute of Standards and Technology, 2023.

U. Gasser and V. A. Almeida. A layered model for ai governance. IEEE Internet Computing, 21(6):58–62, 2017.

T. Gillespie. The relevance of algorithms. Media Technologies. : Essays on Communication Materiality and Society, pages 167–194, 2014.

Google AI. Responsible AI Practices. https://ai.google/responsibility/responsible-ai-practices/. Online, acessado em 21/08/2024.

Governo Federal, Ministério da Ciência, Tecnologia e Inovação, and Conselho Nacional de Ciência e Tecnologia. Ia para o bem de todos, 2024. URL https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos/ia_para_o_bem_de_todos.pdf/@@download/file.

GPAI. Global partnership on artificial intelligence. https://gpai.ai, Acesso em: 19/09/2024. 2020.

P. Grother. Demographic differentials in face recognition algorithms. EAB Virtual Event Series-Demographic Fairness in Biometric Systems, 2021.

Gryfo. Benchmarking de reconhecimento facial destaca as métricas utilizadas pelos grandes players do mercado. https://gryfo.com.br/blog/2021/09/14/benchmarking-de-reconhecimento-facial/, 2021.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems, 2016a.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, 2016b.

J. Howard, E. Laird, Y. Sirotin, R. Rubin, J. Tipton, and A. Vemury. Evaluating Proposed Fairness Models for Face Recognition Algorithms. In International Workshops and Challenges (ICPR), 2022.

A. Jacovi. Trends in explainable ai (xai) literature, 2023.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst., 33(1):1–33, oct 2012.

F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In International Conference on Data Mining, 2012.

T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops, pages 643–650, 2011. doi: 10.1109/ICDMW.2011.83.

C. Katzenbach and L. Ulbricht. Algorithmic governance. Internet Policy Review, 8(4): 1–18, 2019.

S. Kijewski, E. Ronchi, and E. Vayena. The rise of checkbox ai ethics: a review. AI Ethics, 2024. doi: 10.1007/s43681-024-00563-x. URL https://doi.org/10.1007/s43681-024-00563-x.

N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL https://arxiv.org/abs/2009.07896.

J. Kroll. Accountability in Computer Systems. Oxford University Press New York, 2020.

X. Li, M. Xia, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo. Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation. Medical Image Analysis, 88:102862, 2023.

Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3):31 – 57, 2018.

P. Lohia, K. Ramamurthy, M. Bhide, D. Saha, K. Varshney, and R. Puri. Bias mitigation post-processing for individual and group fairness. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)), 2019.

S. Lundberg and S. Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017.

J. Lunter. Beating the bias in facial recognition technology. Biometric Technology Today, 2020(9):5–7, 2020.

M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen. Defining organizational ai governance. AI and Ethics, 2(4):603–609, 2022.

R. Marcinkevics and J. E. Vogt. Interpretability and Explainability: A Machine Learning Zoo Mini-tour. arXiv preprint arXiv:2012.01805, 2020.

M. D. Masseno. Das consequências jurídicas da adesão do brasil aos princípios da ocde para a inteligência artificial, especialmente em matéria de proteção de dados. Journal of Law and Sustainable Development, 8(2):113–122, 2020.

S. McLennan, A. Fiske, D. Tigard, and et al. Embedded ethics: a proposal for integrating ethics into the development of medical ai. BMC Medical Ethics, 23(6), 2022.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. ACM Comput. Surv., 54(6), jul 2021.

P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, A. Morales, D. Lawatsch, F. Domin, and M. Schaubert. Synthetic Data for the Mitigation of Demographic Biases in Face Recognition. In Conference on Biometrics (IJCB), 2023.

M. Minkkinen, J. Laine, and M. Mäntymäki. Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks. Digital Society (DISO), 1:21, 2022.

I. Mironov, K. Talwar, and L. Zhang. Rényi differential privacy of the sampled gaussian mechanism, 2019.

C. Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019.

J. Mökander. Auditing of ai: Legal, ethical and technical approaches. DISO, 2:49, 2023.

Netherlands Court of Audit. An Audit of 9 Algorithms used by the Dutch Government. https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government, 2022. Online, Accessed 23-08-2024.

C. Novelli, M. Taddeo, and L. Floridi. Accountability in Artificial Intelligence: What it is and How it works. AI & Society, 39:1871 – 1882, 2023.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453, 2019.

OECD. Advancing accountability in ai: Governing and managing risks throughout the lifecycle for trustworthy ai. OECD Digital Economy Papers, No. 349, 2023.

T. Pagano, R. Loureiro, F. Lisboa, R. Peixoto, G. Guimarães, G. Cruz, M. Araujo, L. Santos, M. Cruz, E. . Oliveira, I. Winkler, and E. Nascimento. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. Big Data and Cognitive Computing, 7(1), 2023.

S. Patricia and H. Ali. IEEE CertifAIEd – Ontological Specification for Ethical Privacy. https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEESTD-2022%20CertifAIEd%20Privacy.pdf?mkt_tok=MjExLUZZTC05NTUAAAGETQHvhbqJ92qFvqon29PnOA4jYmt9VhwjD6oz0WT2NzwiyjUGtBsO8Q5P3TjdT4NwuDIX5E-yRgoUOAadgENoa8mdUn9Fenk3Zb0JV4m-BQ, 2022. Online, Acessed: 23-08-2024.

T. Pereira and S. Marcel. Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2022.

Personal Data Protection Commission et al. Model artificial intelligence governance framework. Singapore, Jan, 2020.

H. A. Pinto. A utilização da inteligência artificial no processo de tomada de decisões: por uma necessária accountability. Senado Federal, 2019.

T. F. Privacy. Privacidade no machine learning. https://www.tensorflow.org/responsible_ai/privacy/guide?hl=pt-br, 2021.

P. Publica. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2024. Online, Accessed in 23-08-2024.

A. Rai. Explainable ai: From black box to glass box. Journal of the Academy of Marketing Science, 48:137–141, 2020.

A. Rai, P. Constantinides, and S. Sarker. Next generation digital platforms: toward human-ai hybrids. Mis Quarterly, 43(1):iii–ix, 2019.

I. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In Conference on AI, Ethics, and Society (AIES), 2020a.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Conference on Fairness, Accountability, and Transparency, 2020b.

C. Ramos, J. C. Augusto, and D. Shapiro. Ambient intelligence—the next step for artificial intelligence. IEEE Intelligent Systems, 23(2):15–18, 2008.

M. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?"Explaining the predictions of any classifier. In Conference on knowledge discovery and data mining (SIGKDD), 2016.

R. Ross and V. Pillitteri. Protecting controlled unclassified information in nonfederal systems and organizations. Computer Security Division, Information Technology Laboratory, 2024. doi: https://doi.org/10.6028/NIST.SP.800-171r3.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5):206 – 215, 2019.

A. Selbst, D. Boyd, S. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and Abstraction in Sociotechnical Systems. In Conference on Fairness, Accountability, and Transparency, 2019.

I. A. Shah, N. Z. Jhanjhi, and S. N. Brohi. Use of AI-Based Drones in Smart Cities. In Cybersecurity Issues and Challenges in the Drone Industry, pages 362–380. 2024.

L. S. Shapley. A Value for N-person Games. Contributions to the Theory of Games, 28 (2):307 – 317, 1953.

V. Smith, A. Shahin, C. Ashurst, and A. Weller. Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey. arXiv preprint arXiv:2310.01424, 2024.

B. C. Stahl. Embedding responsibility in intelligent systems: from ai ethics to responsible ai ecosystems. Scientific Reports, 13(1):7586, 2023.

Thomson Reuters Practical Law. Accountability principle. https://uk.practicallaw.thomsonreuters.com/w-014-8164?transitionType=Default&contextData=(sc.Default), 2024. Accessed: 2024-05-22.

Y. Tian and Y. Zhang. A comprehensive survey on regularization strategies in machine learning. Information Fusion, 80:146–166, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.11.005. URL https://www.sciencedirect.com/science/article/pii/S156625352100230X.

Toronto Declaration. Toronto declaration. https://torontodeclaration.org, 2018. Acesso em: 19/09/2024.

UNESCO. Recomendação da UNESCO sobre a Ética da Inteligência Artificial. https://unesdoc.unesco.org/ark:/48223/pf0000381137_por, 2021. Acesso em 19/09/2024.

G. Vilone and L. Longo. Explainable Artificial Intelligence: a Systematic Review. arXiv preprint arXiv:2006.00093, 2020.

S. Wei and M. Niethammer. The Fairness-Accuracy Pareto Front. Statistical Analysis and Data Mining, 15(3):287–302, 2022.

J. Wolff, W. Lehr, and C. S. Yoo. Lessons from GDPR for AI Policymaking. Virginia Journal of Law & Technology, 27(4), 2024.

B. Xia, Q. Lu, L. Zhu, S. U. Lee, Y. Liu, and Z. Xing. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In Conference on AI Engineering - Software Engineering for AI, 2024.

S. Yan, D. Huang, and M. Soleymani. Mitigating Biases in Multimodal Personality Assessment. In Conference on Multimodal Interaction, 2020.

M. B. Zafar, I. Valera, M. Rodriguez, and K. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In The Web Conference, 2017.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning, 2018. URL https://arxiv.org/abs/1801.07593.

D. Zhang, B. Xia, Y. Liu, X. Xu, T. Hoang, Z. Xing, M. Staples, Q. Lu, and L. Zhu. Privacy and copyright protection in generative ai: A lifecycle perspective. In Conference on AI Engineering Software Engineering for AI, 2024.

R. D. Zota, I. A. Cîmpeanu, and D. A. Dragomir. Use and design of chatbots for the circular economy. Sensors, 23(11):4990, 2023.