Explainability and Privacy in AI-Based Decision Support Systems: A Structured Literature Analysis

Andressa Girotto Vargas[1][0009−0008−8192−2061] and Edna Dias Canedo[1][0000−0002−2159−339X]

[1] University of Brasília (UnB), Computer Science Department, Brasília–DF, Brazil
[2] andressagirottovargas@gmail.com, ednacanedo@unb.br

Abstract. Artificial Intelligence is increasingly embedded in decision support systems that operate on sensitive or regulated data, intensifying demands for transparency, accountability, and privacy protection. In these contexts, explainable AI mechanisms aim to clarify how models produce outcomes, while privacy-preserving techniques such as differential privacy and federated learning seek to restrict the exposure of sensitive information. Reconciling these objectives has become a central challenge in AI-enabled information systems. This paper presents a structured literature analysis of the relationship between explainability and privacy-preserving mechanisms in decision support systems. The study investigates how prior research characterizes compatibility conditions, trade-offs, and the implications of explainability for trust and accountability under privacy constraints. The results show that explainability and privacy preservation are not inherently incompatible, but their coexistence is highly context-dependent. The literature indicates that global and aggregated explanations are generally more compatible with privacy-sensitive environments, whereas local and instance-level explanations tend to increase privacy risks. The findings also highlight that effective explanations must be aligned with stakeholder needs and governance requirements, rather than simply maximizing transparency. The paper contributes an integrated view of technical and socio-technical trade-offs and identifies directions for the design of trustworthy and privacy-aware decision support systems.

Keywords: Explainable Artificial Intelligence · Privacy-Preserving Systems · Decision Support Systems · Trust and Accountability · Structured Literature Analysis

1 Introduction

The growing adoption of Machine Learning (ML) in information systems has intensified demands for transparency, accountability, and trustworthiness, particularly in decision support contexts involving sensitive data [6, 28]. These demands are reinforced by data protection and AI governance frameworks such

as the General Data Protection Regulation (GDPR) [12], the Brazilian General Data Protection Law (LGPD) [15], and the European AI Act [21], which introduce obligations related to transparency, accountability, and, in some contexts, explanations of automated decisions [16, 25]. Similar expectations are also reflected in soft-law instruments, including the UNESCO Ethical Recommendations on AI [24], the OECD Principles for AI Development [11], and the NIST AI Risk Management Framework [23], all of which recognize explainability as an important element of trustworthy AI systems.

In this context, Explainable Artificial Intelligence (XAI) has been advanced as a key mechanism for enabling stakeholders to understand, contest, and audit automated decisions [6]. At the same time, many contemporary ML-based systems operate under strict privacy constraints and rely on privacy-preserving techniques such as differential privacy, federated learning, secure computation, and data anonymization [2]. This creates a fundamental tension: whereas XAI seeks to reveal information about model behavior, privacy-preserving techniques are designed to restrict disclosure.

Recent studies show that this tension is not merely conceptual. Explanations may themselves become vectors for privacy leakage by exposing information about training data, decision boundaries, or sensitive attributes [3]. In particular, feature importance scores and counterfactual explanations have been associated with risks such as membership inference, model extraction, and inversion attacks when deployed without appropriate safeguards [5, 8]. These findings challenge the assumption that explainability and privacy are naturally compatible and indicate the need to better understand the trade-offs involved in combining them.

The challenge also extends beyond the model level. In distributed and regulated environments, organizations require not only privacy protection, but also traceability, auditability, and policy compliance. Some recent approaches suggest that explainability can support accountability and governance when combined with mechanisms such as federated learning, access control, and auditable infrastructures [13]. At the same time, the literature continues to emphasize explainability as a relevant driver of trust and adoption in decision support systems [19], although its effects depend on factors such as context, explanation type, and stakeholder expertise [28]. In privacy-sensitive domains, such as healthcare and public services, explainability is therefore positioned between two competing demands: increasing transparency and preserving confidentiality [18].

Despite the growing body of work on privacy-aware explainability, the literature remains fragmented. Many studies focus either on the technical compatibility between XAI and privacy-preserving mechanisms, or on trust- and accountability-related outcomes in decision support contexts, with limited integration across these perspectives. As a result, there is still no consolidated view of which XAI mechanisms are more compatible with privacy-preserving techniques, what types of trade-offs are most frequently reported, and how these tensions affect trust and accountability in privacy-preserving decision support systems. Motivated by this gap, this paper presents a structured literature-based analysis of the interplay between XAI and privacy-preserving techniques

in information systems. The study is guided by the following research question: **RQ1. How does the literature characterize the compatibility, trade-offs, and trust/accountability implications of explainability in privacy-preserving decision support systems?**

To address this question, we synthesize representative studies on XAI in privacy-preserving settings and organize the discussion around four analytical dimensions: XAI mechanisms, privacy-preserving techniques, reported trade-offs, and stakeholder implications. By doing so, the paper contributes: (i) a structured synthesis of the literature on explainability in privacy-preserving decision support systems; (ii) an integrated view of the technical and organizational trade-offs reported in prior work; and (iii) a discussion of how these trade-offs shape trust and accountability under privacy constraints.

2 Background and Related Work

This section reviews the theoretical and empirical foundations concerning the interaction between Explainable Artificial Intelligence (XAI) and privacy-preserving mechanisms in information systems. The literature increasingly recognizes that explainability and privacy protection introduce competing design requirements in AI-enabled decision support systems. To structure the discussion, we organize prior research around two complementary perspectives: (i) the compatibility and trade-offs between XAI mechanisms and privacy-preserving techniques, and (ii) the role of explainability in supporting trust and accountability in privacy-sensitive decision support systems.

2.1 XAI and Privacy-Preserving Techniques: Compatibility and Trade-offs

Privacy-preserving machine learning techniques such as Differential Privacy (DP), Federated Learning (FL), homomorphic encryption, secure computation, and data anonymization aim to reduce information leakage while enabling data-driven decision making. In contrast, XAI mechanisms deliberately expose aspects of model behavior to support transparency, interpretability, and accountability. This difference in design objectives creates a structural tension that has been increasingly acknowledged in recent literature.

Several studies demonstrate that explanations themselves may introduce new privacy risks by revealing sensitive patterns, feature contributions, or decision boundaries [3, 20, 22]. In particular, *post-hoc* explanation methods such as feature importance scores and counterfactual explanations may facilitate attacks including membership inference, model extraction, or model inversion when deployed without appropriate safeguards [8]. These findings challenge the assumption that explainability is inherently benign and highlight the need to systematically examine its interaction with privacy-preserving mechanisms.

Recent research has moved beyond conceptual discussions and started to empirically analyze the coexistence of explainability and privacy-preserving techniques. Senevirathna et al. propose a model-agnostic development process that

integrates accountability, resilience, and privacy as first-class design objectives alongside predictive performance [17]. Their results show that explanation techniques such as SHAP, LIME, and Layer-wise Relevance Propagation (LRP) can operate in privacy-preserving environments, including settings using DP or FL, but introduce measurable trade-offs. For example, stronger privacy guarantees may reduce predictive utility, while explanations may still expose abnormal or adversarial behavior through attribution stability and prediction diversity analyses.

Complementary technical studies further investigate the relationship between explanation fidelity and privacy guarantees under formal constraints. These works show that enforcing strong privacy protection can significantly degrade explanation precision, reinforcing the need to explicitly reason about explainability–privacy trade-offs during system design [4, 1]. Overall, existing evidence suggests that integrating XAI with privacy-preserving techniques is feasible but inherently conditional. The nature and magnitude of the resulting trade-offs depend on factors such as explanation type, privacy budget, learning paradigm (e.g., centralized versus federated), and the intended stakeholder audience. Despite these advances, prior work remains fragmented, often focusing on isolated techniques or specific domains. As a result, there is still limited guidance on how to systematically select and integrate XAI mechanisms in privacy-sensitive information systems.

2.2 Explainability, Trust, and Accountability in Privacy-Preserving Decision Support Systems

Beyond technical compatibility, explainability plays a crucial socio-technical role in supporting trust and accountability in AI-enabled decision support systems. A growing body of research suggests that explanations can improve users' confidence in automated decisions, particularly in high-stakes domains. For instance, a systematic literature review reports that XAI techniques generally have a positive influence on human trust in machine learning models, although the magnitude of this effect depends on explanation type, task complexity, and user expertise [28].

In privacy-sensitive domains, explainability is closely connected to accountability rather than mere interpretability. Decision support systems used in contexts such as healthcare, public administration, and finance must provide explanations that allow auditing, contestation, and justification of decisions. At the same time, ethical and regulatory discussions emphasize that excessive transparency may conflict with data protection principles, while insufficient explainability may undermine accountability and legitimacy [18, 14]. Consequently, organizations must balance transparency requirements with confidentiality constraints.

Recent work also frames explainability as part of a broader accountability infrastructure for AI systems. For example, Senevirathna et al. operationalize accountability through measurable properties such as explanation stability, consistency across XAI techniques, and compactness, integrating these metrics into

an iterative development process [17]. Their findings suggest that different levels of explainability serve different accountability objectives: detailed explanations may assist developers and auditors, while more abstract explanations may support monitoring and governance under privacy constraints.

From an organizational perspective, recent studies further emphasize that explainability and privacy should not be maximized independently but strategically balanced. In particular, governance-oriented research conceptualizes privacy-aware explainability as a multi-level construct in which different stakeholders require different degrees of transparency to support trust, accountability, and regulatory compliance [9]. Despite these advances, the literature still lacks an integrated understanding of how varying levels of explainability influence stakeholders' trust and perceived accountability in privacy-preserving decision support systems. Many empirical studies examine either technical robustness or user trust in isolation, rarely addressing both dimensions simultaneously. This gap suggests the need for a more systematic synthesis of how explainability mechanisms, privacy-preserving techniques, and stakeholder expectations interact in operational information systems.

To support this analysis, Table 1 summarizes representative studies according to four analytical dimensions: XAI mechanisms, privacy-preserving techniques, reported trade-offs, and primary stakeholders. The synthesis highlights that prior research tends to emphasize either technical compatibility or trust-related implications, but rarely examines both dimensions in an integrated manner. The synthesis indicates that prior studies consistently report non-trivial trade-offs affecting model utility, explanation fidelity, and system complexity. At the same time, the implications for trust and accountability are often addressed only implicitly or for specific stakeholder groups. These limitations highlight the need for integrative analyses that jointly consider technical compatibility and governance implications when designing explainable AI systems under privacy constraints.

3 Research Method

This study adopts a structured literature analysis approach to investigate the interplay between Explainable Artificial Intelligence (XAI) mechanisms and privacy-preserving techniques in information systems. The objective is to synthesize existing knowledge on explainability privacy trade-offs and their implications for trust and accountability in decision support systems. Rather than conducting a full systematic literature review, this study follows a structured and targeted literature analysis approach commonly adopted in software engineering and information systems research when the objective is to consolidate conceptual insights across emerging research areas. This approach enables the identification and synthesis of representative studies addressing both technical and socio-technical aspects of explainability and privacy in AI-enabled systems.

Figure 1 summarizes the research design adopted in this study. The process involves three main stages: identification of relevant literature, comparative syn-

Table 1. Synthesis of prior work on XAI and privacy-preserving systems

Study	XAI Mechanisms	Privacy-Preserving Techniques	Key Trade-offs Identified	Primary Stakeholders
[8]	Feature importance, counterfactual explanations	Implicit privacy constraints, exposure analysis	Explanations may enable membership inference, model extraction, and inversion attacks; increased transparency can amplify privacy leakage	Service providers, auditors
[17]	SHAP, LIME, LRP; accountability metrics (e.g., stability/consistency)	Federated learning, access control, blockchain audit trails	Improved accountability and auditability at the cost of increased system complexity and governance overhead	System operators, compliance officers
[28]	Model-agnostic and model-specific explanations	Not explicitly addressed	Explainability generally increases trust, but effects depend on explanation type, user expertise, and context; privacy aspects remain underexplored	End-users, decision-makers
[18, 14]	Post-hoc explanations for decision support	Data minimization, consent, regulatory safeguards	Need to balance meaningful explanations with strict privacy and confidentiality requirements	Clinicians, patients, regulators
[10]	Layered and role-aware explanations	Privacy-by-design strategies, access control	Explainability must be tailored to stakeholder roles; organizational and governance trade-offs dominate purely technical considerations	Managers, compliance officers
[27]	Not the primary focus	Differential Privacy	Improved privacy leads to reductions in accuracy and fairness, revealing unavoidable multi-dimensional trade-offs	ML practitioners, policy makers
[26]	Explainability cues, transparency mechanisms	Privacy and security controls	Maximizing transparency, privacy, and efficiency simultaneously is infeasible; design requires explicit prioritization	Managers, system designers
[7]	Confidence scores, uncertainty communication	Applicable in privacy-sensitive settings without exposing raw data	Confidence cues may worsen trust calibration and appropriate reliance under knowledge mismatch	End-users, decision-makers

thesis of the selected studies, and integrated analysis of explainability–privacy trade-offs across technical and stakeholder perspectives.

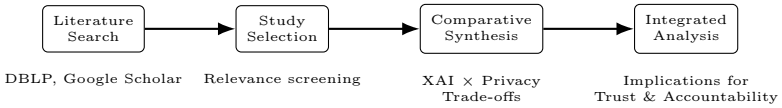

Fig. 1. Overview of the research method.

Relevant studies were identified through targeted searches in major bibliographic databases, including DBLP and Google Scholar. The search strategy combined terms related to explainability and privacy-preserving machine learning, such as "Explainable AI", "XAI", "interpretable machine learning", "privacy-preserving", "differential privacy", "federated learning", and "data protection". The search focused on publications addressing the intersection between explainability mechanisms and privacy constraints in machine learning systems, particularly in the context of decision support systems and information systems. To ensure relevance and quality, the following inclusion criteria were applied during the selection process: 1) Peer-reviewed articles or widely recognized technical reports; 2) Studies published from 2018 onwards, reflecting recent developments in XAI and privacy-preserving machine learning; 3) Research explicitly discussing explainability mechanisms, privacy-preserving techniques, or their interaction in AI-based systems; and 4) Studies addressing technical, organizational, or governance implications of explainability and privacy.

Articles focusing solely on explainability without privacy considerations, or exclusively on privacy-preserving machine learning without discussing transparency or interpretability aspects, were excluded. The selected studies represent a set of influential and representative works covering different perspectives on the explainability privacy relationship. For each selected study, information was extracted along four analytical dimensions: a) the XAI mechanisms employed; b) the privacy-preserving techniques considered; c) the trade-offs reported between explainability, privacy, and system utility; and d) the primary stakeholder groups addressed. These dimensions were chosen to enable a comparative analysis of how prior studies conceptualize the interaction between explainability and privacy in AI-enabled systems. The extracted data were synthesized into the comparative overview presented in Table 1, which highlights key patterns and differences across the literature.

The final stage of the analysis consisted of synthesizing the findings across the selected studies to identify recurring themes, trade-offs, and research gaps. Particular attention was given to understanding how explainability mechanisms interact with privacy-preserving techniques and how these interactions affect system transparency, trust, and accountability. This synthesis supports the identification of patterns across technical and governance perspectives and provides the foundation for the discussion presented in Section 2. The analysis also highlights gaps in the literature regarding the integrated evaluation of explainability, privacy, and stakeholder requirements in decision support systems.

4 Results

This section presents the results of the structured literature synthesis addressing the **RQ1**. To answer this question, the findings from the reviewed studies were organized into three analytical dimensions: (i) compatibility between XAI mechanisms and privacy-preserving techniques, (ii) explainability–privacy trade-offs, and (iii) implications for trust and accountability in decision support systems.

4.1 Compatibility Between XAI and Privacy-Preserving Techniques

The reviewed studies indicate that explainability and privacy-preserving mechanisms can coexist in AI-enabled decision support systems, but their compatibility is highly context-dependent. Compatibility depends primarily on the type of explanation mechanism, the privacy-preserving technique employed, and the level of abstraction used in explanations. Several studies highlight that explanations may unintentionally expose sensitive information. Post-hoc explanation mechanisms such as feature importance scores and counterfactual explanations may reveal patterns related to training data, enabling attacks such as membership inference, model extraction, or model inversion [3, 20, 22, 8]. These findings suggest that explainability mechanisms can act as additional information channels and must therefore be considered within the system's privacy threat model.

Despite these risks, the literature also reports scenarios in which XAI mechanisms can be combined with privacy-preserving approaches. Studies integrating XAI with differential privacy and federated learning demonstrate that explanation techniques such as SHAP, LIME, and Layer-wise Relevance Propagation can still be applied under privacy constraints, although with reduced fidelity or additional system complexity [17]. In general, global and aggregated explanations appear more compatible with privacy constraints than local or instance-level explanations, which tend to reveal more detailed information about specific predictions or data points.

The literature suggests that compatibility between explainability and privacy-preserving techniques is conditional rather than universal. The feasibility of combining both objectives depends on explanation granularity, privacy mechanisms, and the intended stakeholder audience.

4.2 Explainability Privacy Trade-offs

A second major theme identified in the literature concerns the structural trade-offs between explainability, privacy protection, and system utility. Several studies demonstrate that increasing privacy guarantees often reduces explanation precision, predictive accuracy, or the usefulness of explanations for debugging and decision support [4, 1, 27].

These trade-offs arise because privacy-preserving mechanisms frequently restrict access to the information required for detailed explanations. For example, differential privacy introduces noise into model outputs or gradients, which may distort explanation signals, while federated learning limits access to raw data, constraining the ability to generate instance-level explanations. The literature consistently indicates that such trade-offs are inherent to the coexistence of explainability and privacy protection rather than the result of implementation deficiencies. Consequently, several authors argue that explainability should not be maximized independently but instead calibrated according to privacy constraints, regulatory requirements, and operational objectives [26, 10].

This perspective reframes explainability as a bounded design property rather than a universal system attribute. Designers must therefore determine acceptable levels of explanation fidelity while preserving privacy guarantees and maintaining system performance.

4.3 Implications for Trust and Accountability

Beyond technical compatibility and trade-offs, the literature emphasizes that explainability plays an important role in supporting trust and accountability in AI-enabled decision support systems. However, the relationship between explainability and trust is complex and depends strongly on how explanations are designed and presented. Research shows that explanations can increase users' confidence in automated decisions, particularly when they are understandable and aligned with the user's decision-making context [28]. At the same time,

overly complex or poorly calibrated explanations may reduce appropriate reliance on AI systems [7]. These findings suggest that trust is influenced less by the amount of information disclosed and more by the relevance and clarity of explanations.

The literature also highlights the role of explainability in supporting accountability in regulated domains such as healthcare, finance, and public administration. In these contexts, explanations must enable auditing, contestation, and justification of automated decisions [18, 14]. Accountability-oriented explanations therefore require properties such as consistency, traceability, and defensibility over time. Finally, several studies emphasize that explainability requirements vary across stakeholders. Developers, auditors, managers, and end-users require different levels of detail and different forms of explanation [10, 17]. As a result, recent research advocates for layered and role-aware explanation strategies capable of supporting multiple organizational and regulatory needs. Taken together, the reviewed literature indicates that explainability in privacy-preserving decision support systems should be designed as a stakeholder-sensitive capability that balances transparency, privacy protection, and accountability requirements.

RQ1 Summary

The reviewed literature shows that explainability and privacy-preserving mechanisms are compatible only under context-dependent conditions. Global and aggregated explanations are generally reported as more compatible with privacy-sensitive settings, whereas local and instance-level explanations tend to increase privacy risks. The studies also indicate that explainability privacy trade-offs affect not only technical properties such as fidelity, accuracy, and utility, but also organizational concerns related to trust, accountability, and governance. The literature suggests that effective explainability in privacy-preserving decision support systems depends on balancing transparency with stakeholder needs and regulatory constraints rather than maximizing disclosure.

5 Conclusion

The increasing use of Artificial Intelligence in decision support systems has intensified the need to reconcile transparency requirements with the protection of sensitive information. While explainable AI mechanisms aim to provide insight into model behavior, privacy-preserving techniques restrict access to data and internal model details. Understanding how these objectives interact has therefore become a central challenge for the design and governance of AI-enabled information systems. This paper presented a structured literature-based analysis of the relationship between explainability and privacy-preserving mechanisms in decision support systems. Guided by the research question on how the literature characterizes compatibility, trade-offs, and trust and accountability implications

of explainability in privacy-preserving environments, the study synthesized representative works across four analytical dimensions: XAI mechanisms, privacy-preserving techniques, reported trade-offs, and stakeholder implications.

The results indicate that explainability and privacy preservation are not inherently incompatible, but their coexistence is highly context-dependent. The literature consistently shows that explanations can introduce additional privacy risks, particularly when instance-level or highly detailed explanations are used. At the same time, certain explanation strategies especially global or aggregated explanations appear more compatible with privacy-preserving environments. The review also highlights structural trade-offs between explainability, privacy protection, and system utility, suggesting that explainability should be calibrated rather than maximized in privacy-sensitive contexts.

Another key finding concerns the role of explainability in supporting trust and accountability. The literature indicates that effective explanations are those that are understandable, relevant, and aligned with stakeholders' roles and decision-making contexts. Trust is therefore influenced less by the amount of transparency provided and more by the appropriateness of explanations for specific users. In regulated environments, explainability also functions as a mechanism for accountability, enabling auditing, justification, and responsibility attribution for automated decisions. The study contributes a structured synthesis of the literature on explainability in privacy-preserving decision support systems and highlights the importance of considering explainability as a socio-technical design decision shaped by privacy constraints, stakeholder needs, and governance requirements.

This work has some limitations. The analysis focused on representative studies identified through targeted searches rather than a full systematic literature review, which means that some relevant studies may not have been included. Future research could extend this work by conducting systematic mapping studies or empirical investigations exploring how organizations operationalize explainability in privacy-sensitive environments. Additionally, further studies could examine design frameworks and architectural patterns that support layered and role-aware explainability in real-world decision support systems.

6 Acknowledgements

This study was financed in part by the National Council for Scientific and Technological Development (CNPq), Grants No. 300883/2025-0 and No. 406266/2025-5.

References

1. Abbasi, W., Mori, P., Saracino, A.: The explainability-privacy-utility trade-off for machine learning-based tabular data analysis. In: Proceedings of the 20th International Conference on Security and Cryptography, SECRYPT 2023, Rome, Italy, July 10-12, 2023. pp. 511–519. SCITEPRESS (2023), https://doi.org/10.5220/0012137800003555

2. Adnan, M., Syed, M.H., Anjum, A., Rehman, S.: A framework for privacy-preserving in iov using federated learning with differential privacy. IEEE Access **13**, 13507–13521 (2025), https://doi.org/10.1109/ACCESS.2025.3526934
3. Allana, S., Kankanhalli, M., Dara, R.: Privacy risks and preservation methods in explainable artificial intelligence: A scoping review. CoRR **abs/2505.02828** (2025), https://doi.org/10.48550/arXiv.2505.02828
4. Allana, S., Kankanhalli, M., Dara, R.: Privacy risks and preservation methods in explainable artificial intelligence: A scoping review. Trans. Mach. Learn. Res. **2025** (2025), https://openreview.net/forum?id=q9nykJfzku
5. Andrade, R., Filho, G.R., Marques, G., Canedo, E.: Privacy, data protection, risk, and compliance in the age of generative ai systems: A systematic mapping study. In: Anais do XXII Simpósio Brasileiro de Sistemas de Informação. pp. 1181–1200. SBC, Porto Alegre, RS, Brasil (2026), https://sol.sbc.org.br/index.php/sbsi/article/view/41374
6. Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, pportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020), https://doi.org/10.1016/j.inffus.2019.12.012
7. Cabitza, F., Fregosi, C., Vicente, L.: Too sure for trust. the paradoxical effect of calibrated confidence in case of uncalibrated trust in hybrid decision making. In: World Conference on Explainable Artificial Intelligence. pp. 233–254. Springer (2025). https://doi.org/https://doi.org/10.1007/978-3-032-08317-3_11
8. Ezzeddine, F.: Privacy implications of explainable AI in data-driven systems. In: Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), Valletta, Malta, July 17-19, 2024. CEUR Workshop Proceedings, vol. 3793, pp. 361–368. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3793/paper_46.pdf
9. Keaney, S., Berthon, P.: Balancing explainability and privacy in ai systems: A strategic imperative for managers. Business Horizons (2025). https://doi.org/https://doi.org/10.1016/j.bushor.2025.10.004
10. Keaney, S., Berthon, P.: Balancing explainability and privacy in ai systems: A strategic imperative for managers. Business Horizons (2025). https://doi.org/https://doi.org/10.1016/j.bushor.2025.10.004
11. OECD: Recommendation of the council on artificial intelligence. OECD: Paris, France pp. 1–12 (2024), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, originally adopted on 22 May 2019; amended on 3 May 2024
12. Parliament, T.E., Council, T.: General Data Protection Regulation (GDPR): EU Data Protection Rules (2018), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
13. Potluri, S.: Policy-aware secure data governance in distributed information systems using explainable ai models. International Journal of AI, BigData, Computational and Management Studies **6**(3), 1–10 (2025)
14. Radanliev, P.: AI ethics: Integrating transparency, fairness, and privacy in AI development. Appl. Artif. Intell. **39**(1) (2025), https://doi.org/10.1080/08839514.2025.2463722
15. da República, Presidência, N.C.: Brazilian general data protection law (lgpd). Nartional Congress, accessed in April 10, 2022 **1**(1), 1–31 (2018),

https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf
16. Rocha, L.D., Canedo, E.D.: Optimizing compliance: Comparative study of data laws and privacy frameworks. J. Internet Serv. Appl. **16**(1), 431–452 (2025), https://doi.org/10.5753/jisa.2025.5247
17. Senevirathna, T., Sandeepa, C., Siniarski, B., Nguyen, M., Marchal, S., Boerger, M., Liyanage, M., Wang, S.: Enhancing accountability, resilience, and privacy of intelligent networks with XAI. IEEE Open J. Commun. Soc. **6**, 8389–8409 (2025), https://doi.org/10.1109/OJCOMS.2025.3608784
18. Shah, B., Gupta, E.V.: Ethics and privacy in ai-driven healthcare decision support systems. International Journal of Research and Analytical Reviews (IJRAR) **12**, 1–11 (2025)
19. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int. J. Hum. Comput. Stud. **146**, 102551 (2021), https://doi.org/10.1016/j.ijhcs.2020.102551
20. Shokri, R., Strobel, M., Zick, Y.: On the privacy risks of model explanations. In: AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021. pp. 231–241. ACM (2021), https://doi.org/10.1145/3461702.3462533
21. Sonsini, W., Parliament, T.E.: The eu artificial intelligence act. European Union pp. 1–10 (2024), https://www.wsgr.com/a/web/qrkz1SnNzWw6nk7B3oAyDa/10-things-you-should-know-about-the-eu-artificial-intelligence-act_v2.pdf
22. Spartalis, C.N., Semertzidis, T., Daras, P.: Balancing XAI with privacy and security considerations. In: Computer Security. ESORICS 2023 International Workshops - CPS4CIP, ADIoT, SecAssure, WASP, TAURIN, PriST-AI, and SECAI, The Hague, The Netherlands, September 25-29, 2023, Revised Selected Papers, Part II. Lecture Notes in Computer Science, vol. 14399, pp. 111–124. Springer (2023), https://doi.org/10.1007/978-3-031-54129-2_7
23. of Standards, N.I., Technology: Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST Trustworthy and Responsible AI NIST AI 600-1 pp. 1–64 (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
24. United Nations Educational, S., (UNESCO), C.O.: Recommendation on the ethics of artificial intelligence. UNESCO:Paris, France pp. 1–44 (2022), https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en
25. Vicenzi, A.A.F.C., de Cerqueira, J.S., Abrahamsson, P., Canedo, E.D.: Specifying fairness and transparency requirements for public benefit allocation. In: Requirements Engineering: Foundation for Software Quality. pp. 245–253. Springer Nature Switzerland, Cham (2026). https://doi.org/https://doi.org/10.1007/978-3-032-21423-2_17
26. Wang, Y.: Balancing trustworthiness and efficiency in artificial intelligence systems: An analysis of tradeoffs and strategies. IEEE Internet Computing **27**(6), 8–12 (2023). https://doi.org/10.1109/MIC.2023.3303031
27. Wasif, D., Chen, D., Madabushi, S., Alluru, N., Moore, T.J., Cho, J.H.: Empirical analysis of privacy-fairness-accuracy trade-offs in federated learning: A step towards responsible ai. arXiv preprint arXiv:2503.16233 (2025), https://arxiv.org/abs/2503.16233
28. Wiratsin, I., Ragkhitwetsagul, C.: Effectiveness of explainable artificial intelligence (XAI) techniques for improving human trust in machine learning models: A systematic literature review. IEEE Access **13**, 121326–121350 (2025), https://doi.org/10.1109/ACCESS.2025.3575022