Universidade Federal da Paraíba
Centro de Ciências Humanas, Letras e Artes
Departamento de Filosofia
Curso de Pós-Graduação Lato Sensu em
Ética em Inteligência Artificial, EaD

Tiago da Costa Lima

A Caixa-Preta Algorítmica e a Crise do Espaço Público: Opacidade, Explicabilidade e Fragmentação Discursiva

João Pessoa
2026

Tiago da Costa Lima

A Caixa-Preta Algorítmica e a Crise do Espaço Público: Opacidade, Explicabilidade e Fragmentação Discursiva

Trabalho de conclusão de curso apresentado ao Curso de Pós-Graduação Lato Sensu em Ética em Inteligência Artificial, EaD, do Departamento de Filosofia, do Centro de Ciências Humanas, Letras e Artes, da Universidade Federal da Paraíba, como requisito para obtenção do título de Especialista em Ética em Inteligência Artificial

Orientador:. Dr. Danilo Fraga Dantas

João Pessoa
2026

Catalogação na publicação
Seção de Catalogação e Classificação

L732c Lima, Tiago da Costa.
A caixa-preta algorítmica e a crise do espaço público : opacidade, explicabilidade e fragmentação discursiva / Tiago da Costa Lima. - João Pessoa, 2026.
33 f. : il.

Orientador : Danilo Fraga Dantas.
TCC (Especialização) - Universidade Federal da Paraíba/Centro de Ciências Humanas, Letras e Artes,2026.

1. Opacidade algorítmica. 2. Transparência algorítmica. 3. Espaço público. 4. Injustiça epistêmica. 5. Democracia deliberativa. I. Dantas, Danilo Fraga. II. Título.

UFPB/CCHLA CDU 004.81

Elaborado por KARLA MARIA DE OLIVEIRA - CRB-15/485

FOLHA DE APROVAÇÃO

Tiago da Costa Lima

A Caixa-Preta Algorítmica e a Crise do Espaço Público: Opacidade, Explicabilidade e Fragmentação Discursiva

Trabalho de conclusão de curso apresentado ao Curso de Pós-Graduação *Lato Sensu* em Ética em Inteligência Artificial, EaD, do Departamento de Filosofia, do Centro de Ciências Humanas, Letras e Artes, da Universidade Federal da Paraíba, como requisito para obtenção do título de Especialista em Ética em Inteligência Artificial.

Aprovado em: 09 / Julho /2026.

Banca Examinadora

_____

Prof. Dr. Danilo Fraga Dantas

Orientador

(Universidade Federal da Paraíba (UFPB) / Centro de Ciências Humanas, Letras e Artes (CCHLA) / Departamento de Filosofia)

_____

Prof. Dr. Juliano Cordeiro da Costa Oliveira

Membro Interno

(Universidade Federal da Paraíba (UFPB) / Departamento de Filosofia)

_____

Examinador: Prof. Dr. Luiz Paulo da Cas Cichoski

Membro Externo

(Universidade Federal de Mato Grosso (UFMT) / Faculdade de Filosofia)

RESUMO

Este artigo analisa de que maneira a opacidade algorítmica compromete as condições sob as quais argumentos circulam, são avaliados e adquirem validade pública em ambientes digitais mediados por sistemas de recomendação e moderação. Parte-se da hipótese de que a opacidade algorítmica não constitui a origem da fragmentação discursiva nem das injustiças epistêmicas produzidas por esses sistemas, mas atua como o estabilizador estrutural desses danos ao dificultar sua identificação, contestação e correção. O artigo examina casos de moderação assimétrica e a evolução do marco regulatório brasileiro, da Estratégia Brasileira de Inteligência Artificial (EBIA) ao Plano Brasileiro de Inteligência Artificial (PBIA), articulando teoria da argumentação, epistemologia social e ética da inteligência artificial para analisar os efeitos da opacidade dos sistemas. Os resultados indicam que mecanismos de Inteligência Artificial Explicável (XAI) e modelos regulatórios baseados exclusivamente na enunciação de princípios são insuficientes para enfrentar os fatores institucionais e epistêmicos que sustentam essa opacidade. Como contribuição, o artigo propõe seis diretrizes éticas orientadas à transparência, à contestabilidade, à distribuição adequada do controle, à representação plural, à preservação da infraestrutura deliberativa e ao cultivo de virtudes epistêmicas institucionais. Conclui-se que a transparência algorítmica deve ser compreendida como uma condição epistêmica, institucional e democrática indispensável à preservação da legitimidade do espaço público em sociedades mediadas por sistemas de inteligência artificial.

Palavras-chave: opacidade algorítmica; transparência algorítmica; espaço público; injustiça epistêmica; democracia deliberativa.

ABSTRACT
RESUMEN, RÉSUMÉ

This article examines how algorithmic opacity undermines the conditions under which arguments circulate, are evaluated, and acquire public validity in digital environments mediated by recommendation and content moderation systems. It argues that algorithmic opacity is not the source of discursive fragmentation or of the epistemic injustices produced by these systems; rather, it functions as a structural stabilizer of these harms by hindering their identification, contestation, and correction. To this end, the article examines cases of asymmetric content moderation and the evolution of the Brazilian regulatory framework, from the Brazilian Artificial Intelligence Strategy (EBIA) to the Brazilian Artificial Intelligence Plan (PBIA), drawing on argumentation theory, social epistemology, and AI ethics to analyze the effects of system opacity. The findings indicate that Explainable Artificial Intelligence (XAI) mechanisms and regulatory models based exclusively on the articulation of general principles are insufficient to address the institutional and epistemic factors that sustain system opacity. As its main contribution, the article proposes six ethical guidelines oriented toward transparency, contestability, appropriate distribution of control, plural representation, preservation of the deliberative infrastructure, and the cultivation of institutional epistemic virtues. It concludes that algorithmic transparency should be understood as an epistemic, institutional, and democratic condition indispensable for preserving the legitimacy of the public sphere in societies increasingly mediated by artificial intelligence systems.

Keywords: algorithmic opacity; public sphere; discursive fragmentation; epistemic injustice; deliberative democracy.

LISTA DE ILUSTRAÇÕES

Quadro 1 – Taxonomia das injustiças epistêmicas no contexto algorítmico …..... 15
Quadro 2 – Casos empíricos ilustrativos de moderação algorítmica 22
Quadro 3 – Comparação dos marcos regulatórios de IA 24
Quadro 4 – Limites da XAI e diretrizes éticas correspondentes 28

LISTA DE ABREVIATURAS E SIGLAS

AI RMF	Artificial Intelligence Risk Management Framework
EBIA	Estratégia Brasileira de Inteligência Artificial
FAT	Fairness, Accountability and Transparency
IA	Inteligência Artificial
LLM	Large Language Model (Modelo de Linguagem de Grande Porte)
NIST	National Institute of Standards and Technology
PBIA	Plano Brasileiro de Inteligência Artificial
XAI	Explainable Artificial Intelligence (Inteligência Artificial Explicável)

SUMÁRIO

1 INTRODUÇÃO....8
2. ENQUADRAMENTO CONCEITUAL: O ESPAÇO PÚBLICO E A IA....12
2.1 O Auditório Universal e a Crise Contemporânea do Discurso Público....12
2.2 Injustiça Epistêmica, Ignorância Estrutural e em Sistemas de IA....14
2.3 Transparência Algorítmica e os Limites da IA Explicável (XAI)....16
2.4 Transparência, Democracia e Reconstrução do Espaço Público....18
3. ANÁLISE EMPÍRICA E REGULATÓRIA....20
3.1 Casos Empíricos de Moderação Assimétrica....20
3.2 Análise Documental: Da EBIA ao PBIA....22
4. DISCUSSÃO E PROPOSIÇÃO DE DIRETRIZES....25
4.1 Opacidade Algorítmica e Estabilização dos Danos....25
4.2 Avaliação Crítica das Respostas Existentes e Proposição de Diretrizes....25
4.2.1 Insuficiências das Respostas Técnicas....25
4.2.2 Insuficiências das Respostas Regulatórias....26
4.2.3 Diretrizes Éticas para Sistemas de Moderação Algorítmica....27
5. CONSIDERAÇÕES FINAIS....30
REFERÊNCIAS....32

1 INTRODUÇÃO

A crescente incorporação de sistemas de IA na mediação do acesso à informação transformou profundamente o funcionamento do espaço público contemporâneo. Redes sociais como Facebook, YouTube e X organizam seus fluxos informacionais por meio de dois tipos de sistemas distintos. Sistemas de recomendação selecionam e priorizam conteúdos para exibição aos usuários com base em histórico de interação, perfil comportamental e métricas de engajamento. Sistemas de moderação identificam, removem, restringem ou sinalizam conteúdos que violem políticas da plataforma.

Atualmente, ambos os tipos de sistemas costumam empregar técnicas de IA. Os sistemas de recomendação utilizam modelos de aprendizado de máquina para inferir preferências dos usuários e personalizar a distribuição de conteúdos com base em métricas de relevância e engajamento (ZHANG et al., 2019). Esses sistemas analisam históricos de interação, padrões de navegação e comportamentos de consumo para estimar quais conteúdos possuem maior probabilidade de despertar interesse, influenciando diretamente a visibilidade e a circulação de informações nas plataformas digitais.

Por sua vez, os sistemas de moderação recorrem cada vez mais a modelos de linguagem de grande porte (LLMs), capazes de classificar textos, inferir contextos e identificar padrões discursivos em linguagem natural (FRANKLIN et al., 2025). No caso de conteúdos audiovisuais, a moderação automatizada normalmente envolve múltiplas camadas: modelos de visão computacional analisam quadros do vídeo, sistemas de reconhecimento automático de fala geram transcrições do áudio, e modelos de processamento de linguagem natural analisam as transcrições, além de títulos, descrições, comentários e outros metadados. No YouTube, entre 2017 e 2021, aproximadamente 94% das remoções de vídeos foram realizadas de forma automatizada, sendo que a maior parte das exclusões ocorre antes mesmo que o conteúdo alcance ampla circulação (SILVA; CESAR, 2022). Esse processo reconfigurou o papel da revisão humana, que passou a concentrar-se predominantemente em casos controversos, procedimentos de apelação e atividades de treinamento e supervisão dos modelos.

Plataformas como o Facebook seguiram trajetórias semelhantes, incorporando arquiteturas de aprendizado profundo para monitorar proativamente **discurso de ódio, desinformação e outros conteúdos considerados nocivos** (LIBMAN, 2023). Essas

arquiteturas combinam técnicas de **classificação automática, reconhecimento de padrões e mecanismos de correspondência** com bases de conteúdos previamente identificados como violações das políticas da plataforma.

Os avanços recentes em modelos de linguagem de grande porte ampliaram significativamente a capacidade dos sistemas de moderação de analisar linguagem natural, permitindo identificar **contextos, relações semânticas e padrões discursivos** que ultrapassam a simples detecção de palavras-chave. Entretanto, a automação em larga escala também introduz novos desafios, como a dificuldade de capturar nuances contextuais, a reprodução de vieses presentes nos dados de treinamento e a crescente dependência de processos decisórios cuja lógica permanece amplamente inacessível ao escrutínio público. Nesse contexto, os LLMs não constituem um objeto autônomo de análise, mas uma das principais infraestruturas tecnológicas que operacionalizam a filtragem e o silenciamento discursivo nos sistemas contemporâneos de moderação, incorporando aos processos decisórios os vieses presentes em seus dados e procedimentos de treinamento.

A expansão dessas tecnologias, seja das redes neurais profundas em geral, seja dos modelos de linguagem de grande porte (LLMs) em particular, é acompanhada por um elevado grau de opacidade **técnica e institucional,** frequentemente descrito na literatura como o problema da **caixa-preta algorítmica** (PASQUALE, 2015). Em sua dimensão técnica, a opacidade algorítmica refere-se à dificuldade de determinar quais regras ou critérios decisórios foram efetivamente aprendidos pelos modelos durante o processo de treinamento. Em arquiteturas baseadas em redes neurais profundas, embora seja possível conhecer a estrutura geral do modelo e os dados utilizados em seu treinamento, permanece difícil reconstruir de forma inteligível como essas informações são combinadas para produzir decisões específicas.

Em sua dimensão institucional, a opacidade decorre de mecanismos deliberadamente estabelecidos pelas organizações responsáveis pelos sistemas, **como segredo comercial, documentação limitada, termos de serviço vagos e restrições ao acesso externo às informações relevantes para auditoria e contestação** (ANANNY; CRAWFORD, 2018; GILLESPIE, 2018). A combinação dessas dimensões faz da opacidade algorítmica um fenômeno sociotécnico que não pode ser reduzido a um problema estritamente técnico, envolvendo também relações de poder, mecanismos de governança e formas assimétricas de distribuição da informação. Ao longo deste artigo, o termo ***opacidade algorítmica*** designa a

impossibilidade estrutural de escrutínio público dos critérios que orientam as decisões dos sistemas de IA.

Este artigo investiga dois problemas distintos, cada um associado a um tipo de sistema.

O primeiro problema diz respeito à fragmentação discursiva. Sistemas de recomendação personalizam o fluxo informacional com base em preferências inferidas, reduzindo a exposição a perspectivas divergentes. O conceito de auditório universal, formulado por Perelman e Olbrechts-Tyteca (2005), funciona como parâmetro normativo desta análise ao pressupor a possibilidade de justificação de argumentos diante de qualquer interlocutor racional. Contudo, conforme argumenta Sunstein (2017), a personalização promovida pelos sistemas de recomendação tende a ampliar a exposição seletiva e a reforçar crenças previamente existentes. Como consequência, diferentes grupos passam a operar com conjuntos distintos de informações, evidências e critérios de avaliação, fenômeno que Habermas (2023) associa à formação de esferas semipúblicas caracterizadas pelo fechamento discursivo e pela redução da deliberação entre perspectivas divergentes.

O segundo conjunto de efeitos diz respeito ao apagamento de posições e recursos interpretativos produzido pelos sistemas de moderação. Modelos de moderação baseados em LLMs podem **classificar, restringir ou remover conteúdos** segundo critérios cuja operacionalização permanece amplamente opaca aos usuários. Quando tais sistemas reproduzem padrões historicamente presentes em seus dados de treinamento, determinados grupos e formas de expressão tornam-se mais vulneráveis à invisibilização. Mollema (2024) denomina esse fenômeno de apagamento hermenêutico gerativo: a marginalização estatística de **conceitos, experiências e recursos** interpretativos reduz as possibilidades de construção compartilhada de sentido. Esse processo pode ser aprofundado por aquilo que Loock (2025) denomina cognição extraída, isto é, a transferência progressiva de atividades tradicionalmente associadas ao julgamento humano para sistemas automatizados, reduzindo as oportunidades de desenvolvimento das capacidades necessárias para reconhecer e contestar os próprios mecanismos de mediação.

Parte-se da hipótese de que a fragmentação discursiva e as formas de injustiça hermenêutica observadas nesses ambientes não constituem fenômenos independentes, mas efeitos produzidos por arquiteturas técnicas específicas de recomendação e moderação. A opacidade desses sistemas não atua como causa isolada desses danos; sua função principal

consiste em dificultar sua identificação, atribuição e correção, estabilizando estruturalmente condições que, em contextos mais **transparentes, seriam passíveis de contestação pública**. Embora o artigo mobilize os dois tipos de sistemas que organizam as redes sociais, **moderação e recomendação**, o foco analítico central recai sobre os sistemas de moderação devido à sua precedência funcional. A moderação atua de forma mais abrangente, visto que apenas o conteúdo que sobrevive à filtragem editorial pode ser subsequentemente distribuído e priorizado pelos sistemas de recomendação.

A relevância desse problema decorre de seus efeitos sobre a capacidade de responsabilização pública. Quando os critérios que orientam decisões automatizadas permanecem inacessíveis, torna-se difícil **identificar, atribuir e corrigir falhas,** enviesamentos ou práticas discriminatórias. Como observa Veluwenkamp (2025), tal cenário produz uma lacuna epistêmica de responsabilidade, caracterizada pela impossibilidade de determinar de forma adequada quem responde por decisões tomadas por sistemas complexos.

Diante desse cenário, a transparência algorítmica emerge como uma resposta normativa relevante à dimensão institucional da opacidade. Em termos gerais, ela envolve a **disponibilização e a rastreabilidade** de informações sobre o funcionamento dos sistemas de IA, possibilitando **auditoria, supervisão e mecanismos de contestação**. No plano técnico, propostas reunidas sob o campo da **Inteligência Artificial Explicável (XAI)** buscam ampliar a interpretabilidade dos sistemas por meio de modelos **intrinsecamente interpretáveis (ante-hoc)** ou de **técnicas pós-hoc** capazes de produzir explicações aproximadas sobre suas decisões.

Entretanto, tais abordagens frequentemente permanecem limitadas por se apoiarem predominantemente em padrões correlacionais e por não enfrentarem integralmente as dimensões institucionais e políticas da opacidade (BURR et al., 2018). Embora necessárias, **transparência e explicabilidade** não são suficientes para resolver todos os problemas associados aos efeitos epistêmicos produzidos pelos sistemas analisados.

Assim, o objetivo geral do artigo consiste em analisar de que maneira as arquiteturas desses sistemas de IA produzem e consolidam tais efeitos epistêmicos, considerando suas diferenças funcionais e contextuais de operação. Busca-se, **(i)** caracterizar os mecanismos técnicos e institucionais pelos quais esses sistemas afetam a circulação e a visibilidade de conteúdos; **(ii)** avaliar criticamente as potencialidades e limitações da **Inteligência Artificial Explicável (XAI)** como estratégia de mitigação da opacidade; e **(iii)** propor diretrizes

ético-normativas para o desenvolvimento de sistemas de moderação mais transparentes, auditáveis e compatíveis com os princípios da democracia deliberativa.

O artigo contribui para o debate ao integrar perspectivas oriundas da epistemologia social, da ética da IA e da teoria democrática para compreender os efeitos da opacidade algorítmica sobre a estruturação do espaço público contemporâneo. Sustenta-se que a opacidade desempenha um papel de estabilização estrutural dos danos epistêmicos produzidos por sistemas automatizados, razão pela qual a transparência deve ser compreendida como uma condição normativa relevante para a preservação de práticas democráticas em ambientes informacionais mediados por IA.

O texto está organizado em três movimentos principais. Inicialmente, desenvolve-se uma caracterização conceitual dos mecanismos de fragmentação discursiva e das formas de injustiça hermenêutica, associadas aos sistemas analisados. Em seguida, discutem-se casos de moderação assimétrica e aspectos da trajetória regulatória brasileira, relacionados à governança de plataformas digitais. Por fim, são apresentadas diretrizes éticas voltadas à avaliação e ao desenvolvimento de sistemas de moderação mais transparentes, auditáveis e compatíveis com os princípios da democracia deliberativa.

2. ENQUADRAMENTO CONCEITUAL: O ESPAÇO PÚBLICO E A IA

2.1 O Auditório Universal e a Crise Contemporânea do Discurso Público

O ponto de partida para avaliar a legitimidade do discurso mediado por sistemas de IA é o conceito de **auditório universal**, formulado por Perelman e Olbrechts-Tyteca (2005): trata-se da representação ideal de qualquer sujeito racional capaz de avaliar um argumento com base em razões compartilhadas. Esse conceito fundamenta a racionalidade prática, entendida como o exercício da razão em contextos discursivos voltados à persuasão e à tomada de decisão sobre valores e ações, e opera como um parâmetro normativo de validade intersubjetiva: a exigência de que os argumentos sejam passíveis de aceitação racional para além de nichos culturais ou ideológicos específicos. Trata-se de uma regra constitutiva da deliberação democrática, pois, para que uma pretensão de validade aspire à legitimidade pública, ela deve submeter-se a critérios de justificação que possam ser reconhecidos como válidos por qualquer interlocutor racional.

Contudo, a infraestrutura das redes sociais impõe uma barreira sociotécnica a esse ideal. Ao delegar a organização do fluxo informacional a sistemas de IA comerciais opacos que personalizam o ambiente segundo padrões de engajamento, as plataformas favorecem a

emergência de múltiplos auditórios parciais que operam com critérios de verdade e relevância progressivamente incompatíveis. Esse processo corrói as condições de possibilidade do auditório universal ao fragmentar o horizonte comum de justificação necessário à avaliação crítica intersubjetiva, contribuindo para a formação de esferas semipúblicas autorreferenciais (SUNSTEIN, 2017; HABERMAS, 2023).

A relevância desse processo torna-se mais evidente quando as plataformas digitais são compreendidas não apenas como meios de comunicação, mas como instituições epistêmicas. Sob a perspectiva da epistemologia social institucional, sistemas de recomendação e moderação desempenham funções de seleção, organização e validação da informação socialmente disponível, influenciando quais conteúdos circulam, quais perspectivas se tornam visíveis e **quais opiniões adquirem maior credibilidade pública**. Quando essas funções passam a ser exercidas por sistemas de IA operando sob condições de opacidade, a estrutura informacional do espaço público deixa de ser apenas um ambiente de interação e passa a constituir um objeto de governança sociotécnica com implicações diretas para a deliberação democrática e para a **circulação social da informação**.

A opacidade algorítmica, anteriormente apresentada como elemento que estabiliza os danos produzidos por sistemas de recomendação e moderação, possui uma dimensão técnica frequentemente descrita na literatura como o problema da caixa-preta algorítmica (PASQUALE, 2015). Conforme discutido na introdução, essa opacidade limita o escrutínio público dos critérios empregados pelos sistemas de IA e dificulta a atribuição de responsabilidades por seus efeitos. Suas implicações para a responsabilização e para a distribuição do controle decisório tornam-se particularmente evidentes quando analisadas sob a perspectiva da governança sociotécnica.

Devido a essa arquitetura de caixa-preta, torna-se difícil identificar quem deve responder pelos danos produzidos por sistemas de IA. Veluwenkamp (2025) denomina esse fenômeno de l**acuna epistêmica de responsabilidade:** a opacidade dos processos decisórios impede o rastreamento adequado das cadeias de decisão e dificulta a atribuição de responsabilidade pelos efeitos produzidos. O autor identifica ainda um segundo problema, denominado **desalinhamento de controle,** que ocorre quando a capacidade efetiva de influenciar ou corrigir o comportamento dos sistemas encontra-se distribuída de forma inadequada entre desenvolvedores, operadores, empresas e instâncias regulatórias. Nessas circunstâncias, a responsabilização pode tornar-se inviável, mesmo quando existem mecanismos formais de transparência.

As categorias propostas por Veluwenkamp permitem compreender por que a transparência, embora necessária, não é suficiente para enfrentar os problemas associados à governança algorítmica. Tornar os sistemas mais visíveis não elimina automaticamente a lacuna epistêmica de responsabilidade nem garante que os atores com capacidade efetiva de intervenção disponham dos meios necessários para exercer controle sobre suas consequências.

2.2 Injustiça Epistêmica, Ignorância Estrutural e em Sistemas de IA

Os efeitos da opacidade sobre o espaço público não se limitam à fragmentação discursiva e à lacuna de responsabilidade. Ela atua também como um agravante estrutural na reconfiguração da ecologia epistêmica, favorecendo formas específicas de injustiça epistêmica, danos cometidos contra sujeitos em sua capacidade como conhecedores (FRICKER, 2007), cujas manifestações no contexto dos sistemas de IA exigem exame mais detalhado.

Miranda Fricker (2007) distingue duas modalidades fundamentais de injustiça epistêmica. A injustiça testemunhal ocorre quando um falante enfrenta um déficit de credibilidade em razão de preconceitos associados à sua identidade. Já a injustiça hermenêutica manifesta-se quando lacunas nos recursos coletivos de interpretação impedem que sujeitos compreendam ou comuniquem adequadamente aspectos significativos de suas experiências sociais.

No contexto das redes sociais, essa distinção adquire contornos operacionais específicos. Sistemas de IA empregados na moderação podem reproduzir formas de injustiça testemunhal ao reduzir sistematicamente a visibilidade de vozes marginalizadas com base em padrões históricos de descrédito incorporados aos dados de treinamento. Paralelamente, a opacidade institucional dos processos de moderação pode aprofundar formas de injustiça hermenêutica ao dificultar que os usuários compreendam, articulem e contestem as razões pelas quais determinados conteúdos são restringidos ou removidos.

Mollema (2024) expande esse quadro ao propor o conceito de **apagamento hermenêutico gerativo**, definido como a automação do epistemicídio. O autor argumenta que grandes modelos de linguagem (LLMs) podem reproduzir formas de injustiça epistêmica ao inferiorizar ou marginalizar **esquemas conceituais, formas de expressão e opiniões** sub-representadas em seus dados de treinamento. Como esses sistemas aprendem padrões estatísticos a partir de grandes volumes de dados predominantemente produzidos em

contextos ocidentais e anglófonos, tendem a reproduzir essas perspectivas como se fossem universalmente representativas. O resultado é aquilo que Mollema denomina uma visão **de lugar nenhum:** a erosão progressiva de formas alternativas de construção de sentido e dos recursos interpretativos disponíveis aos usuários, produzindo uma forma automatizada de silenciamento de perspectivas no espaço público.

Quadro 1: ***Taxonomia das injustiças epistêmicas no contexto algorítmico.***

Tipo de Injustiça	Problema Produzido	Manifestação em Sistemas de Ia	Fonte
Injustiça Testemunhal Automatizada	Déficit sistemático de credibilidade decorrente de preconceitos incorporados aos dados de treinamento.	**Sistemas de moderação reduzem desproporcionalmente a visibilidade de determinados grupos com base em padrões históricos incorporados aos dados de treinamento.**	**Fricker (2007); Mollema (2024)**
Injustiça Testemunhal Ampliada	Sistemas generativos reproduzem e reforçam estereótipos durante a geração de conteúdos.	**Sistemas generativos reproduzem e reforçam estereótipos durante a geração de conteúdo**	**Mollema (2024); Kay et al. (2024)**
Injustiça Zetética	Restrição da investigação crítica pelo favorecimento sistemático de padrões estatísticos dominantes.	**Sistemas priorizam padrões dominantes e reduzem a exposição a linhas alternativas de investigação.**	**Mollema (2024)**
Ignorância Hermenêutica Gerativa	Empobrecimento dos recursos interpretativos disponíveis para compreender determinadas experiências.	**Sistemas sub-representam experiências e formas de expressão presentes em determinados grupos em razão de sua baixa representatividade nos dados de treinamento.**	**Mollema (2024); Kay et al. (2024)**

Fonte: *Elaborado pelo autor com base em Fricker (2007), Mollema (2024) e Kay, Kasirzadeh e Mohamed (2024).*

Kay, Kasirzadeh e Mohamed (2024) complementam essa taxonomia ao demonstrar que a injustiça testemunhal pode emergir não apenas da exclusão de determinadas vozes, mas também da transferência excessiva de autoridade epistêmica para sistemas de IA. Quando classificações ou respostas produzidas por esses sistemas passam a ser aceitas sem questionamento, mesmo em situações nas quais contradizem avaliações humanas razoáveis, forma-se uma assimetria de credibilidade que favorece o sistema em detrimento dos usuários, reproduzindo padrões de marginalização já existentes.

No contexto brasileiro, a ***Coletânea TIC, Governança da Internet, Gênero, Raça e Diversidade*** (2023) oferece evidências que nacionalizam esse diagnóstico. Xavier et al. (2023), ao analisarem a detecção de discursos violentos contra candidaturas negras e LGBTQ+ nas eleições de 2022, identificam limitações de modelos de IA, como o BERTimbau na identificação de agressões direcionadas a esses grupos. Esses resultados sugerem que sistemas de moderação podem reproduzir padrões compatíveis com formas de injustiça testemunhal automatizada, na medida em que falham em reconhecer adequadamente

ataques dirigidos a grupos historicamente vulnerabilizados, reduzindo a eficácia dos mecanismos de proteção discursiva disponíveis.

A delegação de funções epistêmicas a sistemas de IA favorece ainda a produção de formas de ignorância estrutural. Para Peels (2023), a ignorância não decorre apenas da ausência de informação, mas também de arranjos institucionais que dificultam sistematicamente a compreensão, a verificação e a contestação do conhecimento disponível. Em ambientes mediados por sistemas opacos de **recomendação e moderação,** tais barreiras não apenas dificultam a compreensão dos mecanismos **responsáveis pela circulação, visibilidade e classificação dos conteúdos,** mas também podem restringir o acesso dos usuários a **informações, perspectivas e evidências relevantes**, favorecendo a produção de formas persistentes de ignorância sobre temas socialmente relevantes.

Esse processo pode ser aprofundado pelo fenômeno que Loock (2025) denomina **cognição extraída**. Segundo o autor, a utilização recorrente de sistemas de IA para apoiar processos de decisão e avaliação pode reduzir as oportunidades de exercício de atividades tradicionalmente desempenhadas pelos próprios usuários, como o julgamento, a interpretação e a avaliação crítica. Quando tais funções passam a ser sistematicamente delegadas aos sistemas automatizados, tornam-se mais escassas as oportunidades de desenvolver ou exercer as capacidades necessárias para reconhecer, compreender e contestar os mecanismos de mediação que afetam a experiência informacional.

Os problemas discutidos até aqui evidenciam que a opacidade dos sistemas de IA não se limita a uma dificuldade técnica de interpretação, mas produz consequências **epistêmicas, institucionais e políticas** concretas. Em resposta a esse desafio, propostas de Inteligência Artificial Explicável (XAI) buscam ampliar a transparência dos sistemas de IA. A questão central, contudo, é se tais mecanismos são capazes de responder adequadamente aos danos identificados. A seção seguinte examina criticamente os limites dessa abordagem.

2.3 Transparência Algorítmica e os Limites da IA Explicável (XAI)

Apresentadas como a resposta técnica privilegiada à opacidade, as propostas de Inteligência Artificial Explicável (XAI) buscam ampliar a interpretabilidade dos sistemas de IA por meio de mecanismos como o destaque de termos que ativaram classificadores automáticos ou a atribuição de pesos a variáveis para justificar decisões específicas (BURR et al., 2018). Contudo, o exame desses dispositivos revela quatro limites estruturais que

impedem que a explicabilidade técnica, por si só, supere a função estabilizadora da opacidade algorítmica sobre os danos produzidos por sistemas de recomendação e moderação.

A literatura identifica ao menos quatro limites estruturais das abordagens de XAI quando aplicadas aos problemas de opacidade discutidos neste artigo.

1. **Insuficiência Epistêmica das Explicações** O primeiro limite diz respeito à **suficiência epistêmica** das explicações produzidas. Burr et al. (2018) argumentam que muitas explicações fornecidas por sistemas de XAI, embora tecnicamente corretas ao descrever correlações estatísticas ou padrões de decisão, permanecem epistemicamente insuficientes porque não tornam explícitos os **critérios normativos** que orientam as escolhas do sistema. Para ser epistemicamente adequada, uma explicação deve permitir que o agente afetado compreenda os fundamentos da decisão, avalie criticamente sua justificativa e disponha de elementos para contestá-la. Sem essa capacidade, a explicação reduz-se à descrição de um resultado, sem promover efetivamente a compreensão do processo decisório.
2. **Legitimação Institucional da Opacidade** O segundo limite refere-se ao risco de **legitimação institucional da opacidade**. Quando mecanismos de **XAI** são implementados sem alterações nas estruturas de governança e distribuição de poder, a explicabilidade pode funcionar como um instrumento de legitimação da própria opacidade que pretende combater. Gillespie (2018) demonstra que plataformas digitais frequentemente adotam relatórios de transparência e mecanismos formais de prestação de contas que satisfazem exigências regulatórias, mas **permanecem insuficientes** para revelar os critérios efetivos que orientam decisões de **moderação e visibilidade**. Nesses casos, a explicação produz uma aparência de transparência sem ampliar de maneira significativa a capacidade de escrutínio público.
3. **Perpetuação da Cognição Extraída** O terceiro limite, identificado por Loock (2025), relaciona-se ao fenômeno da **cognição extraída**. Segundo o autor, sistemas de IA não apenas ampliam capacidades cognitivas humanas, mas tendem a substituir progressivamente atividades associadas ao julgamento, à interpretação e à avaliação crítica. Aplicada à XAI, essa crítica sugere que explicações são insuficientes quando não restauram a **capacidade dos usuários de compreender, questionar e resistir** aos mecanismos de mediação algorítmica. Nesse sentido, a transparência não deve ser

concebida apenas como instrumento de prestação de contas institucional, mas também como condição para a preservação da autonomia **cognitiva dos sujeitos.**

4. **Escopo Restrito e Desalinhamento de Controle** O quarto limite diz respeito ao **escopo restrito** das propostas de XAI. Veluwenkamp (2025) argumenta que os problemas de responsabilidade em **sistemas de IA** não se resolvem apenas pela identificação das causas de uma decisão, mas dependem também da forma como o controle sobre esses sistemas é distribuído entre os diferentes atores envolvidos. O conceito de **desalinhamento de controle** descreve situações em que desenvolvedores e operadores concentram capacidade decisória, enquanto usuários, pesquisadores e reguladores permanecem sem instrumentos efetivos de intervenção. Assim, mesmo quando a opacidade epistêmica é parcialmente reduzida, persistem obstáculos institucionais que limitam a responsabilização e a correção dos danos produzidos.

Os limites da **XAI** tornam-se mais evidentes quando contrastados com frameworks de governança de IA, como o **AI Risk Management Framework do NIST** (2023), que associa transparência à visibilidade sobre **modelos, dados e responsabilidades**. Embora relevante, essa abordagem permanece centrada na gestão de riscos e não responde integralmente aos problemas de suficiência epistêmica e desalinhamento de controle discutidos anteriormente.

A transparência algorítmica, portanto, não pode ser reduzida à explicabilidade técnica. Os quatro limites discutidos, **insuficiência epistêmica das explicações, legitimação institucional da opacidade, perpetuação da cognição extraída e desalinhamento de controle**, indicam que os problemas associados aos sistemas de IA não podem ser enfrentados exclusivamente por mecanismos de interpretação de modelos. As respostas exigem intervenções simultâneas nos planos **técnico, institucional e político,** envolvendo mecanismos de **governança, contestabilidade e responsabilização** capazes de enfrentar a função estabilizadora da opacidade algorítmica.

2.4 Transparência, Democracia e Reconstrução do Espaço Público

As seções anteriores permitiram mapear o problema em suas dimensões constitutivas: a erosão do auditório universal pela **fragmentação discursiva (2.1)**, a opacidade algorítmica como **fenômeno sociotécnico deliberado (2.3)**, as **injustiças epistêmicas que ela produz (2.4)** e **os limites das respostas técnicas existentes (2.5)**. O que resta estabelecer, antes de avançar para a análise empírica e regulatória, é o enquadramento normativo que orientará a avaliação crítica desenvolvida nas seções seguintes.

Esse enquadramento parte de uma premissa habermasiana. Para Habermas (2023), a qualidade deliberativa da esfera pública não se constitui espontaneamente; depende de condições institucionais que precisam ser ativamente sustentadas. A manutenção de uma infraestrutura midiática deliberativa não é uma opção política entre outras, mas uma condição para a legitimidade democrática. Quando sistemas de **moderação e recomendação** comerciais substituem **gatekeepers** comprometidos com a mediação pública, essa infraestrutura é reorganizada segundo lógicas orientadas por interesses econômicos e métricas de desempenho que podem entrar em tensão com as exigências da deliberação democrática. Essa reorganização não é neutra, pois altera as condições sob as quais argumentos circulam, são avaliados e adquirem legitimidade pública.

Sob a perspectiva da epistemologia social institucional, as redes sociais podem ser compreendidas como instituições epistêmicas que organizam a **circulação de informações, a distribuição de credibilidade e as condições de acesso ao conhecimento.** Conforme situam Ruivo e Cichoski (2022), esse ramo da epistemologia social desloca o foco da avaliação epistêmica dos indivíduos para sistemas, práticas e instituições, investigando de que forma uma teoria epistêmica pode orientar sua crítica, construção e reforma. Esse enquadramento permite analisar as redes sociais como instituições cuja organização influencia as condições sob as quais informações circulam, opiniões adquirem credibilidade e agentes participam das práticas epistêmicas compartilhadas. Nessa perspectiva, a avaliação dessas redes não depende apenas de sua eficiência operacional, mas também dos efeitos que produzem sobre as condições de deliberação e de avaliação pública das informações, aproximando a epistemologia social institucional do diagnóstico habermasiano acerca da legitimidade do espaço público.

Ananny e Crawford (2018) permitem precisar o alcance normativo dessa constatação. A transparência, entendida como resposta à opacidade algorítmica, não se reduz à **abertura de códigos, à divulgação de parâmetros técnicos ou à explicabilidade de modelos individuais.** Trata-se, fundamentalmente, de garantir que os agentes afetados possam **compreender, contestar e participar** da definição dos critérios que orientam as decisões dos sistemas de recomendação e moderação. Essa distinção é decisiva para o argumento deste artigo porque desloca a transparência do plano estritamente técnico para os planos **político, institucional e epistêmico.** É precisamente esse deslocamento que as abordagens de **XAI** examinadas na seção anterior não conseguem realizar de forma satisfatória.

O que essa perspectiva abre, e que as seções seguintes desenvolvem, é a possibilidade de avaliar sistemas de moderação não apenas por sua e**ficiência operacional ou por sua conformidade regulatória**, mas pelo tipo de espaço público que produzem: se **criam ou destroem condições** para que a divergência ocorra dentro de um horizonte comum de inteligibilidade. É com base nesse critério e no diagnóstico de que a opacidade atua como mecanismo de estabilização dos danos produzidos por sistemas de **recomendação e moderação**, que **a análise empírica, os casos documentados e as diretrizes normativas** propostas neste artigo serão construídos.

3. ANÁLISE EMPÍRICA E REGULATÓRIA

As seções anteriores estabeleceram os fundamentos conceituais para analisar como sistemas de recomendação e moderação produzem fragmentação discursiva e apagamento epistêmico, e como a opacidade algorítmica estabiliza esses danos ao impedir sua identificação e correção. Esta seção examina evidências empíricas desses mecanismos e a trajetória do marco regulatório brasileiro, desde a **Estratégia Brasileira de Inteligência Artificial (EBIA) até o Plano Brasileiro de Inteligência Artificial (PBIA).**

3.1 Casos Empíricos de Moderação Assimétrica

Os casos a seguir ilustram como sistemas de moderação baseados em IA podem produzir efeitos discriminatórios concretos sobre a circulação de conteúdos e a participação discursiva. Conforme sintetizado no Quadro 2, a combinação entre automação em larga escala e opacidade decisória dificulta a identificação e a contestação desses efeitos pelos agentes afetados.

Injustiça Testemunhal Automatizada. Gillespie (2018) documenta situações em que políticas de moderação aparentemente neutras, como filtros de conteúdo restrito ou exigências relacionadas à identificação dos usuários, produziram efeitos desproporcionais sobre grupos historicamente marginalizados, incluindo criadores LGBTQ+. Esses casos evidenciam como sistemas de moderação podem reproduzir assimetrias de credibilidade e visibilidade já presentes no ambiente social, ilustrando empiricamente os mecanismos de injustiça testemunhal discutidos nesta seção.

No contexto brasileiro, Xavier et al. (2023), reunidos na **Coletânea TIC, Governança da Internet, Gênero, Raça e Diversidade (2023)**, analisaram a detecção de discursos violentos contra candidaturas negras e LGBTQ+ nas eleições de 2022 e identificaram limitações do modelo **BERTimbau** na identificação de agressões dirigidas a esses grupos. Os

resultados indicam que sistemas de moderação podem falhar na identificação de ataques direcionados a grupos historicamente vulnerabilizados, comprometendo a efetividade dos mecanismos de proteção discursiva.

Apagamento Hermenêutico Gerativo. O caso do sistema DeepSeek ilustra como a opacidade institucional operacionaliza aquilo que Mollema (2024, p. 2) define como a "automação do epistemicídio, a injustiça cometida contra agentes epistêmicos em sua capacidade de construção coletiva de sentido por meio da supressão de diferenças em epistemologia e conceitualização". Ao responder a questionamentos sobre o ativista Ai Weiwei com mensagens padronizadas de recusa, apresentando-se como um assistente projetado para fornecer respostas "úteis e inofensivas", o modelo executa uma supressão de conteúdos politicamente sensíveis baseada em instruções institucionais incorporadas ao sistema. Esse processo reduz a disponibilidade de recursos conceituais necessários para interpretar e discutir experiências de resistência e dissidência política, produzindo formas de silenciamento que dificultam que diferentes posições sejam submetidas ao escrutínio e à justificação pública, comprometendo o ideal do auditório universal.

Ignorância Hermenêutica e Colonialidade. Andrade e Gobbi (2025) demonstram que sistemas generativos de imagem podem reproduzir representações estereotipadas de populações historicamente marginalizadas e de perspectivas produzidas em contextos periféricos, transformando essas experiências em imagens padronizadas e facilmente consumíveis. Segundo os autores, tais sistemas participam de um "sistema de governança sensível no qual a diferença cultural é transformada em template, recurso e mercadoria" (ANDRADE; GOBBI, 2025, p. 146). Esse processo decorre de vieses presentes nos dados de treinamento e da predominância de referências culturais ocidentais utilizadas na construção dos modelos. Como consequência, perspectivas localizadas passam a ser tratadas como desvios em relação a um padrão implícito, reduzindo as possibilidades de representação adequada dessas experiências no espaço público digital e dificultando sua consideração em processos públicos de discussão e avaliação.

Quadro 2: ***Casos empíricos ilustrativos de moderação algorítmica***

Caso	Mecanismo de Opacidade	Tipo de Injustiça	Dano ao Auditório Universal	Fonte
Moderação assimétrica de grupos marginalizados (YouTube/Facebook)	Aplicação automatizada de critérios opacos de visibilidade e remoção.	**Injustiça Testemunhal Automatizada**	Redução da pluralidade discursiva e assimetrias de credibilidade	**Gillespie (2018); Xavier et al. (2023)**
DeepSeek e supressão de conteúdos politicamente sensíveis	Inistruções institucionais incorporadas ao sistema e inacessíveis aos usuários.	**Apagamento Hermenêutico Gerativo**	Restrição de Recursos conceituais para interpretar experiências de dissidência e resistência	**Mollema (2024)**
Sistemas generativos de imagem e colonialidade do sensível	Universalização de referências ocidentais presentes nos dados de treinamento.	**Ignorância Hermenêutica e Colonialidade**	Invisibilização de perspectivas locais e erosão do reconhecimento intersubjetivo	**Andrade e Gobbi (2025)**

Fonte: *Elaborado pelo autor com base em Gillespie (2018), Mollema (2024) e Kay, Kasirzadeh e Mohamed (2024).*

3.2 Análise Documental: Da EBIA ao PBIA

A trajetória regulatória brasileira evidencia a transição de um modelo que enuncia princípios éticos gerais, sem instituir mecanismos efetivos de governança, fiscalização ou responsabilização, para uma abordagem progressivamente orientada à gestão de riscos, à transparência e à capacidade institucional. Como demonstra o Quadro 3, essa evolução reflete o reconhecimento de que a transparência, enquanto princípio normativo, é insuficiente quando desacompanhada de instrumentos concretos de implementação e controle.

EBIA (2021) e PL 5051/2019: A Estratégia Brasileira de Inteligência Artificial adotou a matriz FAT, **Equidade, Responsabilização e Transparência**, como referência normativa, mas sem operacionalizar esses valores por meio de mecanismos concretos de fiscalização, contestação ou responsabilização. Na mesma direção, o PL 5051/2019, declarado prejudicado em dezembro de 2024, limitava-se à enunciação de princípios gerais, ilustrando o que Veluwenkamp (2025) denomina **lacuna epistêmica de responsabilidade**: a transparência é afirmada como princípio, mas não são instituídos instrumentos capazes de viabilizar o escrutínio público, a atribuição de responsabilidade e a correção de decisões automatizadas.

PL 2338/2023: O projeto representa um avanço substantivo ao adotar uma abordagem regulatória baseada em categorias de risco. Entre suas principais inovações estão o direito à explicação e à revisão humana para decisões automatizadas de alto risco, bem como a definição de mecanismos institucionais de supervisão. Com isso, procura reduzir a lacuna de responsabilidade identificada anteriormente ao transformar princípios gerais em instrumentos concretos de governança e accountability.

PBIA (2025): O Plano Brasileiro de Inteligência Artificial amplia esse movimento ao incorporar uma dimensão estratégica voltada à soberania tecnológica. Por meio do eixo **IA Soberana e Modelos Nacionais**, propõe investimentos no desenvolvimento de modelos de linguagem treinados com corpora nacionais, infraestrutura computacional própria e sistemas abertos à auditoria pública. Sob essa perspectiva, o PBIA oferece uma resposta institucional ao risco de apagamento hermenêutico discutido por Mollema (2024), buscando fortalecer a autonomia epistêmica nacional diante da concentração tecnológica das grandes plataformas internacionais.

Quadro 3: ***Comparação dos marcos regulatórios de IA***

Dimensão Regulatória	PL 5051/2019	EBIA (2021)	PL 2338/2023	AI RMF NIST (2023)	EU AI Act (2024)	PBIA (2025)
Modelo regulatório	Princípios de soft regulation	Estratégia de política pública	Baseado em riscos e responsabilidades	Framework voluntário orientado por resultados	Regulação vinculante baseada em riscos	Plano de investimento estratégico
Natureza jurídica	PL arquivado em 2024	Documento não vinculante	PL em tramitação	Framework não prescritivo	Regulamento em vigor desde julho de 2024	Plano de ação governamental
Categorização de riscos	Ausente	Ausente	Presente (alto, limitado e mínimo)	Orientações qualitativas	Quatro categorias obrigatórias	Ausente, com foco em investimentos
Transparência algorítmica	Princípio geral sem operacionalização	Matriz FAT como referência	Obrigações específicas por nível de risco	Características de IA confiável e voluntária	Obrigações detalhadas e vinculantes	Direito à explicação como diretriz ética
Supervisão humana	Princípio geral	Não especificada	Proporcional ao risco	Função de governança voluntária	Obrigatória para sistemas de alto risco	Princípio ético geral
Accountability (Responsabilização)	Ausente	Matriz FAT sem operacionalização	Prevista para desenvolvedores e operadores	Função de governança voluntária	Obrigatória ao longo de toda a cadeia de valor	Princípio ético geral
Órgão fiscalizador	Não definido	Não definido	Previsto, com competências definidas	Não aplicável (modelo voluntário)	Autoridades nacionais e o Escritório Europeu de IA	CCT/MCTI/ CGEE – coordenação estratégica
Soberania epistêmica	Ausente	Composição diversificada de equipes	Proteção contra discriminação	Gestão voluntária de vieses	Proibição de sistemas de risco inaceitável	LLMs nacionais como pilar estratégico central
Sanções	Ausente	Não aplicável	Previstas em regulamentação complementar	Não aplicável	Multas de até 35 milhões de euros ou 7% do faturamento global	Não aplicável

Fonte: *Elaborado pelo autor com base em MCTI (2021), PL 5051/2019, NIST (2023), PL 2338/2023, União Europeia (2024), MCTI; CGEE (2025) e Veluwenkamp (2025).*

A evolução do marco regulatório brasileiro evidencia que a resposta aos problemas da opacidade algorítmica exige mais do que mecanismos de explicabilidade técnica. À medida que incorpora instrumentos de governança, responsabilização e fortalecimento da capacidade

institucional, a política pública aproxima-se de uma concepção de transparência orientada à preservação das condições de escrutínio, contestação e legitimidade do espaço público digital.

4. DISCUSSÃO E PROPOSIÇÃO DE DIRETRIZES

4.1 Opacidade Algorítmica e Estabilização dos Danos

A análise conjunta dos casos empíricos e do marco regulatório permite compreender como os danos produzidos por sistemas de recomendação e moderação, especialmente a **fragmentação discursiva e o apagamento epistêmico**, tendem a tornar-se persistentes. A opacidade desses sistemas não constitui a causa primária desses danos, mas o mecanismo que dificulta sua identificação, atribuição de responsabilidade e contestação pública, estabilizando seus efeitos ao longo do tempo.

Os casos examinados no capítulo 3 confirmam que os danos produzidos por **sistemas de recomendação e moderação** não decorrem apenas de falhas pontuais de implementação, mas de mecanismos estruturais de organização do espaço público digital. A reprodução de vieses históricos, a supressão de recursos conceituais e a concentração do poder de decisão em sistemas opacos dificultam a identificação, a atribuição de responsabilidade e a contestação dos danos produzidos. Nesse contexto, a opacidade algorítmica atua como o estabilizador estrutural desses processos, evidenciando que respostas baseadas exclusivamente na explicabilidade técnica ou na enunciação de princípios normativos são insuficientes. A mitigação desses problemas exige intervenções que atuem simultaneamente sobre a governança dos sistemas, a distribuição do controle e as condições de contestabilidade pública.

4.2 Avaliação Crítica das Respostas Existentes e Proposição de Diretrizes

4.2.1 Insuficiências das Respostas Técnicas

As evidências examinadas no Capítulo 3 confirmam que os limites das propostas de Inteligência Artificial Explicável (XAI) manifestam-se de forma concreta na mediação algorítmica do espaço público. Embora ampliem a interpretabilidade de decisões específicas, tais mecanismos permanecem incapazes de enfrentar os fatores institucionais e epistêmicos que sustentam os processos de recomendação e moderação. Nesse contexto, a opacidade algorítmica atua como o **estabilizador estrutural** dos danos produzidos por esses sistemas.

A insuficiência epistêmica das explicações (BURR et al., 2018) torna-se evidente nos casos de moderação assimétrica analisados por Xavier et al. (2023). Mesmo que um sistema indique quais atributos influenciam determinada classificação, isso não permite compreender

nem contestar os vieses históricos codificados nos dados de treinamento. Como demonstrado no caso do BERTimbau, a transparência técnica descreve o funcionamento do modelo, mas não torna transparentes os **critérios normativos** incorporados ao desenho e ao treinamento do sistema.

Da mesma forma, a disponibilização de mecanismos de transparência pode atuar como forma de **legitimação simbólica da opacidade** (GILLESPIE, 2018). Relatórios agregados e justificativas padronizadas atendem formalmente às exigências de prestação de contas sem alterar a concentração do poder editorial nem oferecer condições efetivas de contestação pelos usuários.

Os casos analisados também confirmam que a explicabilidade técnica não enfrenta o **apagamento hermenêutico** nem a **cognição extraída.** O exemplo do DeepSeek evidencia que decisões institucionais incorporadas ao sistema permanecem protegidas pela opacidade organizacional e apresentadas sob uma ***"visão de lugar nenhum"*** (MOLLEMA, 2024), mascarando escolhas políticas como se fossem respostas neutras do modelo. Paralelamente, a dependência crescente de processos automatizados reduz a capacidade dos usuários de exercer julgamento independente (LOOCK, 2025). Ampliar a interpretabilidade matemática não restaura, isoladamente, a autonomia epistêmica dos sujeitos.

Por fim, a XAI permanece limitada ao nível do modelo, sem enfrentar o **desalinhamento de controle** (VELUWENKAMP, 2025). Ainda que decisões específicas se tornem explicáveis, as responsabilidades sobre o desenho, a operação e a supervisão dos sistemas permanecem indefinidas. A convergência entre as evidências empíricas e a trajetória regulatória brasileira demonstra que a transparência técnica constitui condição necessária, mas insuficiente, para assegurar accountability e contestabilidade, exigindo respostas institucionais que incluam mecanismos de governança, redistribuição do controle e, no contexto brasileiro, estratégias voltadas à soberania tecnológica.

4.2.2 Insuficiências das Respostas Regulatórias

A análise da trajetória regulatória brasileira demonstra que a incorporação de princípios éticos, embora necessária, não garante sua efetividade institucional. A evolução observada entre a EBIA, o PL 5051/2019, o PL 2338/2023 e o PBIA revela um deslocamento progressivo de uma regulação predominantemente principiológica para instrumentos orientados por **governança, responsabilização e capacidade estatal.** Ainda assim,

permanecem desafios relevantes para **assegurar transparência, contestabilidade e controle efetivo sobre sistemas de IA.**

Em primeiro lugar, a simples previsão normativa de transparência e responsabilização não elimina a **lacuna epistêmica de responsabilidade** identificada por Veluwenkamp (2025). Sem mecanismos de **rastreabilidade, auditoria e atribuição clara de competências**, permanece difícil identificar quem deve responder pelos danos produzidos por sistemas de recomendação e moderação, limitando a efetividade prática dos direitos assegurados.

Em segundo lugar, a proteção da esfera pública exige mais do que a enunciação de direitos individuais. Conforme argumenta Habermas (2023), a deliberação democrática depende da existência de uma **infraestrutura institucional** capaz de sustentar condições compartilhadas de inteligibilidade, publicidade e contestação. A regulação precisa, portanto, criar condições para que os critérios de funcionamento dos sistemas possam ser compreendidos e questionados pelos diferentes atores sociais, e não apenas conhecidos por especialistas ou autoridades reguladoras.

A trajetória legislativa brasileira reflete o reconhecimento progressivo dessas limitações. O PL 2338/2023 representa um avanço substantivo ao introduzir a classificação por categorias de risco, a obrigatoriedade de avaliação de impacto de sistemas de IA e o direito à explicação para decisões automatizadas. Complementarmente, o PBIA (2025) desloca o debate para o plano estratégico por meio do eixo de **IA Soberana e Modelos Nacionais.** Ao incentivar o desenvolvimento de modelos treinados com corpora de dados locais e ampliar a capacidade nacional de auditoria e governança, o plano oferece uma resposta institucional aos riscos de apagamento hermenêutico e colonialidade algorítmica discutidos ao longo deste trabalho. A experiência brasileira demonstra, assim, que a transparência efetiva depende não apenas de conformidade normativa, mas também de capacidade institucional, governança e soberania tecnológica.

4.2.3 Diretrizes Éticas para Sistemas de Moderação Algorítmica

A análise desenvolvida nas seções anteriores evidenciou que as insuficiências das **respostas técnicas e regulatórias** decorrem de uma concepção restrita de transparência, concentrada na explicabilidade de modelos individuais e insuficiente para enfrentar os mecanismos institucionais e epistêmicos que **estabilizam os danos** produzidos por **sistemas de moderação.** Com base nesse diagnóstico, propõem-se seis diretrizes éticas destinadas a

orientar o desenvolvimento, a governança e a avaliação de sistemas de moderação mais transparentes e democraticamente legítimos. O **Quadro 4** sintetiza a correspondência entre os principais limites identificados nas propostas atuais de Inteligência Artificial Explicável (XAI) e as diretrizes éticas formuladas neste trabalho.

Quadro 4: ***Limites da XAI e diretrizes éticas correspondentes.***

Limite da XAI	Natureza do Problema	Diretriz Ética Correspondente	Fundamento Teórico
Injustiça epistêmica das explicações	Explicações tecnicamente corretas, mas socialmente incompreensíveis, aparecem sob a aparência de inteligibilidade sem efetiva compreensão.	**Diretriz 1: Transparência Epistemicamente Suficiente:** as explicações devem criar condições para a formação de crenças verdadeiras por todos os agentes afetados.	**Burt et al. (2018); Dantas e Oliveira et al. (2024)**
Legitimação simbólica	A XAI é utilizada como instrumento de legitimação retórica da opacidade, produzindo uma aparência de responsabilização sem possibilidade real de contestação	**Diretriz 2: Contestabilidade Estrutural**: mecanismos robustos de contestação devem ser tão relevantes quanto a própria explicação.	**Gillespie (2018); Veluwekamp (2025)**
Extração cognitiva	Explicações que não produzem capacidade de julgamento independente aprofundam a substituição progressiva da agência epistêmica.	**Diretriz 1: Transparência Epistemicamente Suficiente:** preservar a autonomia cognitiva, e não apenas cumprir um requisito formal.	**Loock (2025)**
Escopo restrito ao modelo individua	A XAI não evidencia a distribuição inadequada de controle entre desenvolvedores, plataformas, reguladores e usuários.	Diretriz 3: **Distribuição Adequada do Controle:** auditoria externa, acesso regulatório e participação de usuários.	**Veluwekamp (2025)**
Ausência de redes epistêmicas	Nenhuma proposta de XAI aborda explicitamente as injustiças epistêmicas produzidas por sistemas sociotécnicos complexos.	**Diretriz 4: Representação Plural e Redes Epistêmicas:** mecanismos ativos de verificação e correção de injustiças.	**Mollema (2024); Kay et al. (2024)**
Ausência de infraestrutura deliberativa	A XAI não cria condições para o debate público sobre critérios algorítmicos, transferindo a discussão para especialistas e reduzindo a participação democrática.	**Diretriz 5: Infraestrutura Deliberativa:** plataformas reguladas como infraestruturas públicas, com obrigações proporcionais ao poder normativo.	**Habermas (2022); Gillespie (2018)**
Ausência de responsabilização humana cultivada	Sistemas explicáveis continuam operando em ambientes institucionalmente irresponsáveis ou sem mecanismos adequados de prestação de contas.	**Diretriz 6:** Cultivo de Virtudes Epistêmicas Institucionais.	**Constantinescu et al. (2021)**

Fonte: *Elaborado pelo autor com base em Burr et al. (2018), Loock (2025), Veluwenkamp (2025), Mollema (2024), Habermas (2022), Gillespie (2018) e Constantinescu et al. (2021).*

Diretriz 1 – Transparência para a Agência Epistêmica: Sistemas de moderação devem produzir explicações **inteligíveis e utilizáveis** por todos os agentes afetados,

independentemente de sua capacitação técnica. A transparência não deve limitar-se à descrição do funcionamento dos modelos, mas deve fornecer **elementos suficientes** para que os usuários compreendam, avaliem criticamente e contestem as decisões automatizadas que os afetam. Conforme argumentam Burr et al. (2018), explicações eficazes devem **fortalecer a capacidade dos indivíduos de exercer controle e supervisão** sobre sistemas de IA, preservando sua agência epistêmica no espaço público digital.

Diretriz 2 – Contestabilidade Estrutural: Os sistemas devem incorporar mecanismos robustos de contestação que permitam questionar decisões específicas, exigir justificativas fundamentadas e obter revisão por agentes humanos qualificados. A contestabilidade não constitui apenas um recurso remedial, mas uma condição de legitimidade democrática capaz de reduzir os efeitos da **cognição extraída** (Loock, 2025) e de enfrentar as lacunas epistêmicas de responsabilidade identificadas por Veluwenkamp (2025).

Diretriz 3 – Distribuição Adequada do Controle: O desenho institucional dos sistemas deve impedir o monopólio do controle pelas **plataformas digitais**. Isso requer mecanismos concretos, como auditorias independentes, acesso regulatório a dados e parâmetros e canais institucionalizados de participação dos usuários na definição dos critérios de moderação e recomendação, reconhecendo o desalinhamento de controle (Veluwenkamp, 2025) como um problema de governança e de design institucional.

Diretriz 4 – Representação Plural e Redress Epistêmico: As plataformas devem promover representação plural nos **datasets de treinamento** e realizar auditorias periódicas voltadas à identificação e correção ativa de injustiças epistêmicas. O foco deve ser a reparação de formas de apagamento hermenêutico que comprometem a participação de grupos historicamente marginalizados e empobrecem a pluralidade de perspectivas no espaço público.

Diretriz 5 – Preservação da Infraestrutura Deliberativa: Redes sociais que desempenham funções normativas sobre o espaço público devem ser reguladas como infraestruturas deliberativas, assumindo obrigações de proteção à integridade da informação proporcionais ao seu poder editorial e ao impacto de suas decisões sobre a circulação do discurso democrático. Para Habermas (2023, p. 36), a manutenção dessas condições não constitui um aspecto acessório, pois a política deliberativa em sociedades pluralistas representa um "pré-requisito para a existência de qualquer democracia digna deste nome".

Diretriz 6 – Cultivo de Virtudes Epistêmicas Institucionais: A operação dos sistemas deve ser orientada pelo desenvolvimento de virtudes nos agentes humanos responsáveis por seu planejamento, implementação e supervisão. Destaca-se a phronesis aristotélica como a sabedoria prática necessária para o exercício da responsabilidade moral diante de situações concretas. Como argumenta Constantinescu et al. (2021, p. 807), "a escolha moral correta exige a experiência de situações particulares, uma vez que regras gerais não podem ser aplicadas mecanicamente a situações específicas". O cultivo dessa sabedoria prática busca assegurar que a responsabilidade institucional não seja reduzida ao cumprimento formal de requisitos técnicos.

5. CONSIDERAÇÕES FINAIS

A investigação desenvolvida corrobora a hipótese de que a opacidade algorítmica não atua como causa isolada da fragmentação discursiva e das injustiças produzidas por sistemas de recomendação e moderação, mas como o estabilizador estrutural desses danos. A articulação entre o referencial teórico, os casos empíricos analisados e a evolução do marco regulatório brasileiro indica que a opacidade dificulta a identificação, a contestação e a correção dos efeitos decorrentes desses sistemas. Esse resultado desloca a compreensão da transparência algorítmica, que deixa de ser entendida apenas como um requisito de explicabilidade para assumir uma função epistêmica, institucional e democrática indispensável à preservação das condições de escrutínio e responsabilização das decisões automatizadas no espaço público digital.

No plano analítico, o artigo argumenta que a legitimidade do ambiente informacional contemporâneo depende da superação daquilo que Veluwenkamp (2025) denomina lacuna epistêmica de responsabilidade. A articulação entre teoria da argumentação, teoria democrática, epistemologia social e ética da inteligência artificial evidencia que sistemas com poder de organizar a circulação do discurso público não podem permanecer imunes ao escrutínio democrático. Nesse contexto, as seis diretrizes éticas propostas oferecem um quadro orientador para fortalecer a transparência, a contestabilidade, a adequada distribuição do controle, a representação plural e às responsabilidades institucionais na governança da inteligência artificial, respondendo às limitações identificadas tanto nas abordagens de Inteligência Artificial Explicável (XAI) quanto nos modelos regulatórios baseados predominantemente na enunciação de princípios.

Embora este estudo se limite a uma investigação teórica e documental, sem a realização de auditorias técnicas de sistemas de inteligência artificial, seus resultados apontam para a necessidade de pesquisas voltadas à validação empírica das diretrizes propostas e ao aprofundamento das discussões sobre governança, capacidade institucional e desenvolvimento tecnológico em diferentes contextos sociais e culturais. A experiência brasileira, da Estratégia Brasileira de Inteligência Artificial (EBIA) ao Plano Brasileiro de Inteligência Artificial (PBIA), reforça que a transparência efetiva depende não apenas da formulação de princípios, mas também de mecanismos concretos de implementação, responsabilização e fortalecimento institucional. Em última instância, a transparência algorítmica deve ser compreendida como uma condição constitutiva da democracia contemporânea. Sem mecanismos que permitam compreender, contestar e responsabilizar decisões automatizadas que organizam a circulação do discurso público, tornam-se inviáveis a contestação, a deliberação e a própria legitimidade das instituições democráticas.

REFERÊNCIAS

ANDRADE, R.; GOBBI, M. Estéticas computacionais e colonialidade algorítmica. 2025**.**
ANANNY, M.; CRAWFORD, K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. **New Media & Society,** Londres, v. 20, n. 3, p. 973-989, mar. 2018.
BRASIL. Ministério da Ciência, Tecnologia e Inovações. **Estratégia Brasileira de Inteligência Artificial (EBIA).** Brasília: MCTI, 2021.
BRASIL. Ministério da Ciência, Tecnologia e Inovação; CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICO**S. Plano Brasileiro de Inteligência Artificial (PBIA): IA para o Bem de Todos.** Brasília: MCTI; CGEE, 2025.
BRASIL. Senado Federal. **Projeto de Lei n.º 5051, de 2019.** Estabelece os princípios para o uso da Inteligência Artificial no Brasil. Brasília, DF: Senado Federal, 2019.
BRASIL. Senado Federal. **Projeto de Lei n.º 2338, de 2023.** Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Senado Federal, 2023.
BURR, C.; CRISTIANINI, N.; LADYMAN, J. An analysis of the interaction between intelligent software agents and human users**. Minds and Machines,** Dordrecht, v. 28, n. 4, p. 735-774, dez. 2018.
CASSAM, Q. Vices of the Mind: from the intellectual to the political. Oxford: Oxford University Press, 2019.
COLETÂNEA TIC, Governança da Internet, Gênero, Raça e Diversidade. **São Paulo: Comitê Gestor da Internet no Brasil, 2023.**
CONSTANTINESCU, M.; VOINEA, C.; USZKAI, R.; VICĂ, C. Understanding responsibility in Responsible AI: dianoetic virtues and the hard problem of context. **Ethics and Information Technology,** Dordrecht, v. 23, n. 4, p. 803-814, dez. 2021.
DANAHER, J.; SÆTRA, H. S. Technology and moral change: the transformation of truth and trust**. Ethics and Information Technology,** Dordrecht, v. 24, n. 35, ago. 2022.
FRANKLIN, Benjamin Luiz; FERNANDES, Rogerio Paulo Müller; TRIQUES, Maria Lígia. Análise de Conteúdo e Modelos Linguísticos de Larga Escala (LLMs): **Possibilidades e Tensões. ISKO Brasil, Canela, RS, 2025**
FRICKER, M. Epistemic Injustice: power and the ethics of knowing. Oxford: Oxford University Press, 2007.
GILLESPIE, T. Custodians of the Internet: platforms, content moderation, and the hidden decisions that shap**e social media. New Haven: Yale University Press, 2018.**
HABERMAS, J. Reflexões e hipóteses sobre uma renovada mudança estrutural da esfera pública política. In: HABERMAS, J. Uma nova mudança estrutural da esfera pública e a política deliberativa. Tradução de Denilson Luís Werle. São Paulo: Editora Unesp Digital, 2023. p. 13-98.
HOPSTER, J. What are socially disruptive technologies? **Technology in Society,** Amsterdã, v. 67, art. 101750, nov. 2021.
KAY, J.; KASIRZADEH, A.; MOHAMED, S. Epistemic injustice in generative AI. **arXiv,** Ithaca, 21 ago. 2024. Disponível em: https://doi.org/10.48550/ARXIV.2408.11441.
LIBMAN, Juliana. Moderação de conteúdo em redes sociais: por uma regulação que promova a liberdade de expressão. 2023. 137 f. **Dissertação (Mestrado em Direito Civil Contemporâneo e Prática Jurídica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2023**
LOOCK, M. The extracted mind**. Synthese,** Dordrecht, v. 205, art. 133, 2025**.**

LYOTARD, J.-F. The Postmodern Condition: a report on knowledge. Tradução de Geoff Bennington e Brian Massumi. Minneapolis: University of Minnesota Press, 1984**.**

MOLLEMA, W. J. T. A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. **Preprint, 2024.** Disponível em: https://doi.org/10.5281/ZENODO.10155340.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg: U.S. Department of Commerce, jan. 2023.

OLIVEIRA, Rogel Esteves de; ETCHEVERRY, K. M.; RODRIGUES, T. V.; SARTORI, C. A. (org.). Compêndio de Epistemologia. Porto Alegre: Editora Fi, 2022.

PASQUALE, F. The Black Box Society: the secret algorithms that control money and information. Cambridge: Harvard University Press, 2015.

PEELS, Rik. Ignorance: A Philosophical Study. Nova York: Oxford University Press, 2023.

PERELMAN, C.; OLBRECHTS-TYTECA, L. Tratado da Argumentação: a nova retórica. Tradução de Maria Ermantina de Almeida Prado Galvão. 2. ed. São Paulo: Martins Fontes, 2005.

RUIVO, J. L.; CICHOSKI, L. P. Epistemologia social. In: OLIVEIRA, Rogel Esteves de et al. (org.). **Compêndio de Epistemologia.** Porto Alegre: Editora Fi, 2022. p. 445-456.

SUNSTEIN, C. R. #Republic: divided democracy in the age of social media. Princeton: Princeton University Press, 2017.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024. Jornal Oficial da União Europeia, Bruxelas, L 1689, 12 jul. 2024.

VELUWENKAMP, Herman. What responsibility gaps are and what they should be. **Ethics and Information Technology,** Dordrecht, v. 27, n. 14, 26 fev. 2025.

XAVIER, F. et al. Detecção de discurso violento contra candidaturas negras e LGBTQ+ nas eleições de 2022. In: **COLETÂNEA TIC,** Governança da Internet, Gênero, Raça e Diversidade. São Paulo: CGI.br, 2023.

ZHANG, Z.; LUO, L. Hate speech detection: **A solved problem? the challenging case of long tail on twitter. Semantic Web, IOS Press, v. 10, n. 5, p. 925–945, 2019**