PRIVACIDADE POR DESIGN NO DESENVOLVIMENTO DE INTELIGÊNCIAS ARTIFICIAIS: TRANSPARÊNCIA E PRIVACIDADE DIANTE DO DESAFIO DA CAIXA PRETA

PRIVACY BY DESIGN IN THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE: TRANSPARENCY AND PRIVACY IN LIGHT OF THE BLACK BOX CHALLENGE

Deborah Ferreira Maia[1], Luís Felipe Schons Silva[2], Lacordaire Kemel Pimenta Cury[3]

Data de submissão: 05/02/2026
Data de aprovação: 19/02/2026

RESUMO

A expressiva expansão das inteligências artificiais (IA) e a quantidade massiva de dados gerados todos os dias traz alguns problemas. Dentre estes, o problema da "caixa-preta", gerado pela opacidade algorítmica e a falta de compreensão da lógica dos sistemas inteligentes. Esta pesquisa tem como objetivo analisar como os conceitos do *Privacy by Design* (PbD) podem orientar os desenvolvedores de algoritmos inteligentes a garantir transparência e privacidade dos dados desde a concepção do projeto. Este trabalho realizou uma pesquisa exploratória-descritiva qualitativa, analisando 25 documentos da literatura científica relevantes, com objetivo de identificar conexões entre o *Privacy By Design* e a transparência nas IAs. Os resultados indicam que os sete conceitos do PbD podem nortear o desenvolvimento desses algoritmos para que a transparência seja uma funcionalidade intrínseca ao sistema de forma positiva, além do uso de ferramentas como a *explainable* AI (XAI) e a LIME para transformar sistemas opacos em algoritmos explicáveis, tanto para os usuários quanto para seus desenvolvedores. Tudo isso permite reduzir a vulnerabilidade e riscos éticos, estabelecendo confiança na era digital.
Palavras-chave: *Privacy by design*. Caixa-Preta. Inteligência Artificial.

ABSTRACT

The significant expansion of Artificial Intelligence (AI) and the massive volume of data generated daily present certain challenges. Prominent among these is the "black box" problem, resulting from algorithmic opacity and a lack of understanding regarding the logic of intelligent systems. This research aims to analyze how Privacy by Design (PbD) concepts can guide developers of intelligent algorithms (AI) to ensure transparency and data privacy from the project's inception. This study conducted qualitative exploratory-descriptive research, analyzing 25 relevant documents from scientific literature to identify connections between Privacy by Design and transparency in AI. The results indicate that the seven principles of PbD can guide the development of these algorithms so that transparency becomes a positive, intrinsic functionality of the system. Furthermore, the use of tools such as Explainable AI (XAI) and LIME can transform opaque systems into explainable algorithms for both users and developers. Ultimately, this allows for the reduction of vulnerabilities and ethical risks, establishing trust in the digital age.

***Keywords*:** *Privacy by Design. Black Box. Artificial Intelligence.*

1 Email: deborahmaia48@gmail.com
2 Email: luis.schons.silva@gmail.com
3 Email: lacordaire.cury@ifgoiano.edu.br

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

Maia, Deborah
217 PRIVACIDADE POR DESIGN NO DESENVOLVIMENTO DE INTELIGÊNCIAS ARTIFICIAIS: TRANSPARÊNCIA E PRIVACIDADE DIANTE DO DESAFIO DA CAIXA PRETA / Deborah Maia. Catalão 2026.

21f. il.

Orientador: Prof. Dr. Lacordaire Kemel Pimenta Cury.
Coorientador: Prof. Me. Luís Felipe Schons Silva.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0920203 - Bacharelado em Sistemas de Informação - Catalão - Noturno (Campus Catalão).
1. Privacy by design. 2. Caixa-Preta. 3. Inteligência Artificial. I. Título.

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

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

Identificação da Produção Técnico-Científica (assinale com X)

[ ] Tese

[ ] Dissertação

[ ] Monografia – Especialização

[ ] Artigo - Especialização

[X] TCC - Graduação

[ ] Artigo Científico

[ ] Capítulo de Livro

[ ] Livro

[ ] Trabalho Apresentado em Evento

[ ] Produção técnica. Qual: _____

Nome Completo da Autora: Deborah Ferreira Maia

Matrícula: 2022109202030029

Título do Trabalho: **PRIVACIDADE POR DESIGN NO DESENVOLVIMENTO DE INTELIGÊNCIAS ARTIFICIAIS: TRANSPARÊNCIA E PRIVACIDADE DIANTE DO DESAFIO DA CAIXA PRETA**

Restrições de Acesso ao Documento [Preenchimento obrigatório]

Documento confidencial: [X] Não [ ] Sim, justifique: _____

Informe a data que poderá ser disponibilizado no RIIF Goiano: 20/02/2026

O documento está sujeito a registro de patente? [ ] Sim [X] Não

O documento pode vir a ser publicado como livro? [X] Sim [ ] Não

DECLARAÇAO DE DISTRIBUIÇAO NAO-EXCLUSIVA

O/A referido/a autor/a declara que:

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

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

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

Cidade, 20 de fevereiro de 2026

Nome do Autor

Assinado eletronicamente pelo o Autor e/ou Detentor dos Direitos Autorais

Ciente e de acordo:

Nome do(a) orientador(a)

Assinatura eletrônica do(a) orientador(a)

Documento assinado eletronicamente por:

- **Lacordaire Kemel Pimenta Cury**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 20/02/2026 08:27:05.
- **Deborah Ferreira Maia**, **2022109202030029 - Discente**, em 20/02/2026 08:47:47.

Este documento foi emitido pelo SUAP em 20/02/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 790422
Código de Autenticação: 1eba5d9e6f

INSTITUTO FEDERAL GOIANO
Campus Catalão
Rua Salustiano Oliveira da Paz, 1621, Bairro Ipanema, CATALAO / GO, CEP 75.705-040
(64)99212-9907

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

Ata nº 15/2026 - CC-CAT/GE-CAT/CMPCAT/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 19 dia(s) do mês de fevereiro de 2026, às 16:00 horas, reuniu-se a banca examinadora composta pelos docentes: Lacordaire Kemel Pimenta Cury (orientador), Daniel Hilário da Silva (membro) e Eduardo Castilho Rosa (membro), para examinar o Trabalho de Curso intitulado PRIVACIDADE POR DESIGN NO DESENVOLVIMENTO DE INTELIGÊNCIAS ARTIFICIAIS: TRANSPARÊNCIA E PRIVACIDADE DIANTE DO DESAFIO DA CAIXA PRETA da estudante Deborah Ferreira Maia, Matrícula nº 2022109202030029 do Curso de Sistemas de Informação do IF Goiano – Campus Catalão. A palavra foi concedida a estudante para a apresentação oral do TC, houve arguição do candidato pelos membros da banca examinadora. Após tal etapa, a banca examinadora decidiu pela APROVAÇÃO do estudante com a nota de 8,7 pontos. Ao final da sessão pública de defesa foi lavrada a presente ata que segue assinada pelos membros da Banca Examinadora.

Observação:

( ) O(a) estudante não compareceu à defesa do TC.

Documento assinado eletronicamente por:

- **Lacordaire Kemel Pimenta Cury**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 19/02/2026 17:26:50.
- **Eduardo Castilho Rosa**, **COORDENADOR(A) DE CURSO - FUC1 - CCBSI-CAT** , em 19/02/2026 17:30:01.
- **Daniel Hilario da Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 19/02/2026 17:30:04.

Este documento foi emitido pelo SUAP em 19/02/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 790116
Código de Autenticação: 9ec8982f32

INSTITUTO FEDERAL GOIANO
Campus Catalão
Rua Salustiano Oliveira da Paz, 1621, Bairro Ipanema, CATALAO / GO, CEP 75.705-040
(64)99212-9907

1 INTRODUÇÃO

A expressiva expansão das tecnologias digitais possibilitou o desenvolvimento da Inteligência Artificial (IA). A IA é um ramo da ciência que programa os computadores para que sejam inteligentes no modo de agir e pensar. Esta área da ciência deve ser entendida como um campo do conhecimento com múltiplas vertentes (ALVES e ANDRADE, 2022). A expansão e a popularização de modelos computacionais, entre eles os modelos de linguagem generativa *Large Language Models* (LLMs), modelos focados em entender e gerar linguagem humana, a exemplo o ChatGPT e o Gemini, geram diariamente enormes quantidades de dados. Esses dados, muitas vezes sensíveis, podem abranger laudos médicos, processos jurídicos, contratos de trabalho e resultados de pesquisa nos quais a IA é aplicada para identificar padrões e dar previsões precisas (BRASIL, 2025).

A busca pela "Inteligência não humana" é antiga, tendo seu início na época em que filósofos gregos e estudiosos da área já discutiam sobre o tema (RUSSELL; NORVIG 2016). Contudo, os primeiros artigos surgiram apenas em 1943, quando Warren McCulloch e Walter Pitts teorizaram um modelo de neurônio. Outro artigo, publicado em 1950, foi escrito por Marvin Minsky e Dean Edmonds, enquanto alunos de Harvard, no qual criaram o primeiro computador com rede neural (RUSSELL; NORVIG, 2016). No entanto, o principal criador do conceito da IA foi Alan Turing, com seu artigo "*Computing Machinery and Intelligence*" de 1950, onde apresentou as primeiras ideias sobre aprendizagem de máquina, aprendizagem por reforço e algoritmos genéticos por meio do teste de Turing.

O desenvolvimento e a implementação bem-sucedida dessas teorias estudadas na década de 1950, resultaram na popularização das IAs em escala global. Com essa popularização surgiram novos problemas: a opacidade. Conforme define Burrell (2016), podemos observar esse fenômeno de três formas diferentes. Primeiramente, a proteção empresarial, para assegurar o segredo de seus algoritmos; a falta de compreensão total de programação, mesmo que o código seja público sua compreensão é complexa ou inexistente e a diferença de pensamento entre os humanos e as máquinas.

A falta de transparência gerada pela opacidade dos algoritmos não permite que o usuário comum entenda como seus dados são processados. Alves e Andrade (2022) demonstra que alguns tipos de algoritmos, principalmente os de aprendizado profundo, podem ser um grande mistério para os usuários. O cenário intensifica preocupações relacionadas à privacidade de dados e à transparência dos algoritmos, sendo essas as maiores preocupações (DIAS; DIAS; DE PAULA, 2025). Essa questão remete ao problema da caixa-preta, criado pela opacidade dos algoritmos das IAs. Torna-se complexa a compreensão da relação entre os dados de entrada (inputs) e os dados de saída (outputs), o que acaba

comprometendo a confiança nos resultados apresentados, que podem estar enviesados ou conter falhas. A transparência perante agentes inteligentes é uma condição necessária para que possam explicar e argumentar todas as decisões tomadas, (SICHMAN, 2021).

Tentando contornar esse desafio e garantir a segurança e privacidade dos dados, a pesquisadora Ann Cavoukian propôs o conceito de *Privacy by Design* (PbD). O PbD consiste no desenvolvimento de *software* seguindo sete princípios fundamentais, desde o início do desenvolvimento. Estes princípios, descritos no artigo de Cavoukian (2009), funcionam como guia para integrar a proteção de dados já na concepção do projeto, sendo uma alternativa de desenvolvimento relevante num contexto em que a privacidade e a transparência de dados são temas alarmantes.

Os princípios do PbD incluem ser proativo e não reativo, preventivo e não corretivo; adotar a privacidade como configuração padrão; incorporar a privacidade ao design; e buscar sempre a funcionalidade completa, entendida como soma positiva, e não de soma zero. Este trabalho tem como objetivo analisar como os conceitos do PbD podem ser uma alternativa para os desenvolvedores de IAs orientando os algoritmos à transparência e privacidade de seus usuários.

Para alcançar o objetivo proposto, este artigo está organizado da seguinte maneira: a Seção 2 detalha a metodologia aplicada. Na Seção 3, desenvolve-se o conteúdo pesquisado, delineando inicialmente a evolução histórica e em seguida, o texto aprofunda-se nos conceitos de privacidade e opacidade, bem como no problema da caixa-preta, finalizando com a discussão de como os sete princípios do PbD podem ser aplicados. Por fim, a Seção 4 apresenta as considerações finais.

2 METODOLOGIA

Este estudo adotou uma abordagem de pesquisa exploratória-descritiva qualitativa, reconhecida por sua contribuição singular para a compreensão aprofundada de fenômenos complexos. A vertente exploratória possibilita a aproximação inicial do problema proposto, que envolveu a relação entre o conceito de PbD e a transparência em sistemas de inteligência artificial. Essa etapa possibilitou levantar informações preliminares, reconhecer padrões e construir uma base teórica sólida para análises mais aprofundadas. A dimensão descritiva complementou essa estratégia ao organizar e detalhar informações sobre o fenômeno estudado de forma sistemática, permitindo uma compreensão mais ampla e consistente.

Os estudos qualitativos permitem investigar significados, motivações e implicações que extrapolam dados estatísticos ou gráficos, favorecendo uma interpretação mais contextualizada e

próxima das dinâmicas reais envolvidas (CORDEIRO *et al.*, 2023). Essa característica tornou a abordagem especialmente adequada para compreender conceitos, princípios e impactos relacionados ao desenvolvimento e aplicação de tecnologias emergentes.

A coleta de informações ocorreu por meio da leitura e análise crítica da literatura científica, com um total de 25 documentos analisados. A seleção dos trabalhos considerou critérios metodológicos que levaram em conta aspectos como relevância temática e consistência teórica.

Utilizou-se como fonte de busca inteligente a plataforma Perplexity.AI, um mecanismo de busca na web em tempo real, auxiliando na busca de fontes de conteúdo para a pesquisa de forma automatizada. Complementarmente, foi utilizado as bases de conhecimento Google Acadêmico e SciELO, sendo utilizadas para a busca de trabalhos científicos e para a conferência e validação dos artigos retornados pelo agente inteligente. Essas bases foram selecionadas devido ao vasto volume de documentos científicos disponíveis em ambas, utilizando como palavras-chave para a pesquisa: 'Inteligência Artificial', 'Privacidade de Dados', '*Privacy by Design*', 'Caixa-Preta' e 'Transparência de Dados'. As palavras-chave foram definidas por representarem os eixos centrais da pesquisa e garantirem o retorno de publicações com sólida fundamentação teórica. A quantidade de estudos analisados foi definida ao longo do desenvolvimento da pesquisa, garantindo profundidade analítica e representatividade conceitual. Essa etapa possibilitou compreender como diferentes autores e instituições discutiram a relação entre privacidade, transparência e inteligência artificial.

Foi utilizado na ferramenta Perplexity.AI parâmetros de entrada para a busca de conteúdo correlacionados com as palavras-chave que atendessem aos seguintes requisitos de validade: (a) publicações em periódicos científicos, eventos acadêmicos ou fontes institucionais reconhecidas; e (b) exclusão de conteúdos informais postados apenas em páginas comuns da web. Posteriormente, aplicaram-se critérios de inclusão para o refinamento temático do material, sendo eles: (a) apresentação de conceitos e definições claras sobre IA, opacidade algorítmica, transparência e *Privacy by Design*; (b) abordagem específica dos fundamentos e da aplicação prática dos princípios do *Privacy by Design*; e (c) relação direta entre as IAs e a transparência perante os usuários. Como critério de exclusão, foram desconsiderados trabalhos que, mesmo abordando a temática de IA ou privacidade de dados, não estabeleciam correlação direta com o problema central da pesquisa, como os sistemas caixa-preta ou o PbD.

O recorte temporal para a coleta foi entre os anos de 2009 e 2025, mas a pesquisa se concentrou principalmente em trabalhos mais recentes, entre os anos de 2020 e 2025, para garantir alinhamento com as inovações crescentes da área de pesquisa. Contudo, algumas exceções a esse intervalo foram admitidas, com o intuito de incluir obras clássicas publicadas em anos anteriores (2009, 2016 e 2017).

Essa flexibilização justificou-se pela necessidade de embasar conceitos apresentados durante este artigo, como o próprio PbD.

O estudo abordou, de forma integrada, três dimensões fundamentais, a evolução histórica e conceitual da inteligência artificial e sua relação com o tratamento de dados, o problema da opacidade algorítmica e suas implicações para a explicabilidade dos modelos e os princípios do PbD formulados por Ann Cavoukian, com ênfase na incorporação da privacidade desde a concepção dos sistemas. Essa estrutura analítica permitiu examinar como as ideias de transparência e privacidade foram articuladas na literatura e de que forma podem ser aplicadas no desenho de tecnologias mais confiáveis e auditáveis.

A análise identificou conexões conceituais entre os princípios do PbD e os requisitos de transparência e explicabilidade em sistemas de IA, investigando como a adoção dessas diretrizes pode contribuir para mitigar riscos relacionados à privacidade e, ao mesmo tempo, favorecer o desenvolvimento de projetos mais transparentes. O objetivo da pesquisa foi demonstrar que a incorporação dos princípios do PbD desde as etapas iniciais de concepção de sistemas inteligentes não apenas reduz vulnerabilidades, como também estabelece condições que fortalecem a confiança social e institucional nos resultados produzidos por essas tecnologias.

3 DESENVOLVIMENTO

3.1 A história do privacy by design

A história do direito à privacidade teve seu início nos Estados Unidos, em 1890, quando Samuel Warren e Louis Brandeis desenvolveram o artigo intitulado *The Right to Privacy*. Este texto foi o responsável por apresentar o conceito de Privacy pela primeira vez (MARRAFON; COUTINHO, 2020). O artigo discute as características, limitações e funcionalidades desse direito, que passou a focar primariamente na tutela da personalidade humana. Houve um afastamento do antigo conceito de matriz proprietária, que protegia somente as propriedades da pessoa, gerando uma nova preocupação central: a segurança do indivíduo. Essa segurança foi denominada pelo Juiz Cooley como “o direito de estar só”, segundo a explicação de Cancelier (2017).

A partir desse marco fundacional, surgiram novas discussões e preocupações relativas à privacidade em escala internacional. Nações começaram a se preocupar com o direito à vida privada de seus moradores, que se manifestou em documentos internacionais importantes como a Declaração Universal dos Direitos Humanos das Nações Unidas de 1948, pautando que ninguém deve estar

sujeito a interferências em sua vida pessoal. Além disso, o direito à vida privada é tratado na Carta dos Direitos Fundamentais da União Europeia, conforme explicam Melo, Rockembach e Silva (2023).

Contudo, com a aceleração do crescimento da tecnologia, a definição de privacidade, criada como "o direito de ser deixado só", começa a se tornar ineficiente com o passar dos anos. O cenário muda a partir da década de 1960, com o grande desenvolvimento das tecnologias, coleta de dados e a circulação dessas informações, o que acarretou o aumento da capacidade de coletar, processar e utilizar esses dados, mudando a perspectiva sobre privacidade (CANCELIER, 2017). A partir disso o debate a respeito desse direito precisou ser ampliado, passando do pensamento de "estar só" para a ideia de como controlar o uso das informações pessoais de forma correta e segura, conforme explica Genso *et al.* (2023).

O conceito de PbD começou a ser discutido em meados da década de 90 pela Comissária de Informação e Privacidade de Ontário, Ann Cavoukian (GENSO *et al.*, 2023). Apesar disso, o conceito só foi apresentado formalmente em 2009. O conceito foi firmado internacionalmente na 32th *International Conference of Data Protection and Privacy Commissiones* de 2010, evento no qual a *Resolution on Privacy By Design* foi adotada (MELO, ROCKEMBACH E SILVA, 2023). Esta resolução representa um marco para a privacidade de dados, por ter sido considerada um componente essencial na proteção dos dados e da privacidade dos usuários. Essa resolução surge com o avanço da tecnologia e a necessidade de pensar nos novos desafios causados pelo Big Data e a segurança dos dados. Esses fatores geraram uma nova necessidade de controle dos dados e cumprimento da lei frente aos direitos dos cidadãos e o direito à privacidade de dados (MELO, ROCKEMBACH e SILVA, 2023).

A partir disso, a autora Cavoukian (2009) demonstra que o valor da informação e a necessidade de gerenciá-la continuou e continua crescendo, permitindo a liberdade de escolha sobre o que acontece com seus próprios dados e o controle pessoal dessas informações, definindo o conceito de *design-thinking*.

Os princípios do PbD possuem como objetivo definir a privacidade como um padrão já estabelecido nos sistemas, sem a necessidade de correção futura, embutido em toda a operação e planejamento do software (Cavoukian, 2009). O intuito da autora ao publicar a pesquisa foi estabelecer um quadro universal para a proteção e privacidade de dados na era moderna, apresentando vantagens que vão desde a mitigação de riscos relacionados a privacidade até a resolução de problemas como a caixa-preta, impulsionando a transparência no tratamento de dados pessoais (SILVA E DOMINGUES, 2024).

3.2 A inteligência artificial

O panorama da história da Inteligência Artificial é delineado por alguns marcos temporais, sendo o principal deles a publicação do artigo *Computing Machinery and Intelligence* pelo matemático Alan Turing em 1950. Este estudo é um dos marcos fundadores da IA, onde ele propõe saber se um computador conseguiria se passar por um ser inteligente dentro do jogo da imitação, esse conceito se tornou conhecido como o "Teste de Turing".

Apesar do termo "Inteligência Artificial" não ser utilizado por Alan Turing, em 1956 ele aparece pela primeira vez, durante um "*Summer Camp*" proposto por John McCarthy, professor e pesquisador de matemática da Faculdade de Dartmouth. Esse encontro tinha como objetivo juntar dez pesquisadores para que se dedicassem a diferentes problemas de pesquisa durante dois meses. Ao solicitar apoio financeiro, McCarthy utilizou a expressão "Inteligência Artificial" na documentação enviada à Fundação Rockefeller, assim o encontro passou a ser visto como fundador desta área de pesquisa (SOBREIRA, 2025).

O avanço da IA como área de pesquisa, impulsionada inicialmente pela estratégia de abordar problemas específicos, trouxe críticas sobre suas premissas e expectativas. Uma dessas, realizada pelo filósofo Hubert Dreyfus em 1965 em seu trabalho "Inteligência Artificial e Alquimia", questiona a possibilidade de recriar a inteligência humana em códigos e emular a sua capacidade de raciocínio em uma máquina, de forma a desaconselhar o investimento financeiro na área. A segunda crítica surgiu em 1972, divulgada por um relatório escrito pelo matemático inglês James Lighthill, que usou um problema matemático para mostrar que sistemas baseados em regras e possibilidades funcionam somente em ambientes com poucas variáveis, como explicado por Sobreira (2025).

Essas críticas e debates oscilaram o interesse de pesquisadores pela área. Durante esse período, ocorreu o "*AI Winter*", compreendido entre os anos de 1975 e de 1980 e 1987 a 1993, um momento em que não havia interesse em investir recursos financeiros em pesquisas da área de IA (SICHMAN, 2021). Além da falta de recursos financeiros, a falta de recursos computacionais limitava as pesquisas e impedia a realização de testes essenciais (LUDERMIR, 2021). Entretanto, os avanços da tecnologia em *Hardware* reacenderam as pesquisas na área, onde a redução do custo de processamento e memória, permitiu surgir novos paradigmas como as redes neurais e uma enorme quantidade de dados diariamente (SICHMAN, 2021).

Por meio desses estudos, muitas tecnologias se tornaram realidade, por exemplo carros autônomos que podem dirigir sozinhos como os da Google ou da Tesla e robôs, por exemplo o da iFlytek, que foi capaz de passar no teste nacional de médicos da China, analisando pacientes e dando

seus diagnósticos iniciais (LUDERMIR, 2021). Além desses exemplos, temos o reconhecimento de imagens, onde câmeras são capazes de detectar comportamentos humanos suspeitos ou os modelos de linguagens LLMs, que em sua maioria é acessível e gratuito para ser utilizado.

Com o objetivo de investigar o uso das IAs pela população brasileira e as percepções dos usuários, no ano de 2025 a Fundação Itaú em conjunto com o Instituto de Pesquisas Data Folha realizou uma pesquisa sobre o uso da inteligência artificial, entrevistando 2.798 pessoas com mais de 16 anos e de todas as classes sociais. A pesquisa concluiu que 82% dos entrevistados já ouviram falar sobre o assunto, mas apenas 54% dizem entender o termo e 93% utilizam alguma ferramenta de IA no seu cotidiano.

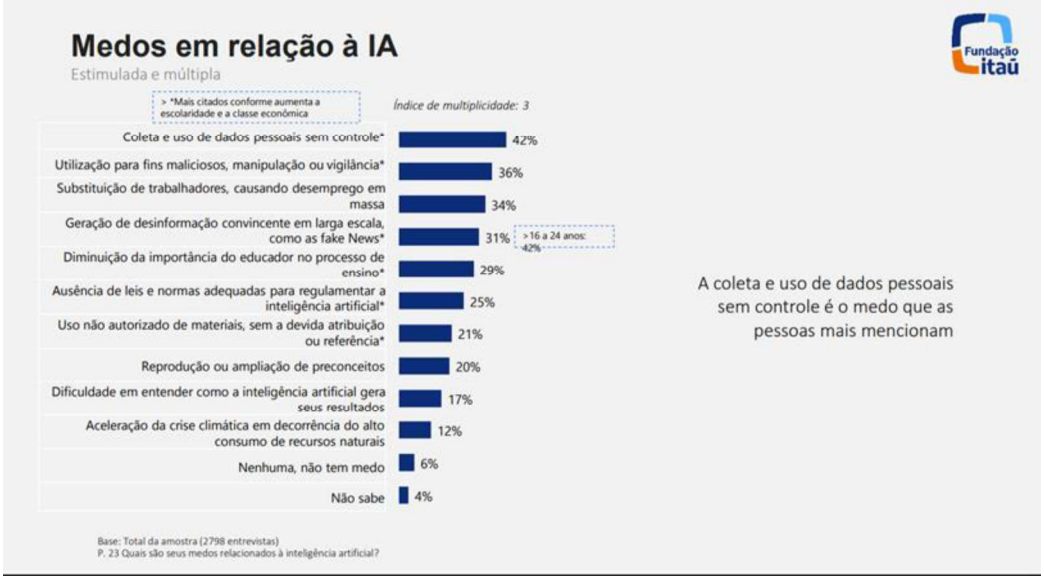

Figura 1 – Gráfico Medos em relação à IA.

Fonte: Fundação Itaú (2025)

Com base na Figura 01, observa-se um grande medo relacionado à coleta e ao uso dos dados de forma indevida e sem controle e a falta de compreensão em como os algoritmos geram os resultados. Neste sentido pode-se analisar que a população utiliza em larga escala as ferramentas mesmo não tendo o total conhecimento do que se trata, mas com receios em relação a coleta de dados e como os resultados foram gerados.

Neste ponto de vista, essas inovações estão transformando por completo o cotidiano das pessoas. Atividades que, para os seres humanos podem ser massivas e rotineiras, para os algoritmos inteligentes são de fácil execução. A realização destas atividades simples pelas IAs pode reduzir erros e até mesmo evitar acidentes de trabalho em áreas de alto risco, exemplificado pelo desarme de bombas ou explosões em minas (Barroso e Mello, 2024). Além disso, os autores ainda afirmam que

as tecnologias inteligentes possuem uma capacidade que nenhum ser humano possui: trabalhar 24 horas por dia, todos os dias, sem se cansar ou adoecer, aumentando significativamente o nível de produção das empresas. Os métodos inteligentes detêm uma capacidade de tomada de decisão muito mais elevada que a de um humano, principalmente por armazenar e analisar milhares de informações em minutos, com correlações e análises desses dados. No entanto, elas ainda enfrentam o problema das “alucinações”, que se trata um desvio onde a ferramenta gera informações absurdas ou as inventa, como explica Barroso e Mello (2024). Portanto, a inteligência artificial está crescentemente se incorporando à rotina de milhões de vidas, desde as tarefas mais simples até as aplicações mais complexas como atendimentos médicos e jurídicos.

3.3 Privacidade e transparência de dados

A privacidade de dados originou-se da necessidade de "ser deixado só", assegurando que o ser humano possa usufruir de liberdade, privacidade e escolha sobre o tratamento das suas informações. Diante da transformação do mundo digital, a economia global foi completamente transformada e os dados pessoais se tornaram um dos bens mais valiosos no século XXI (CANGUSSU *et al.*, 2025). Nessas circunstâncias, tem se tornado cada vez mais complexo para o usuário manter o total controle de seus dados, pois a falta de transparência impede que o ele entenda de fato o que está acontecendo durante o processamento das informações.

O cenário de crescente valorização de dados e a alta complexidade das IAs de coletar, armazenar e analisar grandes quantidades de dados, levanta algumas preocupações que englobam a maneira como essas informações estão sendo utilizadas pelas empresas, o modo como esses dados são processados e, principalmente, como as organizações administram a privacidade desses dados, conforme detalhado por Silva e Domingues (2024). Neste sentido, a Lei Geral de Proteção de Dados Pessoais (LGPD) foi sancionada no Brasil em agosto de 2018, mas entrou em vigor aos poucos ao longo dos anos. Criada com o intuito de garantir os direitos fundamentais da liberdade e privacidade, conforme Art. 1° da LGPD (BRASIL, 2018), a lei é inspirada no Regulamento Geral sobre a Proteção de Dados da União Europeia (GDPR) (CANGUSSU *et al.*, 2025).

A LGPD tem como finalidade garantir o respeito a direitos fundamentais e fomentar o desenvolvimento socioeconômico. Seus objetivos primários incluem assegurar a liberdade, a liberdade de expressão, a formação de opinião e a fundamental autodeterminação informativa dos indivíduos. A legislação busca também impedir a violação da imagem, intimidade e honra, e juntamente ao incentivo do desenvolvimento econômico, tecnológico e inovações. Tais diretrizes são

balanceadas com a garantia da livre iniciativa, a defesa do consumidor, os direitos humanos e o livre exercício da cidadania.

O Artigo 6º da LGPD (BRASIL, 2018) estabelece princípios rigorosos para o tratamento de dados com o intuito de concretizar seus propósitos, destacando-se notadamente o inciso VI, que exige a transparência. Este princípio obriga os agentes de tratamento a disponibilizarem informações claras e de fácil acesso aos titulares, detalhando como os dados estão sendo tratados e quem são os agentes envolvidos. É crucial ressaltar que essa transparência deve ser aplicada, contanto que não revele segredos de negócios das empresas responsáveis pelas plataformas. Além da transparência, o tratamento de dados deve englobar outras garantias fundamentais, como finalidade, adequação, necessidade, livre acesso, segurança, prevenção, não discriminação e a responsabilização.

De acordo com Michener e Bersch (2013), a transparência de dados pode ser definida por dois pilares, a visibilidade da informação e sua inferibilidade. Para a informação ser visível, é essencial que ela seja encontrada com facilidade, caso a informação não seja visível, o conceito de transparência perde o sentido. Além disso, a informação deve ser encontrada por completo, do contrário o nível de visibilidade será baixo, representando a realidade de forma infiel. Os autores também complementam o conceito de visibilidade com a probabilidade de a informação ser encontrada, uma informação pode ser visível, mas nem sempre é de fácil acesso. Desse modo, as informações precisam estar completas, visíveis e de fácil acesso ao seu detentor, sem que este necessite buscá-las ativamente.

O segundo pilar, também proposto por Michener e Bersch (2013), é a inferibilidade. Este termo corresponde ao nível em que a informação permite obter conclusões completas sobre os dados, ou seja, a facilidade em deduzir inferências precisas a partir dessas informações. Esse pilar está ligado diretamente a qualidade dos dados, se eles não são precisos as conclusões provavelmente estarão incorretas ou apresentarão algum tipo de viés. Para aumentar o nível de inferibilidade, os autores apontam três atributos: desagregação, verificabilidade e simplificação. A desagregação refere-se a informações que não sofreram adulteração, mantendo-se o mais próximo possível de sua fonte original. A verificabilidade implica que dados verificados são mais confiáveis e mais fiéis à realidade, gerando inferências mais seguras. Por fim, a simplificação desses dados é a forma de torná-los úteis e de fácil entendimento para os usuários, mesmo que sejam complexos e muito detalhados.

Um dos principais problemas relacionados a inteligência artificial, segundo Fama (2025), é a falta de transparência no processo de tomada de decisão. Isso ocorre porque os algoritmos utilizados são totalmente opacos, dificultando a compreensão do seu raciocínio até mesmo para o desenvolvedor. A transparência nos algoritmos das IAs, de acordo com o autor citado acima, exige que o sistema seja capaz de explicar suas decisões e como chegou a esse resultado, sendo de extrema

importância para o usuário conseguir compreender os passos que o algoritmo utilizou para chegar nessa conclusão. Contudo, a dificuldade em entender por completo como o raciocínio do algoritmo funciona ou qual o caminho percorrido para chegar àquela conclusão torna o sistema completamente opaco para muitos especialistas da área.

De acordo com Burrell (2016), os algoritmos são opacos quando o indivíduo recebe um resultado, mas não possui uma noção clara de como aquela decisão foi tomada ou o porquê daquele resultado ter sido gerado a partir das entradas especificadas. O autor ainda afirma que um computador aprende e constrói sua própria forma de decisão, executando-a sem levar em consideração o entendimento humano. Este fenômeno gera o desafio atual chamado de caixa-preta.

3.4 O problema da caixa-preta

Os autores Guidotti *et al*. (2019) definem a caixa-preta como um modelo obscuro, cujos processos internos são totalmente desconhecidos para quem utiliza. Apesar da arquitetura de alguns sistemas ser conhecida, o seu raciocínio não é interpretável para os seres humanos, se tornando um sistema opaco, que oculta sua lógica interna, mesmo tendo o código fonte conhecido.

A principal característica dessa opacidade é que os mecanismos internos do modelo são desconhecidos para o observador ou são conhecidos, mas ininterpretáveis por humanos, o que representa uma questão crucial tanto prática quanto ética. Entregar decisões a serem tomadas a sistemas de caixa-preta, sem nenhuma possibilidade de interpretação da tomada de decisão pode gerar graves problemas como a discriminação ou problemas de confiança (GUIDOTTI *et al.,* 2019).

O polêmico caso da Amazon em 2014 representa o risco da caixa preta. Nesse caso, a empresa desenvolveu uma inteligência artificial para a seleção dos melhores candidatos as suas vagas, como citado por Borges e Filó (2021). O sistema de IA funcionava processando aproximadamente 50 mil termos presentes nos resumos de candidatos dos 10 anos anteriores, utilizando esses dados históricos para avaliar e classificar os novos candidatos em uma escala de 1 a 5 estrelas. Mas com o passar do tempo a empresa observou que cargos mais técnicos, como desenvolvimento de software, estavam sendo ocupados somente por homens, escancarando a imparcialidade da ferramenta perante candidatas do sexo feminino. Os dados antigos, que foram utilizados para o treinamento da ferramenta, levaram a mesma a discriminar candidatos por seu sexo. O evento leva a crer que este comportamento já estava presente na empresa antes, mas ainda não havia sido observado. A partir disso, em 2015 a empresa reconheceu seu erro e veio a público esclarecer a discriminação de gênero e retirar a ferramenta de funcionamento.

O caso da Amazon ilustra de maneira crucial a complexidade dos vieses algorítmicos em processos de recrutamento. Observa-se que, mesmo quando o sistema é programado de forma aparentemente correta, os desenvolvedores enfrentam dificuldades significativas em discernir o mecanismo pelo qual a Inteligência Artificial chega à conclusão. Essa dificuldade em entender como o sistema tomava sua decisão, especialmente na classificação de certas candidatas como fracas, caracteriza a chamada "caixa-preta" do aprendizado de máquina. A opacidade significava que o processo decisório do algoritmo ignorava fatores que deveriam ser prioritários, como o conhecimento ou a experiência, resultando na atribuição de notas baixas. Esse caso gerou e perpetuou a discriminação de gênero, reforçando um problema social persistente que tem sido observado na sociedade por milhares de anos.

Perante a opacidade dos algoritmos, os autores Alves e Andrade (2022) afirmam que devido à falta de entendimento dos outputs dos algoritmos produzidos, surge o questionamento se decisões tão importantes deveriam ser delegadas a sistemas de IA, nos casos em que é difícil explicar como ela chegou ao resultado. Esse dilema se manifesta claramente em situações de alto impacto, como a seleção de possíveis candidatos para vagas de emprego, citando-se o exemplo da Amazon, onde é questionável se uma decisão tão impactante deveria ser delegada a um sistema desenvolvido por métodos que impedem a justificação e a rastreabilidade de como o resultado foi obtido.

O problema analisado no caso da Amazon configura um viés algorítmico, o qual está relacionado as suas decisões tendenciosas. Conforme exposto no trabalho de revisão bibliográfica de Simões-Gomes, Roberto e Mendonça (2020), o viés algorítmico não apresenta uma definição clara em todos os trabalhos analisados pelos autores. No entanto, eles constatam que há sempre uma relação entre o viés e a discriminação ou violação de outros direitos. Os autores complementam que um sistema pode ser considerado enviesado quando possui características que discriminam indivíduos ou um grupo de indivíduos, seja de forma intencional ou não.

Reconhecendo que essa opacidade resulta em viés algorítmico e que a incapacidade de interpretar a tomada de decisão em sistemas sem justificação e rastreabilidade pode gerar graves problemas, como a discriminação ou a perda de confiança, torna-se imperativo analisar o problema da caixa-preta. Este problema se configura como um modelo obscuro que oculta sua lógica interna, sendo uma questão crucial tanto prática quanto ética, pois torna os processos internos desconhecidos ou ininterpretáveis para os seres humanos, abrangendo desde leigos até os próprios desenvolvedores. Para evitar este problema, a transparência deve ser integrada como requisito essencial no projeto, a fim de garantir a explicabilidade e a confiança.

3.5 Os sete conceitos *do Privacy by Design*

O PbD, como visto anteriormente, é composto por sete conceitos criados por Ann Cavoukian no artigo *The Seven Foundational Principles of Privacy by Design*, de 2009. O intuito da autora com esses princípios é integrar a privacidade de dados desde a concepção de projetos. Os conceitos são: ser proativo, não reativo; preventivo, não corretivo; privacidade como configuração padrão, incorporada ao design; funcionalidade completa, soma positiva, não soma zero; segurança de ponta a ponta; visibilidade e transparência, mantenha-o aberto; respeito pela privacidade do usuário, mantenha-o centrado no usuário. A seguir, os conceitos do PbD serão apresentados na Tabela 1.

Tabela 1 – Os sete princípios do Privacy by Design
Fonte: Autoria própria.

Conceito	Explicação breve
Seja proativo, não reativo, preventivo, não corretivo.	Caracterizada por medidas proativas, e não reativas. Ela antecipa e previne eventos invasivos à privacidade antes que aconteçam.
Privacidade como configuração padrão	Garante que os dados pessoais sejam automaticamente protegidos em qualquer sistema de TI ou prática empresarial. Se a pessoa não fizer absolutamente nada, sua privacidade permanece intacta.
Privacidade incorporada ao design	É incorporado ao design e à arquitetura dos sistemas de TI e das práticas empresariais. Não é acoplado como um complemento, depois que tudo já está pronto.
Funcionalidade completa – soma positiva, não soma zero	Busca acomodar todos os interesses e objetivos legítimos de uma forma positiva, em que todos saem ganhando.
Segurança de ponta a ponta	Privacy by Design, por estar incorporado ao sistema antes mesmo do primeiro dado ser coletado, estende-se de forma segura por todo o ciclo de vida das informações envolvidas.
Visibilidade e transparência	Busca assegurar a todas as partes interessadas que, independentemente da prática empresarial ou da tecnologia envolvida, ela está de fato operando de acordo com as promessas e objetivos declarados, estando sujeita a verificação independente.
Respeito pela privacidade do usuário	Exige que arquitetos e operadores mantenham os interesses do indivíduo em primeiro plano, oferecendo medidas como configurações padrão fortes de privacidade, avisos apropriados e opções práticas que empoderem o usuário

3.6 Privacy by Design X Problema da Caixa-Preta

Como descrito pela criadora do PbD Ann Cavoukian, os sistemas de software devem garantir ao usuário e aos interessados a privacidade dos dados em todo o sistema, desde o início do seu funcionamento e em todo o ciclo de vida dos dados. Sendo necessário garantir a integralidade da informação de forma transparente e com total visibilidade e compreensão de como a informação é processada e a finalidade da coleta desses dados. Mas, como analisado anteriormente, o problema da caixa-preta presente nas IAs hoje em dia não permite que o usuário entenda como as informações estão sendo utilizadas ou como o algoritmo chegou naquele resultado.

A aplicação dos conceitos do PbD pode auxiliar os desenvolvedores das inteligências artificiais a prevenir ou mitigar essa problemática. O primeiro princípio do PbD afirma ser totalmente necessário pensar em como o algoritmo da inteligência artificial irá funcionar e como ele pode ser explicável para o usuário antes mesmo de ser codificado, sendo o objetivo que qualquer usuário comum ou mesmo os desenvolvedores sejam capazes de entender o raciocínio e o resultado gerado pelo *prompt*, instrução ou comando de entrada para a IA, evitando vieses ou respostas corretas acompanhadas de raciocínios equívocos. Reforçando essa busca por transparência, o trabalho de Alves e Andrade (2022) diz que um modelo "transparente" de aprendizado de máquina é aquele que é explicável por si só, sem a ajuda de qualquer ferramenta para entender o que está acontecendo, mas alguns algoritmos opacos precisam de ferramentas extras para a sua compreensão. Neste sentido os autores reforçam o uso da *explainable artificial intelligence* (XAI) que busca interpretabilidade ou explicabilidade, tornando o comportamento do algoritmo mais inteligível para os humanos por meio de estratégias textuais ou visuais que detalham a lógica interna da predição; neste sentido, os autores mostram que para modelos opacos são necessários processos de explicação adicional, como a explicabilidade post-hoc. Esta explicação adicional é, por sua vez, feita por um segundo modelo, com a finalidade de fornecer clareza a um modelo caixa-preta já existente após seu treinamento (VIEIRA E DIGIAMPIETRI, 2022). Problemas como o caso do Google Fotos de 2015, mostrado pelos autores Alves e Andrade (2022), onde o sistema rotulou pessoas negras como "gorilas", poderiam ter sido evitados caso aplicados os conceitos do PbD. A empresa veio a público mostrar que o erro seria corrigido, mas em 2018 a revista estadunidense de tecnologia Wired testou a plataforma e viu que as palavras "gorila", "chimpanzé" e "macaco" tinham sido apenas bloqueadas. Neste sentido, foi observado durante a pesquisa que o problema ainda persiste, enquanto buscas como "gato" e "cachorro" funcionam normalmente, palavras como "gorila" e "macaco" ainda estão bloqueadas. Este

fato retrata o problema da caixa-preta, onde não é possível explicar de fato como o algoritmo funciona e como alcançou aquele resultado, o que teria sido evitado caso a empresa utilizasse os conceitos do PbD e a XAI.

O segundo princípio, denominado Privacidade Como Configuração Padrão, tem o objetivo de garantir a privacidade no software sem a necessidade de nenhuma funcionalidade extra, sendo um ajuste nativo do sistema. Para isso, é preciso assegurar a explicabilidade do algoritmo, sendo um padrão intrínseco ao sistema, sem que isso afete o processo de aprendizado do algoritmo. Conforme exposto, seria necessário que a explicação para o usuário, detalhando como os dados são processados e como o resultado foi alcançado através de ferramentas como a XAI, citada anteriormente, sendo integrada sem impactar qualquer outra funcionalidade. Assim, se empresas como a Google tivessem utilizado a privacidade como padrão e algoritmos transparentes, a correção de erros seria realmente corretiva, sem se limitar apenas em mascarar o verdadeiro raciocínio da ferramenta. Isso permitiria a auditabilidade do sistema, onde é possível identificar de forma mais rápida e de fácil entendimento falhas no sistema, como vazamento de dados ou vieses.

O terceiro princípio criado pela autora Cavoukian (2009), denominado Privacidade incorporada ao design, tem como objetivo incorporar a privacidade de dados no design, na arquitetura do sistema e nas regras de negócios. Para isso, é crucial pensar em como implementar a transparência, não somente no algoritmo em si, mas também em suas regras de negócios. No mesmo sentido do conceito anterior, as empresas devem planejar a integração de algoritmos explicáveis desde a concepção do projeto, estabelecendo essa prática não apenas em código, mas nos princípios da própria organização. Por sua vez, o quarto princípio, o Funcionalidade completa – soma positiva, não soma zero, tem o intuito de sempre agregar a privacidade de forma positiva ao projeto, sem que isso afete qualquer funcionalidade. Neste contexto, a transparência e a explicabilidade dos códigos da IA devem constituir uma soma positiva ao sistema, evitando comprometer as funcionalidades principais, como o aprendizado de máquina ou o processo de raciocínio do algoritmo.

O princípio da segurança de ponta a ponta é apresentado como o quinto item da lista. A autora reforça que as informações do usuário devem ser protegidas em todo o seu ciclo de vida. Ou seja, essa proteção deve ser aplicada antes mesmo de a informação ser criada e estender-se até a sua destruição. No contexto das IAs, é essencial garantir a transparência durante todo o ciclo de vida da informação. Para isso, torna-se necessário que o usuário compreenda como a sua informação está sendo utilizada e para qual finalidade por meio dos Termos de Uso, mas cabe ao usuário ler, interpretar e concordar com o conteúdo do documento. Nessa perspectiva, é indispensável mostrar ao usuário todo o processo que o dado percorreu e confirmar se ele de fato foi deletado. Para garantir

que esse processo aconteça, se faz necessário utilizar métodos da XAI como a técnica *Local Interpretable Model-agnostic Explanations* (LIME). Os autores Levy e Adaniya (2024) citam a técnica LIME em seu trabalho "XAI: esclarecendo o problema da caixa-preta com transparência e interpretabilidade" como uma poderosa ferramenta que oferece explicações locais para modelos de aprendizado de máquina complexos. Esta ferramenta oferece explicações de fácil entendimento dos modelos, destacando quais características dos dados de entrada influenciam mais nas decisões tomadas pelo algoritmo, permitindo não apenas entender o que o modelo previu, mas entender o motivo da sua previsão.

O sexto conceito a ser abordado é o de visibilidade e transparência, o qual reforça a necessidade de garantir a compreensão do usuário em relação ao uso de seus dados. Na era digital, a segurança, a visibilidade e a transparência merecem maior destaque, visto que são considerados direitos fundamentais. Atualmente, os dados são captados a todo momento, o que gera um grande desequilíbrio entre aqueles que controlam e processam as informações e o indivíduo que as possui (MARRAFON e COUTINHO, 2020). Portanto, para que o usuário não seja prejudicado, é imprescindível garantir que os dados estejam visíveis, sejam de fácil acesso e que haja total transparência, por parte das empresas, sobre o funcionamento do processamento e o motivo da utilização dessas informações.

O último conceito fundamental é o respeito pela privacidade do usuário. Este princípio busca colocar o usuário no centro de toda a operação, criando regras de negócio e funcionalidades que priorizam inequivocamente sua privacidade. O conceito exige o desenvolvimento de uma plataforma amigável, mantendo os interesses do usuário em primeiro lugar para evitar qualquer tipo de violação das informações. Para que isso ocorra, é necessário que o desenvolvimento das IAs inclua a concepção de métodos e termos que garantam ao usuário a confiança de que seus dados estão realmente seguros. Além disso, capacitar os titulares dos dados a desempenhar um papel ativo na gestão de suas próprias informações é considerada a verificação mais eficaz contra abusos. O respeito pela privacidade também se estende à necessidade de que as interfaces sejam centradas no ser humano e fáceis de usar. Essa característica é vital para permitir que decisões informadas sobre privacidade sejam exercidas de forma confiável. Por fim, é essencial garantir que o usuário consiga compreender plenamente o que a plataforma realiza com esses dados.

A aplicação do PbD demonstra ser o caminho indispensável para mitigar a problemática da "caixa-preta" presente nas IAs atuais, que impede o usuário de compreender como as informações são utilizadas ou como o algoritmo chegou a determinado resultado. A essência do PbD é a abordagem proativa e preventiva, que exige que a transparência e a explicabilidade sejam

consideradas antes mesmo da codificação, transformando a explicação detalhada sobre o processamento dos dados e a forma como o resultado foi alcançado em uma configuração padrão intrínseca ao sistema. Ao incorporar a privacidade e a explicabilidade no design, na arquitetura e nas regras de negócio da organização, adota-se uma abordagem preventiva que evita problemas notórios, como os vieses de racismo e machismo empregados por algoritmos, observados em casos como o Google Fotos e Amazon. Além disso, a transparência e a explicabilidade devem constituir uma soma positiva ao sistema, sem comprometer as funcionalidades principais, como o aprendizado de máquina, garantindo que a correção de falhas seja realmente corretiva, e não apenas um mascaramento do raciocínio subjacente da IA.

Para que essa transparência se concretize de ponta a ponta, é crucial que ela se estenda por todo o ciclo de vida da informação, exigindo visibilidade e fácil acesso aos dados por parte do usuário. Neste contexto, o uso da XAI se torna fundamental, oferecendo ferramentas como o LIME para detalhar quais características dos dados de entrada influenciam as decisões do algoritmo, tornando seu comportamento mais inteligível para os humanos. Por fim, o princípio do respeito pela privacidade do usuário reforça que o indivíduo deve estar no centro de todas as funcionalidades e regras de negócio, capacitando os titulares de dados a desempenhar um papel ativo na gestão de suas próprias informações. Através da integração sistêmica e proativa dos sete princípios do PbD as empresas poderão construir IAs que garantam a integralidade da informação de forma transparente e com total visibilidade, solidificando a confiança e a prestação de contas na era digital.

4 CONCLUSÃO

A pesquisa analisou como os conceitos do *Privacy by Design* (PbD) podem orientar os desenvolvedores de inteligências artificiais, contribuindo para a transparência e a privacidade dos usuários. O estudo demonstrou que a opacidade algorítmica é um fenômeno complexo que afeta tanto os usuários que não entendem o raciocínio do código, quanto os próprios desenvolvedores que não conseguem prever o comportamento da ferramenta. Este cenário exemplifica o problema da caixa-preta, um desafio que impede a compreensão de como os dados são sendo processados e geram resultados.

A análise proposta pelo trabalho confirmou que a aplicação dos sete conceitos do PbD, formulado por Ann Cavoukian, é um caminho eficaz para mitigar a problemática da caixa-preta dentro do desenvolvimento das IAs. Os conceitos fortalecem a confiança entre o usuário e a plataforma com o intuito de deixá-lo no centro de toda a aplicação, desde a concepção do projeto. Neste sentido, a adoção torna a explicabilidade uma configuração padrão e intrínseca ao sistema,

incorporada ao design e aos princípios da empresa, atuando de forma preventiva em casos de vieses e vazamentos de dados, como observado nos casos Amazon e Google Fotos. O estudo também reforça o uso de métodos como a XAI, para que as IAs sejam de fato explicativas para os usuários, utilizando ferramentas como a LIME para tornar o comportamento do algoritmo mais inteligível para os humanos. Além disso, a privacidade e a transparência de dados devem ser tratadas como uma soma positiva dentro do sistema e não como uma concorrente das outras funcionalidades, sem a necessidade de correções futuras ou ações do usuário para garantir a própria segurança.

Diante disto, sugere-se que investigações futuras se concentrem na aplicação prática dos conceitos citados no trabalho em ambientes reais de desenvolvimento de IA, bem como na avaliação de sua eficácia quando integrados a ferramentas de XAI. Por fim, adotar o PbD mostra-se um passo fundamental no desenvolvimento de sistemas confiáveis e éticos, garantindo a transparência de dados e a privacidade do usuário em todo o ciclo de vida da informação, estabelecendo confiança e segurança para toda a sociedade nesta era digital.

5 REFERÊNCIAS

ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato de. Da “caixa-preta” à “caixa de vidro”: o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, Brasília, v. 18, n. 100, 27 jan. 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 14 out. 2025.

BARROSO, Luís Roberto; MELLO, Patrícia Perrone Campos. Inteligência artificial: promessas, riscos e regulação. Algo de novo debaixo do sol. **Revista Direito e Práxis**, v. 15, n. 4, p. e84479, 2024. Disponível em: https://www.scielo.br/j/rdp/a/n89PjvWXTdthJJKwb6TtYXy/. Acesso em: 12 nov. 2025.

BORGES, Gustavo Silveira; FILÓ, Maurício da Cunha Savino. Inteligência artificial, gênero e direitos humanos: o caso Amazon. **Revista Justiça do Direito**, v. 35, n. 3, p. 218–243, 31 dez. 2021. Disponível em: https://ojs.upf.br/index.php/rjd/article/view/12259. Acesso em: 05 nov. 2025.

BRASIL. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). **A inteligência artificial na pesquisa e no fomento: desafios e oportunidades**. Brasília: Capes, 2025.

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

BURRELL, Jenna. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, p. 2053951715622512, 1 jun. 2016. Disponível em: https://journals.sagepub.com/doi/10.1177/2053951715622512. Acesso em: 16 out. 2025.

CANCELIER, Mikhail Vieira de Lorenzi. O direito à privacidade hoje: perspectiva histórica e o cenário brasileiro. **Sequência Estudos Jurídicos e Políticos**, v. 38, n. 76, p. 213–240, 20 set. 2017. Disponível em: https://periodicos.ufsc.br/index.php/sequencia/article/view/2177-7055.2017v38n76p213. Acesso em: 22 out. 2025.

CAVOUKIAN, Ann**. Privacy by Design: The 7 Foundational Principles**. Toronto: Information and Privacy Commissioner of Ontario, 2009. Disponível em: https://www.ipc.on.ca/en/media/1826/download?attachment. Acesso em: 14 out. 2025.

CANGUSSU, João Pedro Barbosa et al. PRIVACIDADE DE DADOS NA ERA DIGITAL: DESAFIOS E ADAPTAÇÕES DAS EMPRESAS ONLINE À LEI GERAL DE PROTEÇÃO DE DADOS (LGPD). **Revista Multidisciplinar do Nordeste Mineiro**, v. 11, n. 1, p. 1–12, 13 jun. 2025. Disponível em: https://remunom.ojsbr.com/multidisciplinar/article/view/4138. Acesso em: 01 nov. 2025.

CORDEIRO, Fernanda De Nazaré Cardoso Dos Santos et al. Estudos descritivos exploratórios qualitativos: um estudo bibliométrico. Brazilian Journal of Health Review, v. 6, n. 3, p. 11670–11681, 5 jun. 2023. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BJHR/article/view/60412. Acesso em: 14 out. 2025.

DIAS, Ketylen Suelen dos Anjos; DIAS, Cristiane Toniolo; DE PAULA, Marcelo**.** Privacidade e inteligência artificial: uma análise exploratória**. Revista Eletrônica Científica Inovação e Tecnologia**, Medianeira, v. 16, n. 39, 2025. Disponível em: https://periodicos.utfpr.edu.br/recit/article/view/19181. Acesso em: 15 out. 2025.

FAMA, Josué Sá. Inteligência artificial e privacidade: implicações legais e éticas na era digital. **Revista Científica Multidisciplinar Núcleo do Conhecimento**, São Paulo, p. 15–39, 2 jan. 2025. Disponível em: https://www.nucleodoconhecimento.com.br/tecnologia/implicacoes-legais. Acesso em: 03 nov. 2025.

FUNDAÇÃO ITAÚ; DATAFOLHA. **Inteligência Artificial: conhecimento, uso e expectativas**. São Paulo: Fundação Itaú, 2025. Disponível em: https://www.fundacaoitau.org.br/observatorio/biblioteca/pesquisa-consumo-e-uso-da-inteligencia-artificial-no-brasil. Acesso em: 17 dez. 2025.

GENSO, Millena Gabriele; PICOLI, Gabriela Regina Nardi; DA LUZ, Pedro Henrique Machado. Privacy by design como possível medida de efetivação da proteção dos dados pessoais no século XXI. **Revista Contemporânea**, São José dos Pinhais, v. 3, n. 8, p. 11506–11533, 11 ago. 2023. Disponível em: https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/1415. Acesso em: 22 out. 2025.

GUIDOTTI, Riccardo et al. A Survey of Methods for Explaining Black Box Models. **ACM Computing Surveys**, New York, v. 51, n. 5, p. 1–42, 30 set. 2019. Disponível em: https://dl.acm.org/doi/10.1145/3236009. Acesso em: 06 nov. 2025.

LEVY, Gal; ADANIYA, Mario Henrique. XAI: esclarecendo o problema da caixa preta com transparência e interpretabilidade. **Revista Terra & Cultura**: Cadernos de Ensino e Pesquisa, Londrina, v. 40, p. 346-371, 2024. Disponível em: http://periodicos.unifil.br/index.php/Revistateste/article/view/3170. Acesso em: 12 nov. 2025.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, São Paulo, v. 35, n. 101, p. 85–94, abr. 2021. Disponível em: https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd/. Acesso em: 28 out. 2025.

MARRAFON, Marco Aurélio; COUTINHO, Luiza Leite Cabral Loureiro. Princípio da Privacy by Design: fundamentos e efetividade regulatória na garantia do direito à proteção de dados. **Revista Eletrônica Direito e Política**, Itajaí, v. 15, n. 3, p. 955–984, 18 dez. 2020. Disponível em: https://periodicos.univali.br/index.php/rdp/article/view/17119. Acesso em: 22 out. 2025.

MELO, Jonas Ferrigolo; ROCKEMBACH, Moisés; SILVA, Armando Malheiro da. Ciência da Informação e privacy by design: aspectos éticos e possibilidades de pesquisa. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 16, n. 1, p. 177-197, 2023. Disponível em: https://lume.ufrgs.br/handle/10183/256213. Acesso em: 21 out. 2025.

MICHENER, Gregory; BERSCH, Katherine. Identifying transparency. **Information Polity**, Amsterdam, v. 18, n. 3, p. 233-242, 2013. Disponível em: https://journals.sagepub.com/doi/abs/10.3233/IP-130299. Acesso em: 03 nov. 2025.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial intelligence: a modern approach**. Third edition, Global edition ed. Boston Columbus Indianapolis: Pearson, 2016.

SICHMAN, Jaime Simão. Inteligência Artificial e sociedade: avanços e riscos. **Estudos Avançados**, São Paulo, v. 35, n. 101, p. 37–50, abr. 2021. Disponível em: https://www.scielo.br/j/ea/a/c4sqqrthGMS3ngdBhGWtKhh/. Acesso em: 28 out. 2025.

SILVA, Tainã Dias da; DOMINGUES, Patrícia Martinez. Inteligência Artificial e Privacidade: Os desafios do Privacy by Design. **Advances in Knowledge Representation**, [S. l.], v. 4, n. 2, p. 160–194, 2024. Disponível em: https://periodicos.ufmg.br/index.php/advances-kr/article/view/52813. Acesso em: 12 nov. 2025.

SIMÕES-GOMES, Letícia; ROBERTO, Enrico; MENDONÇA, Jônatas. Viés algorítmico – um balanço provisório. **Estudos de Sociologia**, Araraquara, v. 25, n. 48, 24 jul. 2020. Disponível em: https://periodicos.fclar.unesp.br/estudos/article/view/13402. Acesso em: 15 nov. 2025.

SOBREIRA, Victor. Um panorama da História da Inteligência Artificial e suas aplicações na pesquisa histórica. **Varia História**, Belo Horizonte, v. 41, p. e25035, 2025. Disponível em: https://www.scielo.br/j/vh/a/rBFnX4gCZ4N8fcRt6mjxbFQ/. Acesso em: 28 out. 2025.

VIEIRA, Carla Piazzon; DIGIAMPIETRI, Luciano Antonio. Machine Learning post-hoc interpretability: a systematic mapping study. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 2022, Curitiba. **Anais** [...]. New York: ACM, 2022. Disponível em: https://dl.acm.org/doi/10.1145/3535511.3535512. Acesso em: 15 nov. 2025.