DEEPFAKES, REALIDADE SINTETIZADA, CONFIANÇA ABALADA: ANÁLISE DE CONTEÚDO DE CASOS PÚBLICOS E IMPLICAÇÕES PARA GOVERNANÇA ORIENTADA POR RISCO

Deepfakes, synthetic reality, and eroded trust: a content analysis of public cases and their implications for risk-oriented governanc

Wesley Antonio Gonçalves[1]
Lucas Daniel Santana Rosa[2]

RESUMO

A difusão de sistemas de inteligência artificial generativa amplia a produção e a circulação de conteúdos sintéticos hiper-realistas e intensifica riscos à integridade informacional, à confiança pública e à credibilidade de registros audiovisuais. Apesar do avanço dos debates sobre governança da IA e desinformação, permanece limitada a produção de evidências empíricas que sistematizem, sob enfoque ético, como deepfakes operam como mecanismo de manipulação da veracidade informacional e de geração de danos sociais. O estudo tem como objetivo investigar os dilemas éticos e os riscos sociais envolvidos na geração e disseminação de deepfakes, com ênfase em suas implicações para uma governança orientada por risco. Adota-se pesquisa aplicada, de abordagem mista e predominância qualitativa, baseada em análise documental de 20 casos públicos

ABSTRACT

The widespread adoption of generative artificial intelligence increases the production and circulation of highly realistic synthetic content and intensifies risks to information integrity, public trust, and the credibility of audiovisual records. Despite growing debates on AI governance and misinformation, empirical evidence remains limited regarding how deepfakes operate as mechanisms of informational manipulation and social harm. This study investigates the ethical dilemmas and social risks involved in the creation and dissemination of deepfakes, with emphasis on their implications for risk-oriented governance. The research is applied, with a mixed-method design and qualitative predominance, combining documentary analysis of 20 public cases and content analysis supported by thematic coding, double categorical review, and analytical saturation. The findings identify seven recurring thematic axes, with a predominance of high-impact occurrences

[1] Professor Titular no Instituto Federal do Triângulo Mineiro (IFTM). Pós-Doutor em Educação Tecnológica. Doutor, Mestre, Especialista e Bacharel em Administração. Bacharel e Especialista em Computação. Licenciatura em Educação Profissional e Tecnológica.

[2] Graduando em Análise e Desenvolvimento de Sistemas. Pesquisador de iniciação científica e bolsista pelo CNPq (Edital PIBIC 04.2025).

e análise de conteúdo, com codificação temática, dupla revisão categorial e saturação analítica. Os resultados identificam sete eixos temáticos, com predominância de ocorrências de alto impacto associadas à desinformação político-institucional, fraudes e violência reputacional, além de usos ambivalentes condicionados por transparência e salvaguardas. Conclui-se que deepfakes configuram risco sociotécnico de relevância jurídico-institucional e demandam respostas estruturadas de autenticação, responsabilização e mitigação, tendo o estudo proposto um framework normativo-autoral aplicável a políticas públicas, organizações e práticas de verificação.

Palavras-chave: Deepfakes; Integridade informacional; Governança orientada por risco; Manipulação audiovisual; Confiança digital.

linked to political-institutional disinformation, fraud, and reputational violence, as well as ambivalent uses conditioned by transparency and safeguards. The study concludes that deepfakes constitute a sociotechnical risk of legal and institutional relevance and therefore require structured responses based on authentication, accountability, and mitigation. As a contribution, the article proposes an original normative framework to support public policies, organizational practices, and verification strategies in contexts involving synthetic media.

***Keywords*:** *Deepfakes; Information integrity; Risk-oriented governance; Audiovisual manipulation; Digital trust.*

SUMÁRIO

1 INTRODUÇÃO

A difusão recente de sistemas de Inteligência Artificial (IA) generativa amplia a capacidade de produzir e disseminar conteúdos sintéticos com elevado grau de realismo, incluindo voz, imagem e vídeo. Nesse cenário, os deepfakes destacam-se por viabilizarem a manipulação de registros audiovisuais de modo plausível, dificultando a distinção entre conteúdo autêntico e conteúdo artificialmente gerado e, consequentemente, tensionando a confiança pública na informação. Em ecossistemas digitais marcados por viralização, economia da atenção e circulação acelerada de mensagens, conteúdos manipulados tendem a adquirir alcance social ampliado, com efeitos potencialmente relevantes para reputações, decisões organizacionais e processos democráticos (Moraes, 2020; STF, 2024; Vaccari; Chadwick, 2020; Chesney; Citron, 2018; Westerlund, 2019).

No plano científico, embora se reconheça que deepfakes constituem um fenômeno sociotécnico associado à desinformação, a fraudes e a violações de direitos, a lacuna central do estudo reside na escassez de pesquisas que, com base em casos públicos documentados, articulem evidências empíricas, dilemas éticos e implicações normativas para a governança orientada por risco. Permanece limitada, portanto, a produção de evidências que sistematizem como e em quais condições deepfakes operam como mecanismo de manipulação da veracidade informacional, bem como a forma pela qual seus impactos se distribuem entre danos individuais e riscos coletivos. Essa insuficiência é particularmente sensível diante de agendas contemporâneas de governança e regulação de IA orientadas por risco, as quais demandam tipologias, diagnósticos e critérios sustentados por dados para orientar decisões públicas e institucionais (CGEE, 2025; SBC, 2024; NIST, 2023; ISO/IEC, 2023; Unesco, 2021).

Diante dessa lacuna, formula-se o problema de pesquisa de modo direto: em que medida a geração e a disseminação de deepfakes produzidos por IA generativa criam dilemas éticos e riscos sociais capazes de ameaçar a veracidade informacional, a confiança pública e a proteção de direitos em ambientes digitais? A partir dessa pergunta-problema, parte-se, neste estudo, da tese de que a circulação de conteúdos sintéticos

hiper-realistas sem mecanismos mínimos de autenticação, proveniência e transparência compromete a integridade informacional e tensiona, de modo crescente, a tutela de direitos fundamentais e a própria estabilidade das relações institucionais mediadas digitalmente.

Em termos normativos, sustenta-se que deepfakes não podem ser tratadas apenas como manifestações tecnológicas neutras ou como episódios isolados de falsidade, mas como fenômeno que demanda resposta jurídica e regulatória proporcional ao risco, sobretudo quando sua circulação potencializa dano reputacional, fraude, manipulação político-institucional e enfraquecimento da confiança pública. A hipótese argumentativa que orienta o artigo, portanto, é a de que a ausência de deveres mínimos de autenticação e transparência tende a ampliar a assimetria entre capacidade técnica de falsificação e capacidade social de verificação, deslocando para indivíduos e instituições um ônus epistêmico e jurídico excessivo.

A relevância científica do tema decorre da necessidade de converter um debate frequentemente normativo e abstrato em um quadro analítico capaz de identificar padrões recorrentes de finalidade, impacto e dano, favorecendo o acúmulo incremental de conhecimento no campo. No âmbito acadêmico, pretende-se contribuir para a compreensão dos deepfakes como problema de integridade informacional, de confiança digital e de proteção de direitos, articulando desinformação, ética da IA e governança responsável.

No plano gerencial e institucional, os achados podem subsidiar protocolos de verificação, prevenção de fraudes e rotinas de mitigação de riscos em contextos de comunicação digital. Em plano mais assertivo, o estudo também sustenta que a circulação de mídias sintéticas sem salvaguardas mínimas de autenticação e transparência deve ser compreendida como questão de relevância constitucional e regulatória, uma vez que afeta não apenas interesses privados, mas também a confiabilidade do espaço público informacional e das estruturas decisórias mediadas por registros audiovisuais.

Destaca-se, ainda, que este estudo é fruto de projeto de pesquisa desenvolvido em uma instituição pública federal, no âmbito de programa de pesquisa institucional com bolsa do CNPq, o que reforça seu compromisso com rigor metodológico, ética em pesquisa e produção de evidências socialmente relevantes. Nesse enquadramento, o objetivo geral consiste em investigar os dilemas éticos e os riscos sociais envolvidos na geração e disseminação de conteúdos deepfake produzidos por meio de IA generativa, com ênfase na ameaça à veracidade informacional e nas implicações sociais decorrentes. Para alcançar tal objetivo, adota-se análise de casos públicos documentados e tratamento sistemático por análise de conteúdo.

A partir dessa estrutura, o artigo organiza-se em cinco seções. Após esta introdução, descreve-se a metodologia, com delineamento, corpus e procedimentos de análise. Em seguida, apresenta-se o referencial teórico, que discute fundamentos conceituais sobre deepfakes, desinformação e ética da IA. Na sequência, expõe-se a análise e discussão dos resultados, articulando categorias empíricas e literatura. Por fim, as considerações finais sintetizam os achados, limitações e sugestões para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Fundamentos Introdutórios

A compreensão do fenômeno deepfake, sobretudo quando associado à desinformação e à manipulação de evidências, exige abordagem multidimensional que articule fundamentos técnicos, impactos sociopolíticos e parâmetros éticos de governança em inteligência artificial. No plano internacional, a literatura reconhece que as mídias sintéticas não constituem apenas uma inovação tecnológica, mas um problema sociotécnico capaz de reconfigurar relações entre evidência, confiança e deliberação pública. Chesney e Citron (2018) destacam que deepfakes desafiam simultaneamente a privacidade, a segurança e a democracia, ao passo que Westerlund (2019) identifica sua emergência como marco de transformação da circulação informacional digital.

Em complemento, Vaccari e Chadwick (2020) demonstram que vídeos sintéticos de natureza política podem produzir efeitos sobre percepção, engano, incerteza e confiança em notícias, reforçando que o problema não se restringe à falsificação de conteúdo, mas atinge a própria arquitetura de credibilidade do ecossistema comunicacional contemporâneo. Nessa perspectiva, a literatura nacional e internacional converge ao indicar que a facilidade de produção e a crescente sofisticação desses artefatos ampliam desafios para verificação, autenticação e responsabilização, o que justifica a necessidade de uma abordagem ético-jurídica e regulatória mais densa e empiricamente fundamentada (Moraes, 2020; STF, 2024).

Diante disso, este referencial está estruturado em quatro eixos: 1) fundamentos conceituais e técnicos de deepfakes e IA generativa; 2) deepfakes como desinformação e crise de confiança informacional; 3) ética da IA e governança responsável orientada por risco; e 4) implicações aplicadas ao cenário brasileiro e ao debate regulatório, com ênfase em evidências digitais e proteção de direitos.

2.2 Eixo 1: Fundamentos teóricos e conceituais: deepfakes e IA generativa

A expressão "deepfake" deriva da associação entre *deep learning* e *fake*, denotando conteúdos falsos produzidos com elevado grau de elaboração, frequentemente em formatos multimodais (texto, imagem, vídeo e áudio). Em materiais de orientação institucional, deepfakes são apresentadas como "falsidades profundas", isto é, conteúdos sintéticos altamente realistas, cuja característica central reside na dificuldade de distinção, por observadores comuns, entre o que é autêntico e o que foi artificialmente gerado (STF, 2024).

A literatura especializada reforça que deepfakes se consolidaram como técnica de substituição e manipulação, como a troca de *faces* em vídeos, viabilizada por softwares e aplicações que reduzem barreiras técnicas de uso e ampliam a escala de circulação em redes sociais. Nesse sen-

tido, Moraes (2020) enfatiza que ferramentas contemporâneas permitem manipulação com extrema facilidade e com baixa detectabilidade visual, favorecendo a produção de conteúdos potencialmente indistinguíveis dos autênticos, aspecto que se torna estrutural para o problema investigado nesta pesquisa: o enfraquecimento de critérios ordinários de verificação social do real.

Do ponto de vista conceitual, infere-se que, deepfakes não devem ser tratadas apenas como "imagens falsas", mas como fenômeno sociotécnico que reconfigura relações entre evidência, credibilidade e disputa de narrativas. Tal reconfiguração decorre do uso de sistemas de aprendizagem de máquina e do acoplamento entre automação e circulação em plataformas digitais, o que amplia externalidades negativas (p. ex., dano reputacional, fraude, violência política e desinformação). A criticidade do tema decorre, portanto, menos da existência isolada da técnica e mais do seu acoplamento a ecossistemas de atenção, viralidade e polarização, que favorecem efeitos sociais de larga escala.

2.3 Eixo 2: Deepfakes como desinformação: integridade informacional e crise de confiança

No eixo da desinformação, deepfakes constituem modalidade especialmente sensível porque combinam persuasão audiovisual, realismo sintético e alta escalabilidade em plataformas digitais. A literatura internacional sustenta que o principal efeito desses conteúdos não é apenas enganar pontualmente, mas desestabilizar regimes de confiança e ampliar a incerteza epistêmica no espaço público. Vaccari e Chadwick (2020) demonstram que vídeos políticos sintéticos podem aumentar dúvida, hesitação e desconfiança em relação ao conteúdo jornalístico, mesmo quando o observador não adere integralmente à falsidade.

Tal diagnóstico dialoga com a noção de integridade informacional desenvolvida pelas Nações Unidas, segundo a qual ambientes digitais degradados por manipulação e desinformação comprometem a qualidade do debate público e a capacidade coletiva de distinguir o verdadeiro do

falso (United Nations, 2023). Nessa chave, deepfakes deixam de ser compreendidas apenas como "conteúdo falso" e passam a ser vistas como fator de corrosão da confiança pública, com efeitos institucionais e democráticos mais amplos.

A literatura também aponta que a disseminação de deepfakes se beneficia da dificuldade de verificação por usuários e do "ganho de credibilidade" associado ao audiovisual, o que reconfigura o valor probatório socialmente atribuído a imagens e vídeos (Moraes, 2020). Nesse cenário, a confiança pública torna-se variável estratégica: quando conteúdos sintéticos passam a circular com alta plausibilidade, emerge risco sistêmico de "contaminação" da esfera pública por dúvidas generalizadas, fenômeno que pode gerar tanto a aceitação acrítica do falso quanto a negação oportunista do verdadeiro (efeitos que, na prática, favorecem agentes mal-intencionados) (Caldeira, 2025).

No contexto brasileiro recente, observa-se crescente atenção pública e regulatória ao tema. Em obra organizada por Souza e Silva (2025), há registro de que deepfakes podem ser definidas como falsificações realistas geradas por IA e que, pela facilidade de criação e disponibilidade de ferramentas gratuitas e intuitivas, representam risco à integridade da informação e a agravar um "caos informacional", e incidem de forma mais intensa sobre grupos vulneráveis e sobre assimetrias de letramento midiático.

Sob perspectiva crítica, a relevância ética do fenômeno se fortalece porque o dano não se limita ao indivíduo atingido, mas alcança dimensões coletivas, como a erosão da confiança em instituições, a desorganização do espaço de deliberação pública e a ampliação do custo social de verificação, que passa a recair sobre vítimas, imprensa e sistemas de justiça. Nesse sentido, deepfakes configuram problema de infraestrutura informacional, no qual a veracidade deixa de funcionar como pressuposto prático e passa a ser permanentemente disputada. Essa leitura se aproxima da ideia de *digital trust*, entendida como a expectativa social de que os ambientes digitais preservem autenticidade, rastreabilidade e confiabilidade mínimas para sustentar relações comunicacionais e decisórias.

Quando tais condições são corroídas por conteúdos sintéticos de difícil verificação, produz-se não apenas desinformação, mas também uma crise epistemológica de reconhecimento do real (Vaccari; Chadwick, 2020; United Nations, 2023; Partnership on AI, 2023).

2.4 Eixo 3: Ética da IA e correntes de governança orientadas por risco

A análise ética das deepfakes requer enquadramento dentro do debate mais amplo de IA responsável, especialmente diante do avanço de modelos generativos e do seu uso em larga escala. Documentos estratégicos no Brasil explicitam que o desenvolvimento da IA demanda diretrizes públicas e monitoramento contínuo, com participação de múltiplos atores e estabelecimento de padrões para uso responsável (SBC, 2024).

No Plano da Sociedade Brasileira de Computação (SBC), o debate regulatório é situado como tema central e conectado a propostas legislativas em tramitação, bem como à necessidade de segurança e proteção de dados, com referência expressa à Lei Geral de Proteção de Dados - LGPD (Lei n. 13.709/2018) e à relevância de boas práticas no uso de dados em sistemas de IA generativa (SBC, 2024). Essa orientação é particularmente pertinente à pesquisa, pois deepfakes frequentemente dependem de dados pessoais (imagem/voz) e podem envolver tratamento irregular de informações sensíveis, demandando análise ética que inclua privacidade, consentimento e prevenção de danos.

De modo convergente, o Plano Brasileiro de Inteligência Artificial (IA para o bem de todos) assinala que o país discute marco regulatório e enfrenta o desafio de promover confiança pública e padrões para uso ético e responsável de IA, destacando ainda lacunas significativas na governança de dados para IA e na parametrização ética do desenvolvimento e uso da tecnologia (CGEE, 2025). O documento projeta, entre metas, a implementação de diretrizes claras, criação de comitê nacional de ética em IA e padrões nacionais para avaliação de risco e impacto em setores críticos (CGEE, 2025), o que se conecta diretamente ao objetivo desta

pesquisa de mapear riscos sociais e dilemas éticos em casos concretos de deepfake.

A abordagem ética, nesse quadro, tende a operar com princípios e requisitos como transparência, prestação de contas, prevenção de danos, justiça e proteção de direitos; o desafio central, porém, reside em converter tais princípios em critérios operacionais para análise e regulação de casos concretos.

A literatura internacional sobre governança algorítmica tem insistido que princípios abstratos, embora necessários, mostram-se insuficientes quando desacompanhados de procedimentos de avaliação, classificação de risco, mecanismos de auditoria e estruturas de responsabilização. O NIST AI *Risk Management Framework* propõe precisamente essa passagem do plano principiológico ao plano procedimental, ao sugerir que sistemas de IA sejam avaliados por seus contextos de uso, impactos e vulnerabilidades (NIST, 2023).

Em sentido convergente, a ISO/IEC 23894:2023 estrutura a gestão de risco em IA como processo contínuo de identificação, análise, tratamento e monitoramento, enquanto a Unesco (2021) e a OECD (2019) reforçam que ética em IA deve ser compreendida como compromisso institucional com direitos humanos, robustez técnica e governança responsável.

Assim, a pesquisa se beneficia de uma postura epistemológica crítica e orientada por risco, em que a tecnologia é analisada não apenas por sua capacidade técnica, mas por seus efeitos sociais, assimetrias de poder e impactos sobre direitos, instituições e grupos vulnerabilizados.

2.5 Governança algorítmica, confiança digital e integridade informacional

A literatura internacional recente indica que o avanço da IA generativa e das mídias sintéticas exige deslocamento do debate do plano estritamente tecnológico para o plano da governança algorítmica. Nessa perspectiva, o problema central deixa de ser apenas a existência de conteúdos falsificados e passa a envolver a forma como instituições, platafor-

mas e sistemas normativos são capazes de responder à produção, circulação e autenticação de conteúdos sintéticos.

A noção de governança algorítmica, nesse contexto, refere-se ao conjunto de mecanismos regulatórios, institucionais, técnicos e procedimentais destinados a orientar o desenvolvimento, o uso e a mitigação de riscos de sistemas baseados em IA (OECD, 2019; NIST, 2023; ISO/IEC, 2023). Em vez de compreender a tecnologia como elemento neutro, essa abordagem reconhece que algoritmos e modelos generativos operam em ecossistemas de poder, mercado e circulação informacional, exigindo estruturas de *accountability* compatíveis com seus impactos sociais.

Um dos conceitos centrais desse debate é o de digital trust, entendido como a confiança mínima necessária para que sujeitos, organizações e instituições considerem legítimas e confiáveis as interações realizadas em ambientes digitais. No caso das deepfakes, essa confiança abala-se porque a mídia sintética compromete a expectativa social de autenticidade dos registros audiovisuais e enfraquece a relação entre evidência visual e verdade factual. Nessa linha, Vaccari e Chadwick (2020) demonstram que conteúdos políticos sintéticos podem aumentar incerteza, hesitação e desconfiança, mesmo quando o usuário não aceita integralmente a falsidade apresentada.

Em chave semelhante, o documento das Nações Unidas sobre integridade informacional sustenta que a deterioração da confiança em plataformas e conteúdos digitais compromete a capacidade coletiva de deliberação pública e de resposta institucional baseada em fatos (United Nations, 2023).

Associado a isso, emerge a noção de integridade informacional, que ultrapassa a mera preocupação com desinformação factual. Trata-se da preservação das condições estruturais que permitem reconhecer a proveniência, a autenticidade e a confiabilidade dos fluxos informacionais em ambientes digitais. Quando essas condições são corroídas, o problema deixa de ser apenas o conteúdo falso individualmente considerado e passa a atingir a própria infraestrutura de verificação social da verdade. É nesse sentido que a literatura sobre práticas responsáveis para mídias

sintéticas enfatiza a necessidade de mecanismos combinados de transparência, proveniência, rotulagem, rastreabilidade e responsabilização compartilhada entre desenvolvedores, plataformas e instituições (Partnership on AI, 2023; C2PA, 2025).

Do ponto de vista epistemológico, essa transformação pode ser compreendida como uma crise das mediações tradicionais da evidência. Em ecossistemas digitais saturados por simulação realista, o audiovisual deixa de funcionar automaticamente como indício forte de verdade e passa a depender de procedimentos externos de autenticação e validação. A consequência é a transição de uma cultura de confiança presumida para uma cultura de verificação permanente, na qual o custo epistêmico da checagem se eleva e recai desproporcionalmente sobre vítimas, imprensa e instituições públicas.

Portanto, a análise ética das deepfakes exige articulação entre epistemologia digital, governança algorítmica e proteção de direitos, sob pena de reduzir o fenômeno a uma discussão simplificada sobre "conteúdos falsos", quando, na realidade, está em jogo a estabilidade das bases normativas e informacionais da vida social contemporânea.

2.6 Eixo 4: Perspectivas aplicadas: cenário brasileiro, evidências digitais e mitigação

No plano aplicado, deepfakes têm sido associadas a golpes, violência reputacional e riscos eleitorais, o que motivou a produção de guias institucionais com orientação para identificação e denúncia. O guia do STF explicita que, se as falsidades atentarem contra a integridade eleitoral, existem canais específicos de reporte, e que, quando envolverem possibilidade de crime (usurpação de identidade, golpes, atentados contra honra/imagem), a denúncia pode ser encaminhada por canais próprios (STF, 2024). Essa dimensão aplicada reforça que a problemática não é abstrata: ela se manifesta em fluxos institucionais de resposta e em práticas sociais de contenção de danos.

Além disso, a literatura aponta desafios diretos para o campo forense, dado que deepfakes podem ser indistinguíveis de vídeos autênticos para observadores e, em certos contextos, até para métodos usuais de verificação, elevando a complexidade da prova audiovisual (Moraes, 2020; Verdoliva, 2020). Esse ponto é central para o recorte do projeto ("manipulação de evidências"), pois sugere que a crise de confiança informacional atinge também o estatuto social da prova, exigindo atualização de protocolos, capacidades periciais e parâmetros de autenticidade para mídias digitais, inclusive com o desenvolvimento de mecanismos de proveniência, autenticação e rastreabilidade técnica (C2PA, 2025).

No debate público brasileiro, há sinais de consolidação de uma agenda de mitigação que articula educação midiática, discussão regulatória e responsabilização. Sob esse viés, Caldeira (2025) registra que, no Brasil, o tema ganha espaço no âmbito regulatório, com exemplos de projetos legislativos citados, além de tensões relacionadas a direitos autorais e desafios técnicos de comprovação de autoria e intenção maliciosa, elementos que impactam diretamente modelos de responsabilização.

Dessa forma, o cenário aplicado sugere três frentes complementares de enfrentamento: a) aprimoramento de governança e padrões, com avaliação de risco e impacto; b) capacidades técnico-institucionais, envolvendo perícia, verificação, canais de denúncia e resposta; e c) medidas socioculturais, como letramento midiático e proteção de grupos vulneráveis.

A literatura internacional reforça que tais frentes não devem operar isoladamente. O relatório da Partnership on AI (2023) enfatiza a necessidade de práticas responsáveis para mídias sintéticas, combinando transparência, proveniência, políticas de uso e responsabilização compartilhada. De modo semelhante, o desenvolvimento de padrões de autenticação e rastreabilidade, como os propostos pela C2PA (2025), indica que a mitigação técnica deve ser articulada a parâmetros regulatórios e organizacionais, e não tratada como solução autossuficiente.

2.7 Diálogo comparado: Brasil, União Europeia, Estados Unidos e Canadá

A análise comparada revela que o enfrentamento das deepfakes e das mídias sintéticas estrutura-se internacionalmente por caminhos regulatórios distintos, embora convergentes quanto ao reconhecimento de que conteúdos sintéticos hiper-realistas podem comprometer a confiança pública, a autenticidade informacional e a proteção de direitos. No caso brasileiro, o debate encontra-se em processo de consolidação, combinando diretrizes programáticas de governança de IA, recomendações institucionais de enfrentamento à desinformação e preocupação crescente com riscos associados à imagem, à privacidade, à fraude e à manipulação de evidências. Em comparação com jurisdições estrangeiras, nota-se que o Brasil avança mais fortemente no plano principiológico e programático do que na positivação de deveres específicos de transparência, autenticação e diligência aplicáveis à circulação de mídias sintéticas.

Na União Europeia, a abordagem regulatória apresenta grau mais elevado de normatividade e densidade procedimental. O AI Act adota lógica explícita de regulação orientada por risco e prevê, em seu regime de transparência, deveres específicos relacionados a conteúdos gerados ou manipulados por IA, especialmente quando podem induzir terceiros a erro. Além disso, o *Digital Services Act* (DSA) reforça deveres de diligência para plataformas digitais, vinculando moderação, rastreabilidade, transparência e mitigação de riscos sistêmicos. Desse modo, o modelo europeu não trata deepfakes apenas como problema de falsidade pontual, mas como questão de governança da infraestrutura digital e de proteção do espaço público informacional (European Union, 2022).

Nos Estados Unidos, observa-se modelo mais fragmentado e menos centralizado do que o europeu. Em vez de um estatuto geral equivalente ao AI Act, o país opera por instrumentos executivos, guias de direitos e frameworks técnicos. A Executive Order 14110, editada em 2023, enfatiza segurança, confiabilidade e uso responsável da IA, enquanto o NIST AI Risk Management Framework consolida um modelo de gestão de riscos baseado em contexto, impactos e mitigação contínua. Ainda que tal arran-

jo fortaleça parâmetros técnicos e de governança, ele deixa maior espaço à autorregulação, à atuação setorial e a respostas regulatórias difusas, o que pode gerar maior flexibilidade, mas também menor uniformidade na disciplina específica das mídias sintéticas (United States, 2023).

No Canadá, por sua vez, o debate regulatório aparece fortemente associado à segurança informacional, à integridade democrática e à atuação coordenada de órgãos públicos em ambiente digital. Documentos institucionais canadenses alertam que deepfakes e outras tecnologias avançadas de síntese ameaçam a democracia, a confiabilidade de conteúdos oficiais e a capacidade de distinguir informação autêntica de conteúdo falsificado. Essa perspectiva aproxima o caso canadense da noção de integridade informacional, pois enfatiza que a insuficiência estatal para provar a autenticidade de conteúdos oficiais pode ampliar incerteza pública e vulnerabilidade a campanhas de desinformação. Embora o modelo canadense ainda não consolide um estatuto unificado tão estruturado quanto o europeu, ele oferece contribuição importante ao destacar a articulação entre segurança, confiança pública e dever institucional de autenticidade (Canada, 2023, 2025).

Comparativamente, os quatro contextos permitem identificar ao menos três padrões. Primeiro, há convergência internacional no reconhecimento de que deepfakes produzem riscos que ultrapassam a esfera individual e alcançam níveis institucionais e sistêmicos. Segundo, a União Europeia representa o modelo mais normativamente estruturado e explícito quanto a deveres jurídicos de transparência e mitigação. Terceiro, Estados Unidos e Canadá oferecem respostas mais descentralizadas, embora robustas em termos de gestão de risco, segurança e confiança digital (European Union, 2024a, 2024b).

O caso brasileiro, nesse cenário, encontra-se em posição intermediária: dispõe de documentos orientadores e de debate regulatório em amadurecimento, mas ainda carece de mecanismos mais claros e operacionais de autenticação, transparência e responsabilização aplicáveis à circulação de mídias sintéticas. Essa comparação reforça a necessidade de que o Brasil avance de um plano predominantemente programático para

um modelo mais integrado de governança orientada por risco, articulando direitos fundamentais, integridade informacional, deveres institucionais e salvaguardas tecnológicas (STF, 2024).

Quadro 3: Síntese comparada dos modelos regulatórios sobre IA, mídias sintéticas e confiança informacional

JURISDIÇÃO	TRAÇO PREDOMINANTE	FOCO REGULATÓRIO	CONTRIBUIÇÃO PARA O DEBATE SOBRE DEEPFAKES
Brasil	Modelo em consolidação	Diretrizes programáticas, governança de IA, proteção de direitos, combate à desinformação	Oferece base principiológica, mas ainda demanda maior operacionalização normativa para transparência, autenticação e responsabilidade
União Europeia	Modelo normativo estruturado	Regulação orientada por risco, deveres de transparência, diligência de plataformas, mitigação de riscos sistêmicos	Representa o paradigma mais denso de governança jurídica para conteúdos sintéticos e riscos informacionais
Estados Unidos	Modelo técnico-executivo descentralizado	Gestão de riscos, segurança, confiabilidade, frameworks técnicos e orientações federais	Contribui com instrumentos operacionais de gestão de risco e governança, embora com menor uniformidade normativa
Canadá	Modelo institucional focado em integridade democrática	Segurança informacional, autenticidade de conteúdos oficiais, confiança pública	Reforça a centralidade da integridade informacional e da autenticação como dever institucional em contextos de desinformação

Fonte: *elaborado pelos autores com base em CGEE (2025), STF (2024), European Union (2022; 2024a; 2024b), United States (2023), NIST (2023) e Canada (2023, 2025).*

2.8 Considerações finais do debate teórico

Dessa forma, o debate internacional indica que deepfakes devem ser analisadas simultaneamente como problema de desinformação, de confiança digital e de governança algorítmica. Essa compreensão amplia

o horizonte teórico do estudo, pois permite interpretar os casos empíricos não apenas como eventos isolados de fraude ou manipulação, mas como expressões de uma transformação estrutural das condições de produção, circulação e validação da informação em ambientes digitais. Assim, a revisão da literatura sustenta a necessidade de uma análise empírica que articule riscos, danos, autenticidade e responsabilidade institucional, fornecendo base conceitual para a etapa metodológica e para a interpretação dos resultados.

Nesse contexto, evidencia-se a pertinência de investigar dilemas éticos e riscos sociais por meio da análise de casos públicos documentados por fact-checking, com categorização temática, visando produzir uma tipologia analítica aplicável a políticas, educação midiática e práticas institucionais de mitigação. Assim, o referencial teórico apresentado fundamenta a transição para a Metodologia, na qual a análise de conteúdo permite operacionalizar categorias e critérios éticos a partir de evidências empíricas documentadas.

3 ASPECTOS METODOLÓGICOS (METODOLOGIA)

A pesquisa caracteriza-se como aplicada, por buscar produzir conhecimento com utilidade social e científica para enfrentar riscos éticos e informacionais associados a deepfakes, e adota uma abordagem mista (quali-quantitativa), com predominância qualitativa. A adoção do desenho misto justifica-se pela necessidade de integrar: a) evidências documentais públicas sobre deepfakes; e b) procedimentos de codificação e síntese descritiva do corpus empírico, permitindo triangulação descritiva e interpretativa. Tal opção metodológica é coerente com a orientação de que projetos de pesquisa articulam pressupostos filosóficos, estratégias e procedimentos metodológicos conforme o problema e os objetivos do estudo (Creswell, 2013). Do ponto de vista analítico, privilegiou-se uma lógica de validação interna fundada na explicitação dos critérios de seleção, codificação, revisão categorial e consolidação interpretativa do corpus, de modo a assegurar rastreabilidade, consistência e coerência entre dados, categorias e inferências.

Quanto ao gênero, trata-se de uma investigação exploratória-descritiva, pois explora um fenômeno contemporâneo em expansão (deepfakes como desinformação) e descreve padrões, impactos percebidos e dilemas éticos emergentes a partir de fontes e depoimentos. No que se refere ao método, o estudo combina: 1. Pesquisa de revisão sistemática da literatura e do estado da arte, voltada à consolidação conceitual e crítica do tema; 2. Pesquisa documental, baseada em casos públicos documentados por fontes abertas e verificáveis; 3. Estudo de casos múltiplos, na medida em que os deepfakes analisados são tratados como ocorrências representativas que permitem inferências sobre situações análogas, observados rigor de coleta, registro e análise (Severino, 2014). Logo, a pesquisa documental mostra-se pertinente porque incide sobre documentos em sentido amplo (incluindo conteúdos digitais e registros audiovisuais), cuja matéria-prima pode demandar tratamento analítico, alinhando-se ao escopo do objeto investigado (Severino, 2014).

O universo da pesquisa compreende casos públicos de deepfakes registrados e descritos por plataformas de verificação de fatos (*fact-checking*) e fontes abertas confiáveis, no recorte temporal de 2021 a 2025. A amostragem é intencional (não probabilística), adotando-se como critério mínimo a seleção de ao menos 10 casos com documentação suficiente para: I) descrição do conteúdo manipulado; II) contexto de circulação; III) estratégia de engano; IV) evidências de verificação/checagem; e V) repercussões sociais observáveis.

O corpus empírico da etapa aplicada é constituído por 20 casos identificados e tratados exclusivamente por codificação alfanumérica (DFK001 a DFK020) e por síntese contextual, sem inclusão de dados pessoais adicionais. O uso do corpus observou finalidade acadêmico-científica e cautelas de minimização, em conformidade com a Lei nº 13.709/2018 - LGPD, com destaque para a hipótese legal do art. 7º, II (cumprimento de obrigação legal ou regulatória) quando aplicável em contextos institucionais, além de boas práticas de ética em pesquisa ao lidar com conteúdo potencialmente sensíveis (Brasil, 2018).

A coleta é operacionalizada por instrumentos complementares. Em primeiro lugar, utiliza-se protocolo de extração documental dos casos, estruturado em planilha padronizada para registrar variáveis descritivas e analíticas, tais como: fonte de checagem, tipo de deepfake, finalidade aparente, canal de disseminação, indícios técnicos relatados e consequências observáveis. Em segundo lugar, adota-se procedimento de codificação temática aplicado ao corpus, orientado pela análise de conteúdo (Bardin, 2011). As unidades de registro são definidas como os casos individualmente identificados (DFK001 a DFK020), enquanto as unidades de contexto correspondem às descrições, finalidades, impactos, observações e categorias éticas atribuídas a cada ocorrência. Esse procedimento favorece a rastreabilidade do corpus e a consistência do tratamento analítico, e permite que a passagem da descrição à inferência fosse realizada com base em critérios explícitos de categorização.

A definição do corpus e o encerramento da etapa de categorização observam também o critério de saturação analítica, entendido, neste estudo, como o ponto em que a inclusão de novos casos deixa de produzir categorias substantivamente novas e passou apenas a reiterar padrões já identificados. Assim, embora o recorte inicial preveja um número mínimo de 10 casos, a ampliação para 20 ocorrências permite confirmar recorrências temáticas, estabilizar a tipologia construída e aumentar a densidade interpretativa dos resultados. A saturação, portanto, não foi tratada como mera quantidade de casos, mas como evidência de recorrência, redundância temática e suficiência analítica para sustentar inferências coerentes com os objetivos da pesquisa.

A organização, classificação, codificação e consolidação dos dados são realizadas em planilhas eletrônicas elaboradas pelos autores, assegurando: a) rastreabilidade de cada unidade de registro; b) controle de versões; c) reprodutibilidade do procedimento; d) geração de tabelas, quadros e sínteses descritivas; e e) visualização gráfica dos padrões identificados. No plano analítico, as planilhas permitem comparar categorias, revisar codificações, controlar inconsistências e documentar as decisões interpretativas adotadas ao longo do processo, conferindo maior transparência ao percurso metodológico. Desse modo, o uso do instrumento

não se limitou à organização operacional dos dados, mas constituiu parte integrante da estratégia de validação interna da análise.

Ferramentas de IA são utilizadas exclusivamente como suporte instrumental aos pesquisadores, sem substituição do julgamento acadêmico e sem automação decisória sobre resultados. Para isso, utilizam-se: Perplexity no apoio à redação e checagem de coerência textual e terminológica; Litmaps no apoio ao mapeamento de conexões bibliográficas e aprofundamento do estado da arte; PDF.ai no apoio à leitura orientada e extração assistida de trechos para organização do referencial; ChatGPT5 versão paga - pro (modo "Estudar e aprender") no apoio para compreensão de conceitos específicos, elaboração de sínteses de trabalho e revisão de coesão argumentativa.

Em todas as situações, a decisão final sobre categorias, inferências e redação permanece sob responsabilidade dos pesquisadores, preservando-se o princípio de que a análise qualitativa envolve leitura, codificação contínua e interpretação crítica orientada por critérios científicos (Creswell, 2013).

Em síntese, o percurso metodológico adotado procura assegurar consistência entre problema, objetivos, corpus e procedimento analítico, combinando seleção intencional de casos, codificação temática, dupla revisão categorial e saturação analítica. Tal estrutura permite interpretar os deepfakes analisados não apenas como eventos isolados, mas como manifestações recorrentes de riscos éticos e sociais identificáveis, fornecendo base empírica suficiente para a construção da tipologia apresentada na seção seguinte.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

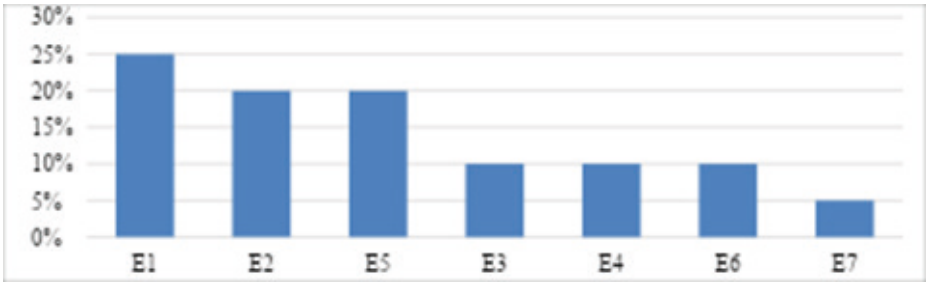
A codificação desta pesquisa resulta em sete eixos temáticos. Observa-se predominância de casos com impacto social alto e predominância de modalidades audiovisuais (vídeo e multimodal), seguidas por áudios/voz. A Tabela 1 sintetiza a distribuição dos casos por eixo temático, e as Tabelas 2 e 3 descrevem modalidade e impacto social.

Tabela 1: Distribuição dos casos por eixo temático (n=20)

EIXO TEMÁTICO	FREQUÊNCIA	PERCENTUAL (%)
E1: Desinformação político-eleitoral/militar	5	25,0
E2: Fraudes financeiras e golpes (incl. corporativos)	4	20,0
E5: Usos criativos/educacionais e entretenimento	4	20,0
E3: Violência sexual/dano à imagem e direitos da personalidade	2	10,0
E4: Publicidade/marketing e uso indevido de imagem	2	10,0
E6: Saúde pública e causas humanitárias	2	10,0
E7: Caos social e efeitos sistêmicos	1	5,0

Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Figura 1: Distribuição dos casos por eixo temático (n=20).



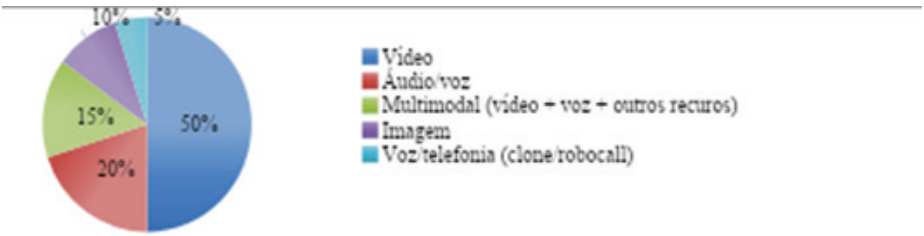
Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Tabela 2: Modalidade do deepfake nos casos analisados (n=20)

MODALIDADE	FREQUÊNCIA	PERCENTUAL (%)
Vídeo	10	50,0
Áudio/voz	4	20,0
Multimodal (vídeo + voz + outros recursos)	3	15,0
Imagem	2	10,0
Voz/telefonia (clone/robocall)	1	5,0

Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Figura 2: Modalidade do deepfake nos casos analisados (n=20).



Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Tabela 3: Impacto social reportado (n=20)

IMPACTO SOCIAL	FREQUÊNCIA	PERCENTUAL (%)
Alto	15	75,0
Médio	5	25,0

Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Figura 3: Impacto social reportado nos casos analisados (n=20).

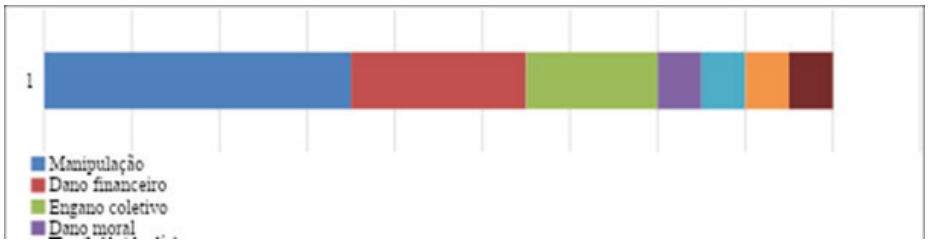
Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Quadro 1: Categorias éticas mais recorrentes (termos declarados no banco)

CATEGORIA ÉTICA (TERMO)	FREQUÊNCIA	PERCENTUAL (%)
Manipulação	7	35,0
Dano financeiro	4	20,0
Engano coletivo	3	15,0
Dano moral	2	10,0
Manipulação da verdade	2	5,0
Engano	2	5,0
Fraude democrática	2	5,0
Uso indevido de imagem	2	5,0

Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

Figura 4: Frequência das categorias éticas mais recorrentes no corpus (n=20).



Fonte: *elaborado pelos autores com base no Banco de Casos Deepfake, PIBIC/CNPq (2025).*

A concentração de impacto alto (75%) sugere que, no recorte analisado, deepfakes tendem a alcançar maior visibilidade e documentação quando associadas a dano efetivo (financeiro, reputacional, político ou sistêmico). Tal achado é convergente com a literatura que associa deepfakes à erosão de confiança pública e à intensificação de riscos informacionais e institucionais (Moraes, 2020; STF, 2024; CGEE, 2025; SBC, 2024).

4.1 Discussão por eixos temáticos emergentes

Eixo E1: Desinformação político-eleitoral/militar

Os casos classificados no eixo E1 evidenciam deepfakes como instrumentos de influência e desorganização deliberativa, com potencial de afetar segurança, eleições e percepção pública de autoridades. Em DFK001, por exemplo, registra-se a circulação de um vídeo manipulado em contexto de conflito, associado à indução de comportamento coletivo. Em termos analíticos, o eixo E1 concentra marcadores éticos como manipulação, fraude democrática e engano coletivo, indicando que o dano extrapola o indivíduo e incide sobre bens públicos informacionais, como integridade do debate público e confiança institucional (STF, 2024; CGEE, 2025).

A presença recorrente de finalidades político-eleitorais e militares no corpus sustenta a interpretação de que deepfakes podem operar como risco macroinstitucional quando associadas a contextos de alta sensibilidade pública, deslocando o fenômeno do plano individual para o plano sistêmico, conforme também sugerem diagnósticos nacionais sobre governança de IA orientada por risco (SBC, 2024; CGEE, 2025).

Eixo E2: Fraudes financeiras e golpes (incl. corporativos)

O eixo E2 reúne casos nos quais deepfakes são mobilizadas como mecanismo de engenharia social e fraude. Em DFK003, descreve-se a ocorrência de transferência de grande monta após interação em videoconferência com supostas lideranças corporativas, posteriormente reconhecidas como simulações sintéticas. O padrão observado sugere que a

deepfake atua como amplificador de credenciais: não substitui o golpe, mas eleva sua eficácia ao simular sinais de autoridade (voz, face e contexto).

A recorrência de dano financeiro como categoria ética associada ao eixo E2 evidencia que o prejuízo é mensurável e, simultaneamente, depende de ambientes de confiança operacional (plataformas de videoconferência e comunicação remota). Assim, os achados reforçam a necessidade de protocolos de verificação e governança de risco em organizações, convergindo com diretrizes nacionais que destacam segurança, responsabilidade e mitigação no ciclo de vida de sistemas de IA (SBC, 2024; CGEE, 2025).

Eixo E3: Violência sexual, dano à imagem e direitos da personalidade

Os casos do eixo E3 evidenciam deepfakes como tecnologia de violência reputacional e potencial violação de direitos da personalidade. Em DFK005, registra-se a associação indevida de identidade a conteúdo íntimo falso, vinculada a dano moral e violação de direitos. Ainda que numericamente menos frequente no corpus, este eixo apresenta gravidade elevada, pois o dano reputacional tende a ser persistente, com assimetria entre facilidade de produção/difusão e dificuldade de reparação. Em chave ética, os dados reforçam a centralidade de consentimento, dignidade e proteção de direitos em políticas de mitigação (STF, 2024; CGEE, 2025).

Eixo E4: Publicidade/marketing e uso indevido de imagem

No eixo E4, deepfakes aparecem como mecanismo de persuasão comercial, com recorrência do marcador "uso indevido de imagem". Os achados sinalizam que o fenômeno não se restringe ao domínio criminal, podendo ser incorporado como estratégia de marketing, o que tensiona fronteiras entre criatividade, consentimento e transparência. Sob perspectiva crítica, a normalização de usos comerciais sem padrões claros pode ampliar a zona cinzenta ética e favorecer degradação cumulativa da confiança informacional (SBC, 2024).

Eixo E5: Usos criativos/educacionais e entretenimento

O eixo E5 reúne ocorrências classificadas como entretenimento, educação, arte e acessibilidade. A presença desse eixo no corpus é analiticamente relevante por evidenciar a ambivalência do fenômeno: a síntese realista pode gerar valor social quando orientada por finalidades legítimas e por parâmetros éticos adequados. Assim, os dados sustentam que a abordagem normativa mais consistente tende a ser proporcional ao risco, distinguindo usos maliciosos de usos consentidos e transparentes, em consonância com perspectivas de governança responsável (CGEE, 2025; SBC, 2024).

Eixo E6: Saúde pública e causas humanitárias

No eixo E6, identificam-se usos vinculados a campanhas e comunicação em saúde. Em DFK020, registra-se campanha com recurso de síntese para ampliar alcance e compreensão pública. Ainda que envolva manipulação audiovisual, a finalidade humanitária desloca a análise para critérios de ética aplicada: transparência, autorização, prevenção de confusão e avaliação de impacto. O resultado reforça a necessidade de salvaguardas específicas em temas sensíveis como saúde, de modo a proteger confiança pública e reduzir efeitos colaterais.

Eixo E7: Caos social e efeitos sistêmicos

O eixo E7 condensa efeito sistêmico mensurável. Em DFK008, descreve-se a viralização de imagem sintética associada a evento crítico, seguida de repercussões econômicas de curto prazo. O caso sustenta a noção de risco sistêmico: a verossimilhança de conteúdos sintéticos pode acionar respostas automáticas em redes e mercados diante de sinais visuais plausíveis, reforçando a centralidade do tema para protocolos de verificação e governança pública.

Complemento: acessibilidade tecnológica (ranking de softwares)

O ranking de softwares evidencia empiricamente a redução de barreiras operacionais para criação de conteúdos sintéticos. A maior parte

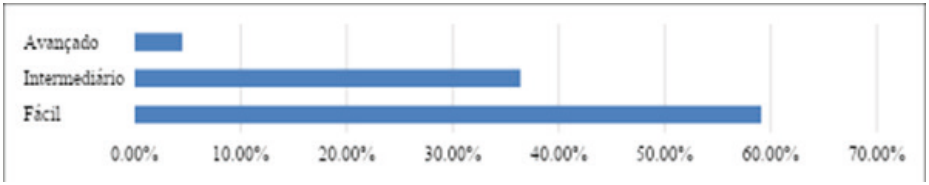
das ferramentas listadas foi classificada como de dificuldade fácil ou intermediária, o que reforça a hipótese de banalização operacional e amplia a relevância de medidas de mitigação orientadas por risco (CGEE, 2025; SBC, 2024).

Tabela 4: Nível de dificuldade das ferramentas listadas (n=22)

NÍVEL DE DIFICULDADE	FREQUÊNCIA	PERCENTUAL (%)
Fácil	13	59,1
Intermediário	8	36,4
Avançado	1	4,5

Fonte: *elaborado pelos autores com base no Banco de Dados - Ferramentas, PIBIC/CNPq (2025).*

Figura 5: Nível de dificuldade das ferramentas listadas (n=22).



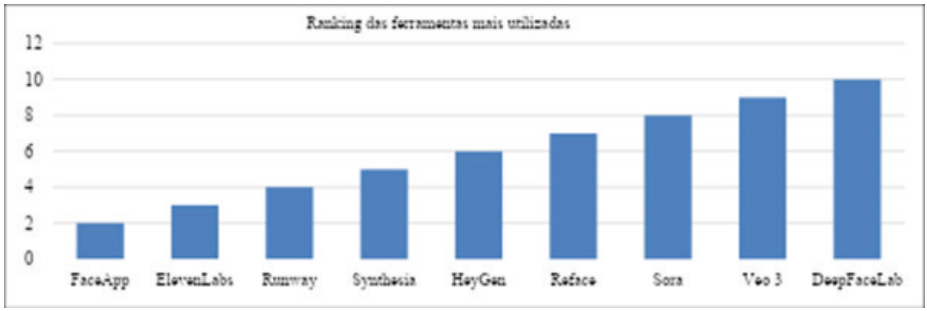
Fonte: *elaborado pelos autores com base no Banco de Dados - Ferramentas, PIBIC/CNPq (2025).*

Tabela 5: Top 10 ferramentas por classificação de uso (síntese)

RANKING	FERRAMENTA	CLASSIFICAÇÃO DE USO	PLATAFORMA	DIFICULDADE
1	FaceApp	Edição de foto por IA	Android, iOS	Fácil
2	ElevenLabs	Síntese de voz (texto para fala) e clonagem de voz	Web	Fácil
3	Runway	Geração/edição de vídeo e troca de rosto/corpo	Web + iOS	Intermediário
4	Synthesia	Vídeo com avatares IA (produção profissional)	Web	Intermediário
5	HeyGen	Geração de vídeos com IA (avatar falante)	Web	Intermediário
6	Reface	Inserção de rosto em vídeos, GIFs e fotos	Android, iOS	Fácil
7	Sora	Geração de vídeos por texto/ imagem	Web	Fácil
8	Veo 3	Geração de vídeos por texto/ imagem	Web	Fácil
9	DeepFaceLab	Troca completa de rostos em vídeos	Windows	Avançado
10	Grok	Geração multimodal (conforme uso)	Web, App	Fácil

Fonte: *elaborado pelos autores com base no Banco de Dados - Ferramentas, PIBIC/CNPq (2025).*

Figura 6: Top 10 ferramentas por classificação de uso: nível de dificuldade (escala ordinal).



Fonte: *elaborado pelos autores com base no Banco de Dados - Ferramentas, PIBIC/CNPq (2025).*

A predominância de ferramentas classificadas como fáceis (59,1%) sugere que o risco social não depende exclusivamente de agentes altamente especializados. Ao reduzir custo, tempo e conhecimento exigido, o ecossistema tecnológico amplia a superfície de uso malicioso e eleva a necessidade de padrões de transparência, verificação e responsabilização, conforme recomendações presentes em documentos nacionais (SBC, 2024; CGEE, 2025; STF, 2024).

Sob perspectiva teórica ampliada, os achados desta pesquisa corroboram o argumento de que deepfakes constituem ameaça à integridade informacional e ao *digital trust,* uma vez que desorganizam expectativas sociais de autenticidade e deslocam o ônus da verificação para atores vulneráveis e para instituições de resposta.

Nessa direção, a contribuição do estudo não reside apenas em classificar casos de uso malicioso ou ambivalente, mas em demonstrar que a problemática ética das mídias sintéticas deve ser interpretada como problema de governança algorítmica aplicada, no qual riscos informacionais, dano reputacional, autenticação e *accountability* precisam ser tratados de forma integrada. Essa leitura é consistente com as proposições internacionais de governança orientada por risco e de responsabilização multissetorial na gestão de IA generativa (NIST, 2023; Unesco, 2021; Partnership on AI, 2023).

4.2 Framework normativo-autoral para governança orientada por risco de deepfakes

Com base na tipologia empírica construída a partir dos sete eixos temáticos identificados, propõe-se um framework normativo-autoral de governança orientada por risco para deepfakes, estruturado em quatro dimensões complementares: risco informacional, responsabilidade institucional, salvaguardas tecnológicas e proteção de direitos fundamentais. O objetivo do modelo é converter os achados empíricos em uma arquitetura analítica e propositiva capaz de orientar respostas regulatórias, organizacionais e sociotécnicas ao fenômeno investigado. Em vez de tratar deepfakes apenas como conteúdo falso ou ilícito isolado, o framework parte da premissa de que a circulação de mídias sintéticas afeta ecossistemas de confiança, autenticidade, deliberação pública e tutela de direitos, exigindo abordagem integrada e proporcional ao risco.

Na primeira dimensão, denominada risco informacional, situam-se os casos em que deepfakes incidem diretamente sobre a confiança pública, a estabilidade do debate democrático e a capacidade coletiva de distinguir o verdadeiro do falso. Essa dimensão emerge, sobretudo, dos eixos E1 e E7, nos quais se verificaram ocorrências de desinformação político-eleitoral, manipulação institucional e efeitos sistêmicos associados à circulação de conteúdos sintéticos. O principal fundamento normativo dessa dimensão consiste na necessidade de reconhecer a integridade informacional como bem jurídico e social a ser protegido, o que implica priorizar mecanismos de resposta em contextos de alto impacto público e risco de desorganização deliberativa.

A segunda dimensão, relativa à responsabilidade institucional, decorre dos achados que evidenciaram a participação ou omissão de múltiplos atores na produção, circulação, monetização e mitigação de deepfakes. Os eixos E2, E4 e parte do E1 demonstram que organizações, plataformas, agentes econômicos e instituições públicas podem ser afetados ou corresponsáveis pela amplificação de danos quando não adotam protocolos mínimos de verificação, resposta e contenção. Nessa perspectiva, o framework propõe que a governança de deepfakes não

recaia apenas sobre o usuário final ou sobre a figura abstrata do "criador do conteúdo", mas seja compartilhada entre desenvolvedores, plataformas, instituições e agentes organizacionais, em lógica de *accountability* distribuída e orientada por contexto de uso.

A terceira dimensão corresponde às salvaguardas tecnológicas, compreendidas como mecanismos de autenticação, proveniência, rastreabilidade e transparência aptos a reduzir o potencial de engano e aumentar a verificabilidade dos conteúdos sintéticos. Os dados da pesquisa mostram que a acessibilidade tecnológica e a redução das barreiras operacionais ampliam a superfície de uso malicioso, o que torna insuficiente uma resposta exclusivamente repressiva. Por isso, o framework sugere que a mitigação de risco seja acompanhada do uso progressivo de instrumentos como rotulagem, metadados de origem, padrões de proveniência e autenticação técnica, sem, contudo, pressupor que a tecnologia, isoladamente, resolva o problema normativo e institucional.

A quarta dimensão é a proteção de direitos fundamentais, especialmente relevante nos eixos E3, E4, E5 e E6, nos quais apareceram situações envolvendo imagem, honra, privacidade, consentimento, dignidade e uso socialmente ambivalente da síntese realista. Essa dimensão parte do reconhecimento de que o enfrentamento jurídico e regulatório das deepfakes não pode ser orientado apenas por lógica securitária ou de combate à desinformação, devendo preservar a centralidade da pessoa humana e o equilíbrio entre tutela de direitos, liberdade de expressão, inovação tecnológica e usos legítimos de IA generativa. Assim, o framework não propõe proibição absoluta da tecnologia, mas um modelo de diferenciação normativa por grau de risco, dano potencial e contexto de aplicação.

Quadro 2: Framework normativo-autoral para governança orientada por risco de deepfakes

DIMENSÃO	FINALIDADE REGULATÓRIA	RISCOS CENTRAIS IDENTIFICADOS	RESPOSTA NORMATIVA/ INSTITUCIONAL SUGERIDA
Risco informacional	Preservar a integridade informacional e a confiança pública	Desinformação político-eleitoral, manipulação da verdade, caos informacional, instabilidade sistêmica	Classificação de risco por contexto; prioridade de resposta em casos de alto impacto público; protocolos de verificação e contenção rápida
Responsabilidade institucional	Distribuir *accountability* entre os atores do ecossistema digital	Fraudes corporativas, engenharia social, monetização indevida, omissão de plataformas	Deveres de diligência, compliance digital, verificação reforçada, responsabilização proporcional de plataformas, organizações e agentes
Salvaguardas tecnológicas	Aumentar autenticidade, rastreabilidade e transparência dos conteúdos	Uso massivo de ferramentas acessíveis, simulação hiper-realista, dificuldade de detecção	Rotulagem, proveniência, autenticação, metadados verificáveis, registros de origem, mecanismos técnicos de identificação
Proteção de direitos fundamentais	Assegurar tutela da dignidade, imagem, privacidade e consentimento	Violência reputacional, pornografia não consensual, publicidade enganosa, usos ambivalentes	Proteção reforçada da personalidade, transparência, consentimento, diferenciação por risco, balanceamento entre inovação e direitos

Fonte: *elaborado pelos autores com base nos resultados da pesquisa (2026).*

O framework proposto permite deslocar a discussão de um plano meramente classificatório para um plano normativo-operacional, no qual os achados empíricos passam a sustentar critérios de intervenção regulatória e institucional. Sua principal contribuição reside em demonstrar que deepfakes exigem respostas graduadas, e não uniformes: casos de alto impacto público e manipulação político-institucional reclamam mecanismos céleres de contenção; ocorrências de fraude corporativa e comercial exigem reforço de deveres institucionais de diligência; usos massivos da tecnologia demandam salvaguardas técnicas mínimas; e situações que envolvam imagem, intimidade e dignidade requerem tutela reforçada de direitos fundamentais.

Sob essa perspectiva, o framework aqui proposto não é meramente classificatório, mas normativo em sentido material: ele parte da premissa de que riscos diferenciados exigem respostas diferenciadas, porém

não arbitrárias. Isso significa reconhecer que a circulação de conteúdos sintéticos sem autenticação, sem transparência quanto à origem e sem mecanismos mínimos de rastreabilidade não pode ser tratada como simples contingência tecnológica, sobretudo quando produz efeitos sobre direitos da personalidade, confiança institucional ou estabilidade democrática. A tese que orienta esta subseção é a de que a governança de deepfakes deve se estruturar sobre um núcleo mínimo de deveres preventivos e procedimentais, aptos a reduzir opacidade, ampliar verificabilidade e redistribuir adequadamente a responsabilidade entre os atores do ecossistema.

Assim, a tipologia empírica deixa de ser apenas descritiva e passa a funcionar como base para um modelo autoral de governança orientada por risco, coerente com o problema de pesquisa, com os resultados identificados e com o referencial teórico mobilizado.

Logo e em síntese, o framework normativo-autoral aqui proposto representa a principal contribuição teórico-aplicada do estudo, pois sistematiza a passagem entre evidência empírica e proposição normativa. Ao articular risco informacional, responsabilidade institucional, salvaguardas tecnológicas e proteção de direitos fundamentais, o modelo oferece uma chave interpretativa e regulatória capaz de orientar futuras pesquisas, políticas públicas, práticas organizacionais e estratégias de mitigação em contextos de circulação de mídias sintéticas. Dessa forma, o estudo avança do diagnóstico dos danos para a formulação de um esquema analítico aplicável ao debate contemporâneo sobre IA generativa e governança orientada por risco.

5 CONSIDERAÇÕES FINAIS

O estudo tem como objetivo investigar os dilemas éticos e os riscos sociais envolvidos na geração e disseminação de conteúdos deepfake produzidos por meio de IA generativa, a partir da análise de casos públicos documentados entre 2021 e 2025. À luz do referencial teórico mobilizado, os resultados permitem demonstrar que o impacto das deepfakes

extrapola a dimensão da falsificação técnica e incide diretamente sobre estruturas de credibilidade, processos decisórios, proteção de direitos e estabilidade do espaço público informacional. Em termos analíticos, a pesquisa evidencia que deepfakes devem ser compreendidas como fenômeno sociotécnico de elevada complexidade, que tensiona simultaneamente confiança pública, governança algorítmica, integridade informacional e tutela de direitos fundamentais.

A análise de conteúdo resulta na identificação de sete eixos temáticos recorrentes, revelando a predominância de ocorrências de alto impacto associadas à desinformação político-institucional, a fraudes financeiras, à violência reputacional e a efeitos sistêmicos sobre confiança pública e estabilidade informacional. Também se observa usos ambivalentes da síntese realista, especialmente em contextos de educação, acessibilidade, saúde pública e comunicação institucional, o que reforça a inadequação de respostas normativas uniformes ou exclusivamente proibitivas. Os achados permitem concluir que o problema das deepfakes não reside apenas na falsidade do conteúdo produzido, mas na capacidade de tais artefatos de corroer as condições estruturais de autenticação, verificação e reconhecimento social da verdade.

No plano científico, a principal contribuição do estudo consiste na construção de uma tipologia empírico-analítica e, a partir dela, na proposição de um framework normativo-autoral de governança orientada por risco, estruturado em quatro dimensões: risco informacional, responsabilidade institucional, salvaguardas tecnológicas e proteção de direitos fundamentais. Essa contribuição permite deslocar o debate de uma abordagem meramente descritiva para uma perspectiva mais operacional, na qual os dados empíricos passam a sustentar critérios de resposta regulatória e institucional graduada. Assim, o artigo agrega valor teórico ao articular ética da IA, desinformação, confiança digital e governança algorítmica, oferecendo um modelo interpretativo aplicável a futuras pesquisas e a contextos decisórios concretos.

No plano regulatório, os resultados indicam que o enfrentamento das deepfakes exige avanço para além de respostas pontuais, casuísticas

ou exclusivamente repressivas. Os achados sugerem que modelos regulatórios mais consistentes devem incorporar, de forma articulada, deveres de transparência, mecanismos de autenticação e proveniência, protocolos de diligência institucional, critérios de classificação de risco e estruturas proporcionais de responsabilização. A análise comparada desenvolvida no artigo evidenciou que a União Europeia apresenta o desenho mais estruturado nesse campo, ao passo que Estados Unidos e Canadá oferecem aportes importantes em gestão de risco, segurança informacional e confiança pública. O caso brasileiro, por sua vez, ainda se encontra em estágio de consolidação, com densidade programática relevante, mas com necessidade de maior operacionalização normativa e procedimental. Nesse sentido, os resultados reforçam que a agenda regulatória nacional pode beneficiar-se de uma transição de diretrizes principiológicas para mecanismos concretos de autenticação, rastreabilidade, rotulagem, verificação reforçada e *accountability* institucional em contextos de circulação de mídias sintéticas de alto risco.

No plano gerencial e institucional, os achados oferecem implicações práticas relevantes para organizações públicas e privadas. Deepfakes exigem revisão de protocolos de autenticação de identidade, validação de ordens e comunicações, controle de risco em ambientes digitais e critérios mais robustos de prevenção a fraudes e engenharia social. Instituições que dependem intensamente de registros audiovisuais, videoconferências, comunicação digital e provas digitais tendem a se tornar mais vulneráveis quando operam sob pressupostos tradicionais de autenticidade. Desse modo, o estudo sugere que organizações adotem mecanismos mínimos de diligência, trilhas de auditoria, autenticação em múltiplas camadas e práticas de verificação compatíveis com o aumento da capacidade de simulação sintética. Além disso, reforça-se a necessidade de que plataformas e agentes do ecossistema digital assumam responsabilidade compartilhada na redução de riscos informacionais, evitando deslocar integralmente para usuários e vítimas o ônus da checagem.

A partir dessas limitações, sugerem-se quatro direções para futuras pesquisas. A primeira é a ampliação do corpus em termos temporais, geográficos e linguísticos, permitindo comparações mais finas entre con-

textos sociopolíticos e plataformas. A segunda é o aprofundamento da interface entre deepfakes e prova digital, com exame de autenticidade, cadeia de custódia, admissibilidade e valor probatório de mídias sintéticas em ambientes jurídicos e institucionais. A terceira consiste na avaliação empírica da eficácia de mecanismos de mitigação, como marcas d'água, rotulagem, metadados de proveniência e protocolos de autenticação. A quarta envolve o estudo dos usos socialmente benéficos de síntese realista, buscando delimitar condições éticas e regulatórias que preservem inovação, acessibilidade e utilidade pública sem comprometer confiança e direitos.

De modo específico, a contribuição científica do estudo explicita-se em quatro dimensões complementares. No plano teórico, o artigo articula deepfakes, integridade informacional, confiança digital e governança orientada por risco, oferecendo leitura integrada de um fenômeno frequentemente tratado de forma fragmentada. No plano metodológico, apresenta-se uma aplicação sistematizada da análise de conteúdo a casos públicos documentados, com codificação temática, dupla revisão categorial e saturação analítica, permitindo a construção de uma tipologia empírico-analítica replicável. No plano aplicado e institucional, os resultados fornecem subsídios para protocolos de verificação, autenticação, prevenção de fraudes, gestão de riscos e educação midiática. No plano normativo, o framework autoral proposto converte os achados empíricos em quatro dimensões regulatórias, risco informacional, responsabilidade institucional, salvaguardas tecnológicas e proteção de direitos fundamentais, oferecendo instrumento analítico útil para políticas públicas, organizações e futuras pesquisas sobre mídias sintéticas.

Por fim, responde-se objetivamente à pergunta-problema da pesquisa: os dilemas éticos e os riscos sociais implicados na geração e disseminação de deepfakes produzidos por IA generativa concentram-se na erosão da confiança informacional, na amplificação de fraudes e manipulações, na intensificação de violações de direitos da personalidade e na ambivalência de usos legítimos que exigem salvaguardas específicas. A partir dos achados, sustenta-se que a circulação de deepfakes sem mecanismos mínimos de proveniência, autenticação e transparência não constitui apenas proble-

ma técnico ou comunicacional, mas risco sociotécnico de relevância jurídico-institucional, capaz de comprometer direitos fundamentais, confiança pública e estabilidade do espaço informacional. Diante disso, conclui-se que o enfrentamento das deepfakes deve avançar para um modelo de governança orientada por risco, com implicações regulatórias, institucionais e operacionais mais densas, aptas a combinar proteção de direitos, autenticação informacional, responsabilização proporcional e resposta coordenada à complexidade das mídias sintéticas contemporâneas.

REFERÊNCIAS

BARDIN, Laurence. **Análise de conteúdo.** Lisboa: edições 70, 2011.

BRASIL. **Lei nº 13.709**, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (marco civil da internet). **Diário Oficial da União: seção *1*,** Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 17 out. 2025.

CALDEIRA, Júlia. Deepfakes no cenário brasileiro: efeitos e estratégias de mitigação. In: Souza, Gustavo; Silva, Tarcízio (org.). **Enfrentando deepfakes**. 1. ed. Brasília, DF: desvelar, 2025. p. 23-41. Disponível em: https://desvelar.org/enfrentando-deepfakes. Acesso em: 17 out. 2025.

CANADA. Canadian Security Intelligence Service. **The evolution of disinformation:** a deepfake future. Ottawa: CSIS, 2023. Disponível em: https://www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future/the-evolution-of-disinformation-a-deepfake-future.html. Acesso em: 11 out. 2025.

CANADA. Canadian Security Intelligence Service. **Deepfakes:** a real threat to a Canadian future. Ottawa: CSIS, 2025. Disponível em: https://www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future/deepfakes-a-real-threat-to-a-canadian-future.html. Acesso em 11.03.2026

C2PA. **C2PA technical specification (content credentials).** Coalition for content provenance and authenticity, 2025. Disponível em: https://spec.c2pa.org/specifications/specifications/2.2/specs/_attachments/C2PA_Specification.pdf. Acesso em: 17 mar. 2026.

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS (CGEE). MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO (MCTI)**. IA para o bem de todos:** plano brasileiro de inteligência artificial. Brasília, DF: MCTI; CGEE, 2025. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/ia/plano-brasileiro-de-inteligencia-artificial. Acesso em: 17 out. 2025.

CHESNEY, Robert; CITRON, Danielle Keats. **Deep fakes:** a looming challenge for privacy, democracy, and national security. Rochester, NY: SSRN, 2018. Disponível em: https://papers.ssrn.com/abstract=3213954. Acesso em: 17 mar. 2026.

CRESWELL, John W. **Projetos de pesquisa:** métodos qualitativo, quantitativo e misto. 3. ed. Porto Alegre: penso, 2013.

EUROPEAN UNION. REGULATION (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (digital services act). **Official Journal of the European Union**, Brussels, 2022. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32022R2065. Acesso em: 17 mar. 2026.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). **Official Journal of the European Union**, Brussels, 2024a. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/. Acesso em: 17 mar. 2026.

EUROPEAN UNION. **AI act – article 50 (transparency obligations).** Brussels: artificialintelligenceact.eu, 2024b. Disponível em: https://artificialintelligenceact.eu/article/50/. Acesso em: 17 mar. 2026.

GONÇALVES, Wesley Antônio; ROSA, Lucas Daniel Santana. **Banco de casos deepfake - PIBIC/CNPq 2025**. Patrocínio, MG: IFTM, 2025. Base de dados em planilha eletrônica.

ISO/IEC. **ISO/IEC 23894:2023 – information technology – artificial intelligence – guidance on risk management**. Geneva: iso, 2023. Disponível em: https://www.iso.org/standard/77304.html. Acesso em: 23 nov. 2025.

MORAES, Cristiane Pantoja de. "Deepfake" como ferramenta de manipulação e disseminação de "fakenews" em formato de vídeo nas redes sociais. **Biblios,** Pittsburgh, n. 79, p. 1-15, 2020. DOI: 10.5195/biblios.2020.864. Disponível em: https://biblios.pitt.edu/ojs/biblios/article/view/864. Acesso em: 12 out. 2025.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). **Artificial intelligence risk management framework (AI RMF 1.0)**. Gaithersburg,

MD: NIST, 2023. Disponível em: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. Acesso em: 17 mar. 2026.

OECD. **Artificial intelligence in society.** Paris: oecd, 2019. Disponível em: https://www.oecd.org/content/dam/oecd/en/publications/reports/2019/06/artificial-intelligence-in-society_c0054fa1/eedfee77-en.pdf. Acesso em: 17 mar. 2026.

PARTNERSHIP ON AI. **Responsible practices for synthetic media:** a framework for collective action. San Francisco, CA: partnership on ai, 2023. Disponível em: https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf. Acesso em: 17 mar. 2026.

SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 24. ed. São Paulo: cortez, 2014.

SOCIEDADE BRASILEIRA DE COMPUTAÇÃO (SBC). **Plano de inteligência artificial da sociedade brasileira de computação**. Porto Alegre: SBC, 2024. DOI: 10.5753/sbc.rt.2024.141. Disponível em: https://doi.org/10.5753/sbc.rt.2024.141. Acesso em: 14 out. 2025.

SOUZA, Gustavo; SILVA, Tarcízio (org.). **Enfrentando deepfakes**. 1. ed. Brasília, DF: desvelar, 2025. Disponível em: https://desvelar.org/enfrentando-deepfakes. Acesso em: 14 out. 2025.

SUPREMO TRIBUNAL FEDERAL (STF). **Guia ilustrado contra as deepfakes**. Brasília, DF: STF, 2024. Disponível em: https://www.stf.jus.br/arquivo/cms/cddConteudoTextual/anexo/Guia_Deepfakes.pdf. Acesso em: 17 out. 2025.

UNITED NATIONS. **Information integrity on digital platforms**. New York: united nations, 2023. Disponível em: https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/our-common-agenda-policy-brief-information-integrity-on-digital-platforms-en.pdf. Acesso em: 17 mar. 2026.

UNITED STATES. Executive order 14110: safe, secure, and trustworthy development and use of artificial intelligence. **Federal Register**, Washington, DC, 2023. Disponível em: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Acesso em: 17 mar. 2026.

UNESCO. **Recommendation on the ethics of artificial intelligence**. Paris: unesco, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000380455. Acesso em: 17 mar. 2026.

VACCARI, Cristian; CHADWICK, Andrew. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. **Social media + society**, London, v. 6, n. 1, 2020. DOI: 10.1177/2056305120903408. Disponível em: https://journals.sagepub.com/doi/10.1177/2056305120903408. Acesso em: 17 mar. 2026.

VERDOLIVA, Luisa. Media forensics and deepfakes: an overview. **arXiv**, 2020. Disponível em: https://arxiv.org/abs/2001.06564. Acesso em: 17 mar. 2026.

WESTERLUND, Mika. The emergence of deepfake technology: a review. **Technology innovation management review**, Ottawa, v. 9, n. 11, p. 40-53, 2019. DOI: 10.22215/timreview/1282. Disponível em: https://timreview.ca/article/1282. Acesso em: 17 mar. 2026.

Recebido em 29 de janeiro de 2026.
Aprovado em 19 de maio de 2026.