FGV DIREITO SP

MESTRADO PROFISSIONAL

FRAMEWORK PARA EXPLICABILIDADE E TRANSPARÊNCIA EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL

Camila de Araújo Guimarães

Projeto de pesquisa apresentado ao
Mestrado Profissional da FGV Direito SP – Direito e Tecnologia.

Versão de 31.01.2025

1. Tema, contexto e delimitação de escopo

O desenvolvimento e a adoção da Inteligência Artificial (IA) apresentam velocidade e impacto sem precedentes na história da inovação tecnológica. Enquanto tecnologias revolucionárias anteriores, como a eletricidade e a internet, levaram décadas para atingir ampla adoção e transformar significativamente as relações sociais e econômicas, a IA e principalmente a IA generativa tem demonstrado uma curva de evolução e usuários exponencialmente mais acelerada. Esta velocidade peculiar de desenvolvimento e utilização traz consigo desafios igualmente únicos, especialmente no que tange à compreensão, controle e regulamentação destes sistemas.

Dentre os desafios deste cenário encontra-se a necessidade de garantir que sistemas de IA sejam não apenas eficientes ao propósito pelo qual foram desenvolvidos, mas também compreensíveis e auditáveis para que a ética em IA perpasse a idealização e seja materializada em ações concretas para a garantia de direitos fundamentais.

Neste contexto, reside a discussão fundamental sobre os conceitos interrelacionados, porém distintos, de transparência e explicabilidade em IA. Embora estes termos sejam frequentemente utilizados de forma intercambiável, apresentam nuances significativas que são relevantes para o desenvolvimento de sistemas confiáveis. Esta compreensão tem evoluído significativamente nos últimos anos, com ênfase na necessidade de adoção de abordagem multidisciplinar para alcançar a verdadeira transparência e suficiente explicabilidade, superando, tanto quanto possível, a aparente opacidade de sistemas dotados de aprendizado de máquina.

Transparência e explicabilidade, constam, inclusive, como um dos princípios éticos da IA e a Cooperação e Desenvolvimento Econômico (OCDE)[1]. Como destaca a organização internacional, a transparência em IA possui múltiplas dimensões e implica não apenas na divulgação do uso de sistemas de IA, mas também na compreensão de como estes sistemas são desenvolvidos, treinados, operados e implementados e enquanto a transparência se relaciona com a divulgação apropriada ao contexto sobre o uso de IA e suas capacidades e limitações gerais. Já a explicabilidade foca especificamente em tornar compreensível o processo decisório de um sistema para as pessoas afetadas, possibilitando que elas contestem resultados desfavoráveis a partir da compreensão dos fatores e da lógica que os determinaram.

O cenário regulatório tem evoluído em busca de acompanhar a velocidade das inovações e endereçar estes desafios. A aprovação do AI Act na União Europeia, em 13 de junho de 2024, representa marco significativo neste aspecto e a relevância da transparência no regulamento é evidenciada por sua menção explícita em mais 41 ocasiões, tanto nos considerandos como nos artigos, estabelece requisitos e obrigações de documentação técnica, manutenção de registos dos sistemas de IA e fornecimento direto de informações.

Da fundamentação contida nos considerandos, compreende-se que para o regulador europeu a transparência exige que o desenvolvimento e a implementação dos sistemas de IA sejam conduzidos de maneira que permita a rastreabilidade e explicabilidade de seu funcionamento, oportunidade em que é possível extrair três objetivos principais: primeiro, garantir que as pessoas estejam cientes quando interagem com um sistema de IA; segundo, assegurar que os implementadores compreendam tanto as potencialidades quanto as limitações do sistema; e terceiro, comunicar claramente às pessoas afetadas quais são seus direitos em relação ao uso destes sistemas e permitir que sejam capazes de entender como operam [2].

No Brasil, o artigo 20 da Lei Geral de Proteção de Dados (LGPD) prevê a possibilidade de revisão da decisão automatizada e seu parágrafo primeiro possibilita a solicitação de informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada e até o momento não há clareza aos controladores de dados pessoais, tampouco aos titulares, de qual seria o melhor formato e detalhamento para cumprimento de aludido direito.

Em sua atuação regulatória sobre decisões automatizadas, a Autoridade Nacional de Proteção de Dados (ANPD) promoveu duas importantes consultas públicas em 2024. Em fevereiro, a Autoridade realizou uma tomada de subsídios visando à elaboração de norma sobre direitos dos

---

[1] ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD). OECD Framework for the Classification of AI Systems: AI Principles. Paris: OECD update in 2024. Disponível em: https://oecd.ai/en/dashboards/ai-principles/P7. Acesso em: 14 jan. 2025

[2] UNIÃO EUROPEIA. Regulamento (UE) 2024/1689, de 13 de junho de 2024. Considerando 10. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L_202401689. Acesso em 14 jan. 2025.

titulares de dados pessoais, buscando, dentre outros aspectos, identificar as melhores práticas no fornecimento de informações sobre decisões automatizadas.

Posteriormente, em novembro, a ANPD aprofundou o tema com consulta pública específica sobre inteligência artificial e revisão de decisões automatizadas. Esta segunda iniciativa foi acompanhada pela Nota Técnica nº 230/2024/CON1/CGN/ANPD, na qual a Autoridade reconheceu a necessidade de abordagem regulatória mais ampla, que combine regulamentação com documentos orientativos e estudos técnicos. Esse posicionamento considera tanto a relevância do tema nos debates entre os agentes de tratamento quanto a vulnerabilidade dos titulares que não possuem conhecimentos técnicos aprofundados sobre o assunto.

Ainda no Brasil, o Projeto de Lei nº 2338/2023 – que possivelmente se tornará o futuro Marco Regulatório da IA –, elenca transparência e explicabilidade como princípio legal, além de dispor sobre providências de governança com vistas à transparência.

Contudo, apesar do recente avanço regulatório com a aprovação do AI Act e do amadurecimento dos debates no Brasil, persiste significativa lacuna entre as expectativas normativas e éticas, sua efetiva implementação prática e a produção dos efeitos esperados. Assim, nasce a seguinte questão central de pesquisa: Como estruturar um framework com requisitos que permita aos agentes de IA implementarem, de forma sistemática, a transparência e explicabilidade de seus sistemas de acordo com os melhores padrões técnicos internacionais e em conformidade com o AI Act, LGPD e PL 2338/2023, que possibilite a efetiva compreensão das pessoas afetadas em relação ao sistema de IA para que tenham condições de reclamar eventuais violações a seus direitos?

Este projeto de pesquisa pretende propor solução para a lacuna existente na aplicação prática da transparência e explicabilidade em sistemas de IA. Por meio de investigação teórica, baseada em pesquisa bibliográfica, regulatória e legislativa, serão identificados os parâmetros comuns para a implementação da explicabilidade e transparência, respondendo às seguintes perguntas fundamentais: (i) quais informações devem ser registradas ao longo de cada etapa do ciclo de vida da IA? (ii) qual a melhor forma de documentá-las? (iii) quem são os responsáveis pela documentação? (iv) como equilibrar a necessidade de transparência com a proteção da propriedade intelectual e segredos comerciais? e (v) qual o formato, linguagem, nível de detalhamento e aspectos técnicos ideais para apresentação dessas informações aos usuários impactados, que não possuem literacia em IA?

As respostas a essas questões viabilizarão a criação de um framework pragmático, capaz de guiar os agentes de IA na implementação de práticas de explicabilidade e transparência em todas as fases do ciclo de vida de seus sistemas, diminuindo a assimetria informacional.

2. Quesitos, fontes de pesquisa e formas de acesso

2.1. Quesito 1.

2.1.1. Contextualização fática

- O que é explicabilidade e transparência de sistemas de IA?
- Como se apresenta o cenário atual de regulamentação da explicabilidade e transparência em IA: Mapeamento do AI Act, PL 2338/2023 e artigo 20 da LGPD.

2.1.2. Fontes de pesquisa:

- Documentos oficiais: Texto do regulamento europeu sobre IA, projeto de lei brasileiro sobre IA, pareceres de comissões, texto da LGPD e manifestações da ANPD sobre o tema.
- Vídeos de audiências públicas e sessões do Senado.
- Publicações acadêmicas: Artigos científicos, livros e capítulos de livros que abordam a regulamentação baseada em risco.

2.1.3. Formas de Acesso:

- Bases de dados acadêmicas: Scielo, Scopus.
- TV Senado no Youtube.
- Repositórios de documentos oficiais**:** Portal da União Europeia, Diário Oficial da União do Brasil, Portal da Câmara dos Deputados, Portal do Senado Federal ANPD.

2.2. Quesito 2

2.2.1. Abordagem analítica

- Como os requisitos regulatórios e padrões técnicos internacionais se traduzem em requisitos de implementações práticas?
- Há convergências e divergências entre eles?
- Existe algum grau de hierarquia de relevância em relação às informações que mais impactarão no resultado da transparência e explicabilidade?

2.2.2. Fontes de pesquisa

- Artigos científicos sobre implementação de explicabilidade
- Frameworks existentes
- Padrões técnicos internacionais
- Materiais publicados por autoridades de proteção de dados

2.2.3. Formas de Acesso:

- Bases de dados acadêmicas: Scielo, Scopus.
- Repositórios de documentos oficiais**:** Portal da União Europeia, Diário Oficial da União do Brasil, Portal da Câmara dos Deputados, Portal do Senado Federal ANPD.

2.3. Quesito 3

2.3.1. Conclusão propositiva

- Sistematização de requisitos para materialização da explicabilidade e transparência.
- Proposição de metodologia para aplicação framework

2.3.2. Fontes de pesquisa

- Artigos científicos sobre implementação de explicabilidade
- Frameworks existentes
- Padrões técnicos internacionais
- Materiais publicados por autoridades de proteção de dados

2.3.3. Formas de Acesso:

- Bases de dados acadêmicas: Scielo, Scopus.
- Repositórios de documentos oficiais**:** Portal da União Europeia, Diário Oficial da União do Brasil, Portal da Câmara dos Deputados, Portal do Senado Federal ANPD.

3. Relevância prática, caráter inovador e potencial de impacto

O presente projeto de pesquisa se destaca pela sua significativa relevância prática no contexto atual de crescente regulamentação e adoção de sistemas de IA. Com a recente aprovação do AI Act na União Europeia, a tramitação do PL 2338/2023 no Brasil e as exigências já estabelecidas pela LGPD, organizações enfrentam o desafio complexo de implementar requisitos de explicabilidade e transparência para seus sistemas de IA. Este cenário demanda uma abordagem estruturada que possa traduzir efetivamente as exigências regulatórias e melhores práticas em implementações técnicas verificáveis.

A relevância prática da pesquisa se manifesta em múltiplas dimensões. Para as organizações, o framework proposto oferecerá um guia sistemático para implementação da explicabilidade e transparência, reduzindo significativamente a insegurança jurídica e otimizando a alocação de recursos no processo de conformidade regulatória. Os desenvolvedores de sistemas de IA se beneficiarão da tradução de requisitos legais em providências efetivas. Reguladores e auditores terão à disposição critérios objetivos de avaliação e uma metodologia padronizada para verificação de conformidade, enquanto os usuários finais experimentarão maior transparência e compreensibilidade nas decisões automatizadas que afetam suas vidas.

O caráter inovador da pesquisa reside em sua abordagem integrativa, que busca estabelecer uma ponte metodológica entre os requisitos regulatórios e melhores práticas internacionais e sua implementação técnica. O potencial de impacto da pesquisa pretendido consiste na contribuição aos agentes de IA, na facilitação da implementação de mecanismos para a garantia efetiva da transparência e explicabilidade. Espera-se também, alcançar impacto social na contribuição para efetivação da transparência e explicabilidade em decisões automatizadas e do fortalecimento da proteção efetiva das garantias individuais das pessoas afetadas.

4. Familiaridade com objeto da pesquisa

A afinidade com o tema de transparência e explicabilidade em sistemas de IA deriva de trajetória acadêmica e profissional consistente no campo do direito e tecnologia. O interesse inicial pela intersecção entre direito e IA se materializou no trabalho de conclusão do curso de especialização em Direito e Tecnologia da Informação na Escola Politécnica da Universidade de São Paulo, em 2018, onde foi desenvolvida pesquisa sobre a responsabilidade civil e penal nos acidentes envolvendo carros autônomos. Este trabalho oportunizou a compreensão dos desafios jurídicos relacionados à tomada de decisão automatizada e suas implicações práticas.

Desde então, a busca pelo aprofundamento no campo da IA tem sido constante, com particular atenção aos aspectos de governança e conformidade regulatória. A experiência profissional em escritório especializado em direito digital tem sido fundamental para compreender as necessidades práticas das organizações que desenvolvem e implementam sistemas de IA. A participação direta em projetos de governança de IA têm proporcionado visão privilegiada dos desafios concretos enfrentados pelas organizações na implementação de requisitos regulatórios. Esta combinação de experiência acadêmica e prática oferece perspectiva para o desenvolvimento do framework proposto.

5. Bibliografia preliminar

ARRIETA, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, v. 58, p. 82-115, 2020. https://arxiv.org/abs/1910.10045v2.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Tomada de Subsídios Norma de direitos dos titulares de dados pessoais. 2024.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Tomada de Subsídios Inteligência Artificial e revisão de decisões automatizadas. 2024. Disponível em: https://www.gov.br/participamaisbrasil/tomada-de-subsidios-inteligencia-artificial-e-revisao-de-decisoes-automatizadas.

BHATT, U. et al. Explainable Machine Learning in Deployment. In: Conference on Fairness, Accountability, and Transparency, 2020. Disponível em: https://arxiv.org/abs/1909.06342.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

SENADO FEDERAL. Projeto de Lei nº 2338/2023. 2024. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=9858510&ts=1735605226782&rendition_principal=S&disposition=inline.

FELZMANN, H. et al. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, v. 7, n. 1, 2019. https://doi.org/10.1177/2053951719860542.

GILPIN, L. H. et al. Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018. Disponível em: https://arxiv.org/abs/1806.00069.

GOOGLE AI RESEARCH TEAM. What-If Tool: An Interactive Visual Interface for Model Understanding. Technical Documentation, 2024. Disponível em: https://pair-code.github.io/what-if-tool.

INFORMATION COMMISSIONER'S OFFICE (ICO). Explaining decisions made with AI. Londres: ICO, 2022. Disponível em: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS (IEEE). IEEE 7001-2021 - IEEE Standard for Transparency of Autonomous Systems. New York: IEEE, 2021.

NTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system requirements. Geneva: ISO, 2023

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION. ISO/IEC 23894:2023 Information technology — Artificial intelligence — Risk management. Geneva: ISO, 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION. ISO/IEC 38507:2023 Information technology — Governance implications of the use of artificial intelligence by organizations. Geneva: ISO, 2023.

MICROSOFT RESEARCH. InterpretML: A Unified Framework for Machine Learning Interpretability. Technical Report. 2024. Disponível em: https://interpret.ml/docs/.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019. https://doi.org/10.1016/j.artint.2018.07.007

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). AI Risk Management Framework (AI RMF 1.0). 2024. Disponível em: https://www.nist.gov/itl/ai-risk-management-framework

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. 2017. Disponível em: https://academic.oup.com/idpl/article-abstract/7/2/76/3860948?redirectedFrom=fulltext

6. **Cronograma de execução**

	2025											
Atividade	1	2	3	4	5	6	7	8	9	10	11	12
Revisão bibliográfica												
Desenvolvimento do framework												
Redação da dissertação												
Redação final e submissão												