

ETHOS-SE: UM FRAMEWORK SOCIO-TÉCNICO PARA MITIGAÇÃO DA DESINFORMAÇÃO POR DEEPPAKES

Gabriel Juceni Pinheiro de Lima¹
Gleisson Amaral Mendes²

RESUMO: O avanço da Inteligência Artificial generativa tem ampliado a produção e disseminação de conteúdos sintéticos, tornando os *deepfakes* um dos principais desafios para a confiabilidade da informação em ambientes digitais. Embora a literatura apresente avanços significativos no desenvolvimento de técnicas de detecção automatizada, observa-se que aspectos relacionados à proveniência dos conteúdos, à transparência das decisões algorítmicas, à acessibilidade digital e à integração dessas dimensões ao desenvolvimento de software ainda permanecem pouco explorados de forma conjunta. Nesse contexto, este estudo tem como objetivo propor o framework socio-técnico Ethical Technical Handling of Synthetic Entities (ETHOS-SE), concebido sob a perspectiva da Engenharia de Software para apoiar a mitigação da desinformação baseada em *deepfakes*. Metodologicamente, a pesquisa caracteriza-se como qualitativa, exploratória e propositiva, fundamentada em revisão narrativa estruturada da literatura e análise bibliométrica exploratória da produção científica indexada na base Scopus. A análise permitiu identificar tendências da literatura, evidenciando a predominância de abordagens centradas na detecção automatizada e a existência de lacunas relacionadas à proveniência, explicabilidade e acessibilidade. Com base nessas evidências, foi desenvolvido o framework ETHOS-SE, estruturado em quatro camadas complementares: Proveniência, Detecção e Análise, Explicabilidade e Acessibilidade com Fricção Cognitiva. A principal contribuição do estudo consiste na proposição de uma arquitetura conceitual que integra requisitos de rastreabilidade, transparência, interpretabilidade e acessibilidade, ampliando as abordagens tradicionais de mitigação da desinformação e fornecendo diretrizes para o desenvolvimento de sistemas mais confiáveis, transparentes e centrados no usuário.

Palavras-chave: *Deepfakes*. Desinformação. Engenharia de Software. Inteligência Artificial Explicável. Acessibilidade Digital. Framework Socio-Técnico.

¹Graduando em Engenharia de Software, Universidade do Estado do Pará – UEPA.

²Professor Orientador, Universidade do Estado do Pará – UEPA.

ABSTRACT: The advancement of generative Artificial Intelligence has expanded the production and dissemination of synthetic content, making deepfakes a major challenge to information reliability in digital environments. Although the literature reports significant progress in automated detection techniques, content provenance, algorithmic transparency, digital accessibility, and the integration of these dimensions into software development remain insufficiently explored as a unified problem. This study proposes the socio-technical framework Ethical Technical Handling of Synthetic Entities (ETHOS-SE), conceived from a Software Engineering perspective to support the mitigation of deepfake-based disinformation. Methodologically, the research is qualitative, exploratory, and propositional, grounded in a structured narrative literature review and an exploratory bibliometric analysis of scientific production indexed in Scopus. The analysis identified a predominance of approaches centered on automated detection and gaps related to provenance, explainability, and accessibility. Based on these findings, ETHOS-SE was organized into four complementary layers: Provenance, Detection and Analysis, Explainability, and Accessibility with Cognitive Friction. The main contribution is a conceptual architecture that integrates traceability, transparency, interpretability, and accessibility requirements, extending traditional approaches to disinformation mitigation and providing guidelines for the development of more reliable, transparent, and user-centered systems.

Keywords: Deepfakes. Disinformation. Software Engineering. Explainable Artificial Intelligence. Digital Accessibility. Socio-Technical Framework.

1 INTRODUÇÃO

A consolidação da Inteligência Artificial generativa como tecnologia de amplo acesso tem redefinido as dinâmicas de produção e circulação de informação no ambiente digital. Sistemas capazes de gerar textos, imagens, áudios e vídeos sintéticos com elevado grau de realismo passaram, em poucos anos, de experimentos laboratoriais para ferramentas amplamente disponíveis ao público (WESTERLUND, 2019). Entre essas aplicações, destacam-se os deepfakes, vídeos manipulados ou totalmente gerados por modelos de aprendizado profundo, capazes de reproduzir expressões faciais, entonações de voz e movimentos corporais com precisão cada vez maior (KORSHUNOV; MARCEL, 2018).

A evolução técnica dessas ferramentas reduziu significativamente as barreiras de entrada para a criação de conteúdos sintéticos. O que antes exigia conhecimento especializado em edição audiovisual e processamento de imagens hoje pode ser realizado por meio de interfaces automatizadas e serviços online. Esse processo de democratização tecnológica, embora represente avanços criativos e industriais, também amplia os riscos associados ao uso indevido da tecnologia. A facilidade de geração de vídeos hiper-realistas cria um cenário no qual conteúdos falsos podem ser produzidos em escala, distribuídos rapidamente e consumidos sem verificação adequada (CHESNEY; CITRON, 2019).

No campo da desinformação digital, os deepfakes introduzem uma mudança qualitativa relevante. Diferentemente de textos manipulados ou imagens editadas de forma rudimentar, vídeos sintéticos exploram a confiança histórica depositada no registro audiovisual como evidência da realidade. Ao simular falas e ações de figuras públicas ou cidadãos comuns, essas mídias comprometem a percepção de autenticidade e fragilizam os mecanismos tradicionais de validação informacional. O resultado é a intensificação de um ambiente marcado por incerteza, polarização e erosão da confiança institucional (VACCARI; CHADWICK, 2020).

Esse fenômeno é particularmente crítico quando considerado sob a perspectiva dos usuários finais. A assimetria entre a capacidade técnica de geração de mídia sintética e a capacidade humana de detecção cria uma condição de vulnerabilidade informacional. A maioria dos indivíduos não dispõe de ferramentas técnicas, conhecimento especializado ou tempo suficiente para analisar a veracidade de cada conteúdo consumido. Os públicos com baixo letramento digital ou com limitações cognitivas podem ser ainda mais suscetíveis à aceitação acrítica de vídeos manipulados (LAZER *et al.*, 2018).

Grande parte das iniciativas atuais foca no desenvolvimento de algoritmos de detecção automática baseados em aprendizado de máquina e análise forense digital. Apesar de essenciais, essas abordagens apresentam limitações: há uma constante corrida tecnológica entre técnicas de geração e de detecção, e muitos sistemas operam como caixas-pretas, sem explicações claras. Essa opacidade reduz a confiança do usuário e dificulta o uso consciente das ferramentas (ADADI; BERRADA, 2018; RÖSSLER *et al.*, 2019).

Torna-se insuficiente tratar a redução da desinformação por deepfakes exclusivamente como um problema algorítmico. A complexidade do fenômeno exige uma abordagem socio-técnica, na qual soluções tecnológicas sejam articuladas com princípios éticos, mecanismos de transparência e estratégias de comunicação acessível. Em vez de atuar apenas na identificação posterior de conteúdos manipulados, essa abordagem propõe incorporar mecanismos de autenticidade, rastreabilidade e transparência diretamente ao ecossistema de desenvolvimento e distribuição de software. Trata-se de deslocar o foco da reação isolada para a construção preventiva de ambientes digitais mais confiáveis.

Nesse contexto, o objetivo deste artigo é propor o framework socio-técnico *Ethical Technical Handling of Synthetic Entities* (ETHOS-SE), concebido para integrar mecanismos de proveniência de conteúdo, explicabilidade e acessibilidade ao ciclo de vida de desenvolvimento de software. O framework busca reduzir a vulnerabilidade informacional dos usuários,

promover maior transparência na detecção de mídias sintéticas e fortalecer a confiança na autenticidade de conteúdos digitais por meio de uma abordagem estruturada de Engenharia de Software.

2 METODOLOGIA

2.1 CARACTERIZAÇÃO DA PESQUISA

Esta pesquisa caracteriza-se como qualitativa, de natureza exploratória e propositiva, tendo como objetivo desenvolver um framework socio-técnico para mitigação da desinformação por *deepfakes* sob a perspectiva da Engenharia de Software e da acessibilidade digital. O estudo não se concentra na proposição de novos algoritmos de detecção, mas na concepção de um modelo arquitetural capaz de integrar mecanismos de autenticação de conteúdo, análise automatizada, explicabilidade e acessibilidade em uma abordagem unificada.

Quanto aos objetivos, a pesquisa é classificada como exploratória, por buscar compreender um fenômeno recente e em constante evolução, identificando lacunas na literatura relacionadas à integração entre soluções técnicas de detecção de *deepfakes* e aspectos voltados à transparência, comunicação compreensível e interação com os usuários. O caráter propositivo decorre da elaboração do framework Ethical Technical Handling of Synthetic Entities (ETHOS-SE), concebido como uma arquitetura conceitual destinada a orientar o desenvolvimento de sistemas mais transparentes, confiáveis e acessíveis.

A pesquisa fundamenta-se em uma abordagem qualitativa baseada na interpretação crítica da literatura científica e na sistematização conceitual de conhecimentos provenientes das áreas de Inteligência Artificial, desinformação digital, Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI), acessibilidade digital e Engenharia de Software. Como complemento, foi realizada uma análise bibliométrica exploratória da produção científica indexada na base Scopus, permitindo identificar tendências de pesquisa, temas recorrentes e lacunas que subsidiaram a construção do framework proposto.

O percurso metodológico foi estruturado em etapas sequenciais, compreendendo: (i) revisão da literatura; (ii) levantamento bibliográfico estruturado na base Scopus; (iii) análise bibliométrica exploratória; (iv) identificação das principais lacunas científicas; e (v) modelagem conceitual do framework ETHOS-SE. Essa organização metodológica permitiu fundamentar a proposição do framework em evidências da literatura e em uma análise

sistemática do estado da arte, conferindo maior consistência teórica à arquitetura proposta. A Figura 1 apresenta uma visão geral do percurso metodológico adotado nesta pesquisa, evidenciando a sequência lógica das etapas desenvolvidas desde a identificação do problema até a proposição do framework socio-técnico.

Figura 1 – Percurso metodológico adotado para o desenvolvimento do framework ETHOS-SE.



Fonte: Elaborado pelos autores (2026).

2.2 REVISÃO DA LITERATURA

A revisão da literatura constituiu a etapa inicial do percurso metodológico e teve como finalidade compreender o estado da arte sobre a mitigação da desinformação por *deepfakes*, identificando os principais avanços, desafios e lacunas que fundamentaram a proposição do framework ETHOS-SE. Assim, foi conduzida uma revisão narrativa estruturada, orientada por critérios previamente definidos de busca, seleção e análise das publicações científicas.

A revisão concentrou-se em estudos relacionados à Inteligência Artificial generativa, *deepfakes*, desinformação digital, detecção automatizada, Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI), proveniência de conteúdo, acessibilidade digital e Engenharia de Software. Esses temas foram selecionados por representarem os principais eixos conceituais necessários para compreender o problema investigado e subsidiar a construção da arquitetura proposta.

A seleção das publicações considerou como critérios de inclusão: (i) estudos publicados em periódicos e anais de conferências revisados por pares; (ii) trabalhos diretamente relacionados à geração ou detecção de *deepfakes*, desinformação digital, transparência algorítmica ou acessibilidade; e (iii) publicações disponibilizadas entre 2018 e 2026, período que compreende a consolidação das tecnologias de Inteligência Artificial generativa e o crescimento

das pesquisas sobre mídias sintéticas. Foram priorizados estudos amplamente citados na literatura e pesquisas recentes que abordassem avanços tecnológicos, desafios metodológicos e perspectivas de aplicação.

Após a seleção, os estudos foram submetidos à análise qualitativa, buscando identificar conceitos recorrentes, convergências teóricas e limitações das abordagens existentes. Essa etapa permitiu mapear tendências relacionadas à evolução das técnicas de geração e detecção de *deepfakes*, bem como evidenciar a predominância de soluções centradas em algoritmos de classificação, em contraste com a reduzida atenção dedicada à transparência dos resultados, à acessibilidade das interfaces e à compreensão por diferentes perfis de usuários.

Como resultado dessa análise, foram identificados quatro eixos estruturantes que orientaram o desenvolvimento do framework proposto: (i) evolução das tecnologias de geração de *deepfakes*; (ii) métodos contemporâneos de detecção e suas limitações; (iii) transparência e explicabilidade em sistemas baseados em Inteligência Artificial; e (iv) acessibilidade digital e interação humano-sistema na comunicação de conteúdos potencialmente manipulados. Esses eixos serviram como base conceitual para a definição das camadas que compõem o framework ETHOS-SE e para a integração dos princípios socio-técnicos adotados neste estudo.

2.3 ESTRATÉGIA DE BUSCA BIBLIOGRÁFICA

Com o objetivo de complementar a revisão da literatura e caracterizar o panorama científico relacionado à mitigação da desinformação por *deepfakes*, foi realizado um levantamento bibliográfico estruturado na base de dados Scopus, reconhecida por sua ampla cobertura de periódicos e conferências internacionais nas áreas de Ciência da Computação, Engenharia de Software, Inteligência Artificial e Sistemas de Informação.

A estratégia de busca foi planejada para recuperar estudos que contemplassem simultaneamente três dimensões centrais desta pesquisa: (i) o fenômeno dos *deepfakes*; (ii) sua relação com a desinformação digital; e (iii) mecanismos de detecção, verificação ou autenticação de conteúdos sintéticos. Assim, foi utilizada a seguinte expressão de busca: *deepfake** AND (disinformation OR misinformation OR "information disorder") AND (detection OR verification OR authentication).

A consulta foi realizada diretamente na interface da Scopus, considerando exclusivamente documentos classificados como Article, publicados entre 2020 e 2026. A

delimitação temporal foi definida para contemplar o período de maior crescimento das pesquisas sobre mídias sintéticas e Inteligência Artificial generativa.

Após a recuperação dos registros, os metadados foram exportados em formato CSV e processados na linguagem R, possibilitando a realização da análise bibliométrica exploratória apresentada na seção seguinte. O protocolo metodológico adotado para a obtenção da amostra é apresentado na Figura 2, sintetizando as etapas de busca, filtragem e preparação dos dados para análise.

Figura 2 – Protocolo metodológico utilizado para obtenção e preparação da amostra bibliográfica.



Fonte: Elaborado pelos autores (2026).

7

O Quadro 1 apresenta os principais parâmetros utilizados na estratégia de busca.

Quadro 1 – Parâmetros da estratégia de busca bibliográfica

Parâmetro	Especificação
Base de dados	Scopus
Tipo de documento	Article
Período	2020–2026
Idioma	Todos
String de busca	deepfake* AND (disinformation OR misinformation OR "information disorder") AND (detection OR verification OR authentication)
Formato de exportação	CSV
Ferramenta de análise	R

Fonte: Elaborado pelos autores (2026).

2.4 ANÁLISE BIBLIOMÉTRICA EXPLORATÓRIA

Os registros bibliográficos recuperados na base Scopus foram exportados em formato CSV e submetidos a procedimentos de organização e padronização dos metadados, de modo a possibilitar sua análise computacional. O processamento dos dados foi realizado na linguagem R, utilizando ferramentas específicas para análise bibliométrica e estatística descritiva, permitindo a exploração quantitativa da produção científica relacionada aos *deepfakes* e à desinformação digital.

A análise bibliométrica teve como finalidade caracterizar o panorama da produção científica recente e identificar tendências temáticas capazes de subsidiar a construção do framework ETHOS-SE. Para isso, foram empregados indicadores bibliométricos voltados à avaliação da evolução temporal das publicações e à identificação das relações entre os principais conceitos abordados na literatura.

Inicialmente, realizou-se a análise descritiva da produção científica, considerando a distribuição anual das publicações recuperadas na estratégia de busca. Essa etapa possibilitou identificar o comportamento temporal das pesquisas, evidenciando o crescimento do interesse científico pela temática ao longo do período analisado.

Em seguida, foi realizada a análise de coocorrência de termos extraídos dos títulos, resumos e palavras-chave dos artigos selecionados. Essa técnica permitiu identificar os conceitos mais recorrentes, bem como as relações semânticas estabelecidas entre eles, possibilitando o mapeamento da estrutura temática do campo de pesquisa. A partir dessas relações, tornou-se possível reconhecer áreas de maior concentração científica e identificar tópicos ainda pouco explorados na literatura.

Os resultados obtidos por meio da análise bibliométrica serviram como base para a identificação de lacunas científicas e para a fundamentação da arquitetura conceitual do framework ETHOS-SE. Dessa forma, a análise bibliométrica não foi utilizada apenas como instrumento descritivo da literatura, mas também como suporte metodológico para a proposição da solução apresentada neste estudo.

A Figura 3 apresenta o fluxo metodológico adotado para a realização da análise bibliométrica exploratória, desde o processamento dos registros bibliográficos até a identificação das evidências que subsidiaram a modelagem do framework proposto.

Figura 3 – Fluxo metodológico da análise bibliométrica exploratória utilizada na construção do framework ETHOS-SE.



Fonte: Elaborado pelos autores (2026).

2.5 CONSTRUÇÃO DO FRAMEWORK ETHOS-SE

A construção do framework ETHOS-SE foi conduzida a partir da integração dos resultados obtidos na revisão da literatura e na análise bibliométrica exploratória. Em vez de propor uma arquitetura baseada exclusivamente em pressupostos teóricos, buscou-se estruturar um modelo conceitual fundamentado nas principais evidências científicas identificadas ao longo da pesquisa.

Inicialmente, foram analisadas as limitações das abordagens predominantes descritas na literatura, especialmente aquelas centradas exclusivamente na detecção algorítmica de *deepfakes*. Essa análise evidenciou que, embora os avanços em aprendizado de máquina tenham proporcionado melhorias significativas na identificação de conteúdos sintéticos, aspectos relacionados à proveniência da informação, transparência das decisões automatizadas, acessibilidade das interfaces e compreensão dos resultados pelos usuários permanecem relativamente pouco explorados.

Com base nessas evidências, realizou-se um processo de síntese conceitual, no qual as lacunas identificadas foram agrupadas segundo afinidade funcional. Esse processo permitiu organizar os requisitos necessários para uma abordagem socio-técnica de mitigação da desinformação, estabelecendo relações entre problemas recorrentes observados na literatura e possíveis mecanismos de solução.

Na etapa seguinte, os requisitos identificados foram traduzidos em componentes arquiteturais que passaram a compor o framework ETHOS-SE. A definição dessas camadas

buscou assegurar a integração entre mecanismos de autenticação da origem do conteúdo, técnicas de detecção automatizada, recursos de explicabilidade e estratégias de acessibilidade digital, estruturando um fluxo contínuo de processamento e comunicação das informações ao usuário final.

A arquitetura proposta foi organizada em uma sequência lógica de funcionamento, contemplando desde a verificação da proveniência do conteúdo até a apresentação dos resultados por meio de interfaces acessíveis e mecanismos de fricção cognitiva ética. Essa organização permitiu estabelecer um modelo conceitual capaz de integrar diferentes dimensões da mitigação da desinformação em uma única estrutura arquitetural, fornecendo as bases para a proposição do framework apresentado na Seção 4.

3 RESULTADOS DA ANÁLISE BIBLIOMÉTRICA

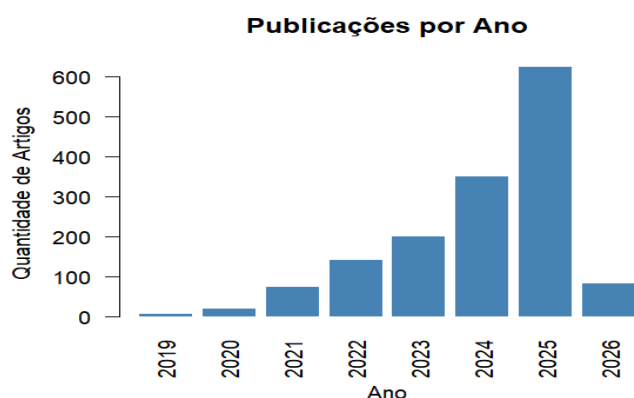
A análise bibliométrica exploratória permitiu caracterizar o panorama recente da produção científica sobre *deepfakes* e desinformação digital, identificando tendências de crescimento, temas predominantes e lacunas de investigação. Os resultados obtidos fornecem evidências quantitativas que complementam a revisão da literatura e fundamentam a construção do framework socio-técnico ETHOS-SE.

10

3.1 EVOLUÇÃO DA PRODUÇÃO CIENTÍFICA

A Figura 4 apresenta a distribuição anual das publicações recuperadas na base Scopus relacionadas aos *deepfakes* e à desinformação digital no período de 2020 a 2026.

Figura 4 – Evolução anual das publicações sobre *deepfakes* e desinformação (2020–2026).



Fonte: Elaborado pelos autores (2026).

Observa-se um crescimento contínuo da produção científica ao longo do período analisado, com aumento mais expressivo a partir de 2022 e expansão acelerada entre 2024 e 2025.

Esse comportamento evidencia a consolidação do tema como uma área emergente de investigação, impulsionada pela rápida evolução dos modelos de Inteligência Artificial generativa e pelo aumento das preocupações relacionadas aos impactos sociais, políticos e econômicos da disseminação de conteúdos sintéticos.

O crescimento observado também reflete o avanço das pesquisas voltadas ao desenvolvimento de técnicas de detecção automatizada, autenticação de conteúdos e mecanismos de combate à desinformação. A redução aparente no número de publicações em 2026 não representa uma diminuição do interesse científico pelo tema, mas decorre do fato de a coleta dos dados ter sido realizada antes do encerramento do ano, resultando em uma base de dados ainda incompleta para esse período.

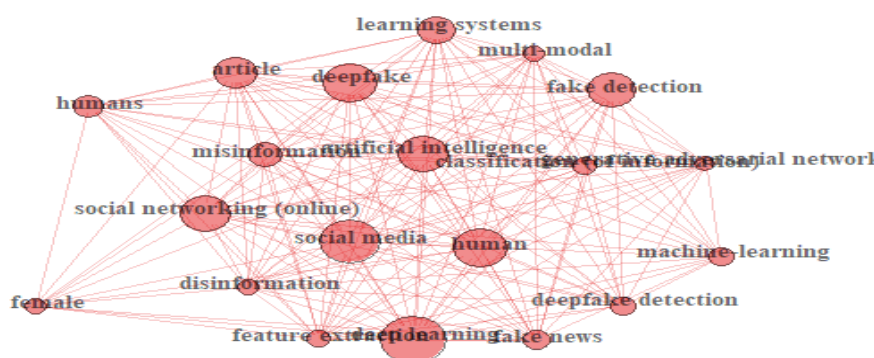
De maneira geral, os resultados evidenciam que a mitigação da desinformação baseada em *deepfakes* vem se consolidando como um dos principais desafios contemporâneos da Inteligência Artificial e da Engenharia de Software, justificando a necessidade de abordagens que integrem aspectos tecnológicos, sociais e humanos na concepção de soluções computacionais.

3.2 ESTRUTURA TEMÁTICA DA LITERATURA

11

A análise de coocorrência de termos foi realizada a partir das palavras extraídas dos títulos, resumos e palavras-chave dos artigos selecionados, permitindo identificar os principais eixos temáticos e as relações estabelecidas entre os conceitos recorrentes na literatura sobre *deepfakes* e desinformação digital. A Figura 5 apresenta a rede de coocorrência construída a partir dos registros bibliográficos recuperados na base Scopus.

Figura 5 – Rede de coocorrência dos principais termos na literatura sobre *deepfakes* e desinformação digital (2020–2026).



Fonte: Elaborado pelos autores (2026).

A visualização da rede evidencia que o termo *deepfake* ocupa posição central, apresentando elevado grau de conectividade com conceitos relacionados à detecção automatizada, aprendizado de máquina (*machine learning*), Inteligência Artificial (*artificial intelligence*) e mídias sociais (*social media*). Essa configuração demonstra que a literatura recente concentra seus esforços principalmente no desenvolvimento de modelos computacionais capazes de identificar conteúdos sintéticos, evidenciando o predomínio de abordagens orientadas por técnicas de Inteligência Artificial.

Observa-se ainda forte associação entre os termos *fake detection*, *deepfake detection* e *machine learning*, indicando que grande parte das pesquisas tem priorizado o aprimoramento de algoritmos de classificação e análise forense digital. Esses estudos buscam aumentar a precisão dos sistemas de detecção diante da evolução contínua das técnicas de geração de mídias sintéticas.

Por outro lado, conceitos relacionados aos impactos sociais da desinformação, como *misinformation* e *disinformation*, embora presentes na rede, apresentam menor centralidade e menor densidade de conexões quando comparados aos termos associados às soluções algorítmicas. Esse comportamento sugere que a literatura ainda dedica menor atenção à compreensão do fenômeno sob uma perspectiva socio-técnica, privilegiando o desempenho técnico dos modelos de detecção em detrimento da interação entre tecnologia e usuário.

12

Observa-se a baixa representatividade de conceitos relacionados à explicabilidade, proveniência, transparência, acessibilidade digital e experiência do usuário. A reduzida presença desses elementos evidencia que aspectos voltados à comunicação dos resultados das análises automatizadas e ao suporte à tomada de decisão pelos usuários permanecem pouco explorados nas pesquisas atuais.

Em conjunto, os resultados da análise de coocorrência indicam que o avanço científico na área tem sido impulsionado principalmente pelo desenvolvimento de soluções de detecção automática, enquanto dimensões relacionadas à confiança, transparência e acessibilidade ainda representam oportunidades de investigação. Essas lacunas constituem o principal fundamento para a proposição do framework socio-técnico ETHOS-SE, concebido para integrar mecanismos de autenticação, análise automatizada, explicabilidade e acessibilidade em uma arquitetura única.

3.3 LACUNAS CIENTÍFICAS E FUNDAMENTAÇÃO DA PROPOSTA DO FRAMEWORK ETHOS-SE

A análise integrada da evolução da produção científica e da rede de coocorrência de termos evidencia que a pesquisa sobre *deepfakes* tem avançado de forma significativa nos últimos anos, impulsionada principalmente pelo desenvolvimento de técnicas de Inteligência Artificial voltadas à detecção automática de conteúdos sintéticos.

Entretanto, esse avanço ocorre de maneira predominantemente orientada ao aperfeiçoamento dos modelos computacionais, com menor atenção à integração de aspectos relacionados à transparência, confiança e interação entre os sistemas automatizados e seus usuários.

Os resultados obtidos indicam que conceitos associados ao aprendizado de máquina, à detecção de *deepfakes* e às mídias sociais ocupam posição central na estrutura temática da literatura, enquanto temas como proveniência do conteúdo, explicabilidade dos modelos, acessibilidade digital e apoio à tomada de decisão apresentam menor representatividade. Esse cenário evidencia que a mitigação da desinformação tem sido tratada, em grande parte, como um problema estritamente tecnológico, concentrando esforços no aumento da capacidade de classificação dos algoritmos.

13

A efetividade de sistemas destinados à identificação de conteúdos sintéticos depende não apenas da precisão dos modelos computacionais, mas também da capacidade de comunicar seus resultados de forma transparente, compreensível e acessível aos diferentes perfis de usuários. A ausência dessa integração limita o potencial das soluções existentes, especialmente em contextos nos quais as decisões automatizadas influenciam diretamente a percepção de credibilidade das informações compartilhadas.

Sob a perspectiva da Engenharia de Software, observa-se ainda a inexistência de modelos arquiteturais que articulem, de forma integrada, mecanismos de autenticação da origem do conteúdo, detecção automatizada, explicabilidade e acessibilidade digital. Em geral, essas dimensões são tratadas como componentes independentes ou como funcionalidades adicionais, sem serem incorporadas como requisitos estruturantes da arquitetura dos sistemas.

Com o objetivo de sintetizar as principais lacunas identificadas na literatura e evidenciar sua relação com a proposta deste estudo, o Quadro 2 apresenta os desafios predominantes observados na análise bibliométrica e os respectivos componentes do framework ETHOS-SE concebidos para responder a essas limitações.

Quadro 2 – Relação entre as lacunas identificadas na literatura e os componentes do framework ETHOS-SE

Lacunas identificadas	Evidências observadas na literatura	Componente do ETHOS-SE
Predomínio de soluções centradas na detecção automatizada	Alta frequência de estudos sobre <i>deepfake detection</i> e <i>machine learning</i>	Integração entre mecanismos preventivos e reativos
Baixa utilização de mecanismos de autenticação da origem dos conteúdos	Pouca ênfase em proveniência e rastreabilidade digital	Camada de Proveniência
Limitada transparência das decisões automatizadas	Escassez de estudos sobre explicabilidade e interpretação dos resultados	Camada de Explicabilidade
Reduzida consideração da acessibilidade e da experiência do usuário	Pouca ocorrência de conceitos relacionados à acessibilidade e usabilidade	Camada de Acessibilidade e Fricção Cognitiva
Ausência de modelos arquiteturais integrados	Soluções desenvolvidas de forma isolada e específica	Arquitetura socio-técnica integrada do ETHOS-SE

Fonte: Elaborado pelos autores (2026).

4. PROPOSIÇÃO DO FRAMEWORK SOCIO-TÉCNICO ETHOS-SE

4.1 ARQUITETURA CONCEITUAL DO ETHOS-SE

As lacunas identificadas na análise bibliométrica evidenciaram que as abordagens atuais para mitigação da desinformação por *deepfakes* concentram-se predominantemente no desenvolvimento de técnicas de detecção automatizada, enquanto aspectos relacionados à proveniência da informação, à transparência das decisões algorítmicas, à acessibilidade e à integração desses elementos ao desenvolvimento de software permanecem pouco explorados de forma conjunta. Em resposta a essas limitações, propõe-se o Ethical Technical Handling of Synthetic Entities (ETHOS-SE), um framework socio-técnico concebido para integrar essas diferentes dimensões em uma arquitetura única, orientada pelos princípios da Engenharia de Software.

O ETHOS-SE foi concebido para apoiar o desenvolvimento de sistemas capazes de reduzir a vulnerabilidade dos usuários à desinformação baseada em mídias sintéticas. Diferentemente das abordagens centradas exclusivamente na classificação automática de conteúdos, o framework organiza mecanismos preventivos e reativos de autenticação, análise, interpretação e comunicação das evidências produzidas pelos sistemas de Inteligência Artificial. Dessa forma, busca ampliar a confiabilidade das soluções computacionais e favorecer uma interação mais transparente entre os sistemas automatizados e seus usuários.

Sob a perspectiva da Engenharia de Software, o framework incorpora requisitos de rastreabilidade, transparência, explicabilidade e acessibilidade como elementos estruturantes da

arquitetura do sistema. Essa abordagem permite que tais requisitos sejam considerados desde as fases iniciais do desenvolvimento, favorecendo a construção de aplicações mais confiáveis, compreensíveis e inclusivas.

A arquitetura conceitual do ETHOS-SE está organizada em quatro camadas complementares e interdependentes: Proveniência, responsável pela verificação da autenticidade da origem do conteúdo; Detecção e Análise, destinada à identificação automatizada de possíveis indícios de manipulação; Explicabilidade, voltada à tradução dos resultados técnicos em informações compreensíveis; e Acessibilidade e Fricção Cognitiva, responsável pela comunicação das evidências ao usuário por meio de interfaces acessíveis e mecanismos que estimulem uma análise crítica antes do compartilhamento de conteúdos potencialmente manipulados. A Figura 6 apresenta a arquitetura conceitual do framework ETHOS-SE e o fluxo de processamento proposto entre suas camadas.

Figura 6 – Arquitetura conceitual do framework socio-técnico ETHOS-SE.



Fonte: Elaborado pelos autores (2026).

Conforme ilustrado na Figura 6, o processamento inicia-se pela verificação da proveniência do conteúdo. Quando a autenticidade não pode ser confirmada, o framework aciona os mecanismos de detecção automatizada. Os resultados produzidos por esses algoritmos são então submetidos à camada de explicabilidade, responsável por transformar evidências técnicas em informações compreensíveis para os usuários. Finalmente, a camada de

acessibilidade apresenta essas informações por meio de recursos visuais e mecanismos de fricção cognitiva ética, favorecendo uma tomada de decisão mais consciente antes da disseminação do conteúdo.

Essa organização evidencia que o principal diferencial do ETHOS-SE não reside apenas na utilização de algoritmos de detecção, mas na integração de diferentes componentes técnicos e humanos em uma arquitetura orientada à transparência, à confiabilidade e à inclusão digital.

4.2 CAMADA DE PROVENIÊNCIA

A camada de Proveniência constitui o primeiro componente operacional do framework ETHOS-SE e tem como objetivo estabelecer um mecanismo inicial de autenticação e rastreabilidade dos conteúdos digitais. Diferentemente das abordagens que iniciam diretamente pela análise automatizada, o framework prioriza a verificação da origem e da integridade do conteúdo, buscando confirmar sua autenticidade antes da aplicação de modelos de Inteligência Artificial.

A Figura 7 apresenta o fluxo de funcionamento da camada de Proveniência. Conforme ilustrado, o processo inicia-se com o recebimento do conteúdo digital e a verificação das informações de proveniência disponíveis. Quando são encontradas evidências suficientes para comprovar a autenticidade do conteúdo, o processo de validação é concluído nessa própria camada. Caso contrário, o conteúdo é encaminhado para a camada de Detecção e Análise, onde são empregados modelos de Inteligência Artificial e técnicas de análise forense para investigar possíveis indícios de manipulação.

Figura 7 – Fluxo de funcionamento da camada de Proveniência.



Fonte: Elaborado pelos autores (2026).

Como ilustrado na Figura 7, a camada de Proveniência atua como um mecanismo preventivo dentro da arquitetura do ETHOS-SE. Sua função é reduzir a dependência exclusiva dos algoritmos de detecção por meio da utilização de evidências relacionadas à origem, autoria e histórico do conteúdo digital. Para isso, podem ser empregados mecanismos como metadados assinados digitalmente, certificados criptográficos, registros de proveniência e outros recursos capazes de garantir a rastreabilidade do conteúdo ao longo de seu ciclo de vida. Essa orientação converge com o padrão C2PA, que utiliza credenciais vinculadas criptograficamente para registrar e verificar informações de proveniência e autenticidade de ativos digitais (C2PA, 2026).

Quando essas evidências são suficientes para validar a autenticidade do conteúdo, evita-se a execução de procedimentos computacionalmente mais complexos, contribuindo para a otimização do processamento e para o aumento da confiabilidade do sistema. Na ausência dessas evidências, o framework prossegue para a etapa seguinte, preservando um fluxo progressivo de verificação no qual diferentes mecanismos de autenticação e análise atuam de forma complementar.

Sob a perspectiva da Engenharia de Software, essa camada incorpora requisitos de rastreabilidade, integridade e auditabilidade como elementos estruturantes da arquitetura do framework. Dessa forma, a autenticação da origem do conteúdo deixa de ser tratada como uma funcionalidade isolada e passa a constituir a primeira etapa do processo de mitigação da desinformação proposto pelo ETHOS-SE.

4.3 CAMADA DE DETECÇÃO E ANÁLISE

A camada de Detecção e Análise é acionada quando a etapa de proveniência não reúne evidências suficientes para confirmar a autenticidade do conteúdo digital. Sua função é realizar uma análise automatizada utilizando modelos de Inteligência Artificial e técnicas de análise forense digital, com o objetivo de identificar indícios de manipulação compatíveis com *deepfakes* ou outras formas de conteúdo sintético.

A Figura 8 apresenta o fluxo de funcionamento dessa camada. Após o recebimento do conteúdo proveniente da etapa anterior, são executados modelos computacionais especializados na extração e análise de características visuais, sonoras ou textuais. As evidências produzidas por esses modelos são então consolidadas para gerar uma avaliação quanto à probabilidade de manipulação do conteúdo.

Figura 8 – Fluxo de funcionamento da Camada de Detecção e Análise.



Fonte: Elaborado pelos autores (2026).

Conforme ilustrado na Figura 8, essa camada não se restringe à aplicação de um único algoritmo de classificação. Sua arquitetura permite integrar diferentes técnicas de análise, incluindo modelos de *deep learning*, métodos de análise forense digital e outros mecanismos especializados que possam ser incorporados ao sistema conforme a evolução das tecnologias de geração e detecção de *deepfakes*. Essa característica confere maior flexibilidade ao framework, favorecendo sua adaptação a diferentes cenários e requisitos tecnológicos.

Ao final do processamento, o sistema produz um conjunto de evidências técnicas acompanhado de um nível de confiança associado à classificação realizada. Em vez de encaminhar diretamente esse resultado ao usuário, as informações são direcionadas para a camada de Explicabilidade, responsável por traduzir os resultados computacionais em informações compreensíveis e passíveis de interpretação.

Sob a perspectiva da Engenharia de Software, a camada de Detecção e Análise foi concebida de forma modular, permitindo a substituição ou atualização dos modelos computacionais sem comprometer o funcionamento das demais camadas do framework. Essa abordagem favorece a manutenção, a evolução tecnológica e a interoperabilidade do ETHOS-SE, possibilitando a incorporação de novos algoritmos à medida que surgem avanços na área de Inteligência Artificial.

4.4 CAMADA DE EXPLICABILIDADE

A camada de Explicabilidade é responsável por transformar os resultados produzidos pelos modelos de Inteligência Artificial em informações compreensíveis para os usuários, reduzindo a complexidade técnica inerente aos processos automatizados de detecção. Seu objetivo é ampliar a transparência do sistema, permitindo que as decisões computacionais sejam apresentadas de forma clara, interpretável e fundamentada.

A Figura 9 apresenta o fluxo de funcionamento dessa camada. Após o recebimento das evidências geradas pela camada de Detecção e Análise, o framework organiza e interpreta essas informações, convertendo indicadores técnicos em explicações acessíveis que auxiliam o usuário na compreensão do processo de classificação realizado pelo sistema.

Figura 9 – Fluxo de funcionamento da Camada de Explicabilidade.



Fonte: Elaborado pelos autores (2026).

Conforme ilustrado na Figura 9, essa camada consolida os resultados provenientes dos modelos de Inteligência Artificial e os apresenta por meio de indicadores de confiança, justificativas textuais, evidências visuais e outros recursos que favoreçam a interpretação da análise realizada. Em vez de disponibilizar apenas uma classificação binária, o framework busca

comunicar os fatores que contribuíram para a decisão do sistema, aumentando a transparência do processo e fortalecendo a confiança do usuário.

Essa abordagem está alinhada aos princípios da Inteligência Artificial Explicável (Explainable Artificial Intelligence – XAI), cuja finalidade é tornar os sistemas inteligentes mais compreensíveis, auditáveis e confiáveis (ADADI; BERRADA, 2018). Estudos recentes de detecção explicável de deepfakes também reforçam a relevância de combinar desempenho de classificação com evidências interpretáveis para apoiar a análise humana (BUDATI; JADAM; MALLEBOYINA, 2025; BHARATI et al., 2026). No contexto do ETHOS-SE, a explicabilidade constitui um elemento essencial para apoiar a tomada de decisão, permitindo que os usuários compreendam as evidências apresentadas antes de avaliar a autenticidade do conteúdo.

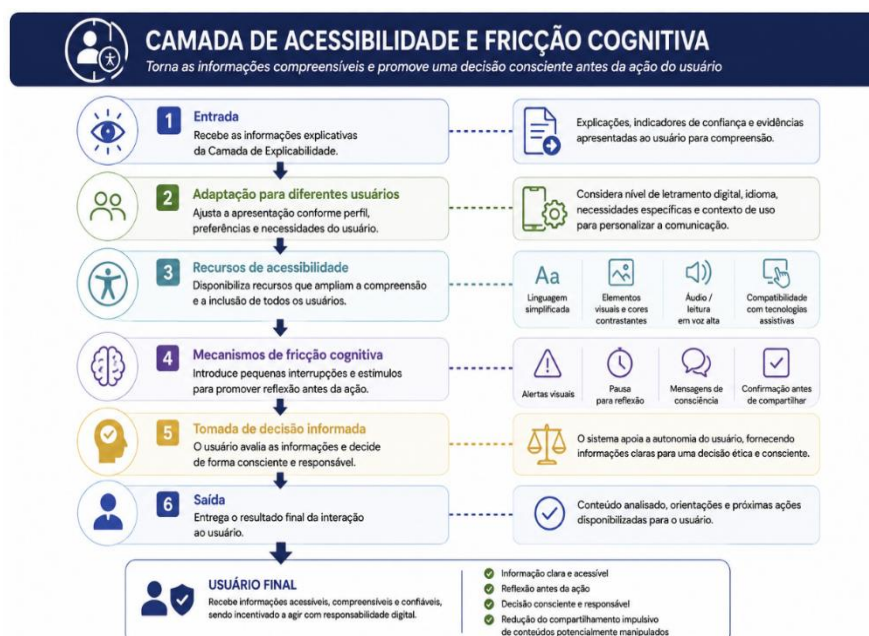
Sob a perspectiva da Engenharia de Software, essa camada atua como uma interface entre o processamento técnico realizado pelos algoritmos e a interação humana, promovendo a comunicação dos resultados de forma estruturada e compreensível. Ao priorizar a interpretação das evidências produzidas pelos modelos computacionais, o framework amplia a transparência das decisões automatizadas e fortalece a confiança dos usuários nos processos de verificação.

4.5 CAMADA DE ACESSIBILIDADE E FRICÇÃO COGNITIVA

A camada de Acessibilidade e Fricção Cognitiva representa a etapa final do framework ETHOS-SE, sendo responsável por disponibilizar ao usuário as informações produzidas pelas camadas anteriores por meio de interfaces acessíveis, compreensíveis e orientadas ao processo de tomada de decisão. Seu objetivo é garantir que os resultados das análises automatizadas possam ser interpretados por diferentes perfis de usuários, promovendo maior inclusão digital e fortalecendo a confiança nas informações apresentadas.

A Figura 10 apresenta o fluxo de funcionamento dessa camada. Após o recebimento das explicações geradas pela etapa de Explicabilidade, o framework adapta a apresentação das informações utilizando recursos de acessibilidade e mecanismos de comunicação visual capazes de facilitar a compreensão do conteúdo analisado. Paralelamente, são incorporadas estratégias de fricção cognitiva ética, destinadas a estimular uma reflexão consciente antes da aceitação ou do compartilhamento de conteúdos potencialmente manipulados.

Figura 10 – Fluxo de funcionamento da Camada de Acessibilidade e Fricção Cognitiva.



Fonte: Elaborado pelos autores (2026).

Conforme ilustrado na Figura 10, essa camada integra recursos de acessibilidade, como linguagem simplificada, elementos visuais, sinalizações intuitivas e compatibilidade com tecnologias assistivas, permitindo que usuários com diferentes níveis de letramento digital ou necessidades específicas compreendam os resultados apresentados pelo sistema. Essa orientação é compatível com as recomendações de acessibilidade digital do WCAG 2.2 e com diretrizes complementares do W3C voltadas à compreensão de conteúdo por pessoas com deficiências cognitivas e de aprendizagem (W3C, 2024; W3C, 2021). Dessa forma, a abordagem amplia a usabilidade do framework e favorece uma interação mais inclusiva entre usuários e sistemas de Inteligência Artificial.

Além da acessibilidade, o ETHOS-SE incorpora mecanismos de fricção cognitiva, compreendidos como estratégias de interface que introduzem pequenas interrupções deliberadas no fluxo de interação, incentivando o usuário a refletir antes de realizar ações como compartilhar, encaminhar ou validar um conteúdo. Evidências experimentais indicam que intervenções simples que redirecionam a atenção para a precisão podem reduzir a intenção de compartilhar informações falsas, oferecendo fundamento empírico para mecanismos de reflexão antes do compartilhamento (PENNYCOOK et al., 2021; PENNYCOOK; RAND,

2022). Em vez de restringir a autonomia do usuário, essas intervenções procuram apoiar uma decisão mais consciente, reduzindo comportamentos impulsivos frequentemente associados à propagação da desinformação.

Sob a perspectiva da Engenharia de Software, essa camada materializa os princípios de interação humano-computador e de design centrado no usuário, incorporando requisitos de acessibilidade, usabilidade e transparência como elementos essenciais da arquitetura do framework. Dessa forma, o ETHOS-SE conclui o processo de mitigação da desinformação não apenas com a identificação de possíveis manipulações, mas com o fornecimento de informações que apoiem uma tomada de decisão informada e responsável.

As quatro camadas apresentadas atuam de forma integrada ao longo do processo de mitigação da desinformação por *deepfakes*, estabelecendo um fluxo contínuo que combina autenticação da origem, análise automatizada, interpretação dos resultados e apoio à tomada de decisão do usuário. Essa organização evidencia a natureza socio-técnica do ETHOS-SE, na qual componentes computacionais e aspectos relacionados à interação humano-computador são considerados de maneira complementar. O Quadro 3 sintetiza a estrutura do framework e as principais funções desempenhadas por cada uma de suas camadas.

Quadro 3 – Síntese das camadas, objetivos e mecanismos do framework ETHOS-SE

Camada	Objetivo	Principais mecanismos	Saída
Proveniência	Validar autenticidade	Metadados, certificados, assinaturas digitais	Conteúdo autenticado ou encaminhamento
Deteção e Análise	Detectar manipulações	IA, <i>deep learning</i> , análise forense	Evidências e nível de confiança
Explicabilidade	Traduzir resultados	Indicadores, justificativas, visualizações	Explicações compreensíveis
Acessibilidade e Fricção Cognitiva	Apoiar decisão	Interface acessível, alertas, confirmação	Decisão informada

Fonte: Elaborado pelos autores (2026).

5 DISCUSSÃO

Os resultados apresentados neste estudo evidenciam que a literatura sobre mitigação da desinformação baseada em *deepfakes* tem evoluído de forma expressiva nos últimos anos, impulsionada principalmente pelo desenvolvimento de técnicas de Inteligência Artificial

voltadas à detecção automatizada de conteúdos sintéticos. Entretanto, a análise bibliométrica revelou que esse avanço ocorre de maneira predominantemente centrada no desempenho dos algoritmos, enquanto aspectos relacionados à proveniência, transparência, explicabilidade e acessibilidade permanecem menos explorados ou tratados de forma isolada.

Nesse contexto, o framework ETHOS-SE amplia a perspectiva das abordagens tradicionais ao integrar essas dimensões em uma arquitetura única. Em vez de restringir o processo de mitigação à classificação automática do conteúdo, o framework organiza um fluxo contínuo que inicia com a verificação da proveniência, prossegue com a análise automatizada quando necessária, traduz os resultados por meio de mecanismos de explicabilidade e, por fim, incorpora recursos de acessibilidade e fricção cognitiva para apoiar a tomada de decisão do usuário. Essa integração representa um avanço em relação às soluções fragmentadas identificadas na literatura.

Sob a perspectiva da Engenharia de Software, a principal contribuição do ETHOS-SE consiste em tratar requisitos como rastreabilidade, transparência, interpretabilidade e acessibilidade como elementos arquiteturais do sistema, e não apenas como funcionalidades complementares. Essa abordagem favorece o desenvolvimento de aplicações mais robustas e alinhadas aos princípios da Inteligência Artificial Confiável (Trustworthy AI). O AI Risk Management Framework do NIST caracteriza sistemas de IA confiáveis, entre outros aspectos, como responsáveis e transparentes, explicáveis e interpretáveis, ressaltando ainda a natureza socio-técnica desses sistemas (NIST, 2023). Assim, aspectos técnicos e humanos podem ser considerados de forma integrada durante o processo de desenvolvimento de software.

23

Outro aspecto relevante refere-se ao papel da interação humano-computador na mitigação da desinformação. Enquanto grande parte dos estudos concentra seus esforços na melhoria dos modelos de classificação, o ETHOS-SE reconhece que a efetividade desses sistemas também depende da forma como os resultados são apresentados aos usuários. A incorporação de mecanismos de explicabilidade e fricção cognitiva busca reduzir interpretações equivocadas e desencorajar o compartilhamento impulsivo de conteúdos potencialmente manipulados, fortalecendo uma tomada de decisão mais consciente.

O Quadro 4 apresenta uma síntese comparativa entre as principais características das abordagens predominantes identificadas na literatura e os componentes incorporados pelo framework ETHOS-SE.

Quadro 4 – Comparação entre as abordagens predominantes na literatura e o framework ETHOS-SE

Aspecto	Abordagens predominantes	Framework ETHOS-SE
Detecção automatizada de <i>deepfakes</i>	✓	✓
Verificação de proveniência	Parcial	✓
Explicabilidade dos resultados	Limitada	✓
Recursos de acessibilidade	Pouco explorados	✓
Fricção cognitiva para apoio à decisão	Não identificada	✓
Arquitetura integrada	Predominantemente fragmentada	✓
Perspectiva socio-técnica	Limitada	✓

Fonte: Elaborado pelos autores (2026).

A comparação apresentada no Quadro 4 evidencia que a principal contribuição do ETHOS-SE não está na proposição de novos algoritmos de detecção, mas na integração de diferentes dimensões que, embora discutidas individualmente na literatura, ainda são pouco articuladas em uma arquitetura unificada. Ao combinar autenticação da origem do conteúdo, análise automatizada, explicabilidade e acessibilidade, o framework amplia o escopo das estratégias de mitigação da desinformação e oferece uma perspectiva arquitetural voltada ao desenvolvimento de sistemas mais transparentes, confiáveis e centrados no usuário.

24

Embora a proposta apresentada possua natureza conceitual, ela estabelece diretrizes para futuras implementações em ambientes reais, podendo subsidiar o desenvolvimento de plataformas digitais, sistemas de verificação de conteúdo, aplicações governamentais e ambientes educacionais. Além disso, sua estrutura modular favorece a evolução tecnológica, permitindo a incorporação de novos mecanismos de autenticação, modelos de Inteligência Artificial e recursos de interação à medida que novas demandas surgirem.

Como limitação, destaca-se que o ETHOS-SE ainda não foi implementado ou validado experimentalmente. Dessa forma, sua efetividade prática dependerá de estudos futuros envolvendo prototipação, avaliação de desempenho, testes de usabilidade e investigações sobre o impacto dos mecanismos de explicabilidade e fricção cognitiva na tomada de decisão dos usuários.

6 CONCLUSÃO

A crescente sofisticação das tecnologias de Inteligência Artificial generativa tem ampliado os desafios relacionados à identificação e mitigação da desinformação baseada em *deepfakes*. Embora a literatura apresente avanços significativos no desenvolvimento de algoritmos de detecção, os resultados obtidos neste estudo evidenciam que aspectos relacionados à proveniência dos conteúdos, à transparência das decisões automatizadas, à acessibilidade digital e à integração desses elementos em arquiteturas de software permanecem pouco explorados de forma conjunta.

A análise bibliométrica realizada permitiu identificar tendências da produção científica e mapear lacunas relevantes na literatura, demonstrando que a maioria das pesquisas concentra seus esforços no aprimoramento do desempenho técnico dos modelos de detecção. Em contrapartida, observa-se reduzida atenção à forma como os resultados produzidos pelos sistemas são interpretados, comunicados e utilizados pelos usuários no processo de tomada de decisão.

Como resposta a essas lacunas, este trabalho propôs o framework ETHOS-SE, concebido sob a perspectiva da Engenharia de Software para integrar mecanismos de proveniência, detecção e análise, explicabilidade e acessibilidade em uma arquitetura conceitual única. Ao estruturar essas dimensões de forma complementar, o framework amplia o escopo das abordagens tradicionais de mitigação da desinformação, incorporando requisitos de transparência, rastreabilidade, interpretabilidade e inclusão digital desde as etapas iniciais do desenvolvimento de sistemas.

A principal contribuição deste estudo consiste na proposição de uma arquitetura conceitual que articula componentes técnicos e humanos em um modelo integrado para apoio à mitigação da desinformação por *deepfakes*. Além de oferecer uma visão sistêmica do problema, o ETHOS-SE fornece diretrizes que podem subsidiar o desenvolvimento de aplicações em diferentes contextos, como plataformas digitais, sistemas governamentais, ambientes educacionais e serviços de verificação de conteúdo.

Como limitação, destaca-se que o framework foi desenvolvido a partir de evidências da literatura e de uma análise bibliométrica exploratória, não contemplando, nesta etapa, sua implementação computacional ou validação em ambientes reais de aplicação. Dessa forma, os

resultados apresentados possuem natureza conceitual e constituem uma base para investigações futuras.

Como perspectivas de continuidade, recomenda-se a implementação do framework em um protótipo funcional, sua avaliação em cenários reais de uso e a realização de estudos experimentais que permitam analisar sua efetividade, usabilidade e impacto na redução da disseminação de conteúdos sintéticos. Também se sugere investigar a integração do ETHOS-SE com tecnologias emergentes de autenticação digital, Inteligência Artificial Explicável e padrões internacionais de proveniência de conteúdo, contribuindo para o desenvolvimento de ecossistemas digitais mais transparentes, confiáveis e acessíveis.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, v. 6, p. 52138-52160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

BHARATI, N.; WONG, P.; KOUADRI MOSTÉFAOUI, S.; KBAIER, D.; COLLIE, J. Explainable deepfake detection: a multi-model framework with human-interpretable rationales for legal investigation purposes. *Machine Learning with Applications*, v. 23, art. 100819, 2026. DOI: 10.1016/j.mlwa.2025.100819.

BUDATI, J. L. N.; JADAM, A. B.; MALLEBOYINA, R. Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics. In: 2025 6th International Conference for Emerging Technology (INCET), Belgaum, India, 2025. p. 1-7. DOI: 10.1109/INCET64471.2025.11140952.

C2PA - COALITION FOR CONTENT PROVENANCE AND AUTHENTICITY. C2PA Technical Specification: Content Credentials, version 2.4. 2026. Disponível em: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html. Acesso em: 17 ago. 2026.

CHESNEY, R.; CITRON, D. K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, v. 107, n. 6, p. 1753-1820, 2019.

KORSHUNOV, P.; MARCEL, S. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *arXiv:1812.08685*, 2018. DOI: 10.48550/arXiv.1812.08685.

LAY, T.; BUI, M. Explainable deepfake detection. *Machine Learning with Applications*, v. 22, 2025.

LAZER, D. M. J. et al. The science of fake news. *Science*, v. 359, n. 6380, p. 1094-1096, 2018. DOI: 10.1126/science.aao2998.

NIST - NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: NIST, 2023. DOI: 10.6028/NIST.AI.100-1.

PENNYCOOK, G.; EPSTEIN, Z.; MOSLEH, M.; ARECHAR, A. A.; ECKLES, D.; RAND, D. G. Shifting attention to accuracy can reduce misinformation online. *Nature*, v. 592, p. 590-595, 2021. DOI: 10.1038/s41586-021-03344-2.

PENNYCOOK, G.; RAND, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, v. 13, art. 2333, 2022. DOI: 10.1038/s41467-022-30073-5.

PINDROP. Voice Intelligence + Security Report: The Surge of Synthetic Media and Digital Identity Fraud. Pindrop Security Research, 2025.

RÖSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C.; THIES, J.; NIESSNER, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. p. 1-11. DOI: 10.1109/ICCV.2019.00009.

VACCARI, C.; CHADWICK, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, v. 6, n. 1, 2020. DOI: 10.1177/2056305120903408.

W3C - WORLD WIDE WEB CONSORTIUM. Making Content Usable for People with Cognitive and Learning Disabilities. W3C Working Group Note, 2021. Disponível em: <https://www.w3.org/TR/coga-usable/>. Acesso em: 17 ago. 2026.

W3C - WORLD WIDE WEB CONSORTIUM. Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation, atualização de 12 dez. 2024. Disponível em: <https://www.w3.org/TR/WCAG22/>. Acesso em: 17 ago. 2026.

WESTERLUND, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, v. 9, n. 11, p. 39-52, 2019. DOI: 10.22215/timreview/1282.