O USO DA INTELIGÊNCIA ARTIFICIAL EM FAKE NEWS E DEEP FAKES

Autores: Kauan Mendes Pereira, Valentino Isaiama da silva, Bruno Michel Pera.

Instituto de Ensino Ribeiro Goulart, Rua Isac Newton, 37, Jardim Oriente - 12236-240 - São José dos Campos-SP, Brasil, mendeszyar@gmail.com, valentinoisaiama@gmail.com.

Resumo
O presente artigo analisa os impactos da Inteligência Artificial generativa na proliferação e sofisticada criação de notícias falsas e *deepfakes*, examinando as consequências para a segurança digital e a integridade informacional. Por meio de pesquisa bibliográfica e documental com base em literatura recente, o estudo mapeia a dinâmica do engano industrializado e os desafios técnicos associados à detecção de mídias sintéticas. Os resultados indicam que, embora ferramentas computacionais avançadas como modelos baseados em aprendizado profundo e redes adversárias ofereçam mecanismos promissores de mitigação, a alta precisão dos conteúdos gerados exige uma resposta multifacetada. Conclui-se que o enfrentamento à desinformação contemporânea demanda a articulação entre o aprimoramento tecnológico de defesa e o fortalecimento de políticas de educação midiática para preservar a confiança pública no ecossistema digital.

Palavras-chave: Fake news. Desinformação. Detecção de conteúdo.

Área do Conhecimento: Ciências Sociais Aplicadas (Subárea: *Comunicação*)

Introdução

A Inteligência Artificial (IA) vem ocupando um espaço cada vez maior na sociedade contemporânea, sendo utilizada em áreas como educação, comunicação, entretenimento, saúde e produção de conteúdo digital. O avanço das ferramentas de IA generativa possibilitou a criação e a edição de textos, imagens, áudios e vídeos com elevado grau de realismo, tornando a produção de conteúdos digitais mais acessível. Apesar dos benefícios proporcionados por essas tecnologias, seu desenvolvimento também trouxe novos desafios relacionados à confiabilidade das informações disponíveis na internet. Entre esses desafios estão as fake news e os deepfakes, que podem ser utilizados para manipular informações, enganar usuários e influenciar a percepção da realidade. A UNESCO destaca que a IA generativa pode facilitar a produção de conteúdos falsos e deepfakes cada vez mais difíceis de distinguir de materiais verdadeiros (UNESCO, 2023).

As fake news podem ser compreendidas como informações falsas ou enganosas apresentadas de maneira semelhante a conteúdos jornalísticos ou informativos, podendo alcançar grande quantidade de pessoas principalmente por meio das redes sociais e aplicativos de comunicação. Com o desenvolvimento da IA, tornou-se possível produzir textos e outros conteúdos falsos de maneira mais rápida e em maior escala. Os deepfakes, por sua vez, correspondem a conteúdos audiovisuais criados ou modificados artificialmente para fazer com que uma pessoa pareça dizer ou realizar algo que não aconteceu. Segundo a Agência Nacional de Proteção de Dados (ANPD), essa tecnologia envolve áudios, fotos e vídeos modificados ou artificialmente criados por meio de IA generativa, apresentando elevado grau de realismo e trazendo desafios para a proteção de dados e a segurança dos ambientes digitais (ANPD, 2026).

Nesse contexto, a utilização inadequada da Inteligência Artificial para produzir e disseminar informações falsas pode gerar consequências sociais relevantes. Conteúdos manipulados podem prejudicar a reputação de pessoas, influenciar opiniões, provocar desconfiança em instituições e dificultar a distinção entre informações verdadeiras e falsas. Além disso, a velocidade de circulação dos conteúdos nas plataformas digitais aumenta o potencial de alcance dessas informações. A UNESCO aponta que conteúdos sintéticos, incluindo deepfakes, já representam uma preocupação relacionada à integridade das informações e podem contribuir para a perda de confiança nas fontes de informação. Dessa forma, compreender a relação entre IA, fake news e deepfakes torna-se importante não apenas para identificar os riscos dessas tecnologias, mas também para discutir formas responsáveis de utilização e mecanismos capazes de reduzir os impactos da desinformação.

Diante desse cenário, este trabalho tem como objetivo compreender de que maneira o desenvolvimento da Inteligência Artificial facilita a criação e a disseminação de fake news e deepfakes, analisando seus principais impactos na sociedade e buscando identificar possíveis soluções para reduzir sua ocorrência e seus efeitos. Para isso, serão analisadas estratégias relacionadas à educação e alfabetização digital, à verificação das informações, à atuação das plataformas digitais, ao desenvolvimento de mecanismos de identificação de conteúdos manipulados e à utilização responsável da própria Inteligência Artificial. Assim, busca-se responder ao seguinte problema de pesquisa: como o avanço da Inteligência Artificial contribui para a criação e a propagação de fake news e deepfakes e quais medidas podem ser adotadas para reduzir os impactos desse fenômeno? Dessa forma, a pesquisa pretende contribuir para uma compreensão mais crítica sobre o uso dessas tecnologias e para o desenvolvimento de formas de prevenção e combate à desinformação.

Metodologia

O presente trabalho foi desenvolvido por meio de uma pesquisa de abordagem qualitativa, de caráter exploratório e bibliográfico, com a finalidade de compreender a relação entre Inteligência Artificial generativa, produção de fake news e criação de deepfakes. A escolha da abordagem qualitativa esteve relacionada à necessidade de analisar conceitos, características, impactos e diferentes perspectivas presentes na literatura sobre o fenômeno estudado, sem ter como objetivo principal a obtenção ou o tratamento estatístico de dados. Dessa forma, a pesquisa buscou compreender como o avanço das tecnologias de Inteligência Artificial tem influenciado a produção, a manipulação e a disseminação de conteúdos potencialmente desinformativos, considerando também seus impactos tecnológicos e sociais.

Quanto aos procedimentos de pesquisa, foi realizado um levantamento bibliográfico e documental, utilizando artigos científicos, trabalhos acadêmicos, livros, documentos institucionais e publicações de organizações reconhecidas na área. A pesquisa bibliográfica possibilitou reunir estudos relacionados à Inteligência Artificial generativa, desinformação, fake news, deepfakes, redes sociais, detecção automatizada e alfabetização digital. Também foram consultados documentos e publicações de instituições públicas e organizações especializadas, incluindo materiais produzidos pela UNESCO e por instituições brasileiras relacionadas ao estudo e à regulamentação das tecnologias digitais. A seleção das fontes considerou critérios de relevância, credibilidade, atualidade e relação direta com os objetivos da pesquisa, priorizando trabalhos que apresentassem contribuições sobre a evolução da IA, as técnicas utilizadas na produção e detecção de conteúdos sintéticos, seus impactos e as estratégias propostas para o enfrentamento da desinformação.

Após o levantamento das fontes, foi realizada uma leitura analítica e comparativa dos materiais selecionados. Inicialmente, foram identificados os principais conceitos relacionados à Inteligência Artificial, fake news e deepfakes. Em seguida, foram analisadas as formas pelas quais as tecnologias generativas podem facilitar a produção, alteração e disseminação de conteúdos enganosos, bem como os mecanismos desenvolvidos para sua detecção. Também foram observadas as limitações apresentadas pelos estudos, especialmente aquelas relacionadas à capacidade de generalização dos sistemas de detecção, à evolução das técnicas de geração e à necessidade de atualização constante dos modelos. A análise permitiu estabelecer relações entre os diferentes estudos e identificar convergências e divergências quanto aos impactos e aos desafios associados ao uso da IA na produção e no combate à desinformação.

Por fim, os resultados encontrados na literatura e nos documentos analisados foram organizados de forma temática, relacionando o avanço das tecnologias de Inteligência Artificial aos desafios decorrentes da produção e disseminação de conteúdos falsos ou manipulados. Foram consideradas, nesse processo, estratégias como o desenvolvimento de ferramentas automatizadas de detecção, mecanismos de autenticação e rastreabilidade, verificação das informações, educação midiática e digital e adoção de práticas responsáveis por parte de usuários, plataformas e instituições. A análise também considerou a necessidade de compreender as limitações das ferramentas automatizadas, uma vez que os estudos examinados demonstram que a evolução dos métodos de geração pode acompanhar ou superar a capacidade de determinados sistemas de detecção. Dessa maneira, a metodologia permitiu reunir e comparar evidências da literatura para compreender não apenas os efeitos da IA sobre a desinformação, mas também as possibilidades e limitações das estratégias atualmente utilizadas para seu enfrentamento.

Resultados

O levantamento bibliográfico e documental evidencia que o desenvolvimento de modelos de Inteligência Artificial generativa, especialmente os grandes modelos de linguagem (LLMs), transformers e sistemas multimodais, vem modificando significativamente a dinâmica de produção e circulação de conteúdos desinformativos. A literatura recente demonstra que essas tecnologias ampliaram a capacidade de automatização da produção de conteúdos, permitindo gerar, reformular e adaptar informações em diferentes formatos e para diferentes públicos. Loth, Kappes e Pahl (2024), em uma revisão de estudos sobre os impactos da IA generativa nas fake news, identificam que a tecnologia pode atuar tanto na criação quanto na detecção de conteúdos falsos, abrangendo aplicações relacionadas a textos, imagens, vídeos e redes sociais. Esse cenário é aprofundado pelo estudo Industrialized Deception (2026), que caracteriza a expansão de um ecossistema de produção de conteúdos sintéticos no qual a IA permite aumentar a velocidade e a escala da geração de informações, dificultando processos tradicionais de verificação e moderação. Assim, os resultados da literatura sugerem que a principal transformação não está apenas na capacidade de produzir conteúdos falsos, mas na possibilidade de automatizar diferentes etapas desse processo, tornando a produção de materiais potencialmente enganosos mais rápida, diversificada e adaptável.

No campo audiovisual, os resultados demonstram que os deepfakes representam uma das aplicações mais relevantes desse processo de evolução tecnológica. Gambín et al. (2024) mostram, em sua revisão sobre detecção de vídeos manipulados, que técnicas baseadas em aprendizagem profunda, incluindo autoencoders, redes generativas adversariais e transformers, possibilitam alterações cada vez mais sofisticadas em características faciais, expressões, identidade e movimentos labiais, além da geração e manipulação de elementos sonoros. Heidari et al. (2024), em uma revisão sistemática e abrangente sobre métodos de detecção de deepfakes, também identificam uma ampla utilização de redes neurais profundas para analisar vídeos, imagens, áudios e conteúdos multimídia. Os estudos demonstram que o avanço das técnicas generativas elevou o grau de realismo dos conteúdos sintéticos, fazendo com que determinadas manipulações sejam difíceis de identificar apenas pela observação humana. Entretanto, essa dificuldade não significa que todos os deepfakes sejam necessariamente indistinguíveis, uma vez que a capacidade de identificação varia de acordo com a qualidade da manipulação, o tipo de conteúdo, o conjunto de dados utilizado e o método de detecção empregado.

Paralelamente ao avanço das técnicas de geração, os estudos analisados evidenciam uma evolução dos mecanismos destinados à identificação de conteúdos sintéticos. Contudo, essa área ainda apresenta limitações importantes. Gambín et al. (2024) destacam desafios relacionados à disponibilidade e à qualidade dos conjuntos de dados, ao desequilíbrio entre diferentes classes de exemplos, ao custo computacional e, principalmente, à capacidade de generalização dos modelos. Um detector treinado para reconhecer determinado tipo de manipulação pode apresentar desempenho inferior quando confrontado com conteúdos produzidos por técnicas diferentes ou por modelos generativos mais recentes. Heidari et al. (2024) reforçam essa questão ao observar que a avaliação dos sistemas de detecção envolve não apenas métricas de acurácia, mas também aspectos como robustez, complexidade computacional, velocidade de processamento e capacidade de funcionamento em diferentes condições. Dessa forma, os resultados indicam que existe uma espécie de processo contínuo de adaptação entre geração e detecção: à medida que novas técnicas são desenvolvidas para produzir conteúdos sintéticos mais sofisticados, os sistemas de identificação precisam ser atualizados para reconhecer novos padrões de manipulação.

Essa evolução também pode ser observada na detecção automatizada de fake news textuais, especialmente com a utilização de modelos baseados em transformers, como BERT e suas variantes, e com a incorporação de grandes modelos de linguagem às tarefas de classificação. Raza, Paulen-Patterson e Ding (2024) realizaram uma avaliação comparativa envolvendo modelos BERT-like, como BERT, RoBERTa e DistilBERT, e LLMs, como Llama 2 e Mistral. Para isso, os pesquisadores utilizaram um conjunto de 10 mil notícias inicialmente anotadas com auxílio do GPT-4 Turbo e posteriormente avaliadas por oito revisores humanos, entre especialistas e estudantes. Os resultados apresentados demonstraram desempenho elevado dos modelos BERT-like na classificação convencional, com o RoBERTa alcançando 89,23% de precisão e 90,14% de recall no conjunto analisado. Ao mesmo tempo, os experimentos indicaram que os LLMs apresentaram maior resistência em determinadas situações envolvendo alterações ou perturbações nos textos.

Esses resultados demonstram que diferentes arquiteturas possuem características distintas e que a escolha do modelo de detecção depende das condições da tarefa, não sendo possível considerar uma única arquitetura como solução universal para todos os tipos de desinformação.

Outra evidência da evolução dos mecanismos de defesa é apresentada pelo estudo LLM-GAN: Construct Generative Adversarial Network Through Large Language Models for Explainable Fake News Detection (2024), que propõe a integração entre grandes modelos de linguagem e uma estrutura baseada em redes generativas adversariais para a detecção explicável de fake news. A abordagem foi avaliada em conjuntos de dados como Weibo21 e GossipCop e comparada com diferentes métodos de detecção existentes. Os resultados apresentados pelos autores indicam que o modelo proposto apresentou desempenho competitivo e, em determinadas configurações, superior ao uso direto de LLMs por meio de prompts. Além da classificação, o LLM-GAN procura produzir justificativas para suas decisões, acrescentando uma dimensão de explicabilidade ao processo de detecção. Essa característica é relevante porque sistemas empregados no combate à desinformação precisam, em muitos contextos, oferecer mais do que uma simples indicação de que determinado conteúdo é falso ou verdadeiro; é necessário fornecer elementos que permitam compreender quais características ou evidências contribuíram para a classificação realizada pelo modelo.

Diante desse conjunto de evidências, os resultados indicam que a relação entre Inteligência Artificial generativa e desinformação apresenta uma característica de uso dual: as mesmas tecnologias que aumentam a capacidade de criação e manipulação de conteúdos também podem ser empregadas para sua identificação e análise. A literatura analisada aponta, portanto, para uma transformação estrutural no ambiente informacional, caracterizada pela evolução simultânea das técnicas de geração e dos mecanismos de detecção. O problema passa a envolver não apenas a produção de conteúdos falsos, mas também questões relacionadas à escala de geração, à dificuldade de autenticação, à adaptação dos conteúdos, à capacidade de generalização dos detectores e à necessidade de explicabilidade das decisões automatizadas. Loth, Kappes e Pahl (2024) e os estudos posteriores sobre geração e detecção reforçam, nesse sentido, que o enfrentamento da desinformação produzida ou potencializada por IA não depende exclusivamente de um único modelo ou algoritmo, mas de um conjunto de mecanismos tecnológicos e metodológicos capazes de acompanhar a rápida evolução das técnicas generativas. Os resultados encontrados sustentam, assim, a existência de uma dinâmica contínua entre geração e detecção, na qual o avanço de uma dessas capacidades estimula o desenvolvimento da outra.

Discussão

Os dados obtidos na literatura demonstram que a expansão das notícias falsas, dos conteúdos sintéticos e dos deepfakes produzidos ou potencializados por Inteligência Artificial ultrapassa a dimensão de uma falha técnica ou de um problema restrito à segurança digital, configurando um fenômeno com implicações sociais e informacionais mais amplas. A literatura analisada evidencia que a IA generativa possui um caráter de uso dual, pois as mesmas tecnologias empregadas para produzir textos, imagens, áudios e vídeos sintéticos também podem ser utilizadas para identificar e analisar esses conteúdos. Loth, Kappes e Pahl (2024) demonstram que a IA generativa interfere tanto nos processos de criação quanto nos mecanismos de detecção de fake news, enquanto o estudo Industrialized Deception (2026) evidencia que a utilização de grandes modelos de linguagem pode ampliar a escala e a velocidade de produção de conteúdos, aumentando a quantidade de informações que precisam ser avaliadas. Esse cenário modifica o problema da desinformação porque não se trata apenas da existência de conteúdos falsos isolados, mas da possibilidade de produção contínua de grandes volumes de materiais que podem dificultar os processos tradicionais de verificação, especialmente quando diferentes conteúdos são adaptados para públicos e contextos específicos.

Os resultados relacionados aos deepfakes reforçam essa preocupação e demonstram que a evolução das técnicas generativas vem acompanhada de novos desafios para a autenticação de conteúdos audiovisuais. Gambín et al. (2024) identificam o emprego de diferentes arquiteturas de aprendizagem profunda na criação e manipulação de vídeos, enquanto Heidari et al. (2024) demonstram que os métodos de detecção também vêm evoluindo por meio da utilização de redes neurais e outras técnicas de aprendizagem automática. Entretanto, ambos os estudos apontam limitações, especialmente relacionadas à generalização dos detectores, à qualidade dos conjuntos de dados, ao custo computacional e à capacidade de identificar manipulações.

Dessa forma, os resultados não indicam que os deepfakes sejam necessariamente impossíveis de identificar, mas demonstram que a detecção constitui um processo dinâmico, no qual os sistemas precisam ser constantemente atualizados para acompanhar a evolução das técnicas de geração. Essa característica evidencia uma das principais dificuldades do enfrentamento tecnológico da desinformação: enquanto novos métodos de geração podem ser desenvolvidos rapidamente, os mecanismos de detecção precisam ser treinados e validados para reconhecer padrões que também estão em constante transformação.

A mesma dinâmica é observada na detecção de fake news textuais. Raza, Paulen-Patterson e Ding (2024), ao compararem modelos BERT-like e grandes modelos de linguagem, demonstraram que diferentes arquiteturas apresentam desempenhos distintos conforme a tarefa e as condições de análise. Os modelos BERT-like apresentaram desempenho elevado na classificação convencional, enquanto os LLMs demonstraram maior robustez em determinadas situações envolvendo alterações nos textos. De maneira semelhante, a proposta LLM-GAN, apresentada em 2024, demonstra uma tendência de integração entre modelos generativos e mecanismos especializados de detecção, acrescentando também recursos de explicabilidade para que o sistema possa apresentar justificativas para suas classificações. Esses resultados são relevantes porque demonstram que a utilização de Inteligência Artificial na detecção não elimina a necessidade de avaliação crítica dos resultados produzidos pelos próprios sistemas. Um detector pode apresentar elevada precisão em determinado conjunto de dados e, ainda assim, apresentar limitações diante de novos tipos de manipulação. Portanto, a eficácia dessas ferramentas depende da qualidade dos dados, da atualização dos modelos, da capacidade de generalização e da possibilidade de interpretar adequadamente suas decisões.

Diante desse cenário, os resultados sustentam a necessidade de uma abordagem multidimensional para o enfrentamento da desinformação potencializada pela Inteligência Artificial. As ferramentas automatizadas de detecção são importantes porque permitem analisar volumes de conteúdo que seriam inviáveis para avaliação exclusivamente humana, mas não devem ser consideradas uma solução definitiva ou independente. Sua utilização precisa ser acompanhada por mecanismos de verificação, análise de fontes, transparência sobre as limitações dos sistemas e participação humana nos processos de avaliação. Ao mesmo tempo, o fortalecimento da alfabetização midiática e digital pode contribuir para que os usuários desenvolvam maior capacidade de identificar sinais de manipulação, avaliar a confiabilidade das fontes e compreender que uma classificação automatizada também está sujeita a erros. Assim, a principal implicação dos resultados é que o avanço da IA generativa exige um desenvolvimento paralelo das capacidades de detecção, autenticação e avaliação crítica da informação. A literatura analisada, portanto, não permite afirmar que a Inteligência Artificial necessariamente provocará uma perda generalizada de confiança nas instituições ou na mídia, mas demonstra que a expansão da produção automatizada de conteúdos aumenta os desafios para a manutenção de um ambiente informacional confiável e exige respostas que combinem recursos tecnológicos, verificação humana, educação midiática e mecanismos institucionais de controle.

Conclusão

O estudo demonstrou que a ascensão da Inteligência Artificial generativa vem reconfigurando o cenário da desinformação, ampliando a capacidade de produção e circulação de fake news e conteúdos sintéticos, incluindo os deepfakes. Os resultados da literatura analisada evidenciaram que o avanço dessas tecnologias permite maior automatização, escala e sofisticação na criação de conteúdos, ao mesmo tempo em que impõe novos desafios aos mecanismos tradicionais de verificação e detecção. Embora diferentes pesquisas apresentem soluções baseadas em aprendizado de máquina, modelos de linguagem, redes neurais e arquiteturas generativas para identificar conteúdos potencialmente manipulados, também foram identificadas limitações relacionadas à generalização, à atualização dos modelos e à rápida evolução das técnicas de geração. Dessa forma, o enfrentamento da desinformação não pode depender exclusivamente de soluções computacionais, sendo necessária uma abordagem integrada que combine inovação tecnológica, mecanismos de verificação, acompanhamento científico e participação humana. Conclui-se, portanto, que a mitigação dos impactos da desinformação no ambiente digital requer o desenvolvimento contínuo de ferramentas de detecção e autenticação, o fortalecimento da alfabetização midiática e digital e o aperfeiçoamento de estratégias institucionais capazes de acompanhar a evolução da Inteligência Artificial, contribuindo para a construção de um ambiente informacional mais confiável.

Referências

GAMBÍN, Á. F. et al. Deepfake video detection: challenges and opportunities. **Artificial Intelligence Review**, 2024. Disponível em: https://link.springer.com/article/10.1007/s10462-024-10810-6. Acesso em: 24 ago. 2026.

HEIDARI, A. et al. Deepfake detection using deep learning methods: A systematic and comprehensive review. **WIREs Data Mining and Knowledge Discovery**, 2024. Disponível em: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1520. Acesso em: 24 ago. 2026.

INDUSTRIALIZED DECEPTION: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. **arXiv**, 2026. Disponível em: https://arxiv.org/html/2601.21963v1. Acesso em: 24 ago. 2026.

LLM-GAN: Construct Generative Adversarial Network Through Large Language Models For Explainable Fake News Detection. **arXiv**, 2024. Disponível em: https://arxiv.org/pdf/2409.01787. Acesso em: 24 ago. 2026.

LOTH, A.; KAPPES, M.; PAHL, M-O. Blessing or curse? A survey on the impact of generative AI on fake news. **arXiv**, 2024. Disponível em: https://arxiv.org/pdf/2404.03021. Acesso em: 24 ago. 2026.

RAZA, S.; PAULEN-PATTERSON, D.; DING, C. Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language Models with Generative AI-Annotated Data. **arXiv**, 2024. Disponível em: https://arxiv.org/pdf/2412.14276. Acesso em: 24 ago. 2026.

ROLE OF GENERATIVE AI IN FAKE NEWS CREATION AND DETECTION. **Springer Nature Link**, 2025. Disponível em: https://link.springer.com/chapter/10.1007/978-981-96-9275-0_23. Acesso em: 24 ago. 2026.