Da Caixa-Preta à Explicabilidade: Um Estudo sobre Transparência em Sistemas de Inteligência Artificial

Carlos Eduardo Rehbein de Souza, Letízia Manuella Serqueira Eugênio, Emilli Dias de Oliveira, Nelcileno Virgílio de Souza Araújo

[1]Instituto de Computação - Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Corrêa da Costa, nº 2367, Boa Esperança – UFMT, Cuiabá – MT– Brasil.

{carlos.souza5,letizia.eugenio}@sou.ufmt.br, emillidias.o@outlook.com, nelcileno@ic.ufmt.br

Abstract. *The increasing use of Artificial Intelligence (AI) in high-impact decisions exposes the "black box" problem, characterized by the difficulty of understanding complex models. This opacity undermines trust and raises technical and regulatory challenges under laws such as LGPD and GDPR. Explainable AI (XAI) seeks to address this by balancing performance and interpretability through interpretable models and post-hoc methods like LIME and SHAP. This paper presents a critical review of the field, outlining strengths, limitations, and perspectives. It concludes that algorithmic transparency is crucial to foster trust, auditability, and the responsible adoption of AI.*

Resumo. *A crescente adoção de Inteligência Artificial (IA) em decisões de alto impacto evidencia o problema da "caixa-preta", marcado pela dificuldade de compreender critérios de modelos complexos. Essa opacidade compromete a confiança dos usuários e gera desafios técnicos e regulatórios diante de legislações como a LGPD e o GDPR. A Explainable AI (XAI) surge como resposta, buscando conciliar desempenho e interpretabilidade por meio de modelos explicáveis e técnicas pós-hoc, como LIME e SHAP. Este trabalho realiza revisão crítica do tema, destacando potencialidades, limitações e perspectivas. Conclui-se que a transparência algorítmica é essencial para a confiança, auditabilidade e adoção responsável da IA.*

1. Introdução

A rápida evolução da Inteligência Artificial (IA) e a digitalização crescente das interações humanas têm impulsionado a coleta e o processamento de grandes volumes de dados pessoais, transformando-os em ativos estratégicos para organizações e governos [Lorenzon 2021]. Embora esses sistemas tragam ganhos expressivos em termos de eficiência e automação, muitos operam como verdadeiras *caixas-pretas*, utilizando modelos complexos de aprendizado de máquina cujos critérios de decisão são de difícil interpretação até mesmo para especialistas [Goodfellow et al. 2016, Arbix 2020]. Essa opacidade levanta preocupações quanto à confiança, auditabilidade e impacto social das decisões automatizadas.

Nesse cenário, surge a necessidade de desenvolver métodos que possibilitem maior transparência e interpretabilidade dos modelos. A área de *Explainable Artificial Intelligence* (XAI) vem se consolidando como um campo promissor ao propor tanto mode-

los intrinsicamente interpretáveis quanto técnicas pós-hoc, como LIME e SHAP, que buscam explicar decisões de modelos complexos [Guidotti et al. 2018, Ribeiro et al. 2016, Lundberg and Lee 2017]. Essas abordagens têm sido aplicadas em áreas críticas como saúde, finanças e segurança, onde a falta de explicabilidade pode comprometer a confiança dos usuários e a adoção prática de sistemas inteligentes.

Apesar dos avanços, ainda persiste o dilema entre desempenho e interpretabilidade. Modelos altamente complexos oferecem maior acurácia preditiva, mas reduzem a capacidade de compreensão, enquanto métodos explicáveis, embora mais transparentes, tendem a apresentar limitações em tarefas de alta complexidade [Doshi-Velez and Kim 2017]. Além disso, legislações recentes, como a GDPR na União Europeia e a LGPD no Brasil, reforçam a demanda por maior clareza nos critérios de decisão automatizada [Lorenzon 2021, Caetano 2020, Brasil 2018], tornando a explicabilidade não apenas uma questão técnica, mas também regulatória e social.

Contudo, observa-se que a literatura brasileira ainda carece de estudos que articulem de forma integrada os aspectos técnicos da XAI com as exigências regulatórias da LGPD e os impactos sociais da falta de transparência. A maioria das pesquisas nacionais limita-se a análises jurídicas ou conceituais, sem propor uma síntese crítica que una ciência da computação, ética e governança de dados.

Diante desse contexto, este artigo tem como objetivo oferecer uma análise crítica e integradora sobre a transparência algorítmica em sistemas de IA, com foco nas técnicas de XAI e suas implicações no contexto brasileiro. Busca-se identificar lacunas e tendências na literatura, discutir os principais métodos de explicabilidade e refletir sobre o papel da transparência na construção de sistemas mais confiáveis, auditáveis e socialmente responsáveis.

2. Metodologia

Este artigo configura-se como uma pesquisa de natureza **exploratória e qualitativa**, fundamentada em uma abordagem **bibliográfica**. De acordo com [Wazlawick 2009], a pesquisa em Ciência da Computação pode ser classificada quanto aos objetivos, procedimentos e abordagem. Neste caso, optou-se por uma investigação exploratória, cujo propósito é ampliar a compreensão sobre o problema da opacidade algorítmica e as soluções propostas pela área de *Explainable Artificial Intelligence* (XAI).

As fontes consultadas incluem artigos científicos, livros e relatórios técnicos publicados em português e inglês, no período aproximado de 2015 a 2025. Foram priorizados estudos que abordassem o problema da caixa-preta em modelos de IA, a transparência algorítmica e seus impactos sociais, bem como as principais técnicas e métodos de XAI — como modelos interpretáveis, LIME e SHAP. Também foram incluídas contribuições que relacionam a explicabilidade às exigências regulatórias e às demandas sociais decorrentes da LGPD e do GDPR.

A coleta foi realizada em bases científicas nacionais e internacionais, como **Google Scholar, IEEE Xplore, ACM Digital Library, Scopus e SpringerLink**. Utilizaram-se palavras-chave em português e inglês, tais como: *"explainable artificial intelligence"*, *"black-box models"*, *"algorithmic transparency"*, *"LIME"*, *"SHAP"* e *"interpretable machine learning"*.

Os critérios de seleção consideraram: (i) relevância direta com o tema da explicabilidade e transparência algorítmica; (ii) consistência conceitual e metodológica; e (iii) contribuição para a compreensão técnica ou social da XAI. Trabalhos redundantes, superficiais ou que tratavam apenas de aspectos periféricos foram excluídos.

Para a análise, adotou-se a técnica de **análise temática** descrita por [Braun and Clarke 2006], que consiste em identificar padrões conceituais e recorrências nos estudos revisados, agrupando-os em categorias de sentido. As categorias emergentes que estruturam a discussão deste artigo foram:

1. Problema da caixa-preta em IA;
2. Transparência algorítmica e impactos sociais;
3. Métodos e técnicas de XAI;
4. Desafios e perspectivas futuras.

Essa abordagem possibilitou mapear os principais eixos conceituais que orientam a literatura recente, bem como identificar lacunas relacionadas à integração entre aspectos técnicos, regulatórios e sociais da explicabilidade em Inteligência Artificial.

3. A Caixa-Preta da Inteligência Artificial

O avanço da Inteligência Artificial (IA) nas últimas décadas levou ao desenvolvimento de modelos cada vez mais sofisticados, em especial as redes neurais profundas e os métodos de *ensemble*, que revolucionaram tarefas como visão computacional, processamento de linguagem natural e reconhecimento de padrões em larga escala [Goodfellow et al. 2016]. Esses modelos apresentam desempenho superior em termos de acurácia, porém sacrificam a capacidade de explicação de seus processos internos, resultando no chamado problema da caixa-preta.

O termo refere-se à dificuldade de compreender como entradas específicas influenciam as saídas produzidas, uma vez que a lógica de decisão é distribuída em milhares ou milhões de parâmetros ajustados durante o treinamento. Essa opacidade se torna particularmente problemática em cenários críticos, como saúde, finanças e segurança pública, onde decisões erradas ou enviesadas podem gerar impactos severos [Arbix 2020].

Estudos recentes mostram que a falta de interpretabilidade compromete não apenas a confiança dos usuários, mas também a adoção prática dos sistemas. Markus, Kors e Rijnbeek (2020), ao analisarem a aplicação de IA na área da saúde, apontam que profissionais relutam em utilizar algoritmos que não conseguem explicar, mesmo quando esses apresentam alta performance [Markus et al. 2020]. Da mesma forma, Lehmann et al. (2022) destacam que a percepção da complexidade ou simplicidade do modelo influencia o quanto os usuários aceitam ou rejeitam suas recomendações, revelando que a transparência percebida pode ser tão importante quanto a transparência técnica [Lehmann et al. 2022].

Além disso, o dilema entre desempenho e interpretabilidade permanece como uma barreira teórica e prática. Modelos mais simples, como árvores de decisão ou regressões, fornecem clareza e auditabilidade, mas geralmente não atingem o nível de acurácia necessário em aplicações de alta complexidade. Já os modelos mais avançados oferecem resultados superiores, mas funcionam como caixas-pretas, dificultando a validação de suas escolhas. Essa tensão é amplamente discutida na literatura como um dos principais desafios da ciência da computação contemporânea [Doshi-Velez and Kim 2017].

Portanto, compreender o problema da caixa-preta é essencial para a evolução da IA em contextos críticos. Essa discussão fundamenta a busca por abordagens que conciliem desempenho e interpretabilidade, pavimentando o caminho para o desenvolvimento de técnicas de *Explainable AI* (XAI).

4. Transparência Algorítmica e Aceitação Social

A crescente presença da Inteligência Artificial em processos decisórios de alto impacto evidencia a necessidade de transparência nos sistemas computacionais. Transparência algorítmica refere-se à capacidade de compreender como um modelo opera, quais dados utiliza e de que forma gera suas previsões ou recomendações. Trata-se de um requisito técnico e social fundamental para mitigar riscos, aumentar a confiança dos usuários e garantir que os sistemas estejam em conformidade com princípios éticos e regulatórios [Lorenzon 2021].

Lorenzon (2021) argumenta que a falta de transparência compromete o direito dos cidadãos à autodeterminação informacional, ao dificultar a identificação de falhas ou abusos em processos automatizados. Moura, Corrales e Doneda (2021) reforçam esse ponto, destacando que decisões opacas prejudicam a responsabilização e a possibilidade de revisão, aspectos que são essenciais em uma sociedade orientada pela proteção de dados. De forma semelhante, Rossetti e Angeluci (2021) observam que sistemas de recomendação e classificadores sem explicações claras enfrentam menor aceitação social, pois os usuários tendem a desconfiar de processos que não podem ser auditados ou compreendidos.

No âmbito jurídico, legislações como o Regulamento Geral de Proteção de Dados (GDPR) na União Europeia e a Lei Geral de Proteção de Dados (LGPD) no Brasil estabelecem a necessidade de explicações claras e adequadas sobre os critérios utilizados em decisões automatizadas [Brasil 2018, Caetano 2020]. Ainda que formuladas em termos legais, essas exigências colocam para a ciência da computação desafios técnicos concretos: como construir sistemas capazes de justificar suas decisões de modo compreensível tanto para especialistas quanto para usuários leigos.

Pesquisas recentes também destacam o impacto da transparência na aceitação social da IA. Lehmann et al. (2022) demonstram que a percepção de complexidade ou simplicidade de um algoritmo influencia diretamente a disposição dos usuários em seguir suas recomendações. Quando os modelos são percebidos como "caixas-pretas", mesmo que apresentem alta acurácia, há maior resistência em aceitá-los. Já Markus, Kors e Rijnbeek (2020), ao analisarem aplicações em saúde, mostram que profissionais preferem utilizar sistemas explicáveis, mesmo que menos precisos, a depender exclusivamente de modelos de alto desempenho sem justificativa clara.

Dessa forma, a transparência algorítmica não deve ser entendida apenas como uma exigência regulatória, mas como um elemento estratégico para garantir confiança, auditabilidade e aceitação social dos sistemas de IA. Ela constitui um elo entre os avanços técnicos e a sua aplicação responsável, abrindo caminho para que a Explainable AI (XAI) se consolide como resposta aos desafios impostos pela opacidade dos algoritmos.

5. Técnicas de Explainable AI (XAI)

O campo de *Explainable Artificial Intelligence* (XAI) emergiu como resposta ao problema da caixa-preta e às demandas de transparência algorítmica. Seu objetivo é fornecer

mecanismos que permitam compreender, auditar e validar o comportamento de modelos de IA, sem comprometer de maneira significativa sua performance [Guidotti et al. 2018]. A literatura divide os métodos de XAI em duas grandes categorias: modelos intrinsicamente interpretáveis e métodos pós-hoc.

5.1. Modelos Intrinsicamente Interpretáveis

Modelos considerados "transparentes por design" são aqueles cuja estrutura matemática é naturalmente compreensível. Árvores de decisão, regressões lineares e regras de associação são exemplos clássicos dessa categoria. Esses modelos possibilitam rastrear diretamente a relação entre variáveis de entrada e saída, favorecendo a auditabilidade e a confiabilidade [Molnar 2019]. No entanto, sua aplicabilidade é limitada em tarefas de alta complexidade, nas quais a precisão pode ser significativamente inferior à de redes neurais profundas ou métodos de *ensemble*.

5.2. Métodos Pós-hoc

Quando o modelo utilizado é complexo, como no caso de redes neurais ou florestas de decisão, recorre-se a métodos pós-hoc, que visam gerar explicações aproximadas sem alterar a estrutura original. Entre os mais influentes estão:

- **LIME (Local Interpretable Model-Agnostic Explanations)**: proposto por Ribeiro, Singh e Guestrin (2016), o método cria modelos locais simples para explicar decisões individuais de classificadores complexos, utilizando amostras perturbadas dos dados de entrada.
- **SHAP (SHapley Additive exPlanations)**: desenvolvido por Lundberg e Lee (2017), fundamenta-se na teoria dos jogos cooperativos para atribuir valores de importância a cada característica, garantindo consistência e equidade na interpretação das contribuições.

Esses métodos tornaram-se amplamente utilizados por sua versatilidade, mas enfrentam críticas relacionadas à estabilidade das explicações geradas e ao custo computacional, sobretudo em cenários de grande escala [Molnar 2019].

Para sintetizar as principais características das técnicas discutidas, apresenta-se a Tabela 1, que resume os métodos de XAI mais utilizados, suas vantagens e limitações.

Table 1. Resumo das principais técnicas de XAI

Método	Tipo	Vantagens	Limitações
Árvore de decisão	Intrínseco	Fácil de interpretar	Baixa performance em problemas complexos
LIME	Pós-hoc	Explicações locais, modelo-agnóstico	Instabilidade, alto custo computacional
SHAP	Pós-hoc	Baseado em teoria dos jogos, consistente	Difícil escalabilidade, pesado para grandes modelos
Regras de associação	Intrínseco	Transparência clara	Pouca aplicabilidade em dados não estruturados

5.3. Perspectivas Recentes e Desafios

A literatura contemporânea amplia o escopo da XAI, explorando desde visualizações interativas até métricas para avaliação da qualidade das explicações. Doshi-Velez e Kim (2017) defendem a necessidade de uma ciência rigorosa da interpretabilidade, com padrões objetivos para avaliar fidelidade, clareza e estabilidade. Sun et al. (2024) apontam que, no contexto da IA generativa, a ausência de mecanismos de explicação adequados amplia os riscos de desinformação. Zhou et al. (2024), por sua vez, destacam que a opacidade algorítmica tende a reproduzir vieses e erros sistêmicos, reforçando a urgência de soluções explicáveis.

Apesar dos avanços, persiste o dilema entre desempenho e interpretabilidade. A busca por métodos híbridos, que combinem alta acurácia com explicações compreensíveis, constitui atualmente uma das principais fronteiras de pesquisa em XAI [Guidotti et al. 2018, Doshi-Velez and Kim 2017].

6. Síntese da Revisão

A revisão bibliográfica realizada permitiu identificar um conjunto de padrões e tendências recorrentes na literatura sobre transparência e explicabilidade em sistemas de Inteligência Artificial (IA). Observou-se que o debate sobre a "caixa-preta" tem evoluído de uma discussão puramente técnica para uma abordagem interdisciplinar, que integra aspectos éticos, regulatórios e sociais. Essa transição marca uma mudança significativa no foco da pesquisa contemporânea em XAI.

De modo geral, a literatura analisada evidencia três eixos principais de investigação:

1. **Dimensão técnica:** concentrada no desenvolvimento de métodos de explicabilidade, como LIME e SHAP, e na criação de métricas para avaliar a fidelidade e a clareza das explicações [Guidotti et al. 2018, Ribeiro et al. 2016, Lundberg and Lee 2017]. Essa vertente tem buscado conciliar desempenho e interpretabilidade, ainda que o dilema entre ambos permaneça sem solução definitiva [Doshi-Velez and Kim 2017].
2. **Dimensão regulatória:** impulsionada por legislações como o GDPR e a LGPD, que exigem explicações compreensíveis sobre decisões automatizadas [Lorenzon 2021, Brasil 2018, Caetano 2020]. Essa perspectiva evidencia a necessidade de traduzir princípios legais em soluções computacionais concretas.
3. **Dimensão social e ética:** que relaciona a transparência à confiança e à aceitação social de sistemas de IA. Pesquisas apontam que a percepção de opacidade reduz a disposição dos usuários em aceitar recomendações automatizadas, mesmo quando tecnicamente precisas [Markus et al. 2020, Lehmann et al. 2022, Rossetti and Angeluci 2021].

A análise também revelou **lacunas relevantes**. Há escassez de estudos aplicados no contexto brasileiro, especialmente quanto à implementação prática de técnicas de XAI para atender aos requisitos de clareza e auditabilidade previstos na LGPD. Além disso, poucos trabalhos discutem como adaptar explicações técnicas para públicos não especializados, o que limita a aplicabilidade social da explicabilidade algorítmica.

Por outro lado, identificam-se **tendências emergentes** que indicam novas direções de pesquisa. Entre elas, destacam-se o uso de modelos híbridos que buscam conciliar interpretabilidade e desempenho, o desenvolvimento de visualizações interativas para apoiar a compreensão humana e o avanço das discussões sobre transparência em sistemas de IA generativa [Sun et al. 2024, Zhou et al. 2024].

Em síntese, a revisão demonstra que a XAI se consolidou como um campo interdisciplinar e estratégico para o futuro da IA. Contudo, ainda enfrenta desafios quanto à padronização de métodos, à validação das explicações e à tradução efetiva de seus resultados para contextos sociais e normativos diversos.

A partir dessa síntese, a próxima seção aprofunda a discussão crítica dos achados, relacionando os avanços técnicos às implicações éticas, regulatórias e sociais da Inteligência Artificial explicável.

7. Discussão Crítica

A análise realizada evidencia que os avanços da Inteligência Artificial têm proporcionado ganhos significativos em acurácia e eficiência, mas ao custo da interpretabilidade. Esse dilema, sintetizado pelo problema da caixa-preta, levanta implicações que transcendem a esfera técnica, alcançando dimensões éticas, sociais e regulatórias. A literatura revisada mostra um consenso em torno da necessidade de maior transparência, mas diverge em relação às abordagens mais eficazes para alcançá-la.

Um primeiro ponto crítico refere-se ao *trade-off* entre desempenho e interpretabilidade. Modelos complexos, como redes neurais profundas, apresentam resultados superiores em *benchmarks*, mas sua natureza opaca limita a auditabilidade. Já os modelos intrinsicamente interpretáveis oferecem explicações claras, mas não atingem a mesma precisão em tarefas de alta complexidade [Doshi-Velez and Kim 2017, Molnar 2019]. Essa tensão permanece como uma das barreiras centrais na adoção de IA em cenários críticos.

Outro aspecto diz respeito à aceitação social e confiança. Pesquisas mostram que usuários e profissionais tendem a preferir sistemas que ofereçam algum grau de explicação, ainda que menos precisos, em detrimento de soluções de alto desempenho mas opacas [Markus et al. 2020, Lehmann et al. 2022]. Isso sugere que a percepção de transparência pode ser tão relevante quanto a transparência técnica, o que amplia a responsabilidade dos pesquisadores em desenvolver interfaces de explicação acessíveis a diferentes perfis de usuário.

Do ponto de vista regulatório, legislações como o GDPR e a LGPD estabelecem obrigações relacionadas à explicabilidade de decisões automatizadas [Brasil 2018, Caetano 2020]. No entanto, persiste a lacuna entre a linguagem legal e as soluções técnicas efetivas. A literatura indica que ainda não há consenso sobre quais métodos de XAI atendem de forma robusta aos requisitos de clareza, completude e auditabilidade exigidos por esses marcos normativos [Moura et al. 2021, Rossetti and Angeluci 2021].

Por fim, cabe destacar os desafios emergentes. No contexto da IA generativa, a opacidade dos modelos amplia riscos relacionados à desinformação e à reprodução de vieses [Sun et al. 2024, Zhou et al. 2024]. Ao mesmo tempo, cresce a demanda por métodos híbridos, que conciliem acurácia e explicabilidade de forma equilibrada [Guidotti et al. 2018]. Esses pontos indicam que a área de XAI se encontra em um estágio

de consolidação, mas ainda distante de oferecer soluções universais.

Assim, a discussão crítica aponta que os avanços em XAI representam um passo importante, mas insuficiente diante das múltiplas camadas de impacto da IA na sociedade. O desafio contemporâneo não se limita a desenvolver técnicas interpretáveis, mas a integrá-las de forma efetiva às exigências sociais, regulatórias e éticas que cercam o uso da tecnologia.

8. Conclusão

Este trabalho analisou o desafio da transparência algorítmica em sistemas de Inteligência Artificial, com ênfase no problema da caixa-preta e nas abordagens de *Explainable Artificial Intelligence* (XAI). A revisão da literatura evidenciou que a sofisticação crescente de modelos de aprendizado de máquina, em especial redes neurais profundas, amplia a acurácia preditiva, mas compromete a interpretabilidade, criando um dilema central para a ciência da computação contemporânea.

A análise mostrou que a falta de transparência não é apenas uma limitação técnica, mas também uma barreira à aceitação social e à confiança dos usuários. Pesquisas recentes apontam que a percepção de explicabilidade pode ser decisiva para a adoção de sistemas de IA, mesmo em detrimento de desempenho superior. Esse aspecto conecta-se às exigências regulatórias de legislações como o GDPR e a LGPD, que reforçam a necessidade de explicações claras e auditáveis sobre decisões automatizadas.

As técnicas de XAI, tanto os modelos intrinsicamente interpretáveis quanto os métodos pós-hoc como LIME e SHAP, representam avanços relevantes, mas ainda enfrentam limitações quanto à estabilidade das explicações, à escalabilidade e à clareza para diferentes perfis de usuários. Estudos recentes indicam a necessidade de novas abordagens híbridas que conciliem acurácia e interpretabilidade, bem como métricas padronizadas para avaliação da qualidade das explicações.

A discussão crítica permitiu identificar que os desafios da XAI não se restringem ao desenvolvimento de algoritmos explicáveis, mas abrangem também a integração entre ciência da computação, regulamentação e ética. Em especial, o avanço da IA generativa reforça a urgência de mecanismos de transparência robustos, capazes de mitigar riscos de desinformação e vieses sistêmicos.

Dessa forma, conclui-se que a transparência algorítmica constitui um requisito estratégico para a consolidação da Inteligência Artificial como tecnologia confiável, auditável e socialmente legítima. A evolução da área depende não apenas de progressos técnicos, mas também de um esforço interdisciplinar que alinhe desempenho, interpretabilidade e responsabilidade social no desenvolvimento de sistemas inteligentes.

Referências

Arbix, G. (2020). Inteligência artificial e o desafio das caixas-pretas. *Revista USP*, (127):15–28.

Brasil (2018). Lei nª 13.709, de 14 de agosto de 2018. dispõe sobre a proteção de dados pessoais e altera a lei nª 12.965, de 23 de abril de 2014 (marco civil da internet). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Caetano, J. V. L. (2020). O regulamento geral de proteção de dados (gdpr). *Cadernos Eletrônicos Direito Internacional Sem Fronteiras*, 2(1):45–60.

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

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42.

Lehmann, T., Haubitz, C., Fügener, A., and Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management*, 31(3):1094–1113.

Lorenzon, L. N. (2021). Análise comparada entre regulamentações de dados pessoais no brasil e na união europeia (lgpd e gdpr). *Revista do Programa de Direito da União Europeia*, 1:39–52.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777.

Markus, A., Kors, J., and Rijnbeek, P. (2020). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu Press.

Moura, M., Corrales, M., and Doneda, D. (2021). *Artificial Intelligence and Data Protection in Brazil*. Springer.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Rossetti, A. and Angeluci, A. (2021). Vieses algorítmicos e transparência em sistemas de inteligência artificial. *Revista de Sistemas e Computação*, 11(2):45–61.

Sun, T., Li, W., and Zhang, C. (2024). Safety and trust in generative ai: Challenges and opportunities. *Nature Machine Intelligence*, 6(1):15–28.

Wazlawick, R. S. (2009). *Metodologia de Pesquisa em Ciência da Computação*. Elsevier.

Zhou, Y., Wang, L., and Chen, H. (2024). The risks of algorithmic bias in large-scale ai systems. *IEEE Transactions on Artificial Intelligence*, 5(2):123–138.