INSTITUTO FEDERAL GOIANO - CAMPUS CERES
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
ADAUTO TURÍBIO DE OLIVEIRA FILHO

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): UM REQUISITO DE TRANSPARÊNCIA E GOVERNANÇA PARA SISTEMAS DE INFORMAÇÃO

CERES
2025

ADAUTO TURÍBIO DE OLIVEIRA FILHO

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): UM REQUISITO DE TRANSPARÊNCIA E GOVERNANÇA PARA SISTEMAS DE INFORMAÇÃO

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia Goiano – Câmpus Ceres, como requisito parcial para a obtenção do título de bacharel em Sistemas de Informação.

Orientadora: Dra. Maryele Lazara Rezende

CERES
2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

O48i

Oliveira Filho, Adauto Turíbio de
INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): UM REQUISITO DE TRANSPARÊNCIA E GOVERNANÇA PARA SISTEMAS DE INFORMAÇÃO / Adauto Turíbio de Oliveira Filho. Ceres 2025.

32f. il.

Orientadora: Profª. Dra. Maryele Lazara Rezende.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0320203 - Bacharelado em Sistemas de Informação - Ceres (Campus Ceres).
1. Inteligência Artificial. 2. Deep Learning. 3. Explicabilidade. 4. Caixa-preta. 5. Ética. I. Título.

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor: Adauto Turíbio de Oliveira Filho

Matrícula: 2022103202030020

Título do trabalho: INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): UM REQUISITO DE TRANSPARÊNCIA E GOVERNANÇA PARA SISTEMAS DE INFORMAÇÃO

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 06 / 12 / 2025

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Ceres (Local) 05 / 12 / 2025 (Data)

gov.br Documento assinado digitalmente
ADAUTO TURIBIO DE OLIVEIRA FILHO
Data: 05/12/2025 23:40:30-0300
Verifique em https://validar.iti.gov.br

Assina[illegible]s autorais

Ciente e de acordo:

gov.br Documento assinado digitalmente
MARYELE LAZARA REZENDE
Data: 06/12/2025 08:02:16-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Aos 12 dias do mês de novembro do ano de dois mil e vinte e cinco, realizou-se a defesa de Trabalho de Curso do acadêmico Adauto Turíbio de Oliveira Filho , do Curso Bacharelado em Sistemas de Informações, matrícula 2022103202030020, cujo título é "INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): UM REQUISITO DE TRANSPARÊNCIA E GOVERNANÇA PARA SISTEMAS DE INFORMAÇÃO". A defesa iniciou-se às 19 horas e 34 minutos, finalizando-se às 19 horas e 56 minutos. A banca examinadora considerou o trabalho APROVADO com média 8,1 no trabalho escrito, média 9,7 no trabalho oral, apresentando assim média aritmética final de 8,9 pontos, estando o(a) estudante APTO para fins de conclusão do Trabalho de Curso.

Após atender às considerações da banca e respeitando o prazo disposto em calendário acadêmico, o estudante deverá fazer a submissão da versão corrigida em formato digital (.pdf) no Repositório Institucional do IF Goiano – RIIF, acompanhado do Termo Ciência e Autorização Eletrônico (TCAE), devidamente assinado pelo autor e orientador.

Os integrantes da banca examinadora assinam a presente.

(Assinado Eletronicamente)
Maryele Lázara Rezende

(Assinado Eletronicamente)
Roitier Campos Gonçalves

(Assinado Eletronicamente)
Gilmara Barbosa de Jesus

(Assinado Eletronicamente)
Ricardo Takayuki Tadokoro

Documento assinado eletronicamente por:

- **Maryele Lazara Rezende**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 12/11/2025 21:45:33.
- **Gilmara Barbosa de Jesus**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 12/11/2025 21:50:05.
- **Ricardo Takayuki Tadokoro**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 12/11/2025 21:50:28.
- **Roitier Campos Goncalves**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 12/11/2025 21:51:43.

Este documento foi emitido pelo SUAP em 12/11/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 763071
Código de Autenticação: 8812b84dea

INSTITUTO FEDERAL GOIANO
Campus Ceres
Rodovia GO-154, Km 03, SN, Zona Rural, CERES / GO, CEP 76300-000
(62) 3307-7100

AGRADECIMENTOS

Dedico meus agradecimentos primeiramente a Deus, pois Ele é a razão (*Logos*) de todo o meu esforço e empenho. Ele foi o sustentador de toda a minha caminhada e a motivação final de cada conquista.

Em segundo lugar, agradeço aos meus pais, Adauto Turíbio de Oliveira e Emília Raquel dos Santos Turíbio, e à minha irmã, Raquel dos Santos Turíbio. Eles sempre estiveram ao meu lado, sendo meus maiores incentivadores em todos os momentos. Agradeço também à minha namorada, Milene Débora Alves, que tem sido meu apoio e auxílio constante ao longo de muitos anos e em cada novo ciclo.

Por fim, registro meus sinceros agradecimentos à instituição Instituto Federal Goiano. Desde o Ensino Médio, o Instituto me proporcionou os meios, o apoio e a estrutura necessários para a realização da minha jornada acadêmica e para a concretização deste trabalho de conclusão de curso.

RESUMO

A presente pesquisa aborda a problemática central da opacidade algorítmica, inerente a modelos complexos de *Deep Learning* ("Caixa-Preta"), e seu impacto direto sobre os princípios de Justiça e Responsabilidade em Sistemas de Informação. Diante disso, o trabalho teve como objetivo analisar a Inteligência Artificial Explicável (XAI) como um requisito técnico, ético e legal para a Governança e Transparência em Sistemas de Informação. A metodologia empregada foi de natureza qualitativa e aplicada, utilizando a Revisão Bibliográfica e a Pesquisa Documental como procedimentos principais. Os resultados demonstraram que a XAI oferece as ferramentas *Model-Agnostic* (LIME e SHAP) necessárias para a auditabilidade de vieses e a mitigação da falibilidade. A análise da legislação brasileira (LGPD e PL 2338/2023) comprovou que a XAI é o mecanismo técnico indispensável para fornecer a transparência e a auditabilidade exigidas para a prestação de contas (*accountability*) em Sistemas de Informação que decidem sobre direitos fundamentais.

Palavras-chave: Inteligência Artificial Explicável, Ética, *Accountability,* LIME, SHAP.

ABSTRACT

The present research addresses the central problem of algorithmic opacity, inherent to complex Deep Learning models ("Black Box"), and its direct impact on the principles of Justice and Accountability in Information Systems. Consequently, the study aimed to analyze Explainable Artificial Intelligence (XAI) as a technical, ethical, and legal requirement for Governance and Transparency in Information Systems. The methodology employed was qualitative and applied in nature, utilizing Critical Bibliographic Review and Documentary Research as the main procedures. The results demonstrated that XAI offers Model-Agnostic tools (LIME and SHAP) necessary for bias auditability and the mitigation of model fallibility. The analysis of Brazilian legislation (LGPD and PL 2338/2023) proved that XAI is the indispensable technical mechanism required to provide the transparency and auditability necessary for accountability in Information Systems that make decisions impacting fundamental rights.

Keywords: Explainable Artificial Intelligence; Governance; Transparency; Algorithm; Bias.

SUMÁRIO

1. INTRODUÇÃO.....6
2. METODOLOGIA.....8
2.1. Procedimentos de Coleta de Dados e Análise.....8
3. REFERENCIAL TEÓRICO.....9
3.1. INTELIGÊNCIA ARTIFICIAL: ABRANGÊNCIA, CONCEITOS E PROBLEMÁTICA.9
3.1.1. A Era da Inteligência Artificial (IA) na Sociedade e nos Negócios.....9
3.1.2. Conceitos Fundamentais da IA..... 11
3.1.3. O Desafio da “Caixa-Preta” Algorítmica.....13
3.2. INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): FUNDAMENTOS E METODOLOGIAS.....14
3.2.1. Conceitos, Diferenciação e Tipologias.....14
3.2.2. Principais Técnicas de XAI: LIME, SHAP e o Rastreio de Vieses.....15
3.3. TRANSPARÊNCIA E GOVERNANÇA EM IA.....17
3.3.1. Princípios Éticos da IA..... 17
3.3.2. A Necessidade de Regulamentação e Governança..... 19
4. DISCUSSÃO DOS RESULTADOS.....21
4.1. XAI COMO MECANISMO DE JUSTIÇA ALGORÍTMICA.....21
4.2. O PARADOXO DA TRANSPARÊNCIA.....22
4.3. A CONVERGÊNCIA REGULATÓRIA.....22
4.4. O DESAFIO DA AUDITABILIDADE.....24
5. CONSIDERAÇÕES FINAIS..... 24
6. REFERÊNCIAS..... 26

1. INTRODUÇÃO

A ascensão da Inteligência Artificial (IA) tem se consolidado como principal vetor da transformação digital na sociedade contemporânea, revolucionando os mais variados setores e remodelando ferramentas e serviços em escala global. A IA, como paradigma da inovação, representa um diferencial competitivo crucial, pois sua adoção permite que organizações aprimorem a eficiência operacional, criem modelos de negócio disruptivos e gerem valor em um mercado altamente competitivo. Para empresas que buscam modernizar-se e manter sua relevância mercadológica, aderir a essa revolução tecnológica não é apenas uma opção, mas uma oportunidade incontornável para estabelecer e sustentar liderança no setor (BRAGA, 2024).

Essa relevância da IA ultrapassa a esfera puramente competitiva. Devido sua capacidade de processar dados em um volume e velocidade sem precedentes, a Inteligência Artificial passou a ser aplicada para otimizar processos complexos e auxiliar na tomada de decisões que possuem impacto social direto. Seus resultados rápidos e eficazes são evidentes em diversas aplicações, que vão desde a automação de análises jurídicas (TACCA; ROCHA, 2018) até a precisão de diagnósticos médicos (BRAGA et al., 2019). Dessa forma, a IA deixou de ser aplicada somente como um pilar tecnológico e começou a acumular responsabilidades ao assumir funcionalidades que causam impactos reais na vida das pessoas, influenciando diretamente resultados sociais.

Apesar da crescente responsabilidade social delegada à Inteligência Artificial, a confiança nos seus resultados é comprometida pela falibilidade inerente de seus sistemas. O que se aplica em contextos reais são modelos complexos de IA, como *Deep Learning*, cuja eficácia frequentemente é acompanhada da sua opacidade, característica central conhecida como problema da “Caixa-Preta” algorítmica (GUIDOTTI et al., 2018). Este termo descreve a dificuldade técnica de se compreender o processo interno pelo qual o sistema chega a uma conclusão, tornando a linha de raciocínio indecifrável para o usuário leigos e, não raro até mesmo para aqueles com competências mais avançadas na área.

Em muitos casos, a complexidade é tamanha que a opacidade se torna absoluta, o que inviabiliza a detecção de origem de erros para que haja correção.

Essa falta de transparência não é apenas um desafio técnico, mas uma falha crítica de governança e ética com implicações sociais. Conforme o estudo de Alves e Andrade (2021), a opacidade agrava a falibilidade e o enviesamento de IA, pois pode encobrir falhas que resultam em predições incorretas ou baseadas em raciocínios não desejáveis e vieses algoritmos que reproduzem preconceitos, gerando discriminações sistemáticas. Diante da incapacidade de se depreender o processo preditivo, a opacidade algorítmica mina a legitimidade das decisões automatizadas e impede a auditoria e a apuração de responsabilidade, tornando a transição da “caixa-preta” para a “caixa de vidro” uma necessidade.

Neste contexto, surge a Inteligência Artificial Explicável (XAI), um campo de estudo focado no desenvolvimento de ferramentas e técnicas que permite aos usuários e gestores compreender, interpretar, corrigir e confiar nas decisões tomadas pelos sistemas de IA, estabelecendo-se como uma resposta técnica fundamental ao imperativo ético da transparência (LEVY; ADANIYA, 2024). Desta forma, a problemática central deste trabalho reside em responder: Como a Inteligência Artificial Explicável (XAI) pode ser estabelecida como um requisito fundamental para a implementação da transparência e da governança algorítmica em Sistemas de Informação?

A justificativa para este estudo baseia-se na relevância de garantir confiabilidade no uso da IA. A análise da XAI como requisito de governança organizacional é crucial, pois permite a mitigação de riscos, a auditoria de modelos e o alinhamento de Sistemas de Informação com princípios de justiça e equidade, especialmente em setores mais sensíveis.

O objetivo geral deste trabalho é analisar a relevância da Inteligência Artificial Explicável (XAI) como requisito de transparência e governança em Sistemas de Informação. Especificamente, o trabalho visa: (i) analisar as implicações éticas e sociais decorrentes da opacidade algorítmica e como a XAI pode mitigá-las; (ii) investigar as principais técnicas e ferramentas de XAI, e suas aplicação na detecção de vieses; (iii) avaliar a regulamentação atual regente em torno da XAI e da transparência algorítmica em Sistemas de Informação (SI).

2. METODOLOGIA

Este trabalho fundamenta-se em uma pesquisa qualitativa de natureza aplicada e exploratória, buscando conceituar, problematizar e alertar quanto à necessidade de Transparência e Governança na aplicação de Sistemas de Informação baseados em Inteligência Artificial (IA).

O método de procedimento adotado é a pesquisa bibliográfica, complementada pela análise documental e pelo estudo ex-post-facto, que consiste na análise de fatos e fenômenos já ocorridos e de dados não manipulados pelo pesquisador.

2.1. Procedimentos de Coleta de Dados e Análise

O Referencial Teórico foi construído a partir de uma Revisão Bibliográfica que englobou:

1. **Literatura Acadêmica e Técnica:** Foram analisados artigos científicos, *dossiês* acadêmicos periódicos nas áreas de Ciência da Computação, Direito Digital e Ética da IA.
2. **Ferramentas de Busca:** As bases de dados Google Scholar, IEEE Xplore, ACM Digital Library e Scielo foram as principais ferramentas utilizadas para a busca de referências, sem limitação a bibliografias em língua portuguesa.
3. **Termos-Chave:** Os termos explorados na busca incluíram *Explainable Artificial Intelligence (XAI)*, *Inteligência Artificial Explicável*, *Algorithmic Bias*, *Governance*, *Lei Geral de Proteção de Dados (LGPD)*, e *Ética da IA*.
4. **Análise Documental e Legal:** O mapeamento do campo regulatório brasileiro e global foi realizado através da análise documental de:
    - **Legislação Brasileira Vigente:** Em especial, o texto da Lei nº 13.709/2018 (LGPD), com foco no Art. 20.
    - **Projetos de Lei**: Análise da proposta de Marco Legal da Inteligência Artificial (PL nº 2338/2023), que estabelece requisitos de Governança e classificação de risco para sistemas de IA.
    - **Regulamentações Globais:** Estudo da Recomendação sobre a Ética da Inteligência Artificial (UNESCO, 2021), que forneceu o fundamento

ético universal para justificar a necessidade de Transparência e *Accountability*.

5. **Estudo de Caso:** A pesquisa utilizou a análise *ex-post-facto* de casos emblemáticos e já documentados, como o sistema norte-americano COMPAS, para demonstrar o dano real (discriminação racial) decorrente da opacidade algorítmica (*black box*).

Em suma, a pesquisa se baseou em fontes secundárias para atingir os objetivos propostos, empregando uma análise qualitativa e documental para construir o argumento de que a XAI é um imperativo técnico, legal e ético para Sistemas de Informação de alto risco.

3. REFERENCIAL TEÓRICO

3.1. INTELIGÊNCIA ARTIFICIAL: ABRANGÊNCIA, CONCEITOS E PROBLEMÁTICA

3.1.1. A Era da Inteligência Artificial (IA) na Sociedade e nos Negócios

A aplicabilidade da Inteligência Artificial (IA) tem se tornado amplamente difundida, influenciando o cotidiano das pessoas em diversas áreas e atividades. No campo da saúde, inclui o uso de chatbots e diagnósticos assistidos; na esfera corporativa, estende-se ao comércio online, recursos humanos (RH), marketing e ensino corporativo; e, no domínio legal, alcança o campo do direito (MORAIS; CASTELO BRANCO, 2023). Atualmente, é quase inexistente uma sociedade que não possua influência direta ou indireta da IA em seus mais variados setores e na vida comum de seus indivíduos.

O estabelecimento da IA na sociedade é notável por sua velocidade de adoção. Dentre as tecnologias disruptivas do seu tempo, a Inteligência Artificial obteve uma penetração social em um ritmo significativamente mais acelerado. Tal fenômeno foi evidenciado pelo estudo estudo “The AI Revolution” que constatou que a IA obteve uma adesão de 50% da sociedade estadunidense em menos tempo que outras inovações revolucionárias como Internet e Smartphones.

Figura 1 - Quantidade de anos para uma tecnologia penetrar em 50% dos Estados Unidos

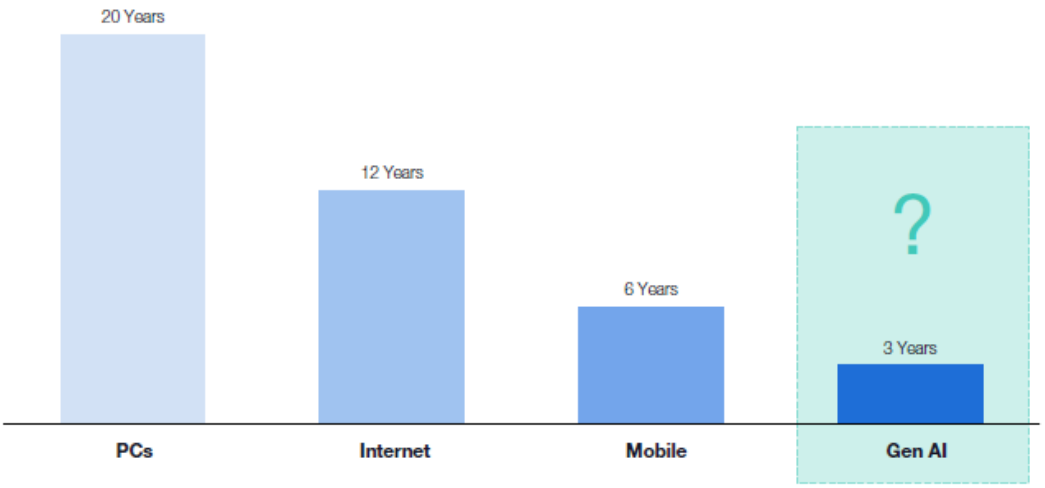

Figura 1. Fonte: VISWANATH et al. (2023)

Conforme exposto no estudo, o tempo de adesão social às últimas grandes tecnologias têm sido reduzido pela metade, sendo a IA a mais veloz a adentrar na sociedade. Essa tecnologia está inevitavelmente presente na rotina dos indivíduos, facilitando tarefas e melhorando a qualidade de vida. Isso se exemplifica nas funcionalidades de assistentes virtuais (*Google Assistente, Siri*), no desbloqueio de smartphones via reconhecimento facial e no algoritmo de sugestão de conteúdo por trás do *feed* das redes sociais.

A influência da IA se manifesta para além do cotidiano social, estabelecendo-se como uma necessidade mercadológica incontornável para empresas que buscam manter sua vantagem no mercado. O tema ganhou tamanha centralidade que grandes consultorias passaram a rastrear sua adoção. Um estudo realizado pela Bain & Company em 2023 demonstrou essa tendência ao constatar que 85% das empresas consideram adotar a Inteligência Artificial em seus processos nos próximos anos.

Figura 2 - Prioridade da IA nas empresas

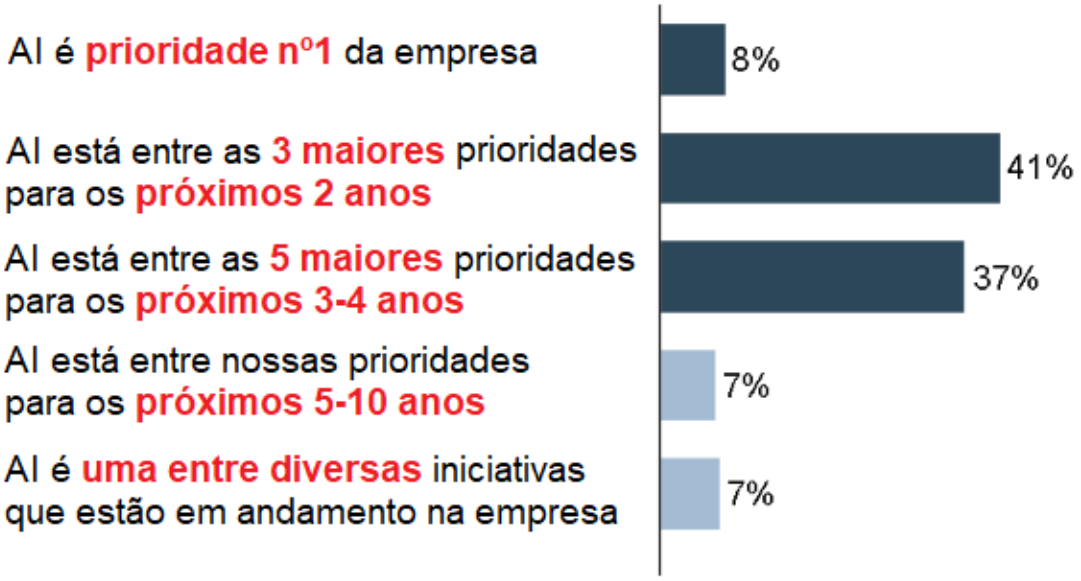

Figura 2. Fonte: Bain & Company (2023)

A pesquisa, que envolveu cerca de 600 empresas de diversos setores, aponta que o principal impulsionador desse desejo coletivo é a perspectiva de reduzir custos e aumentar a produtividade através do potencial da IA. Estima-se que, com o uso desta tecnologia, cerca de 15% de todas as tarefas podem ser concluídas de forma muito mais rápida e com o mesmo nível de qualidade, justificando o interesse massivo em priorizar a integração da IA. A capacidade de otimizar processos em larga escala não só eleva o desempenho, como também aumenta a competitividade dos produtos e serviços oferecidos.

Diante disso, a adesão da Inteligência Artificial aos negócios, sendo este um dos setores mais impactados pela ferramenta, é completamente inviável de ser ignorada. Em um futuro próximo, integrar a IA nas operações tende a deixar de ser apenas uma opção estratégica e passará a ser uma questão de sobrevivência e manutenção da relevância para o negócio (BRAGA, 2024). Dessa forma, a rápida e profunda inserção da IA na vida social e pessoal, aliada à sua utilização em decisões corporativas e governamentais, estabelece o alicerce para uma discussão crucial sobre responsabilidade e governança algorítmica.

3.1.2. Conceitos Fundamentais da IA

O termo “Inteligência Artificial” (IA) foi formalmente criado por John McCarthy (1956), na conferência Dartmouth College, nos Estados Unidos. O conceito

inaugural definia a IA como "A ciência e engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes". Esta definição batizava um novo campo do conhecimento que, desde a década de 1940, buscava produzir modelos matemáticos capazes de simular o funcionamento dos neurônios cerebrais. Em paralelo com as evoluções tecnológicas, esta área cresceu consideravelmente, resultando em um entendimento mais desenvolvido. Atualmente, o conceito de inteligência artificial diz respeito à possibilidade de as máquinas realizarem operações de decisão com um raciocínio que simula o dos humanos, através de algoritmos bem elaborados e complexos (DAMACENO; VASCONCELOS, 2018). Ela possibilita que sistemas aprendam, deliberem, decidam e percebam de forma inteligente de acordo com as situações apresentadas.

A IA não é uma tecnologia única, mas uma área do conhecimento composta por diferentes vertentes e modelos. Dentre os mais conhecidos, destacam-se o *Machine Learning* (ML) e o *Deep Learning* (DL). Embora o *Deep Learning* seja uma derivação do *Machine Learning*, ambos possuem distinções metodológicas relevantes. O *Machine Learning* representa o subcampo que permite aos sistemas de computador aprender e melhorar com a experiência, utilizando algoritmos para analisar e identificar padrões em grandes volumes de dados para a tomada de decisão. Este modelo frequentemente opera a partir de um direcionamento humano nos estágios iniciais, como na rotulação dos dados (BRAGA, 2024). Um exemplo claro e cotidiano da aplicação do ML é o software que identifica e-mails spam: o sistema aprende com a experiência, rotulando novos e-mails a partir de dados que anteriormente já foram classificados como spam (DAMACENO; VASCONCELOS, 2018).

Em contraste com o Machine Learning tradicional, o *Deep Learning* (DL) é o subcampo que alavancou o desempenho da IA em tarefas de alta complexidade. O DL é, essencialmente, um tipo de ML que capacita a máquina a realizar tarefas sofisticadas, como o reconhecimento de fala, a identificação de imagens e a realização de previsões complexas. Metodologicamente, o *Deep Learning* é caracterizado por utilizar redes neurais artificiais com múltiplas camadas de abstração. O seu aprendizado ocorre de forma hierárquica e aprofundada: a saída de uma camada de neurônios matemáticos funciona como a entrada para a camada subsequente (MORAIS; CASTELO BRANCO, 2023). Nesse processo, o DL aprofunda o aprendizado de máquinas com a finalidade de executar atividades que

simulam o comportamento humano. É justamente essa complexidade inerente de sua arquitetura de múltiplas camadas, que processa dados em diversos níveis de abstração, que confere ao DL seu alto desempenho e, simultaneamente, o torna o principal vetor do problema da opacidade algorítmica (ALVES; ANDRADE, 2021).

3.1.3. O Desafio da “Caixa-Preta” Algorítmica

O alto desempenho dos modelos de *Deep Learning* traz consigo o problema inerente da opacidade algorítmica, conhecido como a "Caixa-Preta". No campo da Inteligência Artificial, Guidotti et al. (2018) conceitua esse termo ao fato de que modelos complexos, sobretudo aqueles que utilizam redes neurais profundas, chegam a conclusões cujo processo de raciocínio é muito complexo para que um humano consiga entender a lógica da decisão. Essa opacidade algorítmica se estabelece, portanto, como um déficit de inteligibilidade na tomada de decisões de um sistema. Conforme a definição ampla de Harry Surden (2014), opacidade algorítmica é qualquer momento em que um sistema tecnológico se engaja em um comportamento que, embora apropriado em termos de resultado, é difícil de entender ou prever do ponto de vista humano. Essa falta de transparência na lógica interna do modelo é o ponto de partida para a falibilidade.

É importante destacar que os modelos de Inteligência Artificial não são inerrantes. A falibilidade é uma característica intrínseca a qualquer sistema complexo, mas, no contexto da "Caixa-Preta", ela se torna um risco exponencial. A opacidade algorítmica impede o rastreamento da lógica interna, o que se torna crítico quando o sistema comete um erro. De acordo com Molnar, Casalicchio e Bischl (2018), a falta de transparência e interpretabilidade de modelos de aprendizado de máquina tem sido um problema constante, pois inviabiliza a auditoria e a correção de *outputs* incorretos. Esta falta de rastreabilidade é o fator que agrava o enviesamento, que é a principal manifestação da falibilidade. Modelos opacos, treinados em dados históricos que refletem preconceitos sociais existentes, podem herdar e amplificar essas disparidades, levando a decisões discriminatórias em escala. A falta de transparência em modelos de IA é, portanto, um problema significativo, pois pode levar a decisões injustas ou imprecisas, especialmente

quando esses modelos são utilizados em áreas sensíveis, como saúde e justiça (BENDER et al, 2021).

A falibilidade algorítmica gera um déficit de confiança e legitimidade para os Sistemas de Informação. Delegar decisões de alto impacto social, como diagnósticos médicos ou avaliações judiciais a um sistema incapaz de explicar sua lógica, compromete o princípio de Responsabilidade (*Accountability*) e gera insegurança quanto à validade do resultado. Desse modo, a superação do desafio da “Caixa-Preta” é o principal imperativo para o uso consciente e ético da IA na sociedade. A resolução desta problemática exige uma aplicação de metodologias capazes de revelar e justificar o raciocínio utilizado pelo modelo.

3.2. INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI): FUNDAMENTOS E METODOLOGIAS

3.2.1. Conceitos, Diferenciação e Tipologias

A Inteligência Artificial Explicável (*Explainable Artificial Intelligence - XAI*) é a abordagem de pesquisa e desenvolvimento que visa criar modelos de aprendizado de máquina mais interpretáveis e transparentes, permitindo que humanos compreendam, confiem e auditem as decisões e previsões feitas por sistemas de IA (ARRIETA et al., 2020). A XAI se estabelece como a principal resposta técnica para contornar a opacidade algorítmica, tornando os sistemas mais justos e confiáveis. Isso é alcançado através da utilização de métodos e estratégias, textuais ou visuais, que fornecem uma compreensão qualitativa sobre o processo de predição do modelo (RIBEIRO et al., 2016).

No contexto da XAI, a primeira distinção essencial reside no *trade-off* entre desempenho e transparência. Modelos de IA que são naturalmente interpretáveis por humanos, como Regressão Linear ou Árvores de Decisão, possuem estruturas simples, mas entregam, significativamente, menos desempenho preditivo do que os modelos "Caixa-Preta" (MOLNAR; CASALICCHIO. BISCHL, 2018). É nesse

paradoxo que surge a necessidade da XAI: assegurar a alta acurácia de modelos mais complexos, ao mesmo tempo em que se garante a transparência.

Busca-se, portanto, a explicabilidade, que se estabelece como uma característica ativa do modelo. Diferente da interpretabilidade (que é passiva), a explicabilidade conceitua-se como qualquer ação ou procedimento realizado com o intuito de detalhar suas funções internas, de modo a tornar seu funcionamento claro para uma determinada audiência. A XAI é, essencialmente, a aplicação de métodos *post-hoc*, ou seja, métodos aplicados após o processo de decisão do modelo, com o objetivo de tornar suas escolhas mais compreensíveis, para fornecer essa justificação e transparência (ARRIETA et al., 2020).

Além da distinção conceitual entre interpretabilidade e explicabilidade, as metodologias de XAI são tipologicamente divididas quanto ao escopo da explicação que fornecem, o que é crucial para atender aos diferentes públicos-alvo (audiências) e requisitos de Governança. Essa tipologia é determinada pela abrangência da transparência do modelo, separando-se em Explicação Local e Explicação Global (GUIDOTTI et al., 2018). A Explicação Local é o mecanismo que revela o motivo específico de uma única decisão para uma única instância de dados, permitindo ao usuário entender o porquê de um resultado específico ter sido produzido (MOLNAR; CASALICCHIO; BISCHL, 2018). Este escopo atende diretamente ao requisito de responsabilidade individual e ao "Direito à Explicação", sendo fundamental para a transparência nas interações de um Sistema de Informação com o usuário final.

Em contraste, a Explicação Global busca revelar a lógica geral e o comportamento médio do modelo em sua totalidade, indo além do caso individual. Seu objetivo é identificar quais variáveis são, em média, mais relevantes para todas as decisões do sistema. A Explicação Global é essencialmente voltada para a auditoria de *compliance* e para a detecção sistêmica de vieses que comprometem a justiça do modelo em um nível agregado. Desta forma, o entendimento das tipologias de XAI justifica a sua posição como um requisito fundamental de transparência e Governança para Sistemas de Informação que operam em cenários de alto impacto social (GUIDOTTI et al., 2018).

3.2.2. Principais Técnicas de XAI: LIME, SHAP e o Rastreio de Vieses

A XAI aplica-se primariamente através de técnicas de Explicação *Post-Hoc* (ou *post-modeling explainability*), metodologias externas utilizadas para comunicar informações compreensíveis sobre como um modelo já desenvolvido produz suas predições. Dentre estas, as abordagens Modelo-Agnóstico são as mais relevantes para sistemas que utilizam modelos "Caixa-Preta" complexos. O conceito *Model-Agnostic* significa que a técnica de explicação é independente da arquitetura interna do modelo de *Machine Learning*, atuando como um *wrapper* (invólucro) que interage apenas por meio das entradas (*inputs*) e saídas (*outputs*) (ARRIETA et al., 2020). Esta abordagem é fundamental, pois confere solidez à solução de explicabilidade, permitindo a auditoria e a troca do modelo preditivo sem a necessidade de alterar a metodologia de explicação.

Neste cenário de modelos opacos, as ferramentas mais amplamente utilizadas para obtenção de insights sobre o comportamento preditivo são o LIME (*Local Interpretable Model-agnostic Explanations*) e o SHAP *(SHapley Additive exPlanations*). Estes métodos são frequentemente classificados como geradores de explicações locais para instâncias individualizadas (AMOROSO, 2023). As aplicações práticas destes *frameworks* demonstraram sua eficácia na análise de fatores de maior contribuição para as decisões de modelos complexos.

O LIME (*Local Interpretable Model-Agnostic Explanations)* é uma ferramenta popular de Inteligência Artificial Explicável (XAI), focada em gerar explicações locais para instâncias individuais (CHU, 2022). Seu objetivo é tornar modelos "caixa-preta" mais transparentes, funcionando como uma abordagem *post-hoc*, ou seja, aplicada após o treinamento do modelo. O LIME simplifica modelos complexos por meio de uma aproximação local, com o princípio de que a explicação deve ser interpretável por humanos e fiel à instância analisada (RIBEIRO, 2016). Para isso, ele perturba a instância de interesse, como ao dividir uma imagem em componentes interpretáveis, e gera novas amostras. A partir dessas amostras, o LIME calcula pesos com base na similaridade, ajustando um modelo linear para destacar os componentes mais importantes.

Em complemento, o SHAP (*SHapley Additive Explanations*) é um *framework* (LUNDBERG; LEE, 2017) com ferramentas eficientes para calcular os valores de *Shapley*, um conceito com base matemática sólida na teoria de Jogos Cooperativos. Assim como o LIME, o SHAP é um método *post-hoc* e é classificado como um dos mais populares. Sua principal vantagem é ser capaz de gerar explicações que

podem ser tanto locais quanto globais. Na prática, o SHAP realiza um levantamento de quais características ou atributos foram mais relevantes para o modelo chegar a uma determinada resposta (SALIH et al., 2025). Ele calcula a contribuição exata que cada atributo forneceu à predição final, distribuindo a importância da predição entre todos os fatores de entrada, de forma a quantificar o peso de cada um no resultado. Um exemplo de aplicação do SHAP é a predição de doenças cardíacas, onde o *framework* é utilizado para tornar visíveis os fatores de risco importantes e as contribuições das *features* para a predição do modelo (REZK et al., 2024).

A capacidade do LIME e SHAP de quantificar a influência de cada *feature* é o ponto de convergência com os requisitos de Governança em Sistemas de Informação. Ao quantificar a relevância de cada atributo, essas ferramentas tornam possível o rastreamento de vieses algorítmicos, satisfazendo a necessidade de justiça e transparência (ALVES; ANDRADE, 2021). A Explicação Local, gerada pelo LIME e SHAP, permite que o usuário afetado confronte ou altere o resultado (IBM, 2022). Já a Explicação Global, característica do SHAP, é fundamental para auditorias sistêmicas, onde é avaliado se o modelo usa *features* correlacionadas para gerar um resultado enviesado. Portanto, a aplicação dessas metodologias de XAI é o mecanismo técnico essencial para garantir a auditabilidade e a conformidade regulatória dos Sistemas de Informação em de grande impacto social (GUIDOTTI et al., 2018).

3.3. TRANSPARÊNCIA E GOVERNANÇA EM IA

3.3.1. Princípios Éticos da IA

Diversas organizações relevantes já produziram declarações que abordam valores e princípios no uso e desenvolvimento de sistemas de IA, detalhando suas devidas implicações éticas. Uma síntese abrangente destas declarações, realizada por Floridi et al. (2018), elencou cinco princípios éticos fundamentais que são considerados unânimes entre as iniciativas globais. Estes princípios são: beneficência, não-maleficência, autonomia, justiça, e explicabilidade. Os quatro primeiros são tradicionalmente pilares da bioética, sendo adaptados de forma

coerente para o âmbito da Inteligência Artificial. Complementarmente, a análise das declarações levantou o princípio da explicabilidade, que surge como um pilar essencial e específico para o contexto algorítmico.

O princípio da Beneficência é caracterizado pelo requisito de que os sistemas de IA priorizem e promovam o bem-estar humano. A interpretação deste princípio é frequentemente ampla, relacionando a IA à necessidade de garantir as condições básicas para a vida em nosso planeta, a prosperidade contínua da humanidade e a preservação de um bom ambiente para as gerações futuras. Em contrapartida, o princípio da Não-Maleficência, embora logicamente relacionado, diz respeito à cautela que se deve ter com as consequências negativas do uso indevido ou excessivo de Inteligência Artificial. Essas consequências abrangem desde a violação da privacidade dos dados até o risco de escalada de conflitos (FLORIDI et al., 2018).

Ainda sobre os princípios elencados por Floridi et al. (2018), o conceito de Autonomia está diretamente ligado ao equilíbrio entre o poder de decisão que mantemos para nós mesmos e aquele que delegamos a agentes artificiais. Nesta visão, é necessário que a autonomia das máquinas seja restringida e intrinsecamente reversível, garantindo a possibilidade de restabelecer a autonomia humana em qualquer etapa do ciclo de vida da IA. Complementarmente, o princípio da Justiça (*Fairness*) é caracterizado pela exigência de que a utilização da IA atue na correção de erros do passado, eliminando qualquer tipo de discriminação injusta e a perpetuação de vieses já presentes nos dados de treinamento. Da mesma forma, a Justiça requer a prevenção ativa da criação de novos danos ou a amplificação de desigualdades nas estruturas sociais existentes.

Por fim, o princípio fundamental que faz a ponte entre a ética e a técnica é a Explicabilidade, frequentemente expresso por termos correlatos como: transparência, interpretabilidade e Responsabilidade (Accountability). Este princípio diz respeito à necessidade intrínseca da compreensão dos resultados obtidos pela IA, buscando responder a perguntas essenciais, tais como “como o modelo funciona?” ou “quem é o responsável pelo seu funcionamento?” (FLORIDI et al., 2018). Em termos gerais, a relação entre a humanidade e as máquinas deve ser estabelecida de forma que seja facilmente compreensível para o cidadão comum. É a Explicabilidade que transforma a "caixa-preta" em uma "caixa de vidro" (ALVES; ANDRADE, 2021), permitindo a verificabilidade, a auditoria e, principalmente, a apuração de responsabilidade quando a IA toma decisões potencialmente ilegais ou

enviesadas. Ao gerar *insights* sobre o funcionamento do modelo, a XAI promove a confiança dos usuários e da sociedade na Inteligência Artificial, pois mostra, de maneira geral, quando, como e por que um algoritmo está tomando determinada decisão.

3.3.2. A Necessidade de Regulamentação e Governança

A crescente ocupação da Inteligência Artificial em setores sociais decisivos já é uma realidade no Brasil. No âmbito do Poder Judiciário, o Supremo Tribunal Federal (STF) tem ampliado o uso de diversas ferramentas de IA em apoio à atividade jurisdicional. O STF atualmente opera mais de uma iniciativa de IA, contando com robôs como o Victor (utilizado para análise de repercussão geral), a Rafa (para classificação de processos em relação aos Objetivos de Desenvolvimento Sustentável), a VictórIA (para análise e classificação de processos), e a plataforma Maria (para suporte à redação e revisão gramatical e consulta unificada de precedentes). Cada uma dessas IAs possui funções específicas que visam agilizar o trabalho processual, reforçando o objetivo do Tribunal em aumentar a produtividade e a segurança no tratamento de informações sensíveis (STF, 2025).

A relevância da IA em setores sensíveis, como o Judiciário, torna necessário o debate sobre a falibilidade e os vieses algorítmicos. Um dos exemplos mais emblemáticos do impacto social negativo da opacidade algorítmica é o sistema norte-americano COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Desenvolvido para auxiliar tribunais a estimar o risco de reincidência criminal de réus, o COMPAS elabora uma "classificação de risco" com base em um extenso questionário, histórico criminal e dados da plataforma. Estas classificações influenciam diretamente decisões importantes, como a estipulação de fianças e, em alguns estados, até mesmo a sentença final (Angwin et al., 2016). Contudo, uma investigação da organização *ProPublica* em 2016 demonstrou a perpetuação de vieses raciais: o algoritmo se mostrou mais propenso a classificar réus negros como "prováveis reincidentes" de forma equivocada, enquanto réus

brancos eram classificados de maneira errônea como "indivíduos com baixo risco de reincidência".

A situação foi agravada pela natureza de "caixa-preta" do sistema, que não era disponibilizado ao público o algoritmo, impedindo que os réus soubessem os critérios e procedimentos utilizados para definir seu índice de risco. Este caso se estabeleceu como um marco que ressalta a urgência de exigir Transparência e Explicabilidade (*XAI*) em sistemas que impactam a liberdade e os direitos fundamentais (ALVES; ANDRADE, 2021).

O risco associado à opacidade e ao enviesamento algorítmico, evidenciado no caso COMPAS, impulsionou uma resposta ética e regulatória em nível global. Nesse cenário, a Recomendação sobre a Ética da Inteligência Artificial, aprovada pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) em novembro de 2021, estabeleceu a primeira estrutura normativa global sobre o tema (UNESCO, 2021). O documento consolida um consenso de princípios onde a Transparência e a Explicabilidade são requisitos essenciais para garantir sistemas de IA que sejam auditáveis e justos. É com base nesses princípios que o Brasil avançou na criação de arcabouços legais que buscam minimizar os riscos e proteger os cidadãos da "caixa-preta" algorítmica.

Diante do imperativo ético por Transparência, o Brasil avançou na criação de fundamentos legais para reger a utilização de sistemas de IA. O principal pilar vigente é a Lei Geral de Proteção de Dados (LGPD – Lei nº 13.709/2018), cujo Art. 20 garante ao titular de dados o direito de solicitar a revisão de decisões tomadas unicamente por meios automatizados. Este direito exige que o controlador forneça informações claras sobre os critérios e procedimentos utilizados para o resultado. Em paralelo, o Projeto de Lei nº 2338/2023 (PL 2338/2023) atualmente em tramitação, visa instituir o Marco Legal da Inteligência Artificial, estabelecendo diretrizes de Governança e Responsabilidade, classificando os sistemas por níveis de risco e impondo obrigações mais rigorosas de explicabilidade e auditabilidade para o alto risco. A convergência dessas legislações eleva a Inteligência Artificial Explicável (XAI) a um mecanismo técnico indispensável em Sistemas de Informação que decidem sobre direitos fundamentais.

4. DISCUSSÃO DOS RESULTADOS

4.1. XAI COMO MECANISMO DE JUSTIÇA ALGORÍTMICA

A opacidade algorítmica, característica inerente a modelos complexos como o *Deep Learning*, contribui significativamente para a propagação de vieses sistêmicos. Ao ceder o poder de decisão a sistemas de IA tão complexos que seus próprios desenvolvedores enfrentam dificuldades em definir os critérios utilizados para a tomada de decisão, configura-se um grave problema ético. Este problema transcende o âmbito teórico e manifesta-se em consequências práticas, como a violação do princípio de Justiça. O caso COMPAS, por exemplo, ilustrou como a falta de transparência perpetuou a discriminação racial na esfera judicial. Tais manifestações práticas reforçam, eticamente, a urgência da Transparência em modelos de Inteligência Artificial, que é exatamente o que a Inteligência Artificial Explicável (XAI) se propõe a oferecer.

A resposta técnica para auxiliar a mitigação da opacidade concentra-se na aplicação de metodologias de XAI. A investigação das ferramentas LIME (*Local Interpretable Model-agnostic Explanations*) e SHAP (*SHapley Additive exPlanations*) demonstrou que a XAI oferece as capacidades necessárias para o rastreamento do raciocínio algorítmico. A relevância destas técnicas *Model-Agnostic* reside justamente em sua função de auditoria e conformidade. O LIME cumpre a função de Explicação Local, sendo o mecanismo ideal para justificar a predição a um indivíduo específico. Em contrapartida, o SHAP, que quantifica a contribuição exata de cada atributo para o resultado, atua em dois níveis: provê explicação Local para o cidadão e explicação Global para a auditoria sistêmica. É justamente essa capacidade de quantificação, especialmente em nível Global, que é crucial para a detecção de vieses. Ao revelar quais características são desproporcionalmente mais relevantes para a decisão do sistema, o SHAP funciona como um mecanismo de auditoria para expor características correlacionadas ou enviesadas.

4.1.1. Detalhamento Operacional do LIME

O LIME é uma técnica de explicação local que busca tornar modelos de aprendizado de máquina mais complexos compreensíveis (GUIDOTTI et al., 2018), atendendo a requisitos legais como o Art. 20 da LGPD, que garante aos indivíduos o direito de obter uma explicação sobre decisões automatizadas. A premissa central do LIME é que, embora um modelo de IA seja muitas vezes complexo e de difícil interpretação como um todo, ele pode ser aproximado por modelos mais simples em regiões locais do espaço de dados, permitindo uma explicação clara e intuitiva para decisões específicas. Isso é possível sem sacrificar a eficácia do modelo original, oferecendo uma compreensão do comportamento do modelo complexo em torno de uma instância particular.

A operação do LIME começa com a perturbação da instância original, ou seja, o LIME cria variações artificiais da instância de dados que se deseja explicar. Essas modificações são feitas alterando ligeiramente as características dessa instância, gerando um conjunto de dados perturbados, mas próximos ao ponto original. A seguir, o LIME pondera essas amostras perturbadas com base na proximidade da instância original. Ou seja, quanto mais semelhantes às amostras perturbadas forem à instância original, maior será o peso atribuído a elas. Isso é feito por meio de uma métrica de distância, como a distância euclidiana, para garantir que o modelo simples resultante reflita adequadamente o comportamento do modelo complexo na vizinhança local da instância que está sendo explicada.

Com essas amostras perturbadas e seus respectivos pesos, o LIME então treina um modelo simples (geralmente uma regressão linear) para aproximar as previsões do modelo complexo nas vizinhanças da instância original. O treinamento é feito localmente, ou seja, apenas usando as amostras geradas, o que permite que o modelo simples capture o comportamento do modelo complexo sem a sobrecarga de tentar entender seu funcionamento global. Esse modelo simples, conhecido como modelo substituto ou *proxy*, revela as características (*features*) mais relevantes que influenciaram a decisão tomada pelo modelo complexo para aquela instância específica (RIBEIRO et al., 2016). O modelo *proxy*, por ser simples e transparente, fornece uma explicação clara sobre quais variáveis foram mais determinantes para a

decisão, permitindo que se compreenda a razão por trás do comportamento do modelo.

4.1.1. Detalhamento Operacional do SHAP

Na prática, o SHAP atua de forma eficiente ao calcular a contribuição de cada *feature* para a predição de um modelo. Quando aplicado a uma instância específica, o SHAP começa pela construção de várias combinações de *features* para entender como a inclusão ou exclusão de cada atributo afeta a saída do modelo. Para isso, ele avalia diferentes configurações possíveis de variáveis, o que exige o cálculo do impacto de cada característica em todas as combinações possíveis de entrada. Esse processo, embora computacionalmente intenso, permite que o SHAP determine, com precisão, a contribuição marginal de cada *feature* para a predição final. A partir desse cálculo, o SHAP gera uma pontuação que representa a contribuição específica de cada atributo na decisão do modelo.

Em termos práticos, quando o SHAP é aplicado a uma instância de dados, ele gera uma explicação baseada em um valor de Shapley para cada feature. Esses valores representam a contribuição individual de cada característica para a diferença entre a predição do modelo e a média global do modelo para todas as instâncias. Para exemplificar, em um modelo de previsão de doenças cardíacas, o SHAP pode calcular o impacto de fatores como "idade", "pressão arterial" ou "nível de colesterol" sobre a probabilidade de diagnóstico. Este tipo de aplicação de XAI é largamente utilizado no domínio biomédico, conforme demonstrado em estudos sobre predição de doenças cardíacas (REZK et al., 2024; SALIH et al., 2025).

Esse processo é repetido para cada instância do conjunto de dados, permitindo a geração de uma explicação global ao agregarmos as contribuições de todas as características ao longo de várias instâncias. Por exemplo, ao analisar um modelo preditivo para doenças cardíacas com um grande número de pacientes, o SHAP pode determinar que, em média, características como "colesterol elevado" têm uma contribuição muito mais forte para as predições do que características como "idade" (REZK et al., 2024). Assim, a explicação global do modelo é construída

a partir da média das explicações individuais, proporcionando uma visão clara sobre quais features têm maior influência no comportamento geral do modelo.

Além disso, o SHAP oferece transparência em sua operação, uma vez que a distribuição das contribuições das *features* é clara e acessível. O cálculo dos valores de Shapley garante que a soma das contribuições individuais seja igual à predição final (LUNDBERG; LEE, 2017), permitindo que qualquer auditoria ou revisão do modelo seja feita com base em uma metodologia sólida e comprovada matematicamente. Isso facilita a identificação de possíveis viéses ou características que podem estar sendo usadas de maneira indevida, como variáveis demográficas, o que torna o SHAP especialmente útil em contextos onde a não-discriminação e a justiça algorítmica são cruciais.

4.2. O PARADOXO DA TRANSPARÊNCIA

A transparência em sistemas de Inteligência Artificial frequentemente esbarra na necessidade de proteger os direitos de propriedade intelectual das empresas, que muitas vezes veem seus algoritmos como ativos valiosos e exclusivos. No entanto, a implementação da Inteligência Artificial Explicável oferece uma solução pragmática para esse desafio. A XAI foca em fornecer explicações claras sobre as decisões tomadas pelos sistemas, sem a necessidade de divulgar o código-fonte ou detalhes confidenciais dos algoritmos. Isso significa que as empresas podem continuar a proteger os aspectos essenciais de sua propriedade intelectual, enquanto ainda atendem aos requisitos de transparência e auditoria exigidos por regulamentações como a LGPD e a Lei de Inteligência Artificial. Em vez de tornar o processo de decisão da IA totalmente acessível ou previsível, a XAI permite que as partes interessadas compreendam como a IA chegou a uma conclusão, quais fatores influenciaram essa decisão e de que maneira os dados foram processados.

Além disso, a XAI não depende de revelar detalhes técnicos complexos que poderiam comprometer o segredo comercial. Em vez disso, ela utiliza abordagens como visualizações, métricas interpretáveis e descrições do raciocínio lógico subjacente às decisões automatizadas. Essa abordagem possibilita que reguladores e auditores verifiquem a conformidade dos sistemas de IA com a legislação,

garantindo que as decisões não sejam discriminatórias, enviesadas ou injustas, sem a necessidade de acesso direto ao código-fonte proprietário. Assim, a XAI equilibra a exigência de explicação e auditabilidade com a necessidade de manter o controle sobre a propriedade intelectual, oferecendo um meio-termo entre transparência e proteção comercial.

4.3. A CONVERGÊNCIA REGULATÓRIA

A exigência da transparência em modelos de IA não pode ser confiada há uma movimentação coletiva e espontânea do mercado. Portanto, faz-se necessário medidas regulatórias para atender essa conformidade. No cenário atual, o fundamento legal vigente manifesta-se inicialmente pela Lei Geral de Proteção de Dados (LGPD – Lei nº 13.709/2018). Embora esta lei não seja um marco regulatório específico para a Inteligência Artificial, ela estabelece o direito à revisão de decisões tomadas unicamente por meios automatizados (Art. 20). Contudo, o escopo da LGPD, focado primariamente na proteção de dados, não oferece diretrizes detalhadas necessárias para a governança de modelos algorítmicos complexos, nem define as metodologias técnicas específicas de explicabilidade (XAI). Desse modo, a lei atual se mostra insuficiente para regulamentar de forma completa a necessidade de transparência em Sistemas de Informação de alto risco.

A insuficiência da LGPD para lidar com a complexidade da “Caixa-Preta” justifica a urgência do Marco legal de Inteligência Artificial, atualmente em tramitação no Congresso, por meio do Projeto de Lei nº 2.338/2023. Diferentemente do marco vigente, o PL endereça diretamente a XAI ao garantir explicitamente os direitos dos afetados, como direito a explicação sobre a decisão, recomendação ou previsão tomada por sistemas de inteligência artificial (Art. 5º, II) e o direito à não-discriminação e à correção de vieses discriminatórios (Art. 5º, V). Adicionalmente, o Art. 7º detalha o que é o “Direito a Explicação”, exigindo informação sobre a racionalidade lógica do sistema, os principais fatores que afetam a decisão e o grau de contribuição do sistema na tomada de decisões. Estas exigências demonstram que o Marco Legal, se aprovado, será suficiente para

regulamentar e garantir a transparência e promover a transição da opacidade para um sistema auditável.

Em síntese, a análise crítica dos resultados valida a premissa de que a Inteligência Artificial Explicável (XAI) é o elo indispensável entre a performance tecnológica e as exigências de uma Governança responsável. A investigação provou que a XAI é o mecanismo técnico para mitigar o viés algorítmico, conforme a análise das ferramentas LIME e SHAP, que fornecem a capacidade de justificar decisões individuais e auditar o modelo no agregado. Do ponto de vista ético, este trabalho reforçou que essa transparência é um imperativo de Justiça e Não-Maleficência, como demandado pelos *frameworks* globais (UNESCO, 2021). No contexto brasileiro, o estudo demonstrou que a convergência da LGPD (Art. 20) com o Marco Legal da IA (PL 2338/2023) transforma a XAI em um requisito de *compliance*. A contribuição deste trabalho para a área de Sistemas de Informação reside, portanto, em fornecer um direcionamento teórico que exige a adoção de metodologias de desenvolvimento responsáveis, elevando a XAI de uma mera opção técnica para uma ferramenta essencial de prestação de contas na sociedade.

4.4. O DESAFIO DA AUDITABILIDADE

Por fim, a conclusão de que a XAI é o instrumento de conformidade transfere a complexidade do modelo preditivo para a validade da explicação. Isso levanta o desafio da Meta-Explicabilidade: como garantir que as técnicas de XAI sejam aplicadas corretamente e sem vieses? O risco de enviesamento na explicação é real e pode ocorrer, por exemplo, através da manipulação de features ou da escolha inadequada de parâmetros que favoreçam um resultado desejado pelo agente regulado. Esse é um desafio regulatório ainda não abordado de forma institucional pelo atual marco legal brasileiro. Para que a XAI cumpra plenamente sua função ética e legal de prestação de contas, será necessário que futuras legislações debatam a criação de mecanismos de certificação ou órgãos auditores independentes. Esses organismos teriam a responsabilidade de validar a isenção e a robustez metodológica dos modelos e das explicações, assegurando que a conformidade não seja apenas teórica, mas efetivamente garantida na prática.

5. CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo Geral analisar a Inteligência Artificial Explicável (XAI) como um requisito técnico, ético e legal para a Governança e Transparência em Sistemas de Informação. A pesquisa confirmou a problemática central que motivou este trabalho: a opacidade algorítmica inerente aos modelos complexos como *Deep Learning* e *Machine Learning*, conhecidos como "Caixa-Preta", compromete os princípios fundamentais de justiça e fiscalização regulatória. Entretanto, o achado principal deste trabalho é que a XAI não é apenas uma área de pesquisa promissora, mas também o mecanismo técnico indispensável que traduz as exigências éticas e legais em parâmetros de *compliance*.

O cumprimento dos objetivos específicos deste trabalho demonstrou que a XAI dispõe das ferramentas técnicas necessárias para mitigar a problemática da opacidade. A investigação das técnicas, como LIME (*Local Interpretable Model-agnostic Explanations*) e SHAP (*SHapley Additive exPlanations*), revelou que a Explicabilidade pode ser implementada tanto em nível local (para a justificação de uma decisão individual ao usuário) quanto em nível global (para a auditoria e rastreio de vieses sistêmicos). Essa capacidade de inteligibilidade é essencial para enfrentar as implicações éticas da IA, como os enviesamentos algorítmicos. O estudo do caso COMPAS, por exemplo, ilustrou a falha moral e social da opacidade, onde a falta de transparência perpetuou a discriminação racial. Conclui-se, portanto, que a XAI é a ferramenta primária de justiça, pois sua aplicação permite auditar a contribuição de variáveis sensíveis e assegurar que os Sistemas de Informação atuem conforme os princípios de Justiça e Não-Discriminação preconizados pela UNESCO (2021).

Uma entrega relevante deste trabalho, que complementa a discussão ética e técnica, é a análise dos principais marcos regulatórios brasileiros e sua convergência com o mecanismo da XAI. Demonstrou-se que a Lei Geral de Proteção de Dados (LGPD – Lei nº 13.709/2018), ao garantir no seu Art. 20 o direito de revisão de decisões tomadas unicamente por meios automatizados, impõe um ônus de prova e transparência que não pode ser atendido pelos sistemas opacos. Paralelamente, a proposta de Marco Legal da Inteligência Artificial (PL nº

2338/2023), que está em tramitação no Congresso, reforça essa exigência ao classificar os sistemas por níveis de risco e demandar documentação e auditabilidade para o alto risco. A convergência dessas legislações cria uma exigência de conformidade considerável no Brasil. A Inteligência Artificial Explicável (XAI), nesse contexto, torna-se o mecanismo técnico indispensável para fornecer a transparência e a auditabilidade necessárias para assegurar a prestação de contas em Sistemas de Informação que decidem sobre direitos fundamentais.

A relevância deste trabalho reside em buscar preencher a lacuna entre o conhecimento técnico da área de Sistemas de Informação e as novas exigências da regulamentação jurídica brasileira. Ao mapear a XAI como uma necessidade de prestação de contas e um requisito de governança, o estudo fornece um fundamento teórico essencial para a adoção de metodologias de desenvolvimento responsáveis em sistemas críticos. Contudo, é fundamental reconhecer as limitações presentes neste estudo. Por se tratar de uma pesquisa exclusivamente bibliográfica e documental, de natureza explicativa, o trabalho não incluiu a etapa de coleta de dados primários ou a implementação prática das técnicas LIME e SHAP em um sistema real de IA. Dessa forma, a análise se concentrou na validação do conceito e na justificação da necessidade regulatória, sem explorar o desafio operacional de sua aplicação em escala.

Para avanço contínuo desta área de estudo, a agenda de pesquisas futuras deve concentrar-se na validação prática dos mecanismos de explicabilidade. Sugere-se prioritariamente, a implementação e o teste das ferramentas LIME e SHAP em modelos de *Deep Learning* aplicados a Sistemas de Informação que possuem alto impacto social como análises em áreas financeira ou judicial. Esta abordagem permitiria quantificar a eficácia da XAI na detecção de vieses e compararia as explicações algorítmicas com a capacidade de interpretação humana. Por fim, recomenda-se o desenvolvimento de métricas específicas para analisar e a compreensibilidade das explicações geradas para usuários que não são especialistas em tecnologia, permitindo que a XAI cumpra, na prática, seu papel de transparência e governança.

6. REFERÊNCIAS

ALVES, M. A. S.; ANDRADE, O. M. de. **Da "Caixa-Preta" à "Caixa de Vidro":** o Uso da Explainable Artificial Intelligence (XAI) para Reduzir a Opacidade e Enfrentar o Enviesamento em Modelos Algorítmicos**.** RDP, Brasília, v. 18, n. 100, p. 349-373, out./dez. 2021.

AMOROSO, F. S. **Inteligência Artificial Explicável com LIME e SHAP Aplicada à Rede Neural Convolucional**. 2023. 48 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Bauru, 2023.

ANGWIN, J. et al. Machine Bias: risk assessments in criminal sentencing**.** **ProPublica**, [S. l.], 23 maio 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 01 nov. 2025.

ARRIETA, A. B. et al. Explanable Artificial Intelligence (XAI)**:** Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, [S. l.], v. 58, p. 82–115, jun. 2020. Disponível em: https://arxiv.org/abs/1910.10045. Acesso em: 27 out. 2025.

BAIN & COMPANY. **Cerca de 85% das empresas consideram adotar a inteligência artificial nos próximos anos, mostra pesquisa da Bain**. Disponível em: https://www.bain.com/pt-br/about/media-center/press-releases/south-america/2023/cerca-de-85-das-empresas-consideram-adotar-a-inteligencia-artificial-nos-proximos-anos-mostra-pesquisa-da-bain/. Acesso em: 14 out. 2025.

BENDER, E. M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT), 2021, Virtual. **Proceedings...** New York: ACM, 2021. p. 610–623. DOI: 10.1145/3442188.3445922. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3442188.3445922. Acesso em: 23 out. 2025.

BRAGA, A. V. et al. **Machine learning:** O Uso da Inteligência Artificial na Medicina/ Machine learning: The Use of Artificial Intelligence in Medicine. Brazilian Journal of Development, São José dos Pinhais, v. 5, n. 9, p. 16407-16413, set. 2019. Disponível em: https://doi.org/10.34117/bjdv5n9-190. Acesso em: 19 out. 2025.

BRAGA, T. D. **Impacto da inteligência artificial nos negócios:** uma estratégia inteligente e cada vez menos artificial. 2024. 46 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecânica) - Universidade Federal de Uberlândia, Uberlândia, 2024.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, ano 155, n. 159, p. 59, 15 ago. 2018.

BRASIL. Supremo Tribunal Federal. **STF amplia uso de inteligência artificial em apoio à atividade jurisdicional**. Brasília, 25 set. 2025. Disponível em: https://noticias.stf.jus.br/postsnoticias/stf-amplia-uso-de-inteligencia-artificial-em-apoio-a-atividade-jurisdicional/. Acesso em: 31 out. 2025.

BRASIL. Projeto de Lei nº 2338, de 2023. **Dispõe sobre o uso da Inteligência Artificial**. Senado Federal, Brasília, DF, 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 01 nov. 2025.

CHU, L. **Model Explainability**. 2022. Disponível em: https://pub.towardsai.net/model-explainability-shap-vs-lime-vs-permutation-feature-importance-98484efba066. Acesso em: 28 out. 2025.

DAMACENO, S. S.; VASCONCELOS, R. O. Inteligência Artificial: uma breve abordagem sobre seu conceito real e o conhecimento popular. **Caderno De Graduação - Ciências Exatas E Tecnológicas - UNIT - SERGIPE**, Aracaju, v. 5, n. 1, p. 11, out. 2018. Disponível em: https://periodicos.grupotiradentes.com/cadernoexatas/article/view/5729. Acesso em: 21 out. 2025.

FLORIDI, L. et al. AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. **Minds and Machines**, Dordrecht, v. 28, n. 4, p. 689–707, nov. 2018.

GUIDOTTI, R. et al. A Survey of Methods for Explaining Black Box Models. **ACM Computing Surveys (CSUR)**, New York, v. 51, n. 5, art. 93, p. 1-42, ago. 2018. DOI: 10.1145/3236009. Disponível em: https://doi.org/10.1145/3236009. Acesso em: 20 out. 2025.

IBM. **Explanable AI (XAI)**. 2022. Disponível em: https://www.ibm.com/think/topics/explainable-a. Acesso em: Acesso em: 28 out. 2025.

LEVY, G.; ADANIYA, M. H. XAI: esclarecendo o problema da caixa preta com transparência e interpretabilidade. **Revista Terra & Cultura: Cadernos De Ensino E Pesquisa**, Londrina, v. 40, n. especial, p. 346-371, ago. 2024. Disponível em: http://periodicos.unifil.br/index.php/Revistateste/article/view/3170/2920. Acesso em: 20 out. 2025.

LUNDBERG, S. M.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions.** 2017. Disponível em: https://arxiv.org/abs/1705.07874. Acesso em: 20 nov. 2025.

MORAIS, F. D. B.; CASTELO BRANCO, V. R. **A Inteligência Artificial:** conceitos, aplicações e controvérsias. In: SIMPÓSIO INTERNACIONAL DE CIÊNCIAS INTEGRADAS DA UNAERP - CAMPUS GUARUJÁ, 20., 2023, Guarujá. **Anais [...].** Guarujá: UNAERP, 2023. Disponível em: https://www.unaerp.br/documentos/5528-a-inteligencia-artificial-conceitos-aplicacoes-e-controversias/file. Acesso em: 14 out. 2025.

MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. iml: An R package for Interpretable Machine Learning. **Journal of Open Source Software**, [S.l.], v. 3, n. 26, p. 786,

2018. DOI: 10.21105/joss.00786. Disponível em: https://doi.org/10.21105/joss.00786. Acesso em: 23 out. 2025.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA (UNESCO). **Recomendação sobre a Ética da Inteligência Artificial**. Paris: UNESCO, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137. Acesso em: 01 nov. 2025.

REZK, N. G. et al. XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. **Bioengineering**, [S.l.], v. 11, n. 10, p. 1016, out. 2024. DOI: 10.3390/bioengineering11101016. Disponível em: https://www.mdpi.com/2306-5354/11/10/1016. Acesso em: 28 out. 2025.

RIBEIRO, M. **LIME - Local Interpretable Model-Agnostic Explanations**. 2016. Disponível em: https://homes.cs.washington.edu/~marcotcr/blog/lime/. Acesso em: 28 out. 2025.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York: ACM, 2016. p. 1135-1144. DOI: 10.1145/2939672.2939778.

SALIH, A. M. et al. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. **Advanced Intelligent Systems**, [S.l.], v. 7, p. 2400304, 2025. DOI: 10.1002/aisy.202400304. Disponível em: https://arxiv.org/abs/2305.02012. Acesso em: 28 out. 2025.

SURD EN, H. Machine learning and law. **Washington Law Review**, Seattle, v. 89, n. 1, p. 87-133, 2014. Disponível em: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5. Acesso em: 23 out. 2025.

TACCA, A.; ROCHA, L. S. Inteligência artificial: reflexos no sistema do direito. **NOMOS: Revista do Programa de Pós-Graduação em Direito da UFC**, Fortaleza, v. 38, n. 2, p. 53-68, jul./dez. 2018.

VISWANATH, S.; KHANNA, V.; LIANG, Y. **AI: The Coming Revolution**. Coatue Management, 2023. Disponível em: https://www.coatue.com/blog/perspective/ai-the-coming-revolution-2023. Acesso em: 14 out. 2025.