DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Inteligência Artificial Explicável em Sistemas de Apoio à Decisão com Preservação de Privacidade: Compatibilidade, trade-offs e um Framework Orientado à Governança sob a Perspectiva de Stakeholders

Andressa Girotto Vargas

Brasília, abril de 2026

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Inteligência Artificial Explicável em Sistemas de Apoio à Decisão com Preservação de Privacidade: Compatibilidade, *trade-offs* e perspectivas de *stakeholders*

Andressa Girotto Vargas

ORIENTADORA: PROFESSORA DRA. EDNA DIAS CANEDO

DISSERTAÇÃO DE MESTRADO PROFISSIONAL EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPEE.MP.113
BRASÍLIA/DF, Abril - 2026

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Inteligência Artificial Explicável em Sistemas de Apoio à Decisão com Preservação de Privacidade: Compatibilidade, trade-offs e um Framework Orientado à Governança sob a Perspectiva de Stakeholders

Andressa Girotto Vargas

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Professora Dra. Edna Dias Canedo, Ph.D, FT/UnB _____
Orientadora

Antenor Pereira Madruga Filho, Doutor, Madruga BTW Advogados _____
Examinador Externo

Fábio Lúcio Lopes de Mendonça, Ph.D, FT/UnB _____
Examinador Interno

Georges Daniel Amvame Nze, Ph.D, FT/UnB _____
Membro Suplente

FICHA CATALOGRÁFICA

VARGAS, ANDRESSA GIROTTO
Inteligência Artificial Explicável em Sistemas de Apoio à Decisãocom Preservação de Privacidade: Compatibilidade, trade-offse um Framework Orientado à Governançasob a Perspectiva de Stakeholders [Distrito Federal] 2026.
xvi, p.76., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2026).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Inteligencia Artificial
2. Direito à explicação
3. Transparência algorítmica
4. Proteção de Dados
5. Privacidade

I. ENE/FT/UnB

REFERÊNCIA BIBLIOGRÁFICA

VARGAS, A.G. (2026). *Inteligência Artificial Explicável em Sistemas de Apoio à Decisãocom Preservação de Privacidade: Compatibilidade, trade-offse um Framework Orientado à Governançasob a Perspectiva de Stakeholders*. Dissertação de Mestrado Profissional, Publicação: PPEE.MP. , Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, p.76.

CESSÃO DE DIREITOS

AUTOR: Andressa Girotto Vargas
TÍTULO: Inteligência Artificial Explicável em Sistemas de Apoio à Decisãocom Preservação de Privacidade: Compatibilidade, trade-offse um Framework Orientado à Governançasob a Perspectiva de Stakeholders.
GRAU: Mestre em Engenharia Elétrica ANO: 2026

Andressa Girotto Vargas
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

AGRADECIMENTOS

Agradeço aos meus amados pais, meus maiores mestres na vida, por terem ensinado o valor do ensino, do trabalho e por me apoiarem sempre. São inegavelmente meus maiores exemplos. Devo tudo a vocês! Ao meu querido irmão, um ser humano visionário. Tenho muito orgulho de tua sabedoria e inteligência! Ao meu marido, por ser meu parceiro de todas horas, apoiar minhas decisões e ser tão compreensivo. Amo-te muito! A minha estimada orientadora, professora Drª. Edna Dias Canedo, pelo apoio e contribuição imprescindível para a realização deste estudo. És um grande exemplo de mulher pesquisadora, cuja dedicação à vida acadêmica é inspiradora. Sou muito grata e honrada por tê-la como orientadora. A Deus, acima de tudo.

RESUMO

Contexto: Sistemas de apoio à decisão baseados em Inteligência Artificial operam cada vez mais em contextos regulados e sensíveis à privacidade, nos quais a explicabilidade é demandada tanto para promover confiança quanto para viabilizar responsabilização e conformidade normativa. Entretanto, mecanismos de explicação podem introduzir riscos adicionais de exposição informacional, evidenciando tensões entre explicabilidade e preservação de privacidade. A literatura reconhece essa relação, mas carece de diretrizes estruturadas que apoiem decisões explícitas sobre compatibilidade, limites e *trade-offs* sob uma perspectiva de governança. **Método**: A pesquisa foi conduzida segundo o paradigma da *Design Science Research* (DSR). Inicialmente, realizou-se análise teórica da literatura e um estudo empírico qualitativo por meio de grupo focal com especialistas, visando compreender percepções sobre compatibilidade, *trade-offs* e necessidades de *stakeholders*. A partir dessas evidências, foi proposto um *framework* conceitual de explicabilidade orientada à governança, estruturado em quatro camadas interdependentes. O artefato foi então validado por meio de *survey* com especialistas atuantes em contextos regulados, combinando análise quantitativa (escala Likert) e qualitativa (questões abertas). **Resultados**:Os achados indicam que a compatibilidade entre explicabilidade e preservação da privacidade é dependente de contexto, finalidade e stakeholder, sendo mediada por *trade-offs* estruturais. Explicações globais e agregadas mostraram-se mais adequadas em ambientes sensíveis à privacidade, enquanto explicações locais podem ampliar riscos de exposição informacional. A validação do *framework* apresentou avaliações predominantemente positivas quanto à utilidade, clareza, completude, aderência à governança e flexibilidade, com reconhecimento da estrutura multicamada, da rastreabilidade por artefatos e da diferenciação de responsabilidades como pontos fortes. Emergiram, contudo, recomendações relacionadas à necessidade de maior operacionalização e definição de métricas objetivas. **Conclusão**: A dissertação demonstra que explicabilidade em sistemas sensíveis à privacidade deve ser tratada como decisão estratégica inserida em um ecossistema de governança, e não como requisito técnico isolado. O *framework* proposto contribui ao estruturar decisões sobre o que explicar, para quem, com qual granularidade e sob quais garantias de privacidade, promovendo rastreabilidade, auditabilidade e alinhamento regulatório. Assim, o trabalho avança na integração entre princípios normativos e práticas de engenharia, oferecendo conhecimento prescritivo aplicável a contextos organizacionais regulados.

Palavras-chave: Inteligência Artificial Explicável; Preservação de Privacidade; Sistemas de Apoio à Decisão; Governança de IA; Design Science Research; Trade-offs.

ABSTRACT

Context: Artificial Intelligence based decision support systems increasingly operate in regulated and privacy-sensitive environments, where explainability is required both to foster trust and to enable accountability and regulatory compliance. However, explanation mechanisms may introduce additional risks of information exposure, revealing tensions between explainability and privacy preservation. While the literature acknowledges this relationship, it lacks structured guidelines to support explicit decisions regarding compatibility, boundaries, and trade-offs from a governance perspective. **Method**: The research was conducted under the Design Science Research (DSR) paradigm. Initially, a theoretical literature analysis and a qualitative empirical study were carried out through a focus group with specialists to understand perceptions regarding compatibility, trade-offs, and stakeholder needs. Based on this evidence, a governance-oriented conceptual framework for explainability was proposed, structured into four interdependent layers. The artifact was subsequently validated through a survey with specialists working in regulated contexts, combining quantitative analysis (Likert scale) and qualitative analysis (open-ended questions). **Results**: The findings indicate that compatibility between explainability and privacy preservation is context, purpose, and stakeholder dependent, mediated by structural trade-offs. Global and aggregated explanations proved more suitable in privacy sensitive environments, whereas local explanations may increase risks of information exposure. The framework validation yielded predominantly positive evaluations regarding utility, conceptual clarity, completeness, governance adherence, and flexibility. The multilayer structure, artifact traceability, and differentiation of responsibilities were recognized as key strengths. Nevertheless, recommendations emerged concerning the need for further operationalization and the definition of objective metrics. **Conclusion**: The dissertation demonstrates that explainability in privacy-sensitive systems should be treated as a strategic decision embedded within a governance ecosystem rather than as an isolated technical requirement. The proposed framework contributes by structuring decisions about what to explain, to whom, with what level of granularity, and under which privacy guarantees, promoting traceability, auditability, and regulatory alignment. Thus, this work advances the integration of normative principles and engineering practices, offering prescriptive knowledge applicable to regulated organizational contexts.

Keywords: Explainable Artificial Intelligence (XAI); Privacy-Preserving AI; AI Governance; Decision Support Systems; Trade-offs; Design Science Research.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Este capítulo apresenta o contexto e a motivação desta dissertação, delimitando o problema de pesquisa, a justificativa, os objetivos e os resultados esperados. Ao final, são descritas as publicações decorrentes da pesquisa e a estrutura deste documento.

1.1 CONTEXTUALIZAÇÃO

A crescente adoção de técnicas de Aprendizado de Máquina (*Machine Learning* – ML) em sistemas de informação tem intensificado as demandas por transparência, responsabilização e confiabilidade, especialmente em cenários de apoio à decisão que envolvem dados sensíveis [4, 5]. Esse movimento é impulsionado pela evolução de técnicas de aprendizado profundo (*deep learning*), pela ampla disponibilidade de dados em larga escala (*Big Data*) e pela expansão do poder computacional, fatores que contribuíram significativamente para o avanço de modelos de Inteligência Artificial (IA) [6].

Atualmente, decisões automatizadas baseadas em IA estão disseminadas em diversos domínios da sociedade, incluindo assistentes virtuais, veículos autônomos, sistemas de detecção de fraudes, diagnóstico de enfermidades e mecanismos de apoio à tomada de decisão em ambientes organizacionais e na Administração Pública [7]. Apesar das oportunidades associadas ao uso da IA, também são recorrentes relatos de vieses algorítmicos, práticas discriminatórias [8] e alucinações em sistemas baseados em modelos generativos [9].

Nesse contexto, cresce a preocupação com o tratamento de dados pessoais, especialmente aqueles classificados como sensíveis, e com os impactos à privacidade e à proteção de dados. Em resposta, diferentes ordenamentos jurídicos passaram a estabelecer salvaguardas normativas. No Brasil, a Lei nº 13.709/2018 – Lei Geral de Proteção de Dados (LGPD) [10] prevê o direito à revisão de decisões tomadas exclusivamente com base em tratamento automatizado, bem como a obrigação do controlador de dados quanto ao fornecimento de informações claras e adequadas sobre os critérios e procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial. [10, 11] . Embora nem a LGPD nem o Regulamento Geral de Proteção de Dados da União Europeia (GDPR) prevejam expressamente um "direito à explicação", parte da doutrina interpreta tais disposições como equivalentes a esse direito [12, 13], interpretação que é contestada por outros autores [14, 15].

Mais recentemente, a União Europeia promulgou o Regulamento 2024/1689 (*EU AI Act*), que estabelece explicitamente, em seu artigo 86, o direito à explicação em decisões decorrentes de sistemas de IA de alto risco que afetem significativamente direitos fundamentais [16]. Paralelamente, instrumentos internacionais de *soft law*, como as Recomendações Éticas para IA da

UNESCO [17], os Princípios da OCDE para o desenvolvimento da IA [18] e o *Artificial Intelligence Risk Management Framework* do National Institute of Standards and Technology (NIST) [19], reforçam a explicabilidade como princípio central para o desenvolvimento de sistemas de IA confiáveis.

Nesse cenário, a *Explainable Artificial Intelligence* (XAI) emerge como um conjunto de técnicas voltadas a tornar modelos de IA mais compreensíveis, auditáveis e contestáveis [4]. Contudo, muitos sistemas contemporâneos operam sob fortes restrições de privacidade, empregando técnicas como *differential privacy*, *federated learning*, computação segura e anonimização de dados [20]. A coexistência dessas abordagens introduz uma tensão fundamental: enquanto a XAI busca revelar informações sobre o comportamento dos modelos, mecanismos de preservação da privacidade visam restringir a divulgação de informações.

Estudos recentes demonstram que explicações podem, elas próprias, constituir vetores de vazamento de privacidade, expondo informações sobre dados de treinamento, fronteiras de decisão ou atributos sensíveis [21]. Em particular, mecanismos de explicação como importância de atributos e explicações contrafactuais podem facilitar ataques de inferência de associação, extração ou inversão de modelos quando empregados sem salvaguardas adequadas [22]. Esses achados desafiam a suposição de compatibilidade natural entre explicabilidade e privacidade, evidenciando a necessidade de compreender seus *trade-offs*.

Além do nível do modelo, esses desafios se estendem às camadas arquitetural e de governança dos sistemas de informação. Ambientes distribuídos e domínios regulados demandam não apenas proteção à privacidade, mas também rastreabilidade, auditabilidade e conformidade com políticas organizacionais e legais. Abordagens que integram explicabilidade a arquiteturas federadas e mecanismos de governança indicam que a XAI pode apoiar a responsabilização quando cuidadosamente projetada [23]. Ainda assim, há escassez de entendimentos sistemáticos sobre quais mecanismos de XAI são compatíveis com técnicas específicas de preservação da privacidade e sob quais condições essas combinações permanecem seguras e eficazes.

Diante desse cenário, torna-se necessário avançar para além da análise isolada de técnicas de explicabilidade, propondo abordagens estruturadas que orientem a seleção, o uso e a governança de mecanismos de XAI em contextos regulados e sensíveis à privacidade.

1.2 PROBLEMA DE PESQUISA

Embora as técnicas de XAI sejam frequentemente apresentadas como soluções para aumentar a transparência, a confiança e a legitimidade de decisões automatizadas, persistem desafios significativos relacionados à sua conceituação, aplicação prática e integração com requisitos de privacidade. A literatura apresenta esforços voltados à classificação e taxonomia de métodos de XAI [4, 24], contudo, a polissemia do termo explicabilidade, muitas vezes utilizado como sinônimo de interpretabilidade, dificulta a definição de critérios técnicos e normativos claros [25].

Adicionalmente, explicações frequentemente não são adequadas ao público-alvo: podem ser excessivamente técnicas para usuários leigos ou, ao contrário, simplificadas a ponto de não refletirem o funcionamento real do modelo [26]. Soma-se a isso a dificuldade de equilibrar acurácia, compreensibilidade e proteção à privacidade, especialmente em contextos regulados.

Diante desse cenário, o problema central desta pesquisa consiste em compreender como práticas de explicabilidade podem ser projetadas de modo a conciliar transparência, privacidade e responsabilização, atendendo simultaneamente a requisitos técnicos, legais e éticos.

1.2.1 Questões de Pesquisa

Considerando os desafios associados à integração entre explicabilidade e preservação da privacidade em sistemas de Inteligência Artificial, bem como as exigências técnicas, legais e éticas, esta pesquisa é orientada pelas seguintes Questões de Pesquisa (QPs):

QP1: Quais mecanismos de *Explainable Artificial Intelligence* (XAI) são compatíveis com diferentes técnicas de preservação da privacidade em sistemas de informação, e quais *trade-offs* emergem dessa combinação?

QP2: Como diferentes níveis de explicabilidade influenciam a confiança, a percepção de responsabilização e a possibilidade de contestação por parte dos *stakeholders* em sistemas de apoio à decisão que operam sob restrições de privacidade?

Essas questões derivam diretamente da tensão identificada entre explicabilidade, proteção de dados e governança de sistemas de IA, e orientam a pesquisa tanto no plano técnico quanto no socio-organizacional, permitindo analisar a explicabilidade como um requisito de confiança, responsabilização e conformidade regulatória.

1.3 JUSTIFICATIVA

A relevância desta pesquisa decorre da ampla disseminação de decisões automatizadas e das exigências legais e normativas que demandam maior transparência, auditabilidade e proteção de direitos fundamentais. A explicabilidade é reconhecida como princípio central tanto em legislações vinculantes, como LGPD, GDPR e *AI Act* [10, 11, 16] quanto em referenciais internacionais de *soft law* [17, 18, 19].

Entretanto, a implementação de explicações em sistemas que operam sob restrições de privacidade impõe riscos adicionais, exigindo abordagens cuidadosas que considerem os *trade-offs* envolvidos. Assim, esta pesquisa se justifica pela necessidade de desenvolver diretrizes que permitam a adoção de práticas de explicabilidade centradas no usuário, auditáveis e compatíveis com a preservação da privacidade, contribuindo para o uso responsável da IA.

1.4 OBJETIVOS

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver e validar um *framework* de explicabilidade para sistemas de Inteligência Artificial que apoie a geração de explicações claras, auditáveis e centradas no usuário, compatíveis com técnicas de preservação da privacidade, através de um survey com profissionais da área.

1.4.2 Objetivos Específicos

- Realizar revisão de literatura e dos referenciais normativos relacionados à explicabilidade e à privacidade em sistemas de IA;
- Identificar *trade-offs* entre mecanismos de XAI e técnicas de preservação da privacidade;
- Propor diretrizes de explicabilidade alinhadas a requisitos técnicos, legais e éticos;
- Desenvolver o *framework*, conforme às diretrizes propostas;
- Aprimorar e validar a aplicabilidade do *framework* proposto sob a perspectiva de usuários e desenvolvedores de sistemas de IA, por meio de survey com perguntas objetivas e abertas.

1.5 RESULTADOS ESPERADOS

Espera-se que esta pesquisa contribua para o desenvolvimento de sistemas de IA mais transparentes e responsáveis, capazes de fornecer explicações compreensíveis sem comprometer a proteção de dados pessoais. Adicionalmente, pretende-se subsidiar o exercício do direito de contestação de decisões automatizadas, bem como fortalecer práticas de governança e auditabilidade em sistemas baseados em IA. Por fim, espera-se que o *framework* proposto sirva como referência para o projeto e a avaliação de mecanismos de explicabilidade em sistemas de IA sensíveis à privacidade, apoiando decisões técnicas e de governança.

1.6 PUBLICAÇÕES

Esta dissertação gerou os seguintes artigo científicos:

- *Explainability and Privacy in AI-Based Decision Support Systems: A Structured Literature Analysis*, que foi aceito para publicação na 21ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI), pela Springer em volumes da série *Lecture Notes in Networks and Systems*, https://link.springer.com/series/15179.

- *A Governance-Oriented Framework for Balancing Explainability and Privacy in AI-Based Decision Systems*, que foi aceito para publicação na edição de 2026 da *Americas Conference on Information Systems* (AMCIS).

1.7 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em seis capítulos, além deste capítulo introdutório, conforme descrito a seguir:

- O Capítulo 2 apresenta a fundamentação teórica e os estudos correlatos que embasam a pesquisa, abordando conceitos relacionados à Inteligência Artificial Explicável, técnicas de preservação da privacidade, decisões automatizadas e governança de sistemas de IA;
- O Capítulo 3 descreve o método de pesquisa adotado e apresenta os resultados da análise estruturada da literatura e o estudo qualitativo, discutindo as respostas às questões de pesquisa relacionadas à compatibilidade entre XAI e técnicas de preservação da privacidade;
- O Capítulo 4 apresenta a primeira versão do *framework* de explicabilidade proposto, detalhando sua motivação, estrutura, dimensões e diretrizes de uso;
- O Capítulo 5 apresenta a validação empírica do *framework* por meio de um survey com profissionais da área, bem como a análise de seus resultados e as limitações do estudo;
- O Capítulo 6 apresenta as conclusões da pesquisa, sintetizando as principais contribuições, limitações e direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos e os principais trabalhos correlatos que sustentam esta dissertação. O objetivo é contextualizar os conceitos centrais relacionados à Inteligência Artificial Explicável (XAI), às técnicas de preservação da privacidade e aos desafios associados à confiança e à responsabilização em sistemas de apoio à decisão baseados em IA. A discussão é organizada de modo a evidenciar as tensões, *trade-offs* e lacunas existentes na literatura, que motivam a proposição do *framework* apresentado nos capítulos subsequentes.

2.1 DECISÕES AUTOMATIZADAS, CONFIANÇA E RESPONSABILIZAÇÃO

Sistemas de apoio à decisão baseados em IA vêm sendo amplamente adotados em domínios sensíveis, como saúde [27], finanças [28], justiça [29] e administração pública [7]. Nesses contextos, decisões automatizadas podem impactar diretamente direitos fundamentais, tornando essenciais mecanismos que permitam compreensão, contestação e responsabilização.

A confiança em sistemas de IA tem sido apontada como fator crítico para sua adoção e uso efetivo [30, 18]. Evidências empíricas sugerem que explicações podem aumentar a confiança dos usuários em decisões automatizadas, embora esse efeito dependa fortemente do tipo de explicação, do contexto de uso e do perfil do usuário [5]. Explicações inadequadas podem, inclusive, reduzir a confiança ou induzir a uma confiança indevida.

A responsabilização (*accountability*) está intimamente relacionada à explicabilidade, mas não se confunde com ela. Enquanto a confiança está associada à aceitação e ao uso do sistema, a responsabilização envolve a capacidade de justificar decisões, atribuir responsabilidades e viabilizar auditorias e revisões [31]. Em ambientes regulados, explicações devem ser não apenas compreensíveis, mas também defensáveis, consistentes e passíveis de documentação ao longo do tempo.

Diferentes instrumentos normativos e éticos reconhecem a explicabilidade como elemento central da IA confiável. As Recomendações Éticas para IA da UNESCO [17], os Princípios da OCDE [18] e o *AI Risk Management Framework* do NIST [19] enfatizam a necessidade de explicações proporcionais, contextualizadas e alinhadas aos riscos do sistema.

2.1.1 Aspectos legais e jurisprudenciais

No ordenamento jurídico brasileiro, o direito à explicação assim como o direito à revisão de decisões automatizadas já haviam sido previstos no art.5° da Lei do Cadastro Positivo (Lei 12.414/2011) [32], assegurando-se aos cadastrados "conhecer os principais elementos e critérios

considerados para a análise de risco, resguardado o segredo empresarial", bem como "solicitar ao consulente a revisão de decisão realizada exclusivamente por meios automatizados".

Na mesma linha, a Lei Geral de Proteção de Dados (LGPD) [10], em 2018, concedeu, por meio do seu art. 20, o direito aos titulares de dados de "solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade".

Segundo dados do Relatório LGPD nos Tribunais[33], a maioria dos casos julgados pelas cortes brasileiras que envolvem questionamento sobre decisões automatizadas tratam sobre motoristas de aplicativos e exclusão automática de perfil ou reconhecimento de vínculo junto ao aplicativo e exclusão de perfis em redes sociais ou em jogos.

Por sua vez, o Regulamento Geral de Proteção de Dados Europeu (*General Data Protection Regulation* - GDPR) [11], preconiza que o titular de dados tem o direito de obter informação do controlador quanto à existência de decisão automatizada, inclusive podendo se opor à realização daquelas que sejam baseadas totalmente em tratamento automatizado e que os impactem significativamente, salvo se a decisão fosse necessária por questão contratual; autorizada por lei e sujeita à salvaguardas, ou ainda, se baseada em consentimento explícito do titular.

Vale destacar que o Tribunal de Justiça da União Europeia ao julgar o caso C-203/2022 Dun & Bradstreet Austria, julgado em 2025, ao examinar a ponderação entre o direito de acesso e o direito à proteção do segredo comercial, reforçou que "para permitir que o titular exerça efetivamente os direitos conferidos pelo GDPR e, em particular, pelo Artigo 22(3) do mesmo, essa explicação deve ser fornecida por meio de informações relevantes e de forma concisa, transparente, inteligível e de fácil acesso."e, ainda "esses requisitos não podem ser satisfeitos nem pela simples comunicação de uma fórmula matemática complexa, como um algoritmo, nem pela descrição detalhada de todas as etapas da tomada de decisão automatizada, pois nenhuma delas constituiria uma explicação suficientemente concisa e inteligível."

Ainda na União Europeia, o AI Act (Regulation (EU) 2024/1689) [16], considerado o primeiro marco legal sobre IA [34], trouxe de forma expressa no seu art. 86 que qualquer pessoa afetada por decisão tomada com base em sistemas de IA de alto risco, exceto aqueles relativos à infraestrutura crítica, que produza efeitos jurídicos que afete de forma significativa sua saúde, segurança ou direitos fundamentais, terá o direito de obter explicações claras e significativas sobre o papel do sistema de IA no processo decisório e sobre os principais elementos da tomada de decisão.

No Brasil, o Projeto de Lei n 2338/2023, atualmente em discussão na Câmara dos Deputados, tem por objetivo dispor sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana. Nele, a explicabilidade é prevista como um dos princípios que deverão nortear o desenvolvimento, a implementação e o uso de sistemas de IA. Além disso, de modo semelhante ao AI Act, as pessoas afetadas por sistema de IA de alto risco têm assegurado o direito à explicação sobre a decisão, a recomendação ou a previsão feitas pelo sistema, a qual deverá incluir informações suficientes, adequadas e inteli-

gíveis, bem como o direito à contestação e à revisão, inclusive, quanto a essa última, que esta seja realizada por pessoa humana, uma vez considerado o contexto, o risco e o estado da arte do desenvolvimento tecnológico.

Na proposta legislativa em discussão na Câmara, há previsão de criação de um Sistema Nacional de Regulação e Governança de Inteligência Artificial (SIA), cuja coordenação ser atribuída à Agência Nacional de Proteção de Dados (ANPD). Segundo o texto, ainda em discussão, cabera à ANPD, na qualidade de coordenadora do SIA, entre outras competências, editar regras gerais sobre Inteligência Artificial no país e prestar suporte aos órgãos setoriais, aos quais caberão a edição de regras específicas. A indicação da Agência como coordenadora do SIA é relevante, considerando a sua missão precípua de zelar pela proteção de dados pessoais no país, o que poderá contribuir para conferir maior segurança jurídica e convergência regulatória entre os temas de privacidade, proteção de dados pessoais à regulação futura da IA no Brasil.

2.2 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

A Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) refere-se a um conjunto de métodos e abordagens cujo objetivo é tornar o funcionamento e os resultados de sistemas de IA compreensíveis para seres humanos [4]. A necessidade de explicabilidade emerge, sobretudo, em sistemas baseados em modelos complexos ou opacos, frequentemente denominados *caixas-pretas*, como redes neurais profundas e modelos de aprendizado profundo.

A existência de estudos sobre a explicação de decisões geradas por sistemas não é algo recente [35]. Nas décadas de 70 e 80, pesquisadores já se dedicavam ao tema, sob a perspectiva de sistemas especialistas baseado em regras (*rule-based expert systems*) [36] [37].

No entanto, o termo Inteligencia Artificial Explicável (Explainable Artificial Intelligence - XAI) ganhou maior notoriedade após a sua utilização pelo Departamento de Defesa dos Estados Unidos, quando do lançamento do DARPA's *Explainable Artificial Intelligence Program* em 2017, que teve como objetivo a criação de um conjunto de técnicas de aprendizado de máquina (*Machine Learning* - ML) que produzissem modelos explicáveis (seu processo decisório bem como os *outputs* gerados) [24] com acurácia de previsão, que uma vez combinados com técnicas de explicação, permitissem que os usuários finais compreendessem, confiassem e gerenciassem com eficácia sistemas de IA [38].

A área de pesquisa de XAI se preocupa em mitigar o problema de caixa preta que surge no aprendizado de máquina, ao mesmo tempo em que busca manter a acurácia dos modelos. Para tanto, os métodos de XAI buscam fornecer informações sobre o funcionamento dos modelos com o objetivo de eliminar, ou ao menos, reduzir a opacidade destes [39]. Além disso, a XAI pode ser considerada como uma área de pesquisa interdisciplinar, na medida em que utiliza-se de conhecimentos das Ciências Sociais e também leva em consideração aspectos da Psicologia da Explicação [4].

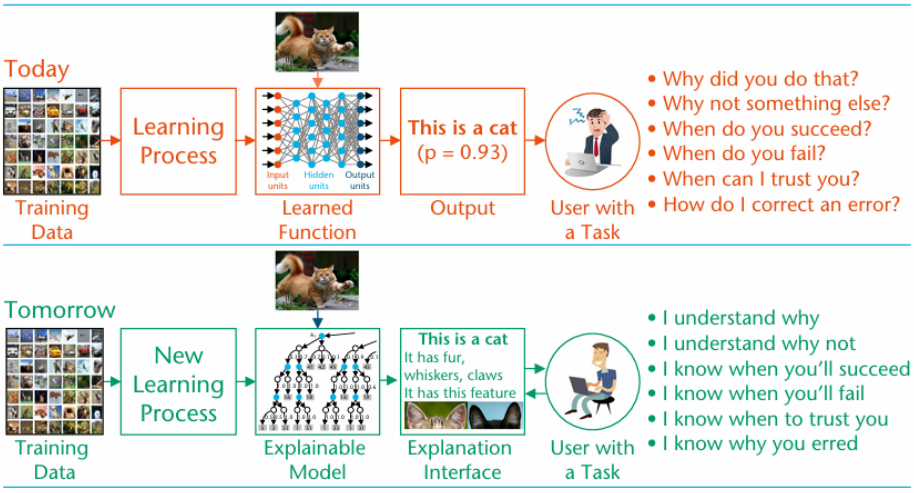

Figura 2.1: Conceito de XAI apresentado no projeto DARPA [38]

A literatura apresenta diferentes definições e interpretações para explicabilidade, frequentemente associadas ou confundidas com conceitos como interpretabilidade, transparência e compreensibilidade. Enquanto a interpretabilidade costuma estar relacionada à capacidade de um humano compreender diretamente a lógica interna de um modelo, a explicabilidade é frequentemente tratada como a capacidade de gerar explicações sobre o comportamento do modelo, mesmo quando sua estrutura interna não é diretamente interpretável [24]. Essa polissemia conceitual dificulta a definição de critérios técnicos e normativos claros, especialmente em contextos regulados.

A explicabilidade pode ser alcançada a partir de modelos que são inerentemente interpretáveis, os quais são chamados de "caixas brancas"ou modelos transparentes, a exemplo de: árvores de decisão, redes bayesianas, regressão linear, K-vizinhos mais próximos (KNN) e sistemas baseados em regras, ou ainda, por meio de técnicas *post-hoc*[21], também conhecida como abordagem de engenharia reversa. Essas se destinam, principalmente, aos modelos que não são naturalmente interpretáveis, ou seja, para aqueles chamados de "caixas pretas", como as redes neurais profundas (*Deep Neural Networks* - DNN) e as florestas de decisão (*Random Forest*). Para tanto, é desenvolvido um segundo modelo (*explainer*), a fim de que forneça explicações aos usuários sobre como o modelo realiza suas predições a partir dos *inputs* recebidos [40] [4].

Como mencionado, embora esses métodos sejam mais comumente empregados em modelos opacos, também podem ser aplicados para modelos originalmente transparentes, que em razão da multiplicidade de regras e parâmetros, acabam por adquirir um comportamento semelhante ao de modelos do tipo caixas pretas [41].

As técnicas de XAI podem ser classificadas segundo diferentes dimensões, como: (i) explicações globais versus locais; (ii) métodos intrínsecos versus *post-hoc*; e (iii) explicações agnósticas ou específicas ao modelo. Explicações globais buscam fornecer uma visão geral do comportamento do modelo, enquanto explicações locais se concentram em decisões individuais. Métodos intrínsecos correspondem a modelos projetados para serem interpretáveis desde sua concepção,

ao passo que métodos *post-hoc* produzem explicações após o treinamento do modelo [4] [21].

Tendo em vista que o objetivo deste trabalho consiste em examinar práticas de explicabilidade aplicáveis a modelos intrinsecamente opacos, atribui-se-á ênfase aos métodos *post-hoc*.

2.2.1 Técnicas *Post-hoc*

Diversos trabalhos tem se dedicado à elaboração de uma taxonomia para métodos de explicação, em razão da variedade de classificações existentes [42] [4] [41]. Há autores que consideram como categorias: escopo, modelo e metodologia dos explicadores. [43]

Com relação ao escopo, é possível a análise de importância dos parâmetros, isto é, das*features* a partir explicadores locais ou globais. No primeiro caso, proporciona-se a compreensão de um *output* específico produzido pelo sistema a partir de determinada entrada. Entre as técnicas utilizadas, destaca-se o *Local Interpretable Model-agnostic Explanations* (LIME) [43], o qual gera explicações a partir de modelos substitutos que são interpretáveis e possuem um menos*features* no intuito de manter baixa a complexidade de interpretação do modelo [44].

Por sua vez, as técnicas que envolvem explicações globais buscam gerar uma compreensão acerca do funcionamento geral do sistema de IA [45]. Uma técnica relevante de explicação global é o SHapley Additive exPlanations (SHAP), baseado na Teoria dos Jogos, que fornece igualmente explicações locais, o que significa que ele possui capacidade de explicar a relevância de diferentes parâmetros para determinado resultado [42].

Quanto à classificação de acordo com o modelo de explicadores, o método de modelo específico, como o próprio nome antecipa, é destinado a modelos específicos de aprendizado profundo. Enquanto que os métodos de modelos agnósticos podem ser utilizados para qualquer modelo de aprendizado de máquina sem que haja os desafios associados à transferência [45].

Em relação à classificação conforme a metodologia dos explicadores, destacam-se os os métodos baseados em retropropagação, gradiente e perturbação [42].

Além dessas categorias, os métodos de explicabilidade *post-hoc* podem ser agrupados a partir da forma da qual as explicações são produzidas, como os métodos de atribuição, visualização (mapas de saliência), exemplos e extração de conhecimento [42].

Apesar de seus benefícios, diversos estudos apontam limitações importantes da XAI. Explicações excessivamente técnicas podem ser inadequadas para usuários leigos, enquanto explicações excessivamente simplificadas podem gerar uma falsa sensação de compreensão, sem refletir o real funcionamento do modelo [26]. Além disso, não há consenso consolidado sobre como avaliar a qualidade de explicações, o que dificulta sua adoção sistemática em ambientes organizacionais e regulados.

2.3 TÉCNICAS DE PRESERVAÇÃO DA PRIVACIDADE EM SISTEMAS DE IA

A preservação da privacidade é um requisito central em sistemas que lidam com dados pessoais, especialmente dados sensíveis. Para tanto, as tecnologias de melhoria de privacidade (*Privacy Enhancing Tecnologies* (PETs)), por vezes referenciadas também como tecnologias de preservação de privacidade (*Privacy Preserving Technologies*), desempenham papel fundamental nessa tarefa [46]. Segundo definição da OCDE (2023, p.4) [47], PETs são um conjunto de tecnologias e abordagens que permitem o tratamento de dados, mantendo a confidencialidade e a privacidade desses dados.

Nesse sentido, técnicas como privacidade diferencial (*Differential Privacy* (DP)), aprendizado federado (*Federated Learning* (FL)), computação multi-partes segura *(Secure Multi-party Computation* (SMPC)), criptografia homomórfica, anonimização de dados e utilização de dados sintéticos são amplamente utilizadas para reduzir riscos de vazamento de informações [20] [48] [49].

A *Differential Privacy* introduz ruído estatístico controlado nos dados ou nos resultados do modelo, fornecendo garantias formais contra a reidentificação de indivíduos. O *Federated Learning*, por sua vez, permite o treinamento descentralizado de modelos, mantendo os dados localmente e compartilhando apenas atualizações de parâmetros [20].

A *Secure Multi-Party Computation* possibilita a interação de várias partes, sem que dados de entrada sejam revelados entre os agentes,garantindo que apenas as saídas sejam conhecidas por ambas partes [49]. De modo semelhante, a criptografia homomórfica permite a manipulação de dados criptografados, sem a necessidade de descriptografia [50].

Por sua vez, a anonimização consiste em um processo que elimina a possibilidade de um dado ser associado de forma direta ou indireta a um indivíduo [10].

Já a utilização de dados sintéticos, permite que replicar propriedades estatísticas da base de dados original, ocultando a identidade destes. [51]

Embora eficazes, essas técnicas introduzem impactos significativos na utilidade dos modelos, na acurácia e na interpretabilidade dos resultados [52]. No documento "*Sharing trustworthy AI models with privacy-enhancing technologies*", a OCDE mapeia as principais limitações associadas a diferentes PETs, sintetizadas na tabela a seguir, traduzida pela autora [53].

Essas abordagens são amplamente incentivadas por legislações como a LGPD e o GDPR, que estabelecem princípios como necessidade, adequação e prestação de contas. Contudo, a aplicação dessas técnicas pode limitar a disponibilidade de informações necessárias para explicar decisões automatizadas de forma significativa.

Tabela 2.1: PETs e suas principais funções, desafios e limitações, segundo OCDE

Tipo de PET	Funções	Desafios e limitações
Ferramentas de ofuscação de dados		
Dados Sintéticos (DS)	- Otimização e teste de modelos sem uso dos dados originais, produzindo dados sintéticos que são anônimos; - Compartilhamento de propriedades estatísticas sem revelar dados reais	- Assegurar que não haja vazamento de informações nas técnicas de geração de dados sintéticos, as quais variam em termos de fidelidade, utilidade e privacidade (risco de reidentificação); - Avaliar a acurácia e a possibilidade de ampliação de vieses nos modelos de IA treinados com base nos DS
Privacidade Diferencial (PD)	- Adição de ruído para reduzir reidentificação no dado de entrada ou de teste, combinado com DS e AF para reduzir o risco de vazamento de dados e de reidentificação, devido à memorização do modelo	- Desafio em determinar o nível adequado de ruído, particularmente para dados não estruturados e multimodais; - Desafio em lidar com efeitos acumulativos quando utilizado diferentes estágios com DS e/ou AF; - Custos computacionais
Processamento de dados criptografados		
Criptografia Homomórfica (CH)	- Treinamento e ajuste fino de modelos de IA com base em dados criptografados; - Atualizações confidenciais de modelos em Aprendizado Federado (AF), como alternativa aos altos custos de comunicação da MPCL	- Alto custo computacional; - Desafios de pré-processamento - Processamento de criptografia pode alterar certas características dos dados, potencialmente afetando os resultados do treinamento ou do ajuste fino do modelo
Computação Multi-Partes Segura (SMPC)	- Reduz o risco de vazamento de dados durante AF graças ao compartilhamento secreto - Agrega confidencialidade ais dados de entrada ou identifica pontos em comum nas diferentes bases de dados antes do treinamento	- Alto custo de comunicação; - Baixa visibilidade, que complica o teste e a correção de falhas - problemas de latência - risco de conluio em modelos semi-honestos
Análise federada e distribuída		
Aprendizado Federado (AD)	- Processamento de dados exclusivamente nos dispositivos dos usuários para o treinamento local de modelos de IA, em que apenas as atualizações do modelo, e não os dados brutos, são agregados para melhorar o modelo global; - Facilita a confidencialidade, preservando a co-criação de modelos por diversas entidades sem centralizar os dados quando combinado com outras PETs para proteger os dados de entrada	- Elevado custo de comunicação não apenas aumenta a necessidade de conectividade confiável, como também impõe limitações significativas à otimização de hiperparametros, o que pode afetar a escalabilidade e o desempenho ao utilizar camadas adiconais de PETs; - Informações sobre os modelos precisam ser disponibilizadas ao controlador de dados, o que é necessário para a otimização refinada - Garantir que não haja vazamento de informações, uma vez que o AF também está sujeito a ataques de reconstrução de dados -Coordenação entre diversos controladores quanto à modelagem de ameaças questões relacionadas ao pipeline de dados e à complexidade da testagem de vulnerabilidades

2.4 COMPATIBILIDADE E TRADE-OFFS ENTRE XAI E PRESERVAÇÃO DA PRIVACIDADE

A coexistência entre XAI e técnicas de preservação da privacidade introduz uma tensão fundamental. Enquanto a XAI busca revelar informações sobre o comportamento do modelo, mecanismos de privacidade têm como objetivo restringir a divulgação de informações. Exemplos de evidências nesse sentido incluem análises de risco de privacidade em explicações e ataques de inferência/extração [54, 21], bem como revisões que sistematizam trade-offs entre explicabilidade, privacidade e segurança e organizam os achados por classes de XAI [55].

Em particular, explicações locais, como importância de atributos e explicações contrafactuais, podem facilitar ataques de inferência de associação, extração ou inversão de modelos [22]. Esses resultados desafiam a suposição de que explicabilidade e privacidade são naturalmente compatí-

veis, evidenciando a necessidade de avaliar cuidadosamente os *trade-offs* envolvidos.

Trabalhos recentes avançam ao analisar empiricamente essa compatibilidade. Senevirathna et al. [56] demonstram que técnicas de XAI podem ser aplicadas em cenários com DP e FL, mas com impactos mensuráveis na fidelidade das explicações, na utilidade do sistema e na complexidade organizacional. Outros estudos reforçam que garantias fortes de privacidade tendem a degradar a precisão e a utilidade das explicações [57, 58].

O estudo conduzido por Allana et al. [57] apresenta uma revisão de escopo abrangente sobre riscos de privacidade associados à explicabilidade em sistemas de IA, bem como métodos de mitigação e características desejáveis de explicações sensíveis à privacidade. O trabalho sistematiza ataques, vetores de vazamento e estratégias de defesa em XAI, reforçando empiricamente a existência de tensões estruturais entre transparência e proteção de dados. Este estudo contribui especialmente para a discussão sobre *trade-offs* entre fidelidade da explicação, utilidade do sistema e privacidade.

Spartalis et al. [55] realizam uma revisão da literatura sobre a interação entre explicabilidade, privacidade e segurança, com motivação em cenários de Infraestruturas Críticas. Os autores destacam que mecanismos de XAI, ao "clarearem" modelos de caixa-preta, podem ampliar a superfície de ataque e viabilizar vazamentos de informação, sobretudo quando explicações são disponibilizadas a atores não confiáveis. Além disso, os autores discutem como diferentes classes de XAI (por exemplo, explicações baseadas em exemplos, gradientes e perturbações) apresentam perfis distintos de risco de privacidade, e argumentam que explicações de maior qualidade tendem a elevar a exposição a ataques, reforçando a necessidade de decisões explícitas de projeto e governança ao combinar XAI com defesas como privacidade diferencial e aprendizado federado.

Esses achados indicam que a compatibilidade entre XAI e privacidade é condicional, dependendo do tipo de explicação, da técnica de preservação adotada, do orçamento de privacidade e do público-alvo das explicações. Em particular, esses resultados sugerem que a escolha do *tipo* e do *nível* de explicação deve ser tratada como decisão de projeto e de governança, pois afeta simultaneamente riscos de privacidade, utilidade do sistema e objetivos de responsabilização.

2.5 EXPLICABILIDADE COMO QUESTÃO SOCIOTÉCNICA E DE GOVERNANÇA

Além dos aspectos técnicos, a literatura recente enfatiza que a explicabilidade deve ser compreendida como uma questão sociotécnica e organizacional. Diferentes stakeholders: desenvolvedores, gestores, profissionais de privacidade, auditores e usuários finais possuem necessidades distintas de explicação [59].

Abordagens orientadas à governança defendem explicações em camadas (*layered explanations*), nas quais diferentes níveis de detalhe são fornecidos conforme o papel e a responsabilidade do stakeholder. Essa perspectiva reconhece que maximizar simultaneamente transparência, privacidade e eficiência é, em geral, inviável, exigindo decisões explícitas de priorização [60].

Nesse sentido, explicabilidade deixa de ser um atributo puramente técnico e passa a ser um elemento de governança, apoiando a responsabilização, a conformidade regulatória e a confiança institucional. A ausência de diretrizes sistemáticas para equilibrar esses fatores representa uma lacuna relevante na literatura, especialmente em contextos de decisão automatizada sensíveis à privacidade.

A Tabela 2.2 sintetiza os principais estudos correlatos analisados, destacando os mecanismos de XAI empregados, as técnicas de preservação da privacidade consideradas, os *trade-offs* identificados e os stakeholders abordados. A análise evidencia que os estudos existentes tendem a tratar isoladamente aspectos técnicos ou organizacionais, havendo escassez de abordagens integradas que considerem simultaneamente compatibilidade técnica, confiança e responsabilização.

Essa lacuna motiva a presente pesquisa, que busca investigar a explicabilidade em sistemas sensíveis à privacidade de forma integrada, culminando na proposição de um *framework* orientado a apoiar decisões técnicas e de governança.

Os trabalhos correlatos evidenciam que a relação entre explicabilidade e preservação da privacidade tem sido abordada sob perspectivas distintas e, em grande medida, fragmentadas. Parte significativa da literatura concentra-se nos riscos técnicos associados à exposição indevida de informações por meio de explicações, enquanto outros estudos enfatizam os efeitos da explicabilidade sobre confiança, responsabilização e governança, raramente integrando essas dimensões.

Estudos com foco em segurança e privacidade demonstram de forma consistente que mecanismos de explicação podem atuar como vetores de ataque. Ezzeddine et al. [22] mostram que explicações locais, como importância de atributos e contrafactuais, podem facilitar ataques de inferência de associação, extração e inversão de modelos, especialmente quando aplicadas a sistemas que lidam com dados sensíveis. Esses resultados desafiam a visão normativa de que a explicabilidade é sempre desejável, indicando que explicações mal projetadas podem ampliar riscos à privacidade.

Em contraste, trabalhos como o de Senevirathna et al. [56] adotam uma perspectiva sistêmica, propondo integrar explicabilidade, privacidade e responsabilização ao longo do ciclo de vida do sistema. Os autores demonstram empiricamente que técnicas de XAI permanecem viáveis em ambientes com aprendizado federado e mecanismos de controle de acesso, mas introduzem *trade-offs* relevantes, como aumento da complexidade arquitetural, custos operacionais e necessidade de governança contínua. Esse tipo de abordagem evidencia que a compatibilidade entre XAI e privacidade é condicional e depende de decisões organizacionais explícitas.

Outros estudos concentram-se nos efeitos da explicabilidade sobre a confiança dos usuários. A revisão sistemática conduzida por Wiratsin e Rivepiboon [5] indica que explicações tendem a aumentar a confiança em sistemas de IA, mas esse efeito varia conforme o tipo de explicação, o contexto e o nível de expertise do usuário. No entanto, tais estudos raramente consideram cenários com fortes restrições de privacidade, tratando explicabilidade e proteção de dados como preocupações independentes.

Tabela 2.2: Síntese dos estudos correlatos sobre XAI e sistemas com preservação da privacidade

Estudo	**Mecanismos de XAI**	**Técnicas de Preservação da Privacidade**	**Principais *trade-offs* identificados**	**Stakeholders**
Ezzeddine et al. [22]	Importância de atributos, explicações contrafactuais	Análise implícita de exposição de dados	Explicações podem viabilizar ataques de inferência de associação, extração e inversão de modelos; maior transparência amplia riscos de vazamento de privacidade	Provedores de serviço, auditores
Spartalis et al. [55]	Taxonomia de XAI (p. ex., LIME, SHAP, LRP, Grad-CAM, *Integrated Gradients*, contrafactuais) e critérios de avaliação (estabilidade, fidelidade, robustez)	Privacidade diferencial, aprendizado federado, criptografia homomórfica, anonimização/dados sintéticos; (discute também ataques de privacidade e segurança)	XAI pode ampliar a superfície de ataque e aumentar riscos; explicações "mais informativas" podem elevar vazamento; privacidade diferencial pode reduzir fidelidade/compreensibilidade; aprendizado federado não elimina risco (p. ex., via gradientes); criptografia homomórfica impõe custo computacional; combinações de defesas mitigam riscos, mas elevam complexidade	Operadores/ analistas de segurança, gestores e profissionais em domínios críticos
Senevirathna et al. [56]	SHAP, LIME, LRP; métricas de responsabilização (estabilidade, consistência)	Aprendizado federado, controle de acesso, trilhas de auditoria	Aumento da auditabilidade e da responsabilização à custa de maior complexidade arquitetural e sobrecarga organizacional	Operadores de sistemas, profissionais de conformidade
Wiratsin e Rivepiboon [5]	Explicações agnósticas e específicas ao modelo	Não abordado explicitamente	Explicabilidade tende a aumentar a confiança, mas os efeitos variam conforme tipo de explicação, contexto e expertise do usuário; privacidade não é tratada	Usuários finais, tomadores de decisão
Shah et al. [61]; Radanliev [31]	Explicações *post-hoc* em sistemas de apoio à decisão	Minimização de dados, salvaguardas regulatórias	Necessidade de equilibrar explicações significativas com requisitos estritos de confidencialidade e proteção de dados	Profissionais de saúde, pacientes, reguladores
Keaney et al. [59]	Explicações em camadas e orientadas ao papel do usuário	*Privacy-by-design*, controle de acesso	Explicabilidade deve ser adaptada aos diferentes stakeholders; *trade-offs* organizacionais superam decisões puramente técnicas	Gestores, responsáveis por governança e conformidade
Wasif et al. [62]	Não é o foco principal	Privacidade diferencial	Garantias mais fortes de privacidade reduzem acurácia, justiça e utilidade do modelo, evidenciando *trade-offs* multidimensionais inevitáveis	Cientistas de dados, formuladores de políticas
Zhang et al. [60]	Pistas de explicabilidade e mecanismos de transparência	Controles de privacidade e segurança	É inviável maximizar simultaneamente transparência, privacidade e eficiência; é necessário priorizar objetivos	Gestores, projetistas de sistemas
Cabitza et al. [63]	Comunicação de incerteza e escores de confiança	Aplicável em domínios sensíveis sem expor dados brutos	Pistas de confiança podem prejudicar a calibração adequada da confiança quando há assimetria de conhecimento	Usuários finais, tomadores de decisão

Em domínios regulados, como saúde e serviços públicos, a literatura enfatiza que explicabilidade está diretamente associada à responsabilização e à legitimidade das decisões. Shah et al. [61] e Radanliev [31] argumentam que explicações devem permitir auditoria e contestação,

mas alertam que transparência excessiva pode violar princípios de confidencialidade e proteção de dados. Essa tensão reforça a necessidade de abordagens proporcionais e contextualizadas.

Uma contribuição relevante surge em trabalhos orientados à governança organizacional. Keaney et al. [59] propõem explicações em camadas, adaptadas aos diferentes papéis dos stakeholders, reconhecendo que desenvolvedores, gestores, auditores e usuários finais possuem necessidades distintas de explicação. Essa visão é reforçada por Zhang et al. [60], que demonstram empiricamente a impossibilidade de maximizar simultaneamente transparência, privacidade e eficiência, exigindo priorização explícita de objetivos. Sob essa perspectiva, a explicabilidade deixa de ser tratada como uma propriedade absoluta do modelo e passa a ser entendida como um recurso organizacional, cujo nível e forma de implementação dependem do domínio de aplicação, do perfil dos stakeholders e do grau de exposição regulatória [55]. Esses estudos reforçam que explicações excessivamente detalhadas podem introduzir riscos adicionais, incluindo vazamentos de informação sensível, aumento da superfície de ataque e dificuldades de conformidade com legislações de proteção de dados.

Por fim, estudos como os de Wasif et al. [62] e Cabitza et al. [63] mostram que tanto a preservação da privacidade quanto mecanismos de explicação e comunicação de incerteza introduzem efeitos colaterais indesejados, como perda de acurácia, redução de justiça algorítmica ou má calibração da confiança. Esses achados reforçam que os *trade-offs* são estruturais e não meramente decorrentes de implementações inadequadas.

Do ponto de vista da engenharia de software, análises de discussões de desenvolvedores sobre conformidade com legislações como GDPR[11] e LGPD[10] indicam que equipes enfrentam dificuldades práticas para operacionalizar requisitos de transparência e explicação em sistemas reais, recorrendo frequentemente a soluções ad hoc [64, 65]. Isso evidencia uma lacuna entre princípios normativos e práticas técnicas, reforçando a necessidade de frameworks que orientem decisões de projeto de forma sistemática.

A literatura converge ao indicar que a compatibilidade entre XAI e preservação da privacidade não é intrínseca, mas depende de trade-offs explícitos, do contexto de uso e de escolhas de governança, o que fundamenta a necessidade de um framework que apoie decisões técnicas e organizacionais de explicabilidade sensível à privacidade.

2.6 SÍNTESE DO CAPÍTULO

A literatura revela três lacunas principais: (i) ausência de abordagens integradas que considerem simultaneamente compatibilidade técnica, confiança e responsabilização; (ii) escassez de diretrizes práticas para seleção de mecanismos de XAI em ambientes sensíveis à privacidade; e (iii) pouca atenção ao papel da governança e à diferenciação entre stakeholders. Essas lacunas fundamentam a necessidade desta pesquisa e orientam a proposição do *framework* apresentado nos capítulos seguintes.

3 CONFIGURAÇÕES DO ESTUDO

3.1 MÉTODO DE PESQUISA

Esta pesquisa adota um desenho qualitativo de natureza mista (*mixed qualitative design*), combinando (i) uma análise estruturada da literatura e (ii) um estudo empírico exploratório por meio de grupo focal [66]. O objetivo é investigar os trade-offs entre explicabilidade e preservação da privacidade sob duas perspectivas complementares: (a) a perspectiva conceitual e técnica, identificando mecanismos, condições de compatibilidade e superfícies de risco; e (b) a perspectiva socio-organizacional, compreendendo como diferentes stakeholders percebem níveis de explicação, confiança e responsabilização em sistemas de apoio à decisão com restrições de privacidade. O desenho metodológico foi alinhado às Questões de Pesquisa definidas na Seção 1, permitindo abordar simultaneamente (QP1) compatibilidade e trade-offs e (QP2) efeitos de níveis de explicabilidade sobre confiança e responsabilização.

A escolha por métodos qualitativos se justifica por duas razões principais. Primeiro, o fenômeno investigado é sociotécnico: envolve escolhas de projeto e governança que não podem ser analisadas apenas por métricas de desempenho de modelos, pois dependem do domínio, do risco, do arcabouço regulatório e do público-alvo das explicações. Segundo, a literatura recomenda métodos qualitativos para explorar percepções, práticas e dilemas em cenários onde há múltiplos atores e conceitos abstratos (por exemplo, "explicação adequada", "confiança" e "responsabilização") [66]. Dessa forma, a análise estruturada da literatura fornece a base conceitual e a consolidação dos achados; e o grupo focal fornece evidências empíricas sobre como tais tensões são interpretadas e negociadas na prática.

A Figura 3.1 sintetiza o encadeamento das etapas. A primeira etapa resultou em uma síntese comparativa dos estudos correlatos (Tabela 2.2), estruturada por mecanismos de XAI, técnicas de preservação da privacidade, *trade-offs* e *stakeholders*. Em seguida, essa síntese foi utilizada como insumo para planejar o grupo focal (papéis convidados, questões norteadoras e temas), permitindo uma análise integrada e diretamente conectada às QPs.

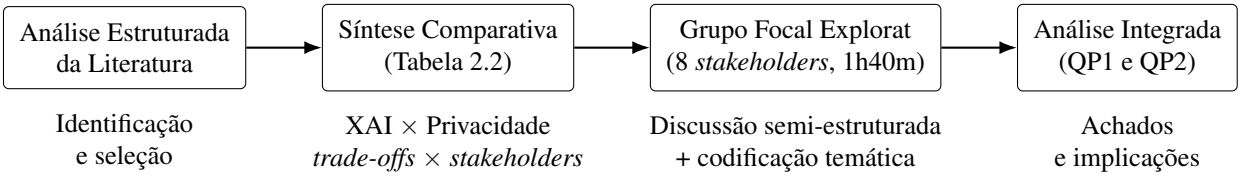

Figura 3.1: Visão geral do método de pesquisa.

3.1.1 Etapa 1: Revisão da Literatura

A primeira etapa consistiu em uma revisão da literatura sobre Inteligência Artificial Explicável (XAI) e técnicas de preservação da privacidade em sistemas de informação. Esta etapa teve como finalidade construir uma base de evidências suficientemente ampla e atual para: (i) identificar mecanismos de XAI relevantes (globais, locais, intrínsecos e post-hoc); (ii) mapear técnicas de preservação da privacidade recorrentes (por exemplo, privacidade diferencial, aprendizado federado, anonimização/dados sintéticos, computação segura); (iii) organizar trade-offs reportados (fidelidade da explicação, utilidade do sistema, complexidade organizacional e risco de vazamento); e (iv) caracterizar os principais grupos de stakeholders discutidos na literatura (por exemplo, usuários finais, tomadores de decisão, desenvolvedores, auditores e responsáveis por conformidade).

3.1.1.1 Estratégia de busca e fontes.

Os estudos foram identificados por meio de buscas direcionadas em bases bibliográficas (por exemplo, DBLP e Google Scholar), combinando termos relacionados à explicabilidade (por exemplo, “Explainable AI”, “XAI”, “interpretable machine learning”) e à privacidade (por exemplo, “privacy-preserving”, “differential privacy”, “federated learning”, “data protection”). Buscas adicionais foram realizadas por snowballing (para trás e para frente), a partir de artigos centrais e revisões recentes, com o objetivo de capturar estudos altamente citados e trabalhos emergentes diretamente relacionados aos trade-offs de explicabilidade sob restrições de privacidade.

3.1.1.2 Critérios de inclusão e exclusão.

Foram priorizados: (i) artigos revisados por pares (conferências e periódicos), (ii) surveys e revisões recentes, e (iii) trabalhos conceituais ou empíricos publicados a partir de 2018 que explicitamente discutem explicabilidade sob restrições de privacidade *ou* discutem confiança, responsabilização e governança em sistemas de apoio à decisão com dados sensíveis. Foram excluídos trabalhos que: (a) tratavam apenas de interpretabilidade em sentido estritamente técnico, sem discutir geração de explicações, transparência para stakeholders ou implicações de privacidade; (b) discutiam privacidade sem relação com explicabilidade; ou (c) eram relatos curtos sem informação suficiente para extração das dimensões analíticas.

3.1.1.3 Extração e síntese dos dados.

Para cada estudo selecionado, foram extraídas informações segundo quatro dimensões analíticas: (i) mecanismos de XAI abordados; (ii) técnica(s) de preservação da privacidade considerada(s); (iii) trade-offs e riscos reportados (por exemplo, degradação de utilidade, perda de fidelidade, aumento de superfície de ataque, sobrecarga de governança); e (iv) stakeholders explicitamente considerados. Essas dimensões orientaram a síntese comparativa apresentada na

Tabela 2.2. A análise também serviu como base para o planejamento do grupo focal, apoiando a seleção de temas, exemplos e dilemas a serem discutidos com participantes de diferentes papéis.

3.1.2 Etapa 2: Estudo Empírico Exploratório por Grupo Focal

Para complementar a síntese conceitual e ancorar os achados em contextos reais de decisão, foi conduzido um grupo focal exploratório. Grupos focais são recomendados como método qualitativo para investigar percepções, experiências e critérios de julgamento em temas complexos, especialmente quando se busca contraste entre perspectivas e construção coletiva de sentido [66]. No escopo desta dissertação, o grupo focal foi utilizado para examinar: (i) como participantes percebem compatibilidade e riscos de diferentes explicações em ambientes sensíveis à privacidade (QP1); e (ii) como níveis de explicabilidade influenciam confiança e responsabilização em diferentes papéis organizacionais (QP2).

3.1.2.1 Participantes e Amostragem

O grupo focal contou com oito participantes, selecionados por amostragem intencional (*purposive sampling*) com base em experiência profissional e familiaridade com, ao menos, um dos seguintes eixos: (i) desenvolvimento/implantação de sistemas intensivos em software; (ii) IA e modelos de ML; (iii) privacidade, proteção de dados e conformidade; e (iv) tomada de decisão e governança organizacional. A estratégia de amostragem buscou assegurar diversidade de perspectivas e papéis, incluindo stakeholders técnicos (por exemplo, desenvolvedor e engenheiro de ML), gerenciais (por exemplo, gerente e product owner) e de governança (por exemplo, profissionais de privacidade e conformidade, além de tomadores de decisão).

A seleção ocorreu por conveniência a partir de redes profissionais das pesquisadoras, prática comum em estudos exploratórios, desde que as limitações sejam explicitadas (Seção de Ameaças à Validade) [66]. Para mitigar vieses, buscou-se manter um perfil heterogêneo e evitar dominância de um único papel (por exemplo, apenas desenvolvedores), favorecendo a comparação entre visões e tensões inter-papéis.

3.1.2.2 Planejamento, Instrumento e Condução

Foi realizada uma única sessão, com duração aproximada de 1h40m, em formato semi-estruturado. Esse formato combina um roteiro orientador (para garantir cobertura dos temas das QPs) com flexibilidade para aprofundar pontos emergentes, permitindo que participantes construam e confrontem argumentos entre si. Antes do início, foram apresentados: (i) objetivos do estudo; (ii) regras de convivência (respeito, não interrupção, direito de não responder); e (iii) termos de confidencialidade e anonimização dos relatos.

O roteiro foi organizado em blocos temáticos alinhados às QPs: (a) noções de "explicação adequada" sob restrições de privacidade; (b) situações em que aumentar explicabilidade eleva risco

de privacidade; (c) necessidades de explicação por papel (*stakeholder-dependent explainability*); (d) confiança e calibração (quando confiar/desconfiar); e (e) responsabilização (defensabilidade, rastreabilidade e justificativa para auditoria/contestação). Um(a) moderador(a) conduziu a sessão com foco em equilibrar participação, evitar indução de respostas e estimular contrapontos, conforme recomendações metodológicas [66]. Para reduzir efeitos de dominância, foram empregadas estratégias como: convites diretos a participantes menos ativos, rodadas rápidas de opinião em perguntas críticas e sínteses parciais para validação coletiva.

3.1.2.3 Coleta e Análise dos Dados

A sessão foi gravada em áudio mediante consentimento e posteriormente transcrita. A análise seguiu abordagem qualitativa orientada por questões (*issue-based analysis*), com codificação temática iterativa: (i) codificação aberta inicial em trechos relevantes; (ii) agrupamento em temas; e (iii) consolidação em categorias analíticas diretamente relacionadas às QPs. A ênfase foi identificar padrões convergentes e divergentes entre papéis, em vez de quantificar opiniões, coerente com o caráter exploratório do estudo.

O processo de codificação foi refinado iterativamente: códigos iniciais (por exemplo, "explicação local vaza informação") foram consolidados em temas de maior nível (por exemplo, "compatibilidade depende da granularidade da explicação") e associados às QPs. Divergências na delimitação de temas foram discutidas entre as pesquisadoras até atingir consenso. A Tabela 3.1 apresenta um exemplo ilustrativo do encadeamento *dado bruto* $\rightarrow$ *código inicial* $\rightarrow$ *tema consolidado* $\rightarrow$ *QP*.

O estudo envolveu participação voluntária de profissionais e reporta resultados de forma anonimizada (P1–P8). Foram adotadas salvaguardas de privacidade e minimização de dados (armazenamento restrito do áudio/transcrição e supressão de informações identificáveis). Quanto à apreciação ética, o procedimento seguiu as orientações institucionais aplicáveis ao tipo de pesquisa e ao nível de risco envolvido.

Tabela 3.1: Exemplo do processo de codificação (*dado* $\rightarrow$ *código* $\rightarrow$ *tema* $\rightarrow$ *QP*)

Dado Bruto	**Código Inicial**	**Tema**	**QP**
"Explicações locais podem ser úteis, mas parecem perigosas em sistemas sensíveis." (P5)	Explicações locais podem vazar informação sensível	Compatibilidade depende da granularidade da explicação	QP1
"Quanto mais forte a privacidade, mais genéricas e menos acionáveis ficam as explicações." (P1)	Privacidade reduz utilidade da explicação	*Trade-offs* entre privacidade, fidelidade e utilidade	QP1
"Para responsabilização, as explicações precisam ser consistentes e defensáveis ao longo do tempo." (P6)	Responsabilização exige defensabilidade	Explicações orientadas à responsabilização	QP2
"Uma única explicação raramente funciona para desenvolvedores e gestores ao mesmo tempo." (P2)	Papéis diferentes exigem explicações diferentes	Necessidades de explicação dependem do *stakeholder*	QP2

3.2 RESULTADOS DO GRUPO FOCAL

Esta seção apresenta os resultados do grupo focal exploratório conduzido para complementar os achados baseados na literatura. Inicialmente, descreve-se o perfil dos participantes, de modo a contextualizar as perspectivas coletadas. Em seguida, são apresentadas as perguntas norteadoras que estruturaram a discussão. Por fim, os achados qualitativos são organizados e analisados à luz das questões de pesquisa (QP1 e QP2), destacando convergências, divergências e implicações para o desenho de explicações em sistemas de apoio à decisão com restrições de privacidade.

3.2.1 Perfil dos Participantes

O grupo focal contou com oito participantes selecionados por amostragem intencional, considerando experiência profissional com sistemas intensivos em software, inteligência artificial e/ou tomada de decisão organizacional. Para preservar a confidencialidade, os participantes foram identificados por rótulos anonimizados (P1–P8). A composição do grupo buscou diversidade de papéis relevantes para sistemas de apoio à decisão com preservação da privacidade, mantendo um nível de expertise suficientemente equilibrado para permitir troca entre perspectivas técnicas, gerenciais e de governança.

Em média, os participantes possuíam aproximadamente sete anos de experiência profissional. As idades variaram de 32 a 48 anos, com formações que incluíram graduação, mestrado completo e estudantes de mestrado. Essa diversidade favoreceu a exploração de visões complementares sobre riscos, necessidades de explicação, requisitos de responsabilização e limitações práticas para operacionalizar explicabilidade sob restrições de privacidade.

Tabela 3.2: Perfil dos participantes do grupo focal

ID	**Papel**	**Idade (anos)**	**Experiência**	**Escolaridade**
P1	Desenvolvedor(a)	32	6 anos	Graduação
P2	Engenheiro(a) de ML	34	7 anos	Mestrando(a)
P3	Gerente de Projetos	38	6 anos	Mestrando(a)
P4	Product Owner	41	9 anos	Mestrado
P5	Profissional de Privacidade	36	7 anos	Mestrado
P6	Profissional de Conformidade	39	6 anos	Mestrado
P7	Tomador(a) de decisão	45	8 anos	Mestrado
P8	Tomador(a) de decisão	48	10 anos	Mestrado

A discussão do grupo focal foi conduzida a partir de perguntas abertas elaboradas para (i) elicitar percepções sobre explicabilidade sob restrições de privacidade, (ii) explorar trade-offs percebidos e (iii) capturar expectativas específicas de diferentes stakeholders quanto à confiança e à responsabilização. As perguntas foram alinhadas às questões de pesquisa e organizadas em blocos temáticos para assegurar cobertura dos principais tópicos, sem limitar a emergência de novos pontos durante a discussão, conforme apresentado na Tabela 3.3.

Tabela 3.3: Perguntas norteadoras do grupo focal

Tema	Pergunta norteadora
Explicabilidade	O que significa uma explicação adequada em sistemas que operam sob restrições estritas de privacidade?
Trade-offs	Em que situações aumentar a explicabilidade pode introduzir riscos à privacidade ou à proteção de dados?
Necessidades dos stakeholders	Diferentes stakeholders precisam de níveis ou tipos distintos de explicação? Por quê?
Confiança	Que tipos de explicação aumentam ou diminuem sua confiança em sistemas de apoio à decisão com preservação de privacidade?
Responsabilização	Que informações ou explicações são necessárias para apoiar responsabilização e justificativa de decisões nesses sistemas?
Implicações de projeto	Como organizações devem definir o nível apropriado de explicabilidade quando há restrições de privacidade?

3.2.2 QP1. Compatibilidade e Trade-offs entre XAI e Técnicas de Preservação da Privacidade

Esta subseção apresenta os resultados relacionados à QP1, que investiga a compatibilidade entre mecanismos de explicabilidade e técnicas de preservação da privacidade, bem como os trade-offs e limitações percebidos pelos participantes. Os achados foram organizados em três temas inter-relacionados: (i) compatibilidade percebida entre explicabilidade e privacidade, (ii) trade-offs entre transparência, privacidade e utilidade, e (iii) limitações técnicas e organizacionais para adoção de explicabilidade sensível à privacidade.

3.2.2.1 Compatibilidade percebida entre explicabilidade e preservação da privacidade

De modo geral, os participantes convergiram na ideia de que explicabilidade e preservação da privacidade não são mutuamente excludentes, mas sua compatibilidade é fortemente dependente do contexto. Participantes com perfil técnico enfatizaram que certos mecanismos de explicação, como explicações agregadas e pós-treinamento, podem ser combinados com técnicas como aprendizado federado e privacidade diferencial, desde que as explicações operem sobre representações abstraídas e evitem expor pontos individuais de dados ou valores sensíveis. Um participante sintetizou esse ponto da seguinte forma:

> *"Ainda é possível explicar como o modelo se comporta no geral, mesmo em um cenário federado ou com preservação de privacidade, desde que se evite expor pontos individuais de dados ou valores específicos de atributos." (P2, Engenheiro(a) de ML)*

Além disso, desenvolvedores e engenheiros de ML relataram que explicações globais e evidências em nível de modelo foram percebidas como mais adequadas a contextos com restrições de privacidade do que explicações em nível de instância. Em contraste, explicações locais e contrafactuais foram vistas como potencialmente mais informativas para diagnóstico e tomada de decisão, porém com maior risco em domínios sensíveis ou regulados. Essa preocupação foi

ilustrada por:

> *"Explicações locais podem ser muito úteis, mas parecem perigosas em sistemas sensíveis, porque você pode acabar revelando mais sobre os dados do que pretendia." (P5, Profissional de Privacidade)*

3.2.2.2 Trade-offs entre explicabilidade, privacidade e utilidade

Um tema central do grupo focal foi a percepção de trade-offs inevitáveis entre explicabilidade, privacidade e desempenho/utilidade do sistema. Os participantes relataram que, ao aumentar garantias de privacidade (por exemplo, por meio de maior ruído, restrições de acesso ou limitação de informações compartilhadas), as explicações tendem a se tornar mais genéricas e menos acionáveis. Por outro lado, explicações mais detalhadas e específicas podem elevar o risco de vazamento de privacidade e de uso indevido. Esse equilíbrio foi expresso por:

> *"Quando aplicamos mecanismos mais fortes de privacidade, as explicações ficam mais genéricas e às vezes menos acionáveis, o que limita sua utilidade para depuração ou apoio à decisão." (P1, Desenvolvedor(a))*

Participantes em funções gerenciais e de privacidade destacaram que tais trade-offs não são apenas decisões técnicas. Em muitos casos, representam escolhas organizacionais que precisam considerar conformidade regulatória, eficiência operacional e expectativas de stakeholders. Nessa linha, explicabilidade foi descrita como uma propriedade negociada, não como característica fixa do sistema:

> *"Decidir o quanto explicar é, no fim, uma decisão de governança. Precisamos ponderar transparência contra risco legal, desempenho e expectativas de diferentes stakeholders." (P4, Product Owner)*

3.2.2.3 Limitações técnicas e organizacionais

Os participantes também identificaram limitações práticas para adoção de explicabilidade sensível à privacidade. Entre os desafios técnicos, foram citados: aumento de sobrecarga computacional, redução de acurácia sob garantias fortes de privacidade e dificuldades para manter consistência das explicações em ambientes distribuídos/federados. Além disso, a depuração e validação do comportamento do modelo podem se tornar mais difíceis quando explicações precisam ser intencionalmente abstratas:

> *"Quando as explicações ficam muito abstratas para preservar privacidade, fica mais difícil rastrear erros ou entender por que o modelo se comporta de forma inesperada, especialmente em ambientes distribuídos." (P3, Gerente de Projetos)*

Do ponto de vista organizacional, emergiu a percepção de ausência de diretrizes e ferramentas padronizadas para implementar explicabilidade em sistemas com preservação de privacidade. Esse cenário favorece decisões ad hoc, aumentando complexidade de desenvolvimento e incerteza, sobretudo em domínios regulados nos quais a responsabilização é requisito explícito:

"Não existe uma referência clara sobre o quanto de explicação é suficiente em sistemas sensíveis à privacidade; aí as equipes tomam decisões isoladas que depois são difíceis de justificar." (P7, Tomador(a) de decisão)

Síntese da QP1

Os resultados indicam que explicabilidade e técnicas de preservação da privacidade são compatíveis apenas sob condições dependentes do contexto. Participantes enfatizaram que explicações globais e agregadas tendem a ser mais adequadas a cenários com restrições de privacidade do que explicações locais por instância. Garantias mais fortes de privacidade reduzem fidelidade e utilidade das explicações, reforçando trade-offs inevitáveis. Tais trade-offs não se restringem ao nível técnico, pois envolvem decisões organizacionais e de governança. Por fim, sobrecarga computacional e ausência de diretrizes e ferramentas padronizadas dificultam a adoção prática de explicabilidade sensível à privacidade.

3.2.3 QP2. Confiança e Responsabilização em Sistemas de Apoio à Decisão com Preservação da Privacidade

Esta subseção descreve os achados relacionados à QP2, que investiga como diferentes níveis de explicabilidade afetam a confiança e as percepções de responsabilização de stakeholders em sistemas de apoio à decisão que operam sob restrições de privacidade. Os resultados foram organizados em três temas: (i) explicabilidade como fator de confiança, (ii) responsabilização e justificativa de decisões e (iii) expectativas dependentes do papel do stakeholder.

3.2.3.1 Explicabilidade como fator de confiança

Os participantes associaram explicabilidade ao aumento de confiança em sistemas baseados em IA, especialmente em contextos sensíveis à privacidade. Entretanto, a confiança não foi vinculada a transparência máxima. Em vez disso, stakeholders enfatizaram a necessidade de explicações compreensíveis, significativas e alinhadas às responsabilidades do papel exercido. Um tomador de decisão descreveu essa relação da seguinte forma:

"Eu não preciso ver todos os dados ou os detalhes internos do modelo. O que constrói confiança para mim é entender a lógica por trás da decisão de um jeito que faça sentido para o meu papel." (P8, Tomador(a) de decisão)

Também emergiu a percepção de que explicações excessivamente detalhadas, sobretudo em sistemas que lidam com dados sensíveis, podem reduzir confiança ao aumentar carga cognitiva e/ou acentuar preocupações com exposição indevida:

"Quando as explicações ficam detalhadas demais, especialmente em sistemas com dados sensíveis, isso me deixa mais desconfortável do que confiante." (P5, Profissional de Privacidade)

3.2.3.2 Responsabilização e justificativa de decisões

Responsabilização foi tratada como conceito relacionado, porém distinto, de confiança. Participantes destacaram que explicações são importantes para justificar decisões, sustentar auditorias e responder a questionamentos externos, sobretudo em contextos regulados. Nesse sentido, explicações foram percebidas menos como mecanismo para entender o "interior" do modelo e mais como recurso para documentar e defender decisões:

> *"Em auditorias ou discussões legais, não perguntam como o modelo funciona por dentro, mas por que uma decisão específica foi tomada e se ela pode ser justificada." (P4, Product Owner)*

Um ponto recorrente foi que requisitos de responsabilização diferem do que é necessário para promover confiança no uso cotidiano. Enquanto confiança pode ser sustentada por explicações mais simples e orientadas à decisão, responsabilização demanda explicações consistentes, reprodutíveis e defensáveis ao longo do tempo:

> *"Para responsabilização, as explicações precisam ser consistentes e defensáveis ao longo do tempo, não apenas intuitivas no momento em que são apresentadas." (P6, Profissional de Conformidade)*

3.2.3.3 Expectativas dependentes do papel do stakeholder

Os participantes reforçaram que explicabilidade é dependente do stakeholder: diferentes papéis demandam diferentes níveis e formatos de explicação, de acordo com responsabilidades, expertise e exposição a risco. Tentativas de padronizar uma explicação única para todos os públicos foram vistas como ineficazes e, em alguns casos, contraproducentes:

> *"O que faz sentido para um desenvolvedor é bem diferente do que um gestor ou usuário final precisa. Forçar um único tipo de explicação geralmente não satisfaz ninguém." (P2, Engenheiro(a) de ML)*

Esse achado sustenta a necessidade de explicações em camadas e orientadas a papéis, em que diferentes interfaces e conteúdos de explicação sejam ofertados conforme o objetivo (por exemplo, depuração técnica, monitoramento operacional, prestação de contas, contestação).

Síntese da QP2

Os resultados mostram que explicabilidade exerce papel central na confiança e na responsabilização em sistemas de apoio à decisão com preservação da privacidade. A confiança é favorecida por explicações compreensíveis e alinhadas ao papel do stakeholder, em vez de transparência máxima. Já a responsabilização requer explicações consistentes, defensáveis e apropriadas para auditoria e justificativa em contextos regulados. As necessidades variam entre stakeholders, reforçando a adoção de estratégias de explicação em camadas e orientadas a papéis.

3.3 DISCUSSÃO

Esta seção discute os resultados do estudo à luz da literatura apresentada no Capítulo 2, integrando a síntese conceitual com os achados empíricos obtidos por meio do grupo focal (Seção 3.2). A discussão é estruturada de acordo com as duas questões de pesquisa, analisando como mecanismos de explicabilidade e técnicas de preservação da privacidade interagem na prática e como diferentes níveis de explicabilidade influenciam a confiança e a responsabilização em sistemas de apoio à decisão baseados em IA sensíveis à privacidade.

Ao articular evidências técnicas, organizacionais e percepções de múltiplos stakeholders, os resultados reforçam a compreensão da explicabilidade como um fenômeno sociotécnico e de governança, e não apenas como uma característica algorítmica isolada.

3.3.1 Compatibilidade e Trade-offs entre XAI e Técnicas de Preservação da Privacidade

Os resultados associados à QP1 indicam que a compatibilidade entre explicabilidade e técnicas de preservação da privacidade é condicional e depende de decisões explícitas de projeto. Essa constatação converge com estudos que demonstram que mecanismos como privacidade diferencial, aprendizado federado e computação segura impõem restrições inerentes à granularidade, à fidelidade e à estabilidade das explicações [56, 20].

Tanto a literatura quanto os participantes do grupo focal enfatizaram que explicações globais e agregadas tendem a ser mais viáveis em contextos sensíveis à privacidade, enquanto explicações locais ou no nível da instância introduzem riscos mais elevados de vazamento de informação. Esse achado é consistente com revisões que mostram que explicações baseadas em importância de atributos, gradientes ou contrafactuais podem expor padrões sensíveis ou facilitar ataques de inferência, extração ou inversão de modelos [54, 22, 57].

Nesse sentido, estudos prévios alertam que mecanismos de explicação podem atuar como superfícies de ataque quando não são governados adequadamente [55]. Essa evidência reforça as preocupações manifestadas pelos participantes quanto ao uso de explicações locais em domínios regulados, nos quais a explicabilidade deve ser concebida a partir de modelos explícitos de ameaça à privacidade e à segurança, e não tratada como um complemento neutro ou universal.

Os resultados também reforçam que os trade-offs entre explicabilidade, privacidade e utilidade não são incidentais, mas estruturais. Evidências empíricas mostram que o aumento das garantias de privacidade frequentemente resulta em degradação mensurável da acurácia, da justiça algorítmica ou da utilidade das explicações [62, 56]. As percepções dos participantes quanto à perda de utilidade das explicações sob fortes restrições de privacidade corroboram esses achados, indicando que tais limitações refletem restrições fundamentais e não falhas pontuais de implementação.

Um aspecto relevante evidenciado pelo grupo focal é que esses trade-offs não são apenas técnicos, mas organizacionais. Essa perspectiva dialoga com trabalhos que defendem que a explicabi-

lidade deve ser tratada como uma escolha sociotécnica inserida em estruturas de governança [59]. Os participantes descreveram a explicabilidade como uma propriedade negociada, influenciada por exposição regulatória, requisitos de conformidade, necessidades operacionais e expectativas dos stakeholders, o que converge com abordagens que propõem modelos multidimensionais de confiança e governança em IA [60, 31].

Por fim, as limitações técnicas e organizacionais identificadas, como sobrecarga computacional, dificuldades de depuração em ambientes distribuídos e ausência de diretrizes padronizadas, são consistentes com revisões recentes sobre aprendizado federado e sistemas de IA com preservação da privacidade [67]. Esses resultados sugerem que o avanço da explicabilidade em ambientes sensíveis à privacidade requer não apenas novos algoritmos, mas também ferramentas, padrões de projeto e orientações institucionais que apoiem decisões de explicabilidade sob restrições informacionais.

3.3.2 Explicabilidade, Confiança e Responsabilização

Os resultados relacionados à QP2 indicam que a explicabilidade exerce papel central na construção da confiança e no suporte à responsabilização em sistemas de apoio à decisão com preservação da privacidade. Entretanto, os participantes não associaram confiança à transparência máxima. Em vez disso, a confiança foi relacionada à oferta de explicações compreensíveis, relevantes e adequadas ao papel organizacional do stakeholder.

Esse achado está alinhado com estudos que mostram que a eficácia da explicabilidade depende do alinhamento com o contexto cognitivo e organizacional dos usuários, e não da maximização da quantidade de informação fornecida [68, 5]. Explicações excessivamente técnicas ou detalhadas podem aumentar a carga cognitiva ou gerar desconforto em contextos sensíveis à privacidade, reduzindo a confiança em vez de fortalecê-la.

Uma contribuição relevante deste estudo é a distinção explícita entre explicações orientadas à confiança e explicações orientadas à responsabilização. Os participantes indicaram que explicações voltadas à confiança priorizam a compreensão e a calibração da confiança do usuário, enquanto explicações voltadas à responsabilização devem sustentar justificativas, auditorias e escrutínio externo. Essa distinção é consistente com evidências empíricas que mostram que certos sinais de explicabilidade, como escores de confiança ou indicadores de incerteza, não necessariamente promovem uso apropriado e podem até degradar a qualidade da decisão quando mal calibrados [63].

Em contextos sensíveis à privacidade, nos quais explicações são frequentemente abstraídas, sinais mal calibrados podem comprometer tanto a confiança quanto a responsabilização. Esses resultados dialogam com o conceito de confiança apropriada, segundo o qual explicações devem apoiar o julgamento crítico dos usuários, ajudando-os a decidir quando seguir ou questionar recomendações automatizadas, em vez de simplesmente aumentar a percepção de transparência [26].

Os achados do grupo focal também reforçam a natureza dependente do *stakeholder* da explicabilidade. Desenvolvedores, gestores, profissionais de privacidade, responsáveis por conformidade e tomadores de decisão expressaram expectativas distintas quanto ao conteúdo, à forma e à profundidade das explicações. Essa diversidade de necessidades é amplamente reconhecida em estudos orientados à governança, que defendem estratégias de explicação em camadas e sensíveis ao papel organizacional [59, 60].

Do ponto de vista da responsabilização, os participantes enfatizaram a necessidade de explicações consistentes, rastreáveis e defensáveis ao longo do tempo. Essa visão converge com trabalhos que enquadram a explicabilidade como parte de uma infraestrutura de responsabilização mais ampla, capaz de sustentar conformidade regulatória, auditoria e aprendizado organizacional [56, 55]. Nessa perspectiva, a explicabilidade deixa de ser apenas um mecanismo de apoio à confiança individual e passa a atuar como um artefato organizacional de governança.

3.3.3 Implicações para o Projeto de Sistemas de Apoio à Decisão Sensíveis à Privacidade

Os resultados indicam que a explicabilidade em sistemas sensíveis à privacidade exige uma mudança de enfoque, da transparência máxima para a transparência apropriada e responsável. Em vez de expor detalhes internos dos modelos, o projeto de sistemas deve priorizar estratégias de explicação que equilibrem interpretabilidade, proteção de dados e relevância para a decisão, conforme o contexto e o *stakeholder* envolvido.

Do ponto de vista técnico, isso implica favorecer explicações globais ou agregadas em ambientes altamente regulados e adotar controles rigorosos sobre explicações locais. Do ponto de vista organizacional, os resultados mostram que decisões sobre explicabilidade devem ser integradas a estruturas formais de governança, que definam critérios, responsabilidades, registros e mecanismos de auditoria.

A ausência de diretrizes sistemáticas identificada pelos participantes reforça uma lacuna prática já apontada na literatura, na qual equipes recorrem a soluções ad hoc para operacionalizar requisitos de transparência e privacidade [64]. Esse cenário dificulta a justificativa posterior das decisões e compromete a consistência organizacional, especialmente em contextos regulados.

A integração entre a síntese conceitual e os achados empíricos evidencia que o avanço da XAI em sistemas com preservação da privacidade é tanto um desafio técnico quanto organizacional e institucional. Assim, a explicabilidade deve ser considerada desde as fases iniciais de projeto, em conjunto com requisitos de privacidade, segurança e conformidade, e não incorporada de forma reativa após a implantação.

Com base na síntese da literatura e nos insights empíricos obtidos com os stakeholders, emergem três implicações práticas centrais para o projeto de sistemas de apoio à decisão sensíveis à privacidade: (i) a adoção de explicações em camadas e orientadas ao papel do stakeholder, reconhecendo que diferentes atores organizacionais possuem necessidades distintas de compreensão,

justificativa e monitoramento; (ii) a documentação explícita dos trade-offs entre explicabilidade e preservação da privacidade, incluindo decisões relativas à granularidade das explicações, ao orçamento de privacidade adotado e ao público-alvo das informações explicativas, de modo a apoiar transparência organizacional e responsabilização; e (iii) o tratamento das interfaces de explicação como potenciais superfícies de ataque, sujeitas a mecanismos de controle de acesso, modelagem de ameaças, monitoramento e auditoria, assim como já ocorre com dados, modelos e infraestruturas críticas.

Essas implicações reforçam que a explicabilidade não deve ser concebida como um atributo técnico isolado, mas como uma capacidade organizacional que emerge da articulação entre decisões técnicas, requisitos regulatórios e estruturas de governança. Nesse sentido, torna-se necessário um instrumento conceitual que auxilie organizações a tomar decisões explícitas e justificáveis sobre como, para quem e em que nível explicar sistemas de IA operando sob restrições de privacidade. Essa necessidade motiva diretamente a proposta do framework apresentada no Capítulo 4.

3.4 AMEAÇAS À VALIDADE

Como toda pesquisa de natureza qualitativa e exploratória, este estudo está sujeito a diferentes ameaças à validade[69]. Nesta seção, discutimos as principais ameaças identificadas e as estratégias adotadas para mitigá-las, visando garantir transparência e apoiar a interpretação adequada dos resultados.

3.4.1 Validade de Construção

A validade de construção refere-se ao grau em que os conceitos investigados refletem adequadamente os fenômenos que se pretende estudar. Neste trabalho, conceitos como explicabilidade, confiança e responsabilização são abstratos, multifacetados e frequentemente interpretados de forma distinta por diferentes stakeholders, o que pode gerar ambiguidades na coleta e análise dos dados.

Para mitigar essa ameaça, as perguntas do grupo focal foram cuidadosamente elaboradas com base na literatura consolidada sobre XAI, privacidade e sistemas de apoio à decisão, garantindo alinhamento conceitual com definições previamente discutidas no Capítulo 2. Além disso, a inclusão deliberada de participantes com diferentes papéis organizacionais permitiu capturar múltiplas interpretações desses conceitos, reduzindo o risco de uma visão restrita ou enviesada. Durante a condução do grupo focal, o moderador também buscou esclarecer termos ambíguos sempre que necessário, promovendo entendimento compartilhado entre os participantes.

3.4.2 Validade Interna

A validade interna diz respeito à relação entre os dados coletados e as interpretações derivadas, incluindo possíveis vieses introduzidos durante a condução do estudo ou na análise dos dados. No contexto de grupos focais, ameaças comuns incluem influência do moderador, efeitos de dominância de determinados participantes e viés de confirmação durante a codificação.

Esses riscos foram mitigados por meio do uso de um formato semi-estruturado de discussão, que equilibra direcionamento e liberdade de expressão. O moderador adotou uma postura neutra, evitando juízos de valor e incentivando explicitamente a participação de todos os envolvidos, inclusive dos participantes menos vocais. Durante a análise dos dados, o foco foi identificar padrões recorrentes e divergências entre diferentes perfis de stakeholders, em vez de enfatizar opiniões individuais isoladas. O processo de codificação foi iterativo e discutido entre os autores, reduzindo a influência de interpretações individuais.

3.4.3 Validade Externa

A validade externa refere-se ao grau em que os resultados podem ser generalizados para outros contextos, populações ou domínios. Como se trata de um estudo qualitativo exploratório com apenas um grupo focal e oito participantes, não é possível realizar generalizações estatísticas dos achados.

No entanto, o objetivo deste estudo não é a generalização estatística, mas a generalização analítica. Os resultados são discutidos à luz da literatura existente, permitindo identificar padrões, tensões e percepções que podem ser relevantes para contextos semelhantes, especialmente em sistemas de apoio à decisão sensíveis à privacidade. A diversidade de papéis representados no grupo focal contribui para ampliar o alcance interpretativo dos resultados, ainda que estes permaneçam dependentes do contexto organizacional e regulatório considerado.

3.4.4 Confiabilidade

A confiabilidade está relacionada à consistência do processo de pesquisa e à possibilidade de que outros pesquisadores compreendam, acompanhem e, quando pertinente, reproduzam o percurso metodológico adotado. Em estudos qualitativos, a replicação exata é limitada pela natureza contextual dos dados, pelo perfil dos participantes e pelas interações sociais estabelecidas durante a coleta. Ainda assim, buscou-se aumentar a confiabilidade por meio da documentação detalhada do desenho de pesquisa, incluindo critérios de seleção dos participantes, roteiro de perguntas, procedimento de condução do grupo focal e abordagem de análise dos dados.

O uso de identificadores anonimizados (P1–P8), a apresentação de trechos ilustrativos das falas e a descrição explícita do processo de codificação contribuem para a transparência do estudo e permitem que pesquisadores interessados repliquem o protocolo em contextos comparáveis, preservando os mesmos objetivos analíticos. Além disso, a análise foi conduzida de forma iterativa,

com definição de códigos iniciais, consolidação em temas e refinamento dos agrupamentos até alcançar coerência interna entre códigos, temas e questões de pesquisa.

Para mitigar riscos de inconsistência interpretativa, os resultados foram organizados em torno de temas diretamente associados às RQs (compatibilidade e trade-offs; confiança e responsabilização), e as sínteses (quadros de resumo ao final de cada RQ) foram elaboradas para manter rastreabilidade entre evidências (falas) e interpretações (temas). Por fim, as decisões metodológicas foram registradas ao longo do processo (por exemplo: ajustes no roteiro, regras de moderação e critérios de agregação de códigos), o que fortalece a auditabilidade do estudo e reduz a dependência de memória ou interpretação posterior dos pesquisadores.

Apesar dessas medidas, reconhece-se que a confiabilidade pode ser afetada por: (i) variações na condução do grupo focal caso outro moderador seja utilizado; (ii) diferenças no repertório técnico dos participantes; e (iii) mudanças no contexto organizacional e regulatório. Assim, a contribuição central do estudo é oferecer um procedimento transparente e replicável em termos de protocolo e lógica analítica, e não a reprodução literal dos mesmos achados empíricos.

3.5 SÍNTESE DO CAPÍTULO

Este capítulo discutiu os resultados do estudo à luz da literatura, evidenciando que a relação entre explicabilidade e preservação da privacidade é marcada por trade-offs estruturais e dependentes do contexto. Os achados demonstram que a compatibilidade entre XAI e técnicas de preservação da privacidade não pode ser assumida como intrínseca, exigindo decisões explícitas de projeto e governança.

A análise revelou que explicações globais e agregadas tendem a ser mais adequadas em ambientes sensíveis à privacidade, enquanto explicações locais introduzem riscos adicionais de vazamento de informação. Além disso, os resultados evidenciaram que explicabilidade exerce papéis distintos na promoção da confiança e no suporte à responsabilização, reforçando a necessidade de diferenciar explicações orientadas à compreensão daquelas orientadas à justificativa e à auditoria.

Outro aspecto central identificado foi a natureza dependente do stakeholder, indicando que explicabilidade deve ser concebida como uma capacidade organizacional adaptativa, e não como uma funcionalidade uniforme. A ausência de diretrizes práticas e sistemáticas para apoiar essas decisões reforça a lacuna existente entre princípios normativos e práticas de engenharia.

Esses achados fundamentam diretamente o Capítulo 4, no qual é proposta uma estrutura conceitual de explicabilidade orientada à governança, destinada a apoiar a tomada de decisão sobre mecanismos de explicação em sistemas de apoio à decisão sensíveis à privacidade, considerando explicitamente trade-offs técnicos, organizacionais e regulatórios.

4 FRAMEWORK PROPOSTO

Este capítulo apresenta a proposta do framework desta dissertação, concebido para apoiar a concepção, a implementação e a governança de explicabilidade em sistemas de apoio à decisão baseados em IA que operam sob restrições de privacidade. O framework foi elaborado a partir da integração de duas fontes de evidência: (i) a síntese conceitual da literatura apresentada no Capítulo 2, incluindo mecanismos de XAI, técnicas de preservação da privacidade e trade-offs reportados; e (ii) os achados empíricos do grupo focal descritos no Capítulo 3.1, que evidenciaram necessidades heterogêneas entre stakeholders e a natureza negociada da explicabilidade em ambientes regulados.

A proposta parte do pressuposto de que explicabilidade e preservação da privacidade não são atributos que podem ser maximizados simultaneamente de forma universal. Em vez disso, são capacidades que precisam ser especificadas em função do contexto, do risco, do público-alvo da explicação e das obrigações regulatórias. Assim, o framework estrutura decisões sobre explicabilidade como decisões de projeto e governança, orientadas por trade-offs explícitos e por mecanismos de controle (por exemplo, controle de acesso e auditoria) aplicados também às interfaces de explicação.

4.1 OBJETIVO E ESCOPO

O objetivo do framework é fornecer uma estrutura prática e justificável para orientar organizações a responderem às seguintes questões durante o ciclo de vida de sistemas de IA sensíveis à privacidade:

1. **O que explicar?** (quais objetos explicativos: modelo, decisão, dados, processo, incerteza, limitações)
2. **Para quem explicar?** (stakeholders e responsabilidades associadas)
3. **Em que nível explicar?** (granularidade e profundidade, explicações globais, locais, híbridas e em camadas)
4. **Com quais garantias de privacidade?** (técnicas adotadas, orçamento de privacidade, controles de acesso)
5. **Como justificar e auditar a explicação?** (rastreabilidade, evidências, consistência temporal, logs e trilhas de auditoria)

O escopo do framework se concentra em sistemas de apoio à decisão em domínios sensíveis (por exemplo, saúde, finanças, justiça e serviços públicos), nos quais há: (i) restrições explícitas

de privacidade e proteção de dados; (ii) múltiplos stakeholders com necessidades distintas; e (iii) exigência de justificativa e responsabilização.

4.2 CONSTRUÇÃO DO FRAMEWORK

A construção do framework foi guiada pelos seguintes fundamentos, derivados diretamente das lacunas e evidências discutidas nos Capítulos 2 e 3.1:

1. **Trade-offs estruturais e explícitos.** A literatura e o grupo focal indicam que aumentar garantias de privacidade tende a reduzir a utilidade e a fidelidade das explicações, enquanto explicações detalhadas podem elevar risco de vazamento e superfície de ataque [55, 57, 62, 22].
2. **Explicabilidade dependente do stakeholder.** Diferentes papéis (desenvolvedores, gestores, conformidade, privacidade, tomadores de decisão) necessitam de explicações com objetivos distintos (depuração, monitoramento, justificativa, auditoria), o que favorece explicações em camadas e orientadas ao papel [59, 60].
3. **Explicabilidade como capacidade de governança.** Em ambientes regulados, explicações precisam ser defensáveis e consistentes ao longo do tempo, exigindo rastreabilidade, documentação de decisões e trilhas de auditoria [31, 56].
4. **Interfaces de explicação como superfície de ataque.** Explicações podem revelar informações sobre dados, atributos sensíveis ou o próprio modelo, exigindo modelagem de ameaças, controle de acesso e minimização de informações nas saídas explicativas [54, 55].

4.3 VISÃO GERAL DO FRAMEWORK

O framework é organizado em quatro camadas complementares que se articulam durante o ciclo de vida do sistema:

1. **Camada 1: Contexto e Riscos** (definição do domínio, criticidade, tipo de decisão e obrigações regulatórias)
2. **Camada 2: Mapeamento de Stakeholders e Necessidades de Explicação** (papéis, responsabilidades e objetivos da explicação)
3. **Camada 3: Projeto de Explicações com Salvaguardas de Privacidade** (seleção de mecanismos XAI e PETs, granularidade e controles)
4. **Camada 4: Governança, Evidências e Auditoria** (documentação de trade-offs, rastreabilidade, consistência e monitoramento)

A Figura 4.1 ilustra a estrutura proposta. O framework pode ser utilizado tanto de forma prescritiva (como roteiro de decisões) quanto avaliativa (como checklist para verificar se um sistema possui explicabilidade apropriada sob restrições de privacidade).

Figura 4.1: Estrutura do framework proposto para explicabilidade em sistemas sensíveis à privacidade.

4.3.1 Camada 1: Contexto e Riscos

A primeira camada define o contexto no qual o sistema opera e estabelece o nível de risco associado à decisão automatizada. Esta etapa é necessária porque, conforme evidenciado na literatura e no grupo focal, a explicabilidade apropriada depende do domínio, do impacto potencial da decisão e das obrigações legais e organizacionais.

Nesta etapa, recomenda-se registrar, no mínimo: 1) **Domínio e finalidade do sistema:** por exemplo, triagem, recomendação, priorização, concessão de benefícios; 2) **Natureza da decisão:** decisória (substitui humanos), recomendatória (apoia humanos) ou híbrida; 3) **Impacto potencial:** efeitos sobre direitos, acesso a serviços, riscos financeiros, riscos reputacionais; e 4) **Partes afetadas:** titulares de dados e populações potencialmente impactadas.

Em seguida, define-se o conjunto de restrições de Privacidade, Segurança e Exposição Reguladora aplicáveis:

- **Tipos de dados processados:** pessoais, sensíveis, dados de terceiros, dados derivados.
- **Princípios de adequação, finalidade, e necessidade:** quais atributos são compatíveis com a finalidade informada ao titular,legítimos, específicos e estritamente necessários para a finalidade pretendida.
- **Regras de acesso e compartilhamento:** quem pode acessar dados, modelos e explicações.
- **Exposição regulatória:** necessidade de justificativa, contestação e auditoria ao longo do tempo.

Como evidenciado por estudos que tratam explicações como superfícies de ataque, o framework inclui explicitamente a avaliação de ameaças associadas às saídas explicativas [55, 54]. Nesta etapa, recomenda-se identificar:

- **Ator adversário:** interno, externo, usuário final, integrador, terceiro.
- **Objetivo do ataque:** inferência de atributos, associação, extração do modelo, reconstrução.
- **Vetor:** explicação local, contrafactual, importância de atributos, exemplos, gradientes.
- **Consequência:** reidentificação, violação de confidencialidade, exploração de vieses.

O resultado da Camada 1 é um artefato de contexto e risco que define limites e critérios mínimos para as próximas camadas. Assim, o **Artefato A1** (Tabela 4.1) operacionaliza a camada de *contexto, risco e ameaças* do framework. Seu objetivo é tornar explícitos os fatores estruturais que condicionam as decisões de explicabilidade, incluindo domínio de aplicação, tipo de decisão apoiada, impacto potencial, natureza dos dados tratados e exposição regulatória. Ao registrar explicitamente ameaças associadas às explicações e limites iniciais de transparência, o artefato permite que a explicabilidade seja concebida desde o início como uma decisão informada por risco e governança, e não como uma funcionalidade adicionada a posteriori. Esse registro atende diretamente às evidências discutidas no Capítulo 2 e aos achados empíricos do grupo focal (3.2), que indicam que a compatibilidade entre XAI e preservação da privacidade depende fortemente do contexto e do nível de risco associado ao sistema.

Tabela 4.1: Artefato A1 – Caracterização de contexto, riscos e ameaças

Dimensão	**Descrição**
Domínio do sistema	Serviço público digital para priorização de solicitações de benefício social
Tipo de decisão	Apoio à decisão humana (recomendação com validação manual)
Impacto potencial	Alto – pode afetar acesso a direitos e benefícios
Tipos de dados	Dados pessoais e dados socioeconômicos
Exposição regulatória	LGPD; exigência de justificativa e possibilidade de contestação
Stakeholders afetados	Usuários finais, gestores públicos, auditores e órgãos de controle
Principais ameaças	Inferência de atributos sensíveis a partir de explicações locais; extração de padrões do modelo
Limites iniciais de explicação	Proibição de explicações por instância para usuários finais; uso restrito a explicações agregadas

4.3.2 Camada 2: Stakeholders e Necessidades de Explicação

A segunda camada operacionaliza a ideia de explicabilidade dependente do stakeholder, evidenciada tanto pela literatura quanto pelo grupo focal [59, 60]. A premissa é que não existe uma explicação única que atenda simultaneamente objetivos de depuração técnica, uso operacional, confiança e auditoria. O *framework* recomenda mapear *stakeholders* em pelo menos cinco grupos:

1. **Equipe técnica:** desenvolvedores, engenheiros de ML, MLOps (depuração, validação, melhoria contínua).

2. **Gestão e produto:** PO, gestores, responsáveis por serviço (monitoramento, decisão sobre adoção e limites).

3. **Privacidade e segurança:** encarregados, profissionais de privacidade, segurança (minimização, riscos e controles).

4. **Conformidade e auditoria:** *compliance*, auditoria interna/externa (defensabilidade e rastreabilidade).

5. **Usuários e afetados:** usuários finais e titulares impactados (compreensão e contestação).

Para cada *stakeholder*, o *framework* define objetivos típicos: **Objetivos orientados à confiança:** compreensão suficiente para uso apropriado e calibração; **Objetivos orientados à responsabilização:** justificativa, evidência, consistência temporal e auditabilidade; e **Objetivos técnicos:** diagnóstico, análise de erro, identificação de viés, robustez e estabilidade.

Essa distinção reflete a diferença apontada no grupo focal entre explicações voltadas ao uso e explicações voltadas à prestação de contas, além de evidências sobre riscos de sinais mal calibrados (por exemplo, escores de confiança) para a tomada de decisão [63].

Com base no mapeamento, recomenda-se especificar requisitos de Explicação (Especificação) para cada *stakeholder*: **Granularidade:** global, local, em camadas; **Forma:** textual, visual, regras, exemplos, contrafactuais, métricas, relatórios; **Acessibilidade e linguagem:** nível técnico e clareza; e **Frequência e momento:** pré-decisão, pós-decisão, auditoria, monitoramento contínuo.

O resultado da Camada 2 é um artefato de requisitos de explicação por papel, que alimenta a seleção de mecanismos na Camada 3. O **Artefato A2** (Tabela 4.2) materializa a camada de *stakeholders e necessidades de explicação*, explicitando que diferentes papéis organizacionais possuem objetivos distintos ao interagir com explicações geradas por sistemas de IA. Ao mapear *stakeholders*, objetivos principais e tipos de explicação requeridos, o artefato apoia o desenho de estratégias de explicabilidade em camadas e orientadas ao papel, conforme defendido na literatura e confirmado pelos resultados da QP2. Esse artefato reduz o risco de abordagens uniformes de explicação, frequentemente ineficazes, e fornece uma base estruturada para alinhar confiança operacional, suporte à decisão e responsabilização institucional.

4.3.3 Camada 3: Projeto de Explicações com Salvaguardas de Privacidade

A terceira camada traduz requisitos de explicação em decisões técnicas, articulando mecanismos de XAI e técnicas de preservação da privacidade. Esta etapa incorpora evidências de que: (i) explicações locais podem aumentar risco de vazamento e facilitar ataques [22]; (ii) explicações em ambientes com privacidade diferencial e aprendizado federado podem sofrer perda de fideli-

Tabela 4.2: Artefato A2 – Mapeamento de stakeholders e objetivos da explicação

Stakeholder	Objetivo principal	Necessidade de explicação
Desenvolvedores	Depuração e melhoria do modelo	Explicações técnicas detalhadas, acesso controlado a explicações locais
Gestores	Monitoramento e decisão estratégica	Explicações globais, métricas de desempenho e vieses
Privacidade e segurança	Mitigação de riscos	Evidências de minimização, justificativa de granularidade
Conformidade/Auditoria	Responsabilização	Explicações consistentes, rastreáveis e versionadas
Usuários finais	Compreensão e contestação	Explicações simples, agregadas e não técnicas

dade e utilidade [56, 62]; e (iii) combinar defesas pode mitigar riscos, mas aumenta complexidade operacional [55].

O framework organiza mecanismos de XAI por classe, com recomendações condicionais: 1) **Explicações globais:** úteis para governança e monitoramento; menor risco de revelar casos individuais; 2) **Explicações locais:** úteis para depuração e contestação, mas com maior risco de vazamento; exigir controles; 3) **Explicações contrafactuais:** informativas, porém potencialmente exploráveis; requerem análise de ameaça; 4) **Explicações baseadas em exemplos:** risco de revelar dados; preferir protótipos agregados ou dados sintéticos; e 5) **Sinais de incerteza:** úteis para calibração, mas podem degradar decisão se mal interpretados [63].

Em relação à seleção de técnicas de preservação da privacidade (PETs), O framework considera, entre outras:

- **Privacidade diferencial:** exige explicações robustas a ruído; documentar orçamento e impacto na fidelidade.
- **Aprendizado federado:** reduzir exposição de dados, mas não elimina risco via gradientes; considerar proteções adicionais.
- **Criptografia/computação segura:** maior custo; pode limitar explicações interativas.
- **Anonimização/dados sintéticos:** risco de utilidade e de reidentificação; avaliar adequação ao objetivo da explicação.

Nesta etapa, o framework recomenda estabelecer: 1) **Matriz de acesso:** quem pode acessar quais tipos de explicação; 2) **Política de granularidade:** quando permitir explicação local e sob quais condições; e 3) **Restrições de interface:** evitar exposição de valores sensíveis, limitar consultas repetidas, rate limiting.

Como resultado do grupo focal, o framework exige que decisões sejam registradas em termos de trade-offs, incluindo: a) **Trade-off esperado:** utilidade vs privacidade; fidelidade vs risco; transparência vs complexidade; b) **Justificativa:** contexto, risco e stakeholders; e c) **Critérios de aceitação:** limites mínimos de utilidade e de proteção.

O **Artefato A3** (Tabela 4.3) representa a camada de *projeto da explicabilidade sob restrições de privacidade*, registrando de forma explícita as decisões técnicas adotadas e os trade-offs assumidos. Esse artefato documenta o mecanismo de XAI selecionado, o tipo de explicação oferecida, a técnica de preservação da privacidade empregada, o orçamento de privacidade e as restrições de acesso associadas. Ao tornar essas decisões rastreáveis e justificáveis, o artefato responde diretamente às preocupações identificadas no grupo focal sobre escolhas implícitas e decisões ad hoc, além de apoiar a transparência organizacional e a defensabilidade das opções de projeto em contextos regulados.

Tabela 4.3: Artefato A3 – Decisões de explicabilidade sob restrições de privacidade

Decisão	**Registro**
Mecanismo de XAI	Importância global de atributos e relatórios explicativos
Tipo de explicação	Global e agregada (sem instância individual)
Técnica de privacidade	Privacidade diferencial aplicada aos relatórios agregados
Orçamento de privacidade	$\varepsilon = 1.0$ para explicações públicas
Trade-off assumido	Redução de fidelidade explicativa em troca de menor risco de inferência
Justificativa	Contexto regulado com dados sensíveis e alto impacto social
Restrições de acesso	Explicações locais disponíveis apenas para equipe técnica autenticada

4.3.4 Camada 4: Governança, Evidências e Auditoria

A quarta camada consolida a explicabilidade como capacidade organizacional, assegurando rastreabilidade, consistência e defensabilidade ao longo do tempo, como requerido por contextos regulados e pelos achados empíricos [31, 56].

O framework propõe um conjunto mínimo de artefatos de governança: 1) **Especificação de explicações por stakeholder** (da Camada 2); 2) **Registro de decisões e trade-offs** (da Camada 3); 3) **Matriz de acesso e política de explicações** (papéis, permissões, limitações); e 4) **Plano de auditoria de explicações** (o que registrar, por quanto tempo, como revisar).

Esta etapa atende diretamente à exigência identificada pelos participantes: explicações orientadas à responsabilização precisam ser consistentes e defensáveis ao longo do tempo (Rastreabilidade e Consistência Temporal). O framework recomenda que seja realizado o versionamento do modelo e do pipeline de dados, o versionamento de mecanismos de explicação, o registro de parâmetros relevantes (incluindo orçamento de privacidade quando aplicável) e o registro das trilhas de auditoria para consultas e respostas explicativas.

Como explicações e riscos evoluem, o framework inclui reavaliação periódica baseada em gatilhos (Monitoramento e Reavaliação):

- mudanças em dados, domínio ou população;
- alteração de técnica de privacidade (por exemplo, novo orçamento de DP);
- incidentes de segurança ou indícios de abuso da interface explicativa;

- mudanças regulatórias ou novas exigências de prestação de contas.

O **Artefato A4** (Tabela 4.4) operacionaliza a camada de *governança, evidências e auditoria da explicabilidade*. Seu foco é tratar explicações como artefatos organizacionais sujeitos a versionamento, controle de acesso, registro de uso e auditoria, assim como ocorre com dados e modelos. Ao definir elementos como logs de acesso às explicações, consistência temporal e gatilhos de reavaliação, o artefato apoia requisitos de responsabilização, conformidade regulatória e aprendizado organizacional. Esse artefato endereça diretamente os achados da QP2, que indicam que explicações orientadas à responsabilização exigem consistência, rastreabilidade e capacidade de defesa ao longo do tempo.

Tabela 4.4: Artefato A4 – Evidências de governança e auditoria da explicabilidade

Elemento	**Descrição**
Versionamento	Modelo, dados, mecanismo de explicação e parâmetros versionados
Registro de decisões	Trade-offs documentados e aprovados por comitê interno
Logs de explicação	Registro de quem acessou explicações, quando e com qual finalidade
Consistência temporal	Capacidade de reproduzir explicação para decisões passadas
Auditoria	Revisão periódica das explicações e controles de acesso
Gatilhos de reavaliação	Mudança regulatória, novo conjunto de dados, incidente de segurança

Os artefatos A1, A2, A3 e A4 apresentados demonstram como o framework proposto pode ser operacionalizado por meio de registros estruturados que conectam decisões técnicas, considerações organizacionais e requisitos regulatórios. Ao explicitar contexto, stakeholders, trade-offs e mecanismos de governança, o framework fornece suporte prático para decisões conscientes sobre explicabilidade em sistemas de apoio à decisão sensíveis à privacidade, superando abordagens puramente normativas ou exclusivamente algorítmicas.

O framework proposto é sintetizado na Figura 4.2, que apresenta de forma integrada suas quatro camadas conceituais, os artefatos associados a cada uma delas e o fluxo de aplicação que conduz à geração de evidências para confiança e responsabilização. A figura explicita que o framework não é apenas um modelo conceitual, mas um instrumento operacional orientado à tomada de decisão: as Camadas 1 e 2 estruturam o entendimento do contexto, dos riscos e das necessidades dos stakeholders; a Camada 3 concentra as decisões técnicas sobre mecanismos de explicação e técnicas de preservação da privacidade, explicitando os trade-offs envolvidos; e a Camada 4 consolida a dimensão de governança, garantindo rastreabilidade, consistência temporal e auditabilidade. A associação direta entre cada camada e seus respectivos artefatos (A1–A4) evidencia como decisões abstratas são traduzidas em registros concretos e verificáveis, enquanto o fluxo vertical indica que escolhas realizadas em camadas superiores condicionam e justificam as decisões subsequentes. Dessa forma, a Figura 4.2 reforça o caráter sociotécnico do framework, ao mostrar como aspectos técnicos, organizacionais e regulatórios são articulados para apoiar explicabilidade responsável em sistemas de apoio à decisão sensíveis à privacidade.

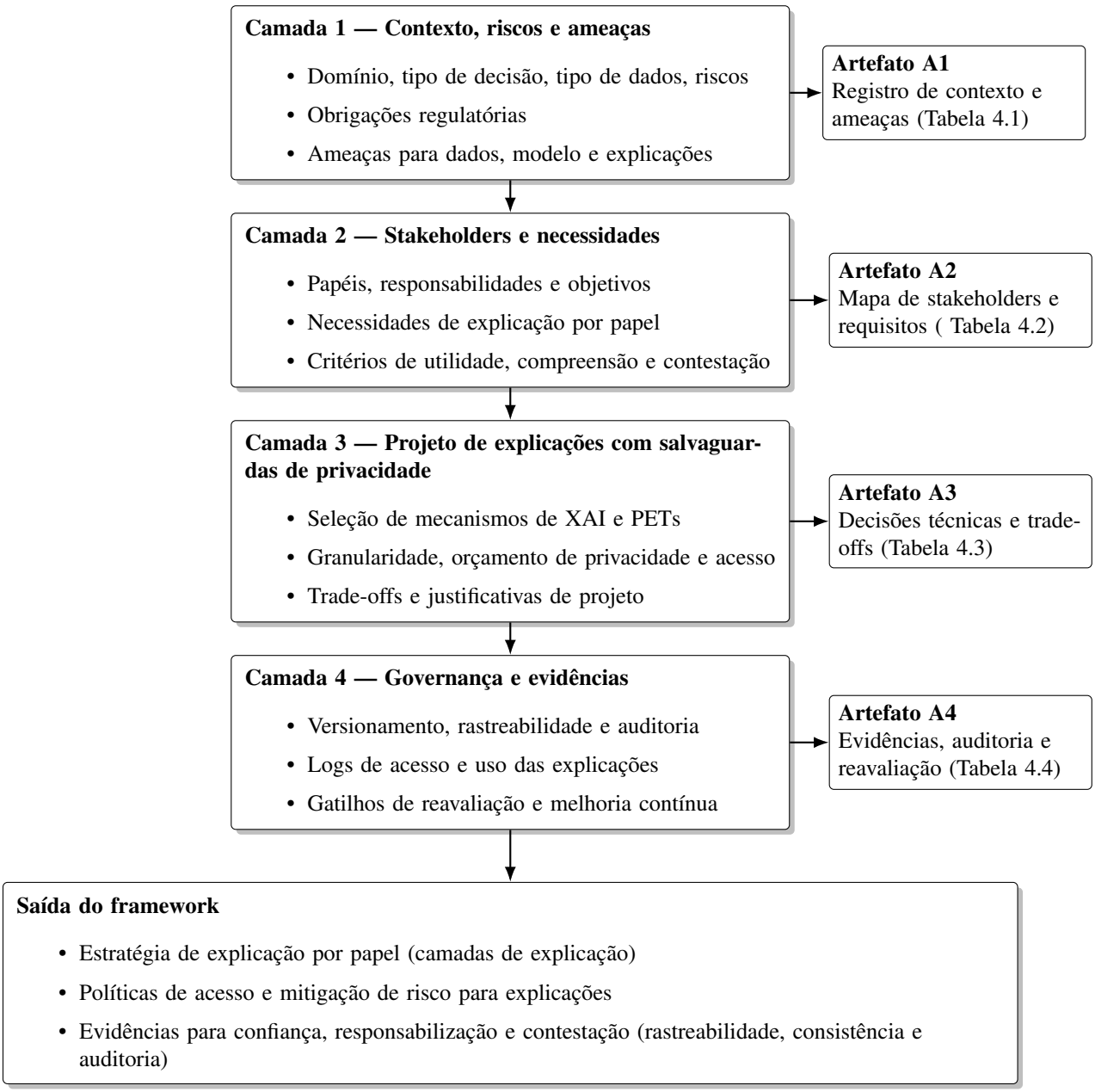

Figura 4.2: Figura-resumo da primeira versão do framework proposto, organizado em quatro camadas e operacionalizado por quatro artefatos (A1–A4).

4.4 PROCEDIMENTO DE APLICAÇÃO DO FRAMEWORK

Para operacionalizar o framework, propõe-se um procedimento em sete passos que conecta decisões de contexto e risco, necessidades de stakeholders, escolhas técnicas e governança. O procedimento foi estruturado para produzir evidências rastreáveis por meio dos artefatos A1–A4 (Tabelas 4.1–4.4) e para reduzir decisões *ad hoc* ao explicitar, em cada etapa, entradas, decisões e saídas verificáveis. A Figura 4.2 sintetiza a relação entre camadas, artefatos e resultado esperado.

4.4.1 Passos do procedimento

1. **Caracterizar contexto e decisão (Camada 1).** Este passo estabelece o enquadramento do sistema: domínio de aplicação, tipo de decisão apoiada, impacto potencial, público afetado, tipo de dados e condições de uso (por exemplo, ambientes regulados ou de alto risco). A saída esperada é um registro mínimo e consistente do contexto, que servirá como base para justificar escolhas de explicação e de privacidade ao longo do processo, consolidado no Artefato A1 (Tabela 4.1).

2. **Identificar restrições e ameaças para dados, modelo e explicações (Camada 1).** A partir do contexto, definem-se restrições legais e organizacionais (por exemplo, princípios de minimização e necessidade) e um conjunto de ameaças plausíveis, incluindo risco de vazamento por explicações, ataques por inferência e uso indevido de saídas explicativas. Este passo explicita que explicações podem ampliar a superfície de ataque, portanto devem ser tratadas como componente sensível do sistema. A saída é um perfil de risco e restrições associado ao Artefato A1 (Tabela 4.1), incluindo pressupostos e limites iniciais de transparência.

3. **Mapear stakeholders e responsabilidades (Camada 2).** Este passo identifica os papéis que interagem com o sistema e com suas explicações (por exemplo, desenvolvimento, operação, privacidade, conformidade, auditoria e tomada de decisão). O objetivo é explicitar responsabilidades e objetivos por papel, distinguindo uso operacional, depuração, governança e prestação de contas. A saída é o mapeamento de stakeholders no Artefato A2 (Tabela 4.2).

4. **Especificar requisitos de explicação por papel (Camada 2).** Para cada stakeholder, definem-se requisitos de explicação (conteúdo, granularidade, formato e frequência), além de critérios mínimos de utilidade e compreensão. Este passo evita a premissa de uma explicação única para todos e antecipa conflitos: requisitos de depuração tendem a demandar maior detalhe, enquanto requisitos de privacidade e conformidade impõem limites de exposição. A saída é uma especificação por papel no Artefato A2 (Tabela 4.2), que alimenta diretamente as decisões técnicas.

5. **Selecionar mecanismos de XAI e PETs compatíveis com os requisitos (Camada 3).** Com base nos requisitos por papel e nas restrições do contexto, selecionam-se mecanismos de explicação e técnicas de preservação da privacidade compatíveis com o nível de risco e com o objetivo de uso. Nesta etapa, a seleção não é apresentada como otimização universal, mas como escolha condicionada: determinados mecanismos podem ser adequados apenas para alguns papéis ou sob controles específicos. A saída é o conjunto de decisões técnicas registradas no Artefato A3 (Tabela 4.3), incluindo justificativas e alternativas descartadas.

6. **Definir controles e políticas de acesso para explicações (Camada 3).** Este passo traduz o princípio de que explicações devem ser governadas como recursos sensíveis. Definem-se controles de acesso por papel, limites de granularidade (por exemplo, permitir explicações

globais amplas e restringir explicações locais), parâmetros operacionais associados à privacidade (por exemplo, orçamento de privacidade quando aplicável) e mecanismos de mitigação (por exemplo, filtragem, agregação e limitação de consultas). A saída é a política de acesso e proteção associada ao Artefato A3 (Tabela 4.3) e preparada para auditoria na Camada 4.

7. **Implantar governança e evidências (Camada 4).** Este passo estabelece como as decisões anteriores serão mantidas e defendidas ao longo do tempo. Define-se como serão registradas evidências de explicação e uso (por exemplo, logs de acesso, versionamento e rastreabilidade), critérios de consistência temporal e gatilhos de reavaliação (por exemplo, mudança de modelo, mudança de dados, incidentes, novas exigências regulatórias). A saída é o pacote de governança e evidências no Artefato A4 (Tabela 4.4), permitindo auditoria e atualização controlada.

O procedimento operacionaliza diretamente as questões de pesquisa ao tornar explícito como decisões técnicas e organizacionais são encadeadas e justificadas.

- **QP1 (compatibilidade e trade-offs):** é endereçada pela Camada 1 ao explicitar contexto, restrições e ameaças, e pela Camada 3 ao registrar escolhas condicionadas de mecanismos e técnicas de preservação da privacidade, incluindo decisões sobre granularidade e controles de acesso (Artefatos A1 e A3; Tabelas 4.1 e 4.3).
- **QP2 (confiança e responsabilização):** é endereçada pela Camada 2 ao diferenciar necessidades por papel e definir requisitos de explicação orientados a objetivos distintos (uso, justificativa, auditoria), e pela Camada 4 ao garantir rastreabilidade, consistência e defensabilidade das explicações ao longo do tempo (Artefatos A2 e A4; Tabelas 4.2 e 4.4).

4.5 SÍNTESE DO CAPÍTULO

Este capítulo apresentou a proposta do framework para explicabilidade em sistemas de apoio à decisão sensíveis à privacidade. A proposta foi construída a partir da integração entre evidências conceituais da literatura e achados empíricos do grupo focal, resultando em uma estrutura em quatro camadas: (i) contexto e risco; (ii) stakeholders e necessidades; (iii) projeto de explicações com restrições de privacidade; e (iv) governança, evidências e auditoria. O *framework* trata explicabilidade como capacidade organizacional e orienta decisões explícitas sobre o que explicar, para quem explicar, com que granularidade e sob quais garantias de privacidade, reconhecendo trade-offs estruturais e a necessidade de controles sobre interfaces de explicação. No próximo capítulo, o *framework* é detalhado quanto ao seu uso em cenários de aplicação e critérios de avaliação.

5 VALIDAÇÃO DO FRAMEWORK PROPOSTO

Este capítulo apresenta a estratégia de validação do *framework* proposto para apoiar decisões de explicabilidade em sistemas de apoio à decisão sensíveis à privacidade. A validação tem como objetivo avaliar a utilidade, a coerência conceitual e a adequação prática do *framework* sob a perspectiva de especialistas, considerando aspectos técnicos, organizacionais e de governança.

A abordagem adotada está alinhada ao paradigma de *Design Science Research* (DSR), no qual artefatos são avaliados com base em sua capacidade de resolver problemas relevantes e produzir conhecimento prescritivo aplicável [70, 2].Segundo Hevner e Chatterjee [2], a DSR enfatiza a construção e avaliação de artefatos com o objetivo de resolver problemas organizacionais relevantes. No contexto desta pesquisa, o framework de explicabilidade orientada à governança constitui o principal artefato, concebido para apoiar decisões explícitas, justificáveis e auditáveis sobre explicações em sistemas de IA sob restrições de privacidade.

A validação do framework corresponde à fase de *evaluation* da DSR, buscando verificar se o artefato atende aos critérios de utilidade, rigor e relevância. Considerando que se trata de um artefato sociotécnico e prescritivo, a avaliação não envolve testes experimentais ou desempenho algorítmico, mas sim a análise da adequação conceitual, clareza estrutural, aplicabilidade prática e capacidade de apoiar responsabilização e governança, conforme recomendado para esse tipo de artefato [2].

5.1 ESTRATÉGIA DE VALIDAÇÃO

A validação foi conduzida por meio de um *survey* estruturado com especialistas da área, abordagem amplamente utilizada na avaliação de *frameworks* conceituais e modelos prescritivos em Engenharia de Software e Sistemas de Informação [70, 71].

O instrumento de coleta foi implementado na plataforma Google Forms e disponibilizado eletronicamente aos participantes. A escolha por um questionário estruturado permitiu coletar tanto avaliações quantitativas (por meio de escalas Likert de cinco pontos) quanto contribuições qualitativas abertas, possibilitando uma análise combinada de percepções objetivas e comentários dos participantes. O tempo estimado para preenchimento completo do questionário foi de 15 a 20 minutos.

A seleção dos participantes seguiu uma estratégia de amostragem intencional [71], buscando especialistas com experiência prévia em desenvolvimento, governança, segurança, privacidade, ciência de dados ou arquitetura de sistemas baseados em IA. Os participantes foram convidados por meio da rede profissional da pesquisadora e de colegas próximos que atuam em órgãos públicos e empresas relacionadas ao tema, bem como a partir de grupos de Whatsapp de pesquisadores

das áreas de Engenharia de Software e de Segurança da Informação. O objetivo foi garantir que os respondentes possuíssem conhecimento prático e contextual suficiente para avaliar o *framework* de forma qualificada.

Antes de prosseguir para a etapa de preenchimento das respostas, o formulário disponibilizou Termo de Consentimento Livre e Esclarecido, no qual foram apresentados os objetivos, a metodologia e as informações relacionadas ao tratamento de dados pessoais no âmbito da pesquisa.

Esclareceu-se, ainda, que a participação seria anônima e voluntária. Ademais, foi disponibilizado endereço de e-mail, para que eventuais dúvidas pudessem ser encaminhadas à pesquisadora e sua professora orientadora.

Ao todo, o *survey* obteve 32 (trinta e duas) respostas completas, provenientes de profissionais com experiência relevante em contextos organizacionais que envolvem sistemas baseados em IA e/ou requisitos de proteção de dados. A avaliação conduzida por meio do questionário estruturado foi organizado em quatro etapas:

1. Apresentação do objetivo da pesquisa e contextualização do problema abordado pelo framework;
2. Disponibilização de material explicativo contendo a visão geral do framework (Figura 4.2), suas camadas, artefatos e relações;
3. Avaliação individual do framework com base em critérios previamente definidos (Seção 5.2), utilizando escala Likert de cinco pontos;
4. Coleta de comentários qualitativos, sugestões de melhoria e percepções críticas por meio de questões abertas.

Os especialistas foram instruídos a avaliar o *framework* considerando sua experiência profissional e sua aplicabilidade em contextos reais de sistemas de apoio à decisão sensíveis à privacidade. O questionário incluiu ainda um bloco inicial de caracterização dos participantes, composto por 9 (nove) questões, permitiu conhecer o perfil dos avaliadores em termos de formação, senioridade, experiência e envolvimento com projetos de IA.

5.2 CRITÉRIOS DE VALIDAÇÃO

A avaliação do *framework* foi estruturada a partir de critérios fundamentados na literatura de *Design Science Research* (DSR) e em estudos sobre avaliação de artefatos conceituais e prescritivos [1, 2, 3]. Esses referenciais orientam a análise de artefatos quanto à sua utilidade, clareza, completude, consistência interna e aplicabilidade prática em contextos organizacionais reais.

Cada critério foi operacionalizado por meio de uma pergunta explícita incluída no *survey* (Tabela 5.2) e avaliado em uma escala Likert de cinco pontos, permitindo mensurar o grau de con-

cordância dos especialistas de forma estruturada e comparável. Essa operacionalização buscou reduzir ambiguidades interpretativas e aumentar a consistência das avaliações.

Tabela 5.1: Critérios conceituais de validação do framework [1, 2, 3]

Critério	Descrição conceitual
Utilidade percebida	Capacidade do framework de apoiar decisões reais relacionadas ao equilíbrio entre explicabilidade e preservação da privacidade em sistemas baseados em IA.
Clareza conceitual	Grau de compreensão da estrutura do framework, incluindo camadas, artefatos, relações e papéis envolvidos.
Completude	Cobertura adequada dos principais aspectos técnicos, organizacionais, regulatórios e de governança necessários à tomada de decisão.
Aplicabilidade prática	Viabilidade de adoção do framework em contextos organizacionais reais, considerando esforço, tempo e integração com processos existentes.
Aderência à governança	Capacidade do framework de apoiar rastreabilidade, auditoria, responsabilização e alinhamento regulatório (por exemplo, sob a LGPD).
Flexibilidade	Capacidade de adaptação a diferentes domínios, níveis de risco, perfis de stakeholders e contextos organizacionais sem perda de coerência conceitual.

Além das questões fechadas, foram incluídas três perguntas abertas com o objetivo de capturar percepções qualitativas sobre pontos fortes, limitações, oportunidades de aprimoramento e contextos de aplicação do *framework*. A combinação de medidas quantitativas e qualitativas permite uma avaliação mais abrangente do artefato, alinhada às recomendações da *Design Science Research* (DSR) para análise de soluções sociotécnicas.

Enquanto a Tabela 5.1 apresenta a estrutura conceitual dos critérios de avaliação e sua fundamentação teórica, a Tabela 5.2 descreve o instrumento completo utilizado na coleta de dados, incluindo as questões de caracterização dos participantes e os itens avaliativos operacionalizados no *survey*.

Tabela 5.2: Instrumento de validação do Framework proposto

ID	Pergunta	Tipo
P1	Aceita participar da pesquisa conforme os termos apresentados?	Múltipla escolha (binária)
Bloco A — Caracterização do(a) Participante		
P2	Qual sua Esfera/Poder de Atuação?	Múltipla escolha
P3	Em qual Unidade Federativa você reside atualmente?	Múltipla escolha
P4	Qual é o seu nível atual de formação?	Múltipla escolha
P5	Qual sua função principal atual?	Múltipla escolha
P6	Indique a senioridade da posição que ocupa.	Múltipla escolha
P7	Quantos anos de experiência você tem em funções relacionadas ao seu papel?	Múltipla escolha

Continuação na próxima página

ID	Pergunta	Tipo
P8	Qual o nível de envolvimento direto com projeto ou governança de sistemas de IA?	Múltipla escolha
P9	Quais são os domínios dos sistemas que você trabalha atualmente?	Múltipla escolha
P10	Qual o tamanho da organização para a qual você trabalha atualmente?	Múltipla escolha
Bloco B — Avaliação do Framework		
P11	Em que medida o framework apoia efetivamente a tomada de decisão sobre como equilibrar explicabilidade e preservação da privacidade em sistemas reais?	Escala Likert
P12	Em que medida a estrutura do framework (camadas, artefatos e relações) é apresentada de forma clara e compreensível?	Escala Likert
P13	Em que medida o framework cobre, de forma suficiente, os aspectos necessários para orientar decisões sobre explicabilidade sob restrições de privacidade?	Escala Likert
P14	Em que medida o framework pode ser aplicado na prática, considerando tempo, esforço e integração com processos organizacionais existentes?	Escala Likert
P15	Em que medida o framework apoia práticas de governança, auditoria e responsabilização em sistemas baseados em IA?	Escala Likert
P16	Em que medida o framework pode ser adaptado a diferentes contextos organizacionais sem perda de coerência conceitual?	Escala Likert
Bloco C — Questões Abertas		
P17	Na sua opinião, quais são os principais pontos fortes e limitações do framework proposto? Há aspectos que poderiam ser aprimorados?	Aberta
P18	Em que contextos organizacionais ou tipos de sistema você considera que este framework seria mais adequado ou menos adequado? Explique brevemente.	Aberta
P19	Na sua avaliação, o framework oferece suporte suficiente para rastreabilidade, auditoria e responsabilização em contextos regulatórios (ex.: LGPD)? Caso contrário, o que deveria ser aprimorado?	Aberta

5.3 ANÁLISE DOS RESULTADOS

Esta seção apresenta os resultados obtidos a partir do *survey* de validação do *framework*. A análise está organizada em duas partes principais: inicialmente, descreve-se o perfil dos especi-

alistas participantes, contextualizando a qualificação da amostra; em seguida, são discutidos os resultados quantitativos e qualitativos referentes aos critérios de avaliação do *framework*.

5.3.1 Perfil dos Especialistas

Os participantes da validação foram selecionados por amostragem intencional, considerando experiência comprovada em ao menos um dos seguintes domínios: (i) inteligência artificial e aprendizado de máquina, (ii) privacidade e proteção de dados, (iii) engenharia de software e arquitetura de sistemas, ou (iv) governança, auditoria ou conformidade regulatória.

Buscou-se garantir diversidade de perfis para capturar diferentes perspectivas sobre utilidade, clareza e viabilidade do *framework*. A validação não teve como objetivo alcançar representatividade estatística, mas sim obter julgamento qualificado e informado, em consonância com as práticas de *Design Science Research*. No total, 32 (trinta e dois) profissionais responderam integralmente ao *survey*.

A Tabela 5.3 apresenta a caracterização detalhada dos participantes, incluindo esfera de atuação (P2), unidade federativa (P3), nível de formação (P4), função principal (P5), senioridade (P6), tempo de experiência (P7), nível de envolvimento com projetos de IA (P8) e porte da organização (P10). Observa-se predominância de profissionais vinculados ao setor público (84.3%), notadamente do Poder Executivo federal (46.9%), com residência no Distrito Federal (59.4%), além de elevado nível de qualificação acadêmica, com predomínio (84%) em nível de mestrado ou doutorado (concluído ou em andamento).

Verifica-se que houve a participação de pessoas de todas as cinco regiões do Brasil, o que configura representatividade regional da pesquisa.

No tocante à função principal atualmente desempenhada pelos respondentes, observa-se que as áreas de Engenharia/Desenvolvimento de Software e Segurança da Informação concentram a maior proporção dos participantes, ambas com 21.9% cada. Em seguida, destaca-se Ciência de Dados, com 15.6%, e Governança de Dados, com 12.5%. A função de Gestão de Projetos foi indicada por 6.3% dos respondentes, ao passo que Privacidade e Proteção de Dados foi assinalada por 3.1%. Outras funções informadas foram OSINT, Infraestrutura, Dirigente, Operador de Sistemas e Tecnologia da Informação, tendo cada uma destas recebido 3.1% de respostas.

Houve, ainda, a presença expressiva de profissionais com senioridade elevada (65.6%) e atuação em organizações de grande porte (56,3%), o que reforça a consistência do julgamento especializado obtido na avaliação do *framework*.

Esfera/Poder de atuação	#	%
Estadual – Poder Executivo	5	15.6
Federal – Empresa Pública	1	3.1
Federal – Poder Executivo	15	46.9
Federal – Poder Legislativo	2	14.2
Federal – Poder Judiciário	3	9.4
Federal – Ministério Público da União	1	3.1
Federal – Sociedade de Economia Mista	1	3.1
Municipal – Poder Executivo	1	3.1
Setor privado –	3	9.4
Unidade Federativa (residência)	#	%
Amazonas (AM)	1	3.1
Bahia (BA)	1	3.1
Distrito Federal (DF)	19	59.4
Goiás (GO)	1	3.1
Mato Grosso do Sul (MS)	3	9.4
Pará (PA)	1	3.1
Paraná (PR)	1	3.1
Rio de Janeiro (RJ)	1	3.1
Rio Grande do Sul (RS)	1	3.1
São Paulo (SP)	2	6.3
Sergipe (SE)	1	3.1
Nível de formação	#	%
Mestrado em andamento	11	34.4
Mestrado	4	12.5
Doutorado em andamento	7	21.9
Doutorado	4	12.5
Pós-Doutorado	1	7.1
Especialização	4	12.5
Superior incompleto	1	3.1
Função principal atual	#	%
Ciência de Dados	5	15.6
Desenvolvimento/Engenharia de Software	7	21.9
Governança de Dados	4	12.5
Gestão de Projetos	2	6.3
Privacidade e Proteção de Dados	1	3.1
Segurança da Informação	7	21.9
Outra função	5	15.6
Senioridade na posição	#	%
Júnior (até 5 anos)	3	15.6
Pleno (6 a 9 anos)	3	18.8
Sênior (10+ anos)	8	65.6
Tempo de experiência em funções relacionadas ao papel	#	%
Menos de 1 ano	0	0.0
Entre 1 e 3 anos	1	15.6
Entre 4 e 6 anos	3	18.8
Entre 7 e 14 anos	5	21.9
Mais de 15 anos	5	43.8
Envolvimento direto com projetos/governança de IA	#	%
Muito alto (trabalho diariamente)	5	15.6
Alto (envolvimento frequente/semanal)	6	18.8
Moderado (participo ocasionalmente)	15	46.9
Baixo (envolvimento raro/superficial)	6	18.8
Tamanho da organização	#	%
Até 10	1	3.1
De 100 a 499	7	21.9
De 500 a 999	5	15.6
Mais de 1000	18	56.3
Não soube informar	1	3.1

Tabela 5.3: Perfil dos participantes do survey de validação do framework ($n = 32$).

Conforme ilustrado na Figura 5.1, observa-se diversidade de domínios de atuação entre os especialistas participantes. Destacam-se os contextos Jurídico/Compliance (31.3%), seguido de Defesa e Segurança (25%), Educação (21.9%) Financeiro/Bancário (15.6%) e Saúde (12.5%), todos caracterizados por forte sensibilidade regulatória e impacto direto sobre direitos individuais.

Engenharia e Telecomunicações representaram ambos 9.4% dos domínios dos sistemas em que os respondentes atuam, enquanto Logística e Comércio tiveram, cada um 6.3%, respectivamente.

Houve, ainda, a sinalização de outros domínios como: Petróleo e Gás; Meio Ambiente; Governamental; Infraestrutura; Sistemas Administrativos; Indústria e; Forense, os quais tiveram menor representatividade.

A presença de domínios como Jurídico/Compliance, Defesa e Segurança, Educação, Financeiro/Bancário e Saúde reforça ainda mais o caráter crítico e regulado dos contextos representados. Essa diversidade contribui para ampliar a validade da avaliação, uma vez que o *framework* foi analisado sob múltiplas perspectivas setoriais.

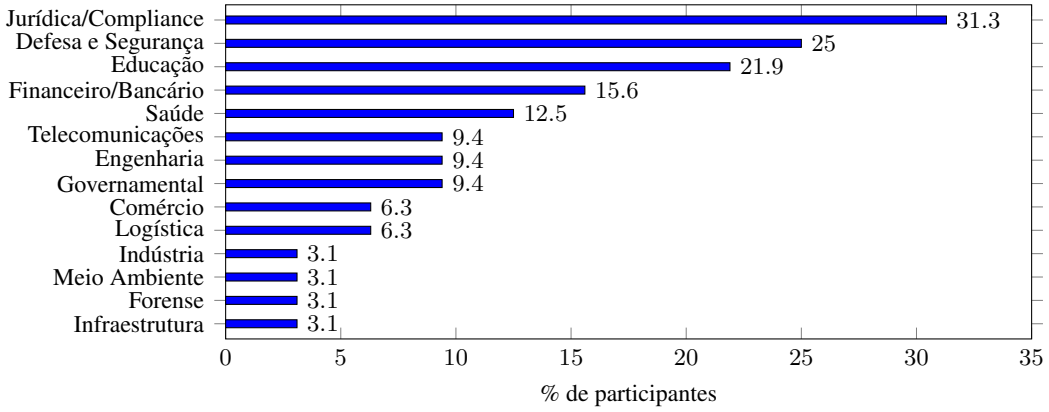

Figura 5.1: P9 — Domínios dos sistemas em que os especialistas atuam ($n = 32$, múltipla seleção).

Uma vez analisados os perfis dos respondentes, passa-se à analise das respostas referentes à avaliação do *framework*.

5.3.2 Avaliação Quantitativa dos Critérios do *Framework*

A Figura 5.2 apresenta a distribuição percentual das respostas nas escalas Likert para os seis critérios centrais de avaliação do *framework*: utilidade (P11), clareza conceitual (P12), completude (P13), aplicabilidade prática (P14), aderência à governança (P15) e flexibilidade (P16). A visualização empilhada permite observar não apenas a tendência central, mas também o padrão de dispersão das respostas entre os níveis da escala.

De modo geral, observa-se uma predominância das categorias 4 e 5 na maioria dos critérios, indicando uma percepção majoritariamente positiva do *framework* pelos respondentes.

O critério de utilidade (P11) apresentou elevada concentração nas categorias 4 (46.9%) e 5

(46.9%), de modo que é possível afirmar que houve 93.8% de avaliações positivas.

Quanto à clareza (P12) da proposta de *framework*, ainda que predominante positiva (78%), houve presença de respostas na categoria 3 (18.8%) e 2 (3.1%), sugerindo, que a despeito de majoritariamente compreensível, a proposta apresenta pontos que demandam maior detalhamento e/ou simplificação.

Por sua vez, no tange à completude (P13), 66% dos especialistas consideraram adequada aos principais aspectos técnicos, organizacionais e de governança envolvidos. Ressalta-se, que 25% das avaliações, foram de categoria 3, ou seja, intermediária, o que provavelmente indica a existência de lacunas e a necessidade de expansão de determinados aspectos do *framework*.

Em relação à aplicabilidade (P14), 69% dos especialistas avaliaram que a proposta é viável na prática, considerando tempo, esforço e integração com processos organizacionais existentes, ao passo que 18.8% entenderam que era moderadamente viável e 9.4%, pouco viável.

No que diz respeito à governança (P15), 88% das respostas foram positivas, entre as quais 68.8% na categoria 4 e 18.8% na categoria 5. A baixa frequência de respostas intermediárias (9.4%) aponta para um elevado grau de concordância entre os participantes quanto à contribuição do *framework* para os aspectos de governança e *accountability*.

No que tange à flexibilidade (P16), 59% dos especialistas a classificaram entre as categorias 4 (37.5%) e 5 (21.9%), enquanto que 37,5% indicaram a categoria 3. Tal avaliação sugere uma percepção mais heterogênea, ainda que a maioria dos respondentes a classifiquem como flexível, de modo que há espaço para melhorias nesse aspecto.

De modo geral, os resultados indicam elevada aceitação conceitual do *framework*, especialmente em relação aos critérios de utilidade e governança. Não obstante, observa-se a ocorrência pontual de avaliações mais baixas, com registros de respostas na categoria 1 para os critérios de clareza (P12) e de aplicabilidade (P14). Ademais, foram identificadas respostas na categoria 2 para utilidade (P11) (6.2%), completude (P13) (9.4%), governança (P15) (3.1%) e flexibilidade (P16) (3.1%), o que demonstra a existência de percepções menos favoráveis, ainda que minoritárias.

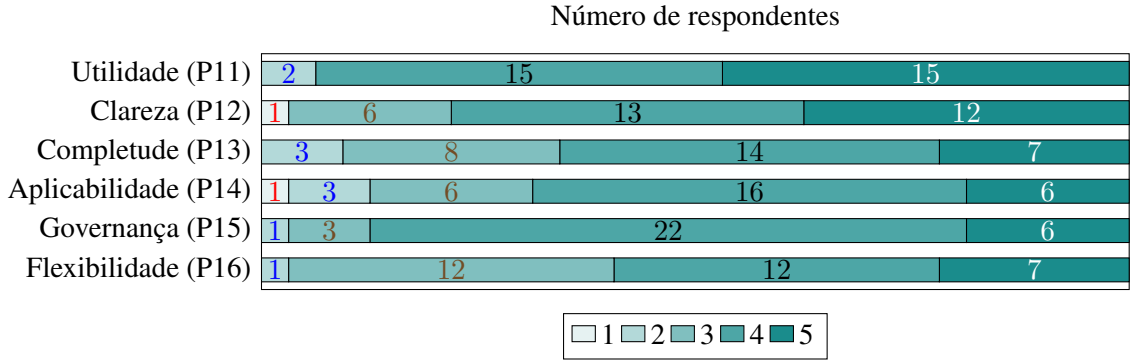

Figura 5.2: Distribuição das respostas Likert para os critérios de avaliação do framework (P11–P16, $n = 32$).

Complementarmente, a Tabela 5.4 apresenta as estatísticas descritivas (média e desvio pa-

drão) para cada critério, permitindo análise quantitativa mais precisa da tendência central e da variabilidade das respostas.

Conforme observado na Tabela 5.4, os critérios Utilidade (P11), Clareza (P12), e Aderência à Governança (P15) obtiveram média superior a 4, indicando percepção majoritariamente favorável. Os critérios de Completude (P13) e Flexibilidade (P16) obtiveram a mesma média (3.78).

A Aplicabilidade prática (P14) apresentou o maior desvio padrão (DP = 0.99), o que sugere maior heterogeneidade nas percepções quanto à viabilidade de implementação do *framework* em diferentes contextos organizacionais. Esse resultado é consistente com as respostas qualitativas apresentadas na seção seguinte, nas quais alguns especialistas mencionaram possíveis desafios relacionados à complexidade ou custo de adoção.

Critério	**Média**	**Desvio Padrão**
Utilidade (P11)	4.34	0.79
Clareza conceitual (P12)	4.09	0.93
Completude (P13)	3.78	0.91
Aplicabilidade prática (P14)	3.72	0.99
Aderência à governança (P15)	4.03	0.65
Flexibilidade (P16)	3.78	0.83

Tabela 5.4: Estatísticas descritivas dos critérios de avaliação do framework (P11–P16, $n = 32$)

5.3.3 Avaliação Qualitativa dos Critérios do *Framework*

A análise qualitativa revelou reconhecimento consistente da estrutura conceitual do *framework*.

Diversos especialistas destacaram, entre os principais pontos fortes da proposta, a abordagem estruturada e multicamadas, a qual favorece auditoria e responsabilização e alinha objetivos técnicos, legais e operacionais, por meio dos artefatos (A1-A4).

A integração explícita entre explicabilidade e proteção da privacidade foi reconhecida como pertinente, especialmente para sistemas de alto impacto e ambientes regulados.

Foi ressaltada a relevância do passo 2 do *framework*, do ponto de vista de cibersegurança, uma vez que ao demonstrar como os ataques de inferência podem ser impedidos, o modelo reduziria decisões *ad hoc*.

Além disso, a diferenciação por papéis, também foi apontada como ponto relevante, na medida em que evita uma explicação única para todos.

Conforme mencionado pelo Participante 4:

> *“Framework proposto apresenta como principal ponto forte a integração explícita entre explicabilidade e proteção da privacidade, tratando mecanismos de explicação como componentes potencialmente sensíveis e sujeitos a governança, o que é especialmente adequado para sistemas de alto impacto e ambientes regulados; a estrutura em camadas, com artefatos rastreáveis e diferenciação de requisitos por papel de stakeholder, favorece auditoria, responsabilização e alinhamento entre objetivos técnicos, legais e operacionais".*

Esse comentário reforça que trata-se de uma abordagem estruturada e rastreável com clara diferenciação de papéis e integração entre explicabilidade e privacidade.

Por outro lado, emergiram preocupações relacionadas à aplicabilidade prática e complexidade. O Participante 1 mencionou que:

“o framework pode apresentar custo de adoção relativamente alto” , enquanto o Participante 5 observou que “sem conhecimento técnico suficiente, nenhum framework funciona” , sugerindo que maturidade organizacional é fator crítico.

Também foram realizadas sugestões de aprimoramento, especialmente quanto à necessidade de maior operacionalização. Conforme indicado pelo Participante 10:

> *“seria importante incluir métricas objetivas de trade-off entre utilidade explicativa e exposição informacional, bem como critérios testáveis para avaliação de risco".*

À título de limitações foram apontadas a ausência de métricas objetivas, exemplos concretos e provas de conceito. Alguns respondentes sinalizaram sobre o potencial custo de adoção, bem como exigência de elevado nível de maturidade organizacional e técnica, o que poderia dificultar sua implementação em equipes menores ou em contextos menos estruturados.

Também foram realizadas sugestões de aprimoramento, especialmente quanto à necessidade de maior operacionalização, isto é, no aumento da aplicabilidade prática do *framework*. Para tanto, foi sugerida a inclusão de exemplos de uso e métricas objetivas de *trade-off* entre utilidade explicativa e exposição informacional, bem como níveis padronizados de transparência e procedimentos de teste adversarial das saídas explicativas, de modo a tornar o *framework* mais mensurável e aplicável em contextos práticos.

Além disso, foi recomendada a automação de determinadas etapas e a utilização de representações visuais, como fluxos BPMN, no intuito de aprimorar a sua usabilidade. Outra sugestão proposta por um dos especialistas foi integrar o Relatório de Impacto à Proteção de Dados Pessoais (RIPD) diretamente aos artefatos e gerir o "Direito à Explicação"do titular de forma automatizada, conectando o Artefato A2 (requisitos) com a interface final do usuário.

A tabela 5.5 sintetiza trechos representativos das respostas obtidas por meio das perguntas abertas do survey, relacionadas a pontos fortes, pontos fracos e propostas de melhoria do *framework*.

Categoria	Exemplos de termos / trechos representativos
Estrutura e Rastreabilidade como ponto forte	“abordagem estruturada e rastreável”; “integração explícita entre explicabilidade e proteção da privacidade”; “artefatos A1–A4 favorecem auditoria”; “diferenciação de requisitos por papel de stakeholder”
Desafios de Aplicabilidade e Complexidade	“custo de adoção relativamente alto”; “dificuldades para startups ou equipes pequenas”; “complexidade e custo”; “depende de maturidade organizacional”
Necessidade de Maior Operacionalização	“ausência de métricas objetivas de trade-off”; “modelagem de ameaças ainda genérica”; “falta de critérios testáveis”; “precisa incluir métricas de risco de explicação e para avaliar a qualidade das explicações.
Contextos Regulatórios e Setores Sensíveis	“especialmente adequado para ambientes regulados”; “saúde, finanças e setor público”; “compatível com exigências como a LGPD”
Necessidade de Exemplos e Guias Práticos	“prova de conceito”; “exemplos de papéis e técnicas XAI e PETs”; “guia explicativo sobre sua utilização”; “melhor amarração ao ciclo MLOps”

Tabela 5.5: Categorias emergentes a partir da análise qualitativa das questões abertas (P17–P19).

5.3.4 Melhorias Realizadas do Framework

Apesar de prevista na proposta inicial na camada 4, a etapa de reavaliação periódica baseada em gatilhos, não estava conectada ao início do *framework*, o que pode ter contribuído para uma percepção dos respondentes de que o *framework* era linear. Com vistas a aprimorar este ponto, de modo a tornar clara a ideia de reavaliação contínua e retroalimentação do processo, buscou-se incluir uma seta conectando a última camada (Governança e evidencias) à primeira camada. A partir dessa modificação, o desenho do *framework* conseguirá expressar de forma mais clara a abordagem cíclica, que tem por objetivo viabilizar o monitoramento e reavaliação contínua, alinhando-se a práticas de MLOps (*Machine Learning Operations*).

Na camada 1 (Contexto, risco e ameaças), foram incluídas métricas de risco de explicação, bem como de medidas de mitigação e salvaguardas. A inclusão de métricas de risco de explicação foi um aspecto sugerido por vários respondentes. De fato, a medida é essencial, razão pela qual alocou-se na camada 1, uma vez que é nela em que é realizado o diagnóstico inicial de contexto e riscos do tratamento de dados. Tal previsão permitirá a quantificação dos riscos associados, fortalecendo, assim uma avaliação mais objetiva na mensuração desse item.

Ainda, na camada 1, foi adicionada previsão sobre o mapeamento de medidas para mitigação e salvaguardas. Muito embora a implementação prática destas medidas ocorra somente na camada 3, propõe-se que nesta primeira camada sejam identificadas, a partir dos riscos que foram apontados, quais serão as medidas aptas a mitigá-los. Esse conjunto de informações, somado àquelas provenientes da camada 2, servirão como subsídio para o Projeto de Explicações com Salvaguardas descrito na camada 3, como também para o Relatório de Impacto à Proteção de Dados (RIPD), que agora, constará de forma expressa como um dos artefatos A2.

Quanto à camada 2 (Stakeholders e necessidades), adicionou-se etapa relativa à definição de níveis padronizados de transparência por papel. A inclusão, alinhada aos achados resultantes do

grupo focal, bem como do *survey*, permite assegurar que cada *stakeholder* receba uma explicação adequada ao seu perfil. Adicionalmente, procedeu-se com um pequeno ajuste substituindo o termo "critérios", por "métricas"de utilidade, compreensão e contestação. A referida alteração teve por objetivo conferir maior operacionalidade ao *framework* e objetividade na mensuração desses critérios. Como supramencionado, no que tange ao artefato A2, foi incluída menção expressa sobre subsídio para o RIPD, bem como a fixação de métricas.

Já na camada 3 (Projeto de explicações com salvaguardas de privacidade), houve o acréscimo de procedimento para explicação e para a contestação pelo titular de dados. Ambos procedimentos operacionalizam os direitos de explicação e de revisão previstos em Leis de Privacidade, como o GDPR, AI Act e LGPD. O procedimento para explicação permite que seja estruturado de modo que as informações sejam fornecidas de forma compreensível para o titular. De modo similar, o procedimento de revisão, definirá fluxos para a contestação das decisões automatizadas. A inserção dessas etapas na camada 3, encontram-se alinhadas, também, aos princípios de transparência e de *privacy by design*.

Por sua vez, na camada 4 (Governança e evidências) foi incluído procedimento de testes adversariais das saídas explicativas. A inclusão tem como objetivo avaliar, de forma empírica, as saídas explicativas frente às ameaças identificadas na camada 1, além de testar outros cenários adversos, de modo a possibilitar que sejam descobertas eventuais vulnerabilidades não localizadas na fase de mapeamento da primeira camada.

Além disso, foi acrescentada uma etapa de monitoramento regulatório. Tal inserção se faz necessária a fim de que o *framework* se mantenha atualizado em razão da superveniência de novas leis ou mesmo alteração daquelas atualmente vigentes, no que se refere às obrigações relativas à explicabilidade de decisões automatizadas. Outra alteração realizada nesta camada foi a inclusão do termo "contínua"à etapa de auditoria, com vistas a explicitar o seu caráter recorrente, de modo a afastar a possível interpretação de que se tratava de uma atividade pontual.

A partir das sugestões coletadas por meio do *survey*, foram realizados ajustes na proposta inicial do *framework*, gerando-se, assim, a versão final daquela descrita na figura 4.2.

5.3.5 Discussão Integrada dos Resultados à Luz da Design Science Research

A análise integrada dos resultados quantitativos e qualitativos fornece evidências consistentes sobre a validade do *framework* enquanto artefato prescritivo no contexto da *Design Science Research* (DSR). Conforme discutido anteriormente, a maioria dos critérios avaliados (P11–P16) apresentaram médias superiores a 4 (P11,P12 e P15), o que indica uma percepção predominantemente positiva quanto à utilidade, clareza e aderência à governança do framework. Enquanto que enquanto os outros critérios (P13,P14 e P16), obtiveram médias acima de 3.72, o que demonstra a necessidade de aperfeiçoamento quanto à aplicabilidade, flexibilidade e completude.

Sob a perspectiva da DSR, tais resultados sugerem que o artefato atende ao critério de *utility*, ao demonstrar capacidade percebida de apoiar decisões reais relacionadas ao equilíbrio entre ex-

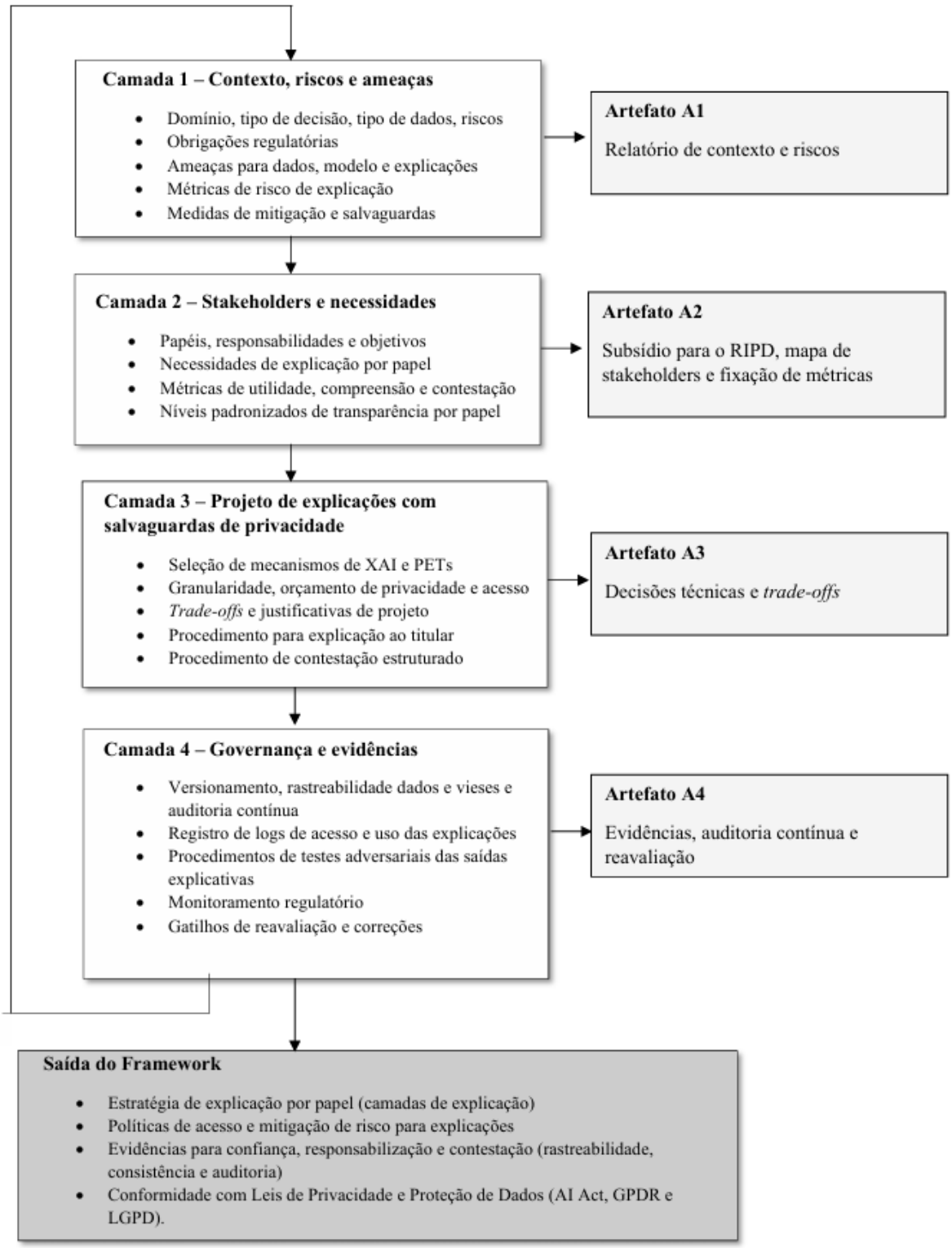

Figura 5.3: Proposta de Framework após validação via survey

plicabilidade e preservação da privacidade. A elevada avaliação nos critérios de Utilidade (P11), Governança (P15) e Flexibilidade (P16) reforça o alinhamento do *framework* com problemas organizacionais relevantes, especialmente em contextos regulados e de alto impacto.

Os resultados qualitativos complementam essa evidência ao destacar como pontos fortes a estrutura multicamada, a rastreabilidade por meio dos artefatos A1–A4 e a diferenciação de responsabilidades por perfil de *stakeholder*. Esses elementos foram reconhecidos como mecanismos que favorecem auditoria, responsabilização e coerência decisória, aspectos centrais para sistemas sociotécnicos sensíveis à privacidade.

Ao mesmo tempo, emergiram críticas construtivas relacionadas à aplicabilidade prática e ao grau de operacionalização do modelo. A maior dispersão observada no critério de Aplicabilidade prática (P14), combinada com menções à complexidade e ao custo de adoção, indica que a implementação do *framework* pode demandar maturidade organizacional e capacidade técnica prévia. Além disso, sugestões de inclusão de métricas objetivas, critérios testáveis e modelagem mais detalhada de ameaças apontam oportunidades de refinamento incremental do artefato.

Sob a ótica da DSR, tais observações não invalidam o artefato, mas contribuem para seu aprimoramento iterativo, característica intrínseca ao ciclo de construção e avaliação. A identificação de oportunidades de melhoria reforça o caráter evolutivo da solução proposta e fornece insumos concretos para futuras versões do *framework*. Em síntese, os resultados obtidos indicam que o *framework* atende aos critérios de relevância e utilidade no domínio investigado, ao mesmo tempo em que evidencia áreas específicas para aprofundamento metodológico e operacional. Assim, a validação conduzida cumpre o papel de avaliar o artefato em contexto especializado, fornecendo evidências empíricas de sua adequação conceitual e aplicabilidade potencial.

5.3.6 Ameaças à Validade

Como em qualquer estudo empírico baseado em julgamento especializado, esta pesquisa está sujeita a limitações que devem ser consideradas na interpretação dos resultados [69]. **Validade de construção**: Os critérios de avaliação foram derivados da literatura de Design Science Research e operacionalizados por meio de perguntas explícitas em escala Likert. Ainda assim, há risco de que diferentes especialistas tenham interpretado os itens de forma levemente distinta, especialmente conceitos como “completude” ou “aplicabilidade prática”. Para mitigar essa ameaça, cada critério foi descrito de forma clara no instrumento, e o *framework* foi previamente apresentado aos participantes antes da avaliação.

Validade interna: Os resultados refletem percepções declaradas dos especialistas e não observações de aplicação prática do *framework* em um ambiente real. Assim, as conclusões indicam aceitação conceitual e viabilidade percebida, mas não medem impacto empírico ou desempenho organizacional após implementação. Essa limitação é inerente a avaliações de artefatos conceituais em estágios iniciais.

Validade externa: A amostra foi composta por 32 (trinta e dois) especialistas selecionados

por amostragem intencional, recrutados a partir da rede profissional da pesquisadora, de colegas atuantes em órgãos públicos e em organizações relacionadas ao tema, bem como a partir de grupos de Whatsapp de pesquisadores da área de Engenharia de Software e da Segurança da Informação Embora os participantes apresentem experiência relevante e diversidade de funções, os resultados não têm pretensão de representatividade estatística. A generalização deve ser entendida como analítica, e não estatística, em consonância com a abordagem da DSR.

Validade de conclusão: O tamanho reduzido da amostra limita análises estatísticas mais robustas. Entretanto, a consistência observada nas respostas (médias elevadas e baixa ocorrência de avaliações negativas) e a convergência entre evidências quantitativas e qualitativas fortalecem a credibilidade dos achados. Ainda assim, estudos futuros com aplicação longitudinal do *framework* poderiam ampliar a robustez das conclusões.

Embora existam limitações inerentes ao desenho metodológico adotado, as estratégias empregadas, fundamentação teórica dos critérios, diversidade de especialistas e triangulação entre dados quantitativos e qualitativos contribuem para mitigar riscos à validade e reforçam a confiabilidade da avaliação realizada.

5.4 SÍNTESE DO CAPÍTULO

Este capítulo apresentou a validação do *framework* proposto sob a perspectiva de especialistas, em consonância com a fase de *evaluation* da Design Science Research. A análise quantitativa evidenciou avaliações predominantemente positivas em todos os critérios investigados, destacando-se a utilidade, a aderência à governança e a clareza conceitual. Complementarmente, a análise qualitativa revelou reconhecimento da estrutura multicamada, da rastreabilidade por meio dos artefatos e da diferenciação de responsabilidades como pontos fortes centrais do modelo. Ao mesmo tempo, emergiram sugestões construtivas relacionadas à necessidade de maior operacionalização, definição de métricas objetivas e apoio à implementação prática. As evidências obtidas indicam que o *framework* apresenta adequação conceitual e potencial aplicabilidade em contextos regulados e sensíveis à privacidade, ainda que sua implementação possa demandar maturidade organizacional e refinamentos incrementais. Assim, a validação realizada fornece suporte empírico à relevância e utilidade do artefato, ao mesmo tempo em que aponta direções claras para seu aprimoramento futuro.

6 CONCLUSÃO

Esta dissertação investigou a relação entre explicabilidade (XAI) e preservação da privacidade em sistemas de apoio à decisão sensíveis, partindo do reconhecimento de que esses dois objetivos, frequentemente apresentados como complementares no discurso normativo, revelam tensões estruturais quando analisados sob a perspectiva técnica, organizacional e regulatória.

Os resultados do estudo empírico conduzido por meio de grupo focal evidenciaram que a compatibilidade entre explicabilidade e privacidade não pode ser assumida como intrínseca. Pelo contrário, trata-se de uma relação mediada por *trade-offs* dependentes de contexto, *stakeholder* e finalidade da explicação. Explicações globais e agregadas mostraram-se mais adequadas em ambientes regulados e sensíveis à privacidade, enquanto explicações locais e altamente detalhadas podem introduzir riscos adicionais de exposição informacional. Além disso, a pesquisa revelou que explicabilidade cumpre funções distintas: pode apoiar confiança operacional (uso adequado do sistema), responsabilização (prestação de contas) ou diagnóstico técnico, exigindo, portanto, decisões explícitas sobre objetivos e limites.

A partir dessas evidências, foi proposta uma estrutura conceitual de explicabilidade orientada à governança, construída segundo o paradigma da Design Science Research. O *framework* organiza-se em quatro camadas interdependentes: (i) contexto e risco; (ii) *stakeholders* e necessidades; (iii) projeto de explicações com restrições de privacidade; e (iv) governança, evidências e auditoria. Essa estrutura trata a explicabilidade não como mera funcionalidade técnica, mas como capacidade organizacional que deve ser planejada, documentada, controlada e auditada.

O modelo propõe decisões explícitas sobre o que explicar, para quem explicar, com qual granularidade e sob quais garantias de privacidade, incorporando mecanismos de rastreabilidade por meio de artefatos estruturados. Ao fazê-lo, contribui para reduzir a lacuna identificada entre princípios normativos como aqueles presentes na LGPD, no GDPR e em diretrizes internacionais de ética em IA e práticas concretas de engenharia e governança.

A validação do *framework*, realizada por meio de *survey* com especialistas atuantes em contextos regulados (jurídico, financeiro, saúde, governamental, entre outros), indicou avaliação predominantemente positiva quanto à utilidade, clareza conceitual e aderência à governança. As análises quantitativas apresentaram médias superiores a 4 na escala Likert na maioria dos critérios avaliados, enquanto as contribuições qualitativas reforçaram o reconhecimento da estrutura multicamada, da rastreabilidade dos artefatos e da diferenciação de papéis como pontos fortes centrais do modelo. Ao mesmo tempo, emergiram sugestões relevantes relacionadas à necessidade de maior operacionalização, definição de métricas objetivas de *trade-off* e apoio à implementação prática.

Sob a ótica da *Design Science Research*, os resultados indicam que o artefato atende aos critérios de utilidade e relevância organizacional, oferecendo conhecimento prescritivo aplicável

a sistemas de apoio à decisão sensíveis à privacidade. O *framework* contribui ao estruturar a tomada de decisão sobre explicabilidade de forma justificável, auditável e alinhada à governança, fortalecendo a defensabilidade institucional em ambientes regulatórios complexos.

Não obstante, o estudo apresenta limitações. A amostra de especialistas, embora qualificada e diversa em termos setoriais, não possui representatividade estatística, sendo a generalização de natureza analítica. A avaliação concentrou-se na percepção especializada sobre o artefato, não envolvendo aplicação longitudinal em cenários reais de implementação. Além disso, a operacionalização de métricas quantitativas para mensurar formalmente os *trade-offs* entre utilidade explicativa e risco de exposição informacional permanece como agenda aberta.

Como trabalhos futuros, recomenda-se: (i) aplicação do framework em estudos de caso reais com monitoramento longitudinal; (ii) desenvolvimento de métricas formais para avaliação de risco de explicação e exposição informacional; (iii) integração explícita com práticas de MLOps e governança de dados; (iv) adaptação do modelo para diferentes níveis de maturidade organizacional e (v) adaptação do framework para ambientes distribuídos, especialmente em cenários de aprendizado federado.

Em síntese, esta pesquisa demonstra que a explicabilidade em sistemas sensíveis à privacidade não deve ser tratada como requisito técnico isolado, mas como decisão estratégica e organizacional inserida em um ecossistema de governança. Ao propor uma estrutura que articula *stakeholders*, riscos, decisões de projeto e mecanismos de auditoria, a dissertação contribui para a consolidação de uma abordagem de explicabilidade responsável, alinhada à proteção de direitos fundamentais e à necessidade crescente de responsabilização em sistemas de inteligência artificial.

REFERÊNCIAS BIBLIOGRÁFICAS

1 WIERINGA, R. J. *Design Science Methodology for Information Systems and Software Engineering*. Springer, 2014. ISBN 978-3-662-43838-1. Disponível em: <https://doi.org/10.1007/978-3-662-43839-8>.

2 HEVNER, A.; CHATTERJEE, S. Design science research in information systems. In: *Design research in information systems: theory and practice*. Springer, 2010. p. 9–22. Disponível em: <https://doi.org/10.1007/978-1-4419-5653-8_2>.

3 PRAT, N.; COMYN-WATTIAU, I.; AKOKA, J. A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, Taylor & Francis, v. 32, n. 3, p. 229–267, 2015.

4 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*, v. 58, p. 82–115, 2020. Disponível em: <https://doi.org/10.1016/j.inffus.2019.12.012>.

5 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*, v. 13, p. 121326–121350, 2025. Disponível em: <https://doi.org/10.1109/ACCESS.2025.3575022>.

6 GOLDSTEIN, J. A.; SASTRY, G.; MUSSER, M.; DIRESTA, R.; GENTZEL, M.; SEDOVA, K. Generative language models and automated influence operations: Emerging threats and potential mitigations. *CoRR*, abs/2301.04246, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2301.04246>.

7 NOORDT, C. van; MISURACA, G. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the european union. *Gov. Inf. Q.*, v. 39, n. 3, p. 101714, 2022. Disponível em: <https://doi.org/10.1016/j.giq.2022.101714>.

8 NTOUTSI, E.; FAFALIOS, P.; GADIRAJU, U.; IOSIFIDIS, V.; NEJDL, W.; VIDAL, M.; RUGGIERI, S.; TURINI, F.; PAPADOPOULOS, S.; KRASANAKIS, E.; KOMPATSIARIS, I.; KINDER-KURLANDA, K.; WAGNER, C.; KARIMI, F.; FERNÁNDEZ, M.; ALANI, H.; BERENDT, B.; KRUEGEL, T.; HEINZE, C.; BROELEMANN, K.; KASNECI, G.; TIROPANIS, T.; STAAB, S. Bias in data-driven AI systems - an introductory survey. *CoRR*, abs/2001.09762, 2020. Disponível em: <https://arxiv.org/abs/2001.09762>.

9 OpenAI. *What is ChatGPT? Commonly asked questions about ChatGPT*. 2023. <https://help.openai.com/en/articles/6783457-what-is-chatgpt>. Acesso em: 4 set. 2025.

10 BRASIL. *Lei nº 13.709, de 14 de agosto de 2018 - Lei Geral de Proteção de Dados Pessoais (LGPD)*. 2018. Disponível em: <https://www.planalto.gov.br/ccivil$_03/_{a}to2015-2018/2018/lei/l13709.htm$>.

11 EUROPEAN PARLIAMENT AND COUNCIL. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2018. Disponível em: <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>.

12 SELBST, A.; POWLES, J. "meaningful information"and the right to explanation. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 48. Disponível em: <http://proceedings.mlr.press/v81/selbst18a.html>.

13 GOODMAN, B.; FLAXMAN, S. R. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.*, v. 38, n. 3, p. 50–57, 2017. Disponível em: <https://doi.org/10.1609/aimag.v38i3.2741>.

14 WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, Oxford University Press, v. 7, n. 2, p. 76–99, 2017.

15 BRKAN, M.; BONNET, G. Legal and technical feasibility of the gdpr's quest for explanation of algorithmic decisions: of black boxes, white boxes and fata morganas. *European Journal of Risk Regulation*, Cambridge University Press, v. 11, n. 1, p. 18–50, 2020.

16 EUROPEAN PARLIAMENT AND COUNCIL. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. 2024. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>.

17 UNESCO. Recommendation on the ethics of artificial intelligence. *UNESCO:Paris, France*, p. 1–44, 2022. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en>.

18 OECD. *Recommendation of the Council on Artificial Intelligence*. [S.l.], 2024. Originally adopted on 22 May 2019; amended on 3 May 2024.

19 NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Artificial intelligence risk management framework (AI RMF 1.0)*. 2023. 1-48 p. Https://doi.org/10.6028/NIST.AI.100-1. Disponível em: <https://nvlpubs. nist.gov/nistpubs/ai/nist.ai>.

20 ADNAN, M.; SYED, M. H.; ANJUM, A.; REHMAN, S. A framework for privacy-preserving in iov using federated learning with differential privacy. *IEEE Access*, v. 13, p. 13507–13521, 2025. Disponível em: <https://doi.org/10.1109/ACCESS.2025.3526934>.

21 ALLANA, S.; KANKANHALLI, M.; DARA, R. Privacy risks and preservation methods in explainable artificial intelligence: A scoping review. *CoRR*, abs/2505.02828, 2025. Disponível em: <https://doi.org/10.48550/arXiv.2505.02828>.

22 EZZEDDINE, F. Privacy implications of explainable AI in data-driven systems. In: LONGO, L.; LIU, W.; MONTAVON, G. (Ed.). *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-WS.org, 2024. (CEUR Workshop Proceedings, v. 3793), p. 361–368. Disponível em: <https://ceur-ws.org/Vol-3793/paper_46.pdf>.

23 POTLURI, S. Policy-aware secure data governance in distributed information systems using explainable ai models. *International Journal of AI, BigData, Computational and Management Studies*, v. 6, n. 3, p. 1–10, 2025.

24 SCHWALBE, G.; FINZEL, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.*, v. 38, n. 5, p. 3043–3101, 2024. Disponível em: <https://doi.org/10.1007/s10618-022-00867-8>.

25 CHUNG, N. C.; CHUNG, H.; LEE, H.; CHUNG, H.; BROCKI, L.; DYER, G. C. False sense of security in explainable artificial intelligence (XAI). *CoRR*, abs/2405.03820, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2405.03820>.

26 RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, v. 1, n. 5, p. 206–215, 2019. Disponível em: <https://doi.org/10.1038/s42256-019-0048-x>.

27 MAJHI, B.; KASHYAP, A.; MOHANTY, S. S.; DASH, S.; MALLIK, S.; LI, A.; ZHAO, Z. An improved method for diagnosis of parkinson's disease using deep learning models enhanced with metaheuristic algorithm. *BMC Medical Imaging*, v. 24, n. 1, p. 156, 2024. Disponível em: <https://doi.org/10.1186/s12880-024-01335-z>.

28 AWOSIKA, T.; SHUKLA, R. M.; PRANGGONO, B. Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. *IEEE Access*, v. 12, p. 64551–64560, 2024. Disponível em: <https://doi.org/10.1109/ACCESS.2024.3394528>.

29 CONSELHO NACIONAL DE JUSTIÇA. *Painel de Projetos de IA no Judiciário Brasileiro*. 2025. Disponível em: <https://paineis.cnj.jus.br/iajud/>.

30 EUROPEAN COMMISSION. *Ethics guidelines for trustworthy AI*. 2019. Disponível em: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

31 RADANLIEV, P. AI ethics: Integrating transparency, fairness, and privacy in AI development. *Appl. Artif. Intell.*, v. 39, n. 1, 2025. Disponível em: <https://doi.org/10.1080/08839514.2025.2463722>.

32 BRASIL. *Lei nº 12.414, de 9 de junho de 2011*. 2011. Disponível em: <https://www.planalto.gov.br/ccivil$_03/_ato2011-2014/2011/lei/l12414.htm$>.

33 CEDIS-IDP; Jusbrasil. *Relatório do Painel LGPD nos Tribunais 2025*. 2025. Disponível em: <https://static.jusbr.com/painel-lgpd/edicoes/relatorio-do-painel-lgpd-nos-tribunais-2025.pdf>.

34 EUROPEAN COMMISSION. *AI Act*. Disponível em: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

35 WICK, M. R.; THOMPSON, W. B. Reconstructive expert system explanation. *Artif. Intell.*, v. 54, n. 1, p. 33–70, 1992. Disponível em: <https://doi.org/10.1016/0004-3702(92)90087-E>.

36 SWARTOUT, W. R. XPLAIN: A system for creating and explaining expert consulting programs. *Artif. Intell.*, v. 21, n. 3, p. 285–325, 1983. Disponível em: <https://doi.org/10.1016/S0004-3702(83)80014-9>.

37 SCOTT, A. C.; CLANCEY, W. J.; DAVIS, R.; SHORTLIFFE, E. H. Explanation capabilities of production-based consultation systems. *American Journal of Computational Linguistics*, p. 1–50, 1977.

38 GUNNING, D.; AHA, D. W. Darpa's explainable artificial intelligence (XAI) program. *AI Mag.*, v. 40, n. 2, p. 44–58, 2019. Disponível em: <https://doi.org/10.1609/aimag.v40i2.2850>.

39 REINHARDT, K.; BUCHHOLZ, O. XAI: on explainability and the obligation to explain. *Digit. Soc.*, v. 4, n. 3, p. 69, 2025. Disponível em: <https://doi.org/10.1007/s44206-025-00215-5>.

40 GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; TURINI, F.; GIANNOTTI, F.; PEDRESCHI, D. A survey of methods for explaining black box models. *ACM Comput. Surv.*, v. 51, n. 5, p. 93:1–93:42, 2019. Disponível em: <https://doi.org/10.1145/3236009>.

41 SPEITH, T. A review of taxonomies of explainable artificial intelligence (XAI) methods. In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 2022. p. 2239–2250. Disponível em: <https://doi.org/10.1145/3531146.3534639>.

42 SALIH, A. M. A.; RAISI-ESTABRAGH, Z.; GALAZZO, I. B.; RADEVA, P.; PETERSEN, S. E.; LEKADIR, K.; MENEGAZ, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.*, v. 7, n. 1, 2025. Disponível em: <https://doi.org/10.1002/aisy.202400304>.

43 ALI, S.; ABUHMED, T.; EL-SAPPAGH, S. H. A.; MUHAMMAD, K.; ALONSO-MORAL, J. M.; CONFALONIERI, R.; GUIDOTTI, R.; SER, J. D.; RODRÍGUEZ, N. D.; HERRERA, F. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, v. 99, p. 101805, 2023. Disponível em: <https://doi.org/10.1016/j.inffus.2023.101805>.

44 CONFALONIERI, R.; COBA, L.; WAGNER, B.; BESOLD, T. R. A historical perspective of explainable artificial intelligence. *WIREs Data Mining Knowl. Discov.*, v. 11, n. 1, 2021. Disponível em: <https://doi.org/10.1002/widm.1391>.

45 RAWAL, A.; MCCOY, J.; RAWAT, D. B.; SADLER, B. M.; AMANT, R. S. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Trans. Artif. Intell.*, v. 3, n. 6, p. 852–866, 2022. Disponível em: <https://doi.org/10.1109/TAI.2021.3133846>.

46 LEMIEUX, V. L.; WERNER, J. Protecting privacy in digital records: The potential of privacy-enhancing technologies. *ACM Journal on Computing and Cultural Heritage*, v. 16, n. 4, p. 83:1–83:18, 2023. Disponível em: <https://doi.org/10.1145/3633477>.

47 OCDE. *Emerging Privacy Enhancing Technologies: current regulatory and policy approaches*. 2023. OECD Digital Economy Papers n. 351. Disponível em: <https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/03/emerging-privacy-enhancing-technologies$_a6bdf3cb/bf121be4 - en.pdf$>.

48 BOEDIHARDJO, M.; STROHMER, T.; VERSHYNIN, R. Privacy of synthetic data: A statistical framework. *IEEE Trans. Inf. Theory*, v. 69, n. 1, p. 520–527, 2023. Disponível em: <https://doi.org/10.1109/TIT.2022.3216793>.

49 MENG, G.; ZHANG, L. A survey on secure multi-party computation techniques based on private set intersection. *Comput. Stand. Interfaces*, v. 96, p. 104067, 2026. Disponível em: <https://doi.org/10.1016/j.csi.2025.104067>.

50 MARCOLLA, C.; SUCASAS, V.; MANZANO, M.; BASSOLI, R.; FITZEK, F. H. P.; AARAJ, N. Survey on fully homomorphic encryption, theory, and applications. *Proc. IEEE*, v. 110, n. 10, p. 1572–1609, 2022. Disponível em: <https://doi.org/10.1109/JPROC.2022.3205665>.

51 RAZI, Q.; DATTA, S.; HASSIJA, V.; CHALAPATHI, G. S. S.; SIKDAR, B. Privacy utility tradeoff between pets: Differential privacy and synthetic data. *IEEE Trans. Comput. Soc. Syst.*, v. 12, n. 2, p. 473–484, 2025. Disponível em: <https://doi.org/10.1109/TCSS.2024.3479317>.

52 RAZI, Q.; PIYUSH, R.; CHAKRABARTI, A.; SINGH, A.; HASSIJA, V.; CHALAPATHI, G. S. S. Enhancing data privacy: A comprehensive survey of privacy-enabling technologies. *IEEE Access*, v. 13, p. 40354–40385, 2025. Disponível em: <https://doi.org/10.1109/ACCESS.2025.3546618>.

53 OCDE. *Sharing Trustworthy AI Models with Privacy-Enhancing Technologies*. 2025. OECD Artificial Intelligence Papers, n.38. Disponível em: <https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/sharing-trustworthy-ai-models-with-privacy-enhancing-technologies$_5df6fd05/a266160b - en.pdf$>.

54 SHOKRI, R.; STROBEL, M.; ZICK, Y. On the privacy risks of model explanations. In: FOURCADE, M.; KUIPERS, B.; LAZAR, S.; MULLIGAN, D. K. (Ed.). *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. ACM, 2021. p. 231–241. Disponível em: <https://doi.org/10.1145/3461702.3462533>.

55 SPARTALIS, C. N.; SEMERTZIDIS, T.; DARAS, P. Balancing XAI with privacy and security considerations. In: KATSIKAS, S. K.; ABIE, H.; RANISE, S.; VERDERAME, L.; CAMBIASO, E.; UGARELLI, R. M.; PRAÇA, I.; LI, W.; MENG, W.; FURNELL, S.; KATT, B.; PIRBHULAL, S.; SHUKLA, A.; IANNI, M.; PREDA, M. D.; CHOO, K. R.; CORREIA, M. P.; ABHISHTA, A.; SILENO, G.; ALISHAHI, M.; KALUTARAGE, H. K.; YANAI, N. (Ed.). *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*. Springer, 2023. (Lecture Notes in Computer Science, v. 14399), p. 111–124. Disponível em: <https://doi.org/10.1007/978-3-031-54129-2_7>.

56 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.*, v. 6, p. 8389–8409, 2025. Disponível em: <https://doi.org/10.1109/OJCOMS.2025.3608784>.

57 ALLANA, S.; KANKANHALLI, M.; DARA, R. Privacy risks and preservation methods in explainable artificial intelligence: A scoping review. *Trans. Mach. Learn. Res.*, v. 2025, 2025. Disponível em: <https://openreview.net/forum?id=q9nykJfzku>.

58 ABBASI, W.; MORI, P.; SARACINO, A. The explainability-privacy-utility trade-off for machine learning-based tabular data analysis. In: VIMERCATI, S. D. C. di; SAMARATI, P. (Ed.). *Proceedings of the 20th International Conference on Security and Cryptography, SECRYPT 2023, Rome, Italy, July 10-12, 2023*. SCITEPRESS, 2023. p. 511–519. Disponível em: <https://doi.org/10.5220/0012137800003555>.

59 KEANEY, S.; BERTHON, P. Balancing explainability and privacy in ai systems: A strategic imperative for managers. *Business Horizons*, 2025. ISSN 0007-6813. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0007681325001715>.

60 WANG, Y. Balancing trustworthiness and efficiency in artificial intelligence systems: An analysis of tradeoffs and strategies. *IEEE Internet Computing*, v. 27, n. 6, p. 8–12, 2023.

61 SHAH, B.; GUPTA, E. V. Ethics and privacy in ai-driven healthcare decision support systems. *International Journal of Research and Analytical Reviews (IJRAR)*, v. 12, p. 1–11, 2025.

62 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. Disponível em: <https://arxiv.org/abs/2503.16233>.

63 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: SPRINGER. *World Conference on Explainable Artificial Intelligence*. [S.l.], 2025. p. 233–254.

64 ROCHA, L. D.; CANEDO, E. D. Optimizing compliance: Comparative study of data laws and privacy frameworks. *J. Internet Serv. Appl.*, v. 16, n. 1, p. 431–452, 2025. Disponível em: <https://doi.org/10.5753/jisa.2025.5247>.

65 KAPITSAKI, G. M.; PAPOUTSOGLOU, M.; TREUDE, C.; THEOPHILOU, I. Analyzing developer discussions on EU and US privacy legislation compliance in github repositories. *CoRR*, abs/2512.10618, 2025. Disponível em: <https://doi.org/10.48550/arXiv.2512.10618>.

66 KONTIO, J.; LEHTOLA, L.; BRAGGE, J. Using the focus group method in software engineering: Obtaining practitioner and user experiences. In: *2004 International Symposium on Empirical Software Engineering (ISESE 2004), 19-20 August 2004, Redondo Beach, CA, USA*. IEEE Computer Society, 2004. p. 271–280. Disponível em: <https://doi.org/10.1109/ISESE.2004.35>.

67 HUANG, R.; SAMARAWEERA, G. D.; CHANG, J. M. Exploring threats, defenses, and privacy-preserving techniques in federated learning: A survey. *Computer*, v. 57, n. 4, p. 46–56, 2024. Disponível em: <https://doi.org/10.1109/MC.2023.3324975>.

68 WIRATSIN, I.-O.; RAGKHITWETSAGUL, C. Effectiveness of explainable artificial intelligence (xai) techniques for improving human trust in machine learning models: A systematic literature review. *IEEE Access*, v. 13, p. 121326–121350, 2025. Disponível em: <https://ieeexplore.ieee.org/abstract/document/11017606>.

69 WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. *Experimentation in Software Engineering*. Springer, 2012. ISBN 978-3-642-29043-5. Disponível em: <https://doi.org/10.1007/978-3-642-29044-2>.

70 PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. *J. Manag. Inf. Syst.*, v. 24, n. 3, p. 45–77, 2008. Disponível em: <https://doi.org/10.2753/mis0742-1222240302>.

71 KITCHENHAM, B. A.; PFLEEGER, S. L. Personal opinion surveys. In: SHULL, F.; SINGER, J.; SJØBERG, D. I. K. (Ed.). *Guide to Advanced Empirical Software Engineering*. Springer, 2008. p. 63–92. Disponível em: <https://doi.org/10.1007/978-1-84800-044-5_3>.