# IA GENERATIVA: implicações do processamento de linguagem natural (PLN) e visão computacional na segurança da comunicação

# GENERATIVE AI: implications of natural language processing (NLP) and computer vision for communication security

Caio Pereira Silva[1]; Wilson Missina Faria[2]

## RESUMO

O avanço da Inteligência Artificial Generativa (IAG) tem redefinido a comunicação digital, especialmente através de modelos capazes de gerar vídeos sintéticos (*deepfakes*) e realizar a clonagem de voz com altíssimo realismo. Essas capacidades, que combinam avanços do Processamento de Linguagem Natural (PLN) e da Visão Computacional, também impulsionam o desenvolvimento de ferramentas capazes de criar textos, imagens e sons de forma autônoma. Entre essas, destacam-se tradutores automáticos e assistentes virtuais, que facilitam interações e superam barreiras linguísticas, enquanto os progressos em Visão Computacional possibilitam o reconhecimento e a manipulação de imagens e vídeos com alto grau de fidelidade. No entanto, essas tecnologias também apresentam riscos significativos à segurança da informação. Conteúdos falsificados, como *deepfakes* de áudio e vídeo, que imitam com precisão vozes e expressões faciais de pessoas reais, podem ser usados para desinformação e fraudes. Além disso, modelos de linguagem de grande porte podem memorizar dados sensíveis durante o treinamento, aumentando o risco de exposição. Este artigo analisa, de forma geral, os benefícios e ameaças da IA generativa no contexto da comunicação segura, com base em uma revisão bibliográfica crítica. Conclui-se que, apesar dos ganhos de eficiência e acessibilidade, é urgente implementar técnicas de mitigação, como criptografia e *watermarking*, além de regulamentações éticas, para preservar a integridade, confidencialidade e autenticidade das informações no ambiente digital.

[1] Aluno do Curso de Ciências da Computação do Centro Universitário do Sul de Minas. Email: caio.silva8@alunos.unis.edu.br
[2] Professor do Curso de Ciências da Computação do Centro Universitário do Sul de Minas. Email: wmfaria@outlook.com

Palavras-chave: IA Generativa, Segurança da Informação, Processamento de Linguagem, Visão Computacional, Deepfake

## *ABSTRACT*

*The advancement of Generative Artificial Intelligence (GAI) has redefined digital communication, especially through models capable of generating synthetic videos (deepfakes) and performing highly realistic voice cloning. These capabilities, which combine progress in Natural Language Processing (NLP) and Computer Vision, also drive the development of tools capable of autonomously creating texts, images, and sounds. Among these tools are automatic translators and virtual assistants, which facilitate interactions and overcome language barriers, while advances in Computer Vision enable the recognition and manipulation of images and videos with a high degree of fidelity. However, these technologies also pose significant risks to information security. Falsified content, such as audio and video deepfakes that accurately imitate the voices and facial expressions of real people, can be used for misinformation and fraud. In addition, large language models can memorize sensitive data during training, increasing the risk of exposure. This article provides a general analysis of the benefits and threats of generative AI in the context of secure communication, based on a critical literature review. It concludes that, despite the gains in efficiency and accessibility, it is urgent to implement mitigation techniques such as encryption and watermarking, as well as ethical regulations, to preserve the integrity, confidentiality, and authenticity of information in the digital environment.*

*Keywords: Generative AI, Information Security, Language Processing, Computer Vision, Deepfake*

# 1 INTRODUÇÃO

A Inteligência Artificial Generativa (IAG) representa uma mudança de paradigma na forma como os seres humanos produzem e consomem informação. Sustentada por avanços em Processamento de Linguagem Natural (PLN)[3] e Visão Computacional (VC)[4], essa tecnologia permite a criação automática de textos, imagens, áudios e vídeos com níveis de realismo e coerência sem precedentes. Ao integrar diferentes modalidades de dados, a IAG tem ampliado as fronteiras da criatividade e da comunicação digital, transformando a maneira como interagimos com máquinas e compartilhamos conhecimento.

No domínio do PLN, modelos de linguagem de grande porte (LLMs)[5] têm aprimorado significativamente a comunicação entre pessoas e sistemas computacionais, possibilitando traduções automáticas mais naturais, síntese de fala com entonação realista e geração de textos contextualizados. De modo complementar, na Visão Computacional, modelos generativos como as Redes Adversariais Generativas (GANs)[6] e as Redes de Difusão permitem criar e manipular imagens e vídeos sintéticos de alta qualidade, com aplicações que vão do entretenimento ao design e à simulação. Ao promover a produção de conteúdo original e multimodal, a IA generativa consolida-se como um dos principais motores da transformação tecnológica contemporânea, redefinindo os limites entre o real e o sintético na comunicação humana.

Contudo, essa capacidade generativa traz riscos proporcionais à segurança da comunicação digital. A fusão entre o Processamento de Linguagem Natural e a Visão Computacional possibilita a criação de *deepfakes* audiovisuais, que combinam clonagem de voz e manipulação facial para produzir representações humanas sintéticas altamente convincentes. Por meio da síntese de conteúdo novo e da transferência de atributos realistas, essas tecnologias podem gerar cenários fraudulentos capazes de enganar tanto pessoas quanto sistemas automatizados de verificação. Em contextos reais, experimentos demonstram que amostras de fala sintética são capazes de iludir assistentes virtuais e

---

[3] Área da Inteligência Artificial que estuda e desenvolve métodos para que computadores possam compreender, interpretar e gerar linguagem humana de forma útil e significativa.
[4] Campo da inteligência artificial que busca permitir que computadores interpretem e compreendam informações visuais do mundo real, como imagens e vídeos, de forma semelhante à percepção humana.
[5] *Large Language Model*: modelo de linguagem de larga escala baseado em redes neurais, treinado com grandes volumes de dados textuais para compreender e gerar texto em linguagem natural.
[6] *Generative Adversarial Networks*: modelo de IA composto por duas redes neurais, uma geradora e outra discriminadora, que competem entre si para produzir dados sintéticos realistas.

mecanismos de autenticação por voz, como o sistema da Alexa, com taxas de sucesso próximas de 100%. Adicionalmente, os próprios modelos de linguagem de grande porte apresentam vulnerabilidades relacionadas à memorização e ao vazamento de dados sensíveis utilizados em seus treinamentos. Pesquisas recentes mostram que informações privadas, como nomes, números de telefone e endereços de e-mail, podem ser extraídas de suas respostas, representando uma ameaça direta à privacidade e à integridade dos dados pessoais.

Atualmente, estamos presenciando uma enxurrada de conteúdos falsos circulando na internet, muitos deles produzidos com o auxílio de tecnologias avançadas de geração de vídeos e imagens realistas, como o *Google Veo* 3 e o *Sora*, da *OpenAI*. Essas ferramentas permitem criar cenas inteiramente sintéticas com aparência quase indistinguível da realidade, o que amplia as possibilidades criativas, mas também facilita a produção de vídeos manipulados, declarações forjadas e imagens adulteradas que se espalham rapidamente nas redes sociais. Diante desse cenário, torna-se fundamental ressaltar o perigo que essas tecnologias representam, pois seu uso indevido pode gerar desinformação, prejudicar reputações e comprometer a credibilidade das fontes digitais.

Nesse contexto, este trabalho propõe-se a discutir as implicações da inteligência artificial generativa, especialmente nas áreas de Processamento de Linguagem Natural e Visão Computacional, para a segurança da comunicação digital. Busca-se compreender de que forma o avanço dessas tecnologias potencializa a criação de conteúdos sintéticos cada vez mais convincentes e quais riscos emergem dessa capacidade, tanto no âmbito técnico quanto ético e social. Ao abordar essa problemática, pretende-se contribuir para o debate sobre a necessidade de políticas, regulamentações e estratégias de mitigação que possam equilibrar o uso responsável e seguro dessas ferramentas.

Metodologicamente, esta pesquisa adota uma abordagem qualitativa, exploratória e descritiva, baseada em revisão bibliográfica e análise crítica dos estudos selecionados. Para isso, serão utilizados artigos científicos, publicações técnicas e relatórios especializados nacionais e internacionais que tratam da IA generativa, com foco em Processamento de Linguagem Natural e Visão Computacional, a fim de construir uma visão abrangente e reflexiva sobre os impactos dessas tecnologias na segurança da comunicação.

# 2 REFERENCIAL TEÓRICO

## 2.1 Tradução Automática e PLN

Ferramentas de PLN baseadas em IA facilitam a comunicação multilíngue e a disseminação de informação técnica. Martínez et al. (2025) mostraram que sistemas convencionais de tradução tendem a entregar maior fidelidade estrutural, enquanto modelos generativos (ex.: ChatGPT) fornecem traduções mais naturais, porém de qualidade variável. Esse potencial de geração natural de linguagem amplia o alcance de alertas de segurança e documentos técnicos a públicos diversos, rompendo barreiras linguísticas. No entanto, a flexibilidade dos modelos generativos acarreta inconsistências que podem comprometer precisão necessária em mensagens críticas. Outros trabalhos ressaltam que grandes modelos de linguagem produzem texto fluente e coerente, mas devido à sua natureza probabilística, podem introduzir erros factuais ou de contexto, o que exige cuidado em aplicações de segurança.

## 2.2 Síntese de Voz e Assistentes Virtuais

IA generativa em áudio tem permitido clonar e sintetizar vozes com alta fidelidade. Embora isso ofereça benefícios, como acessibilidade para deficientes visuais e interações naturais com máquinas, abre caminho para ataques sofisticados. Alali e Theodorakopoulos (2024) demonstram que fragmentos de áudio gerados por IA (*speech deepfakes*) podem passar por sistemas de verificação por voz, sendo quase indistinguíveis da fala real. Seu estudo revela que quase 95–97% dos ataques de voz parcial enganaram sistemas automatizados, enquanto humanos identificaram muito raramente a falsificação. Esses achados evidenciam que sinônimos generativos audíveis podem comprometer mecanismos de segurança baseados em biometria de voz, exigindo novas defesas.

## 2.3 Geração de Mídia Sintética: *Deepfakes* Visuais e Desinformação

Na visão computacional, *GAN*s e outras redes neurais generativas produzem imagens e vídeos falsos cada vez mais convincentes. *Deepfakes* de vídeo, que combinam rostos e movimentos naturais, tornaram-se disseminados em desinformação política e

fraude. Kazim (2025) destaca que conteúdos sintéticos são usados para espalhar *fake news* e realizar roubo de identidade, exigindo métodos forenses de detecção. Estatísticas apontam que o número de *deepfakes* dobra semestralmente, evidenciando escalada dessas ameaças.

### 2.4 Deepfakes e Desinformação Multimodal

A conjunção de áudio e vídeo gerados por IA viabiliza campanhas de desinformação de grande impacto. Em conflitos recentes, agentes estatais exploraram *deepfakes* de vídeo e áudio sintetizados para influenciar narrativas públicas. DFRLab (2024) documenta casos onde deepfakes de líderes políticos foram veiculados com o intuito de desinformar e criar caos, aproveitando a facilidade de produção de conteúdo sintético. Tais ferramentas reduzem drasticamente o custo e o tempo para a criação de material enganoso de alta qualidade.

### 2.5 Fraudes Digitais e Biometria

Além da desinformação, fraudes financeiras e violação de biometria constituem riscos centrais. A clonagem de vozes de executivos de empresas, usada em golpes de *CEO* (Fraud), tem sido viabilizada por IA generativa. Alali et al. relatam casos reais, como ataque que imitava voz de diretor de banco, causando prejuízos monetários substanciais. Do lado das imagens, deepfakes faciais podem facilitar roubo de identidade digital ou contornar sistemas de reconhecimento facial. Esses abusos demonstram que canais de comunicação autenticados via biometria (voz, face) estão ameaçados por conteúdos gerados artificialmente.

### 2.6 Vazamento de Dados e Riscos de Privacidade em LLMs

Modelos generativos treinados em grandes volumes de dados apresentam riscos de privacidade e segurança. Carlini et al. (2021) evidenciam que é possível extrair do *GPT*-2 dados pessoais de seu corpus de treinamento, incluindo nomes, e-mails e

telefones. Em consonância, Feretzakis et al. (2024) revisam que *LLM*s podem memorizar e inadvertidamente revelar *PII*[7] quando geram texto.

## 2.7 Detecção baseada em Redes Neurais Profundas

Para enfrentar a sutileza crescente das falsificações visuais, investiga-se o uso de arquiteturas profundas capazes de modelar simultaneamente padrões espaciais e dinâmicas temporais presentes em vídeos e imagens. Singh et al. (2024) analisam avanços na detecção de *deepfakes* e destacam a eficácia de arquiteturas como redes neurais convolucionais *CNN* e modelos recorrentes *LSTM* na identificação de manipulações sutis em vídeos e imagens. Essas arquiteturas combinam informações espaciais e temporais para identificar inconsistências faciais, texturas artificiais e movimentos labiais desincronizados, permitindo diferenciar conteúdos autênticos de falsificações. Tais abordagens contribuem diretamente para a preservação da autenticidade digital e para o fortalecimento de mecanismos de verificação automatizada.

## 2.8 Abordagens Multimodais de Detecção

Dado que campanhas de desinformação frequentemente articulam texto e imagem de forma coordenada, surgem métodos que avaliam essas modalidades de forma integrada para detectar incoerências entre o conteúdo visual e o textual. Além das técnicas unimodais, métodos multimodais têm ganhado destaque. Shen et al. (2024) propõem uma estrutura de aprendizado contrastivo que analisa simultaneamente texto e imagem para identificar notícias falsas e desinformação em mídias sociais. O modelo combina representações semânticas obtidas de redes de linguagem como *BERT* e redes visuais como *CNN*s ou *Vision Transformers*, utilizando atenção cruzada para correlacionar elementos textuais e visuais. Essa integração melhora a precisão na detecção de *fake news* ao capturar discrepâncias entre legendas e imagens.

[7] *Personally Identifiable Information*: refere-se a qualquer dado que possa identificar direta ou indiretamente uma pessoa, como nome, endereço, telefone, CPF, e-mail ou dados biométricos.

## 2.9 Proteção de Privacidade e Segurança de Dados

No campo da segurança de dados, desenvolvem-se contramedidas voltadas à proteção da privacidade e à rastreabilidade de conteúdos gerados por IA. Feretzakis et al. (2024) descrevem o uso de privacidade diferencial[8] para inserir ruído estatístico nos dados e evitar reidentificação de indivíduos, aprendizado federado para distribuir o treinamento entre múltiplos dispositivos sem centralizar informações sensíveis, e criptografia homomórfica para permitir o processamento de dados cifrados sem descriptografá-los. Essas técnicas reduzem riscos de exposição de dados durante o treinamento e a inferência de modelos generativos.

## 2.10 Rastreabilidade e Responsabilização de Conteúdos gerados por IA

Para além da detecção e das salvaguardas de privacidade, a capacidade de atribuir origem e responsabilizar a produção de conteúdo sintético é elemento central na estratégia de mitigação de abusos. Kuditipudi et al. (2023) propõem *watermarking* digital robusto para modelos de linguagem, inserindo marcas invisíveis nas saídas textuais geradas por IA. Esse tipo de marcação permite rastrear a origem de conteúdos mesmo após edições e redistribuições, oferecendo um mecanismo prático para identificar autoria e responsabilizar o uso indevido de modelos generativos. A rastreabilidade complementa técnicas de detecção e proteção de dados, formando uma estratégia integrada para mitigar abusos associados à IA generativa.

## 2.11 Síntese do Referencial Teórico

Em síntese, as tecnologias centrais da IA generativa em comunicação incluem redes neurais profundas aplicadas à sintaxe e semântica (*transformers*[9], LSTM) e ao processamento de imagens (GANs, CNNs). Essas ferramentas são empregadas em assistentes de voz, chatbots, tradutores automáticos, editores de vídeo e outros. Todos esses aplicativos compartilham a capacidade de aprender padrões complexos de

[8] Técnica que adiciona ruído controlado aos dados ou resultados de um algoritmo para impedir a identificação de indivíduos, mesmo a partir de análises estatísticas.
[9] Arquitetura de rede neural introduzida por Vaswani et al. (2017), projetada para processar sequências de dados, como textos, de forma paralela e eficiente.

linguagem e visão a partir de dados massivos. Esse ponto de convergência entre PLN e visão computacional é duplo: amplia enormemente as possibilidades de interação homem-máquina (por exemplo, legendas automáticas em vídeos, tradução simultânea), mas também pode violar princípios de segurança ao produzir artefatos sintéticos que não distinguem realidade e ficção. Portanto, o referencial mostra que, embora as IAs generativas sejam valiosas para acelerar a difusão de informação e inclusão digital, elas exigem contínua atenção aos mecanismos de segurança e governança.

## 3 METODOLOGIA

Esta pesquisa adotou uma abordagem qualitativa, exploratória e descritiva, fundamentada em revisão bibliográfica. O objetivo foi compreender e analisar criticamente como a Inteligência Artificial Generativa tem sido utilizada em Processamento de Linguagem Natural (PLN) e Visão Computacional, especialmente no que diz respeito à segurança da comunicação, privacidade e contramedidas associadas.

O recorte temporal abrange o período de 2019 a 2025, escolhido por representar a fase em que as técnicas de IA generativa, impulsionadas pelas arquiteturas *transformer* e modelos de grande escala, se consolidaram na literatura científica e técnica. Esse intervalo permite observar o amadurecimento e a popularização das aplicações mais relevantes, embora se reconheça como limitação que trabalhos anteriores a 2019 possam conter discussões importantes que ficaram fora do escopo deste estudo.

As buscas foram realizadas em bases de dados amplamente reconhecidas na área de tecnologia e segurança da informação, a fim de garantir a qualidade e a representatividade das fontes. Foram utilizadas *IEEE Xplore* e *ACM Digital Library*, por reunirem pesquisas de alto impacto sobre ciência da computação e segurança, *SpringerLink*, que fornece revisões teóricas e capítulos de livros de referência, e *Google Scholar*, empregado para ampliar o alcance interdisciplinar e incluir também pré-prints e trabalhos em fase inicial de publicação. Essa combinação permitiu contemplar perspectivas técnicas, conceituais e contextuais sobre o tema.

As consultas utilizaram palavras-chave exclusivamente em inglês, tais como *"generative artificial intelligence"*, *"deepfake"*, *"deepfake detection"*, *"natural language processing"*, *"NLP security"*, *"computer vision"*, *"communication security"*, *"differential privacy"* e *"LLM vulnerabilities"*. Foram aplicadas combinações com

operadores booleanos (*AND*, *OR*) para equilibrar a abrangência e a precisão dos resultados obtidos.

O processo de triagem ocorreu em três etapas: (i) identificação inicial das publicações, (ii) leitura de títulos e resumos para eliminar duplicatas e estudos irrelevantes e (iii) leitura integral dos textos elegíveis, a fim de verificar a aderência aos critérios de inclusão e exclusão. Foram incluídos trabalhos publicados entre 2019 e 2025, redigidos em inglês, com acesso integral e que abordassem de forma direta o uso da IA generativa em PLN ou Visão Computacional, com foco em segurança, privacidade, ética ou contramedidas. Foram priorizados artigos revisados por pares, relatórios técnicos e pré-prints de fontes reconhecidas. Foram excluídos materiais fora do recorte temporal, em outros idiomas ou sem acesso completo, bem como literatura cinzenta sem revisão editorial.

A análise dos materiais selecionados seguiu um processo de codificação e síntese temática. Primeiramente, foi feita uma leitura exploratória para identificar termos e conceitos recorrentes, como *"machine translation"*, *"speech synthesis"*, *"deepfake"*, *"multimodal detection"*, *"differential privacy"*, *"watermarking"*, *"CNN/LSTM"* e *"facial authentication"*. Esses elementos foram transformados em códigos iniciais e, em seguida, agrupados em categorias mais amplas de acordo com sua natureza técnica ou temática:

- Tecnologias, reunindo arquiteturas como *transformers*, *GAN*s, *CNN*s e *LSTM*;
- Aplicações, incluindo tradução automática, síntese e verificação biométrica;
- Contramedidas e Proteção, com técnicas como detecção de *deepfakes*, *watermarking* e privacidade diferencial.

Com base nessas categorias, foram definidos cinco eixos analíticos que estruturaram os resultados:

1. Processamento de Linguagem Natural (PLN) – incluindo tradução, síntese e vulnerabilidades de *LLM*s;
2. Visão Computacional – abordando reconhecimento, geração e autenticação visual;
3. Desinformação Multimodal – integração entre texto e imagem na criação e detecção de *fake news*;
4. Privacidade e Ética – envolvendo vazamento de dados e regulamentações;
5. Detecção e Contramedidas – englobando métodos de detecção de *deepfakes* e técnicas de proteção.

Cada eixo foi sintetizado em matrizes comparativas, relacionando objetivos, métodos, resultados e limitações das fontes analisadas. Esse procedimento permitiu identificar padrões, divergências e lacunas nos estudos revisados, orientando a construção da análise crítica e das conclusões do trabalho.

A avaliação da qualidade das fontes considerou a reputação editorial, a clareza metodológica e a atualidade das publicações. Resultados provenientes de estudos com limitações explícitas foram tratados com cautela e comparados a evidências mais consistentes.

Por fim, as principais limitações deste estudo incluem a possibilidade de exclusão de pesquisas relevantes anteriores a 2019, o recorte linguístico restrito ao inglês e a heterogeneidade metodológica dos estudos analisados, que pode dificultar comparações diretas. Ainda assim, o método adotado proporcionou uma visão ampla, crítica e atualizada sobre o papel da IA generativa na segurança da informação, servindo como base sólida para as discussões e conclusões apresentadas.

## 4 RESULTADOS E DISCUSSÃO

A revisão da literatura evidencia que a Inteligência Artificial generativa redefine profundamente o ecossistema da comunicação digital, apresentando tanto oportunidades inéditas quanto desafios éticos e de segurança sem precedentes. As tecnologias baseadas em modelos generativos, como as arquiteturas *Transformer*, viabilizaram ganhos expressivos em tradução automática, síntese de voz e geração de imagens. Esses avanços promoveram uma democratização do acesso à informação e uma ampliação da acessibilidade digital, especialmente por meio da automação de tarefas linguísticas complexas e da criação de conteúdos multimodais de alta qualidade. No entanto, o mesmo mecanismo que potencializa esses benefícios é também o que sustenta os riscos mais críticos à segurança da informação.

A dualidade da IA generativa se manifesta de forma mais evidente no contraste entre suas aplicações criativas e suas vulnerabilidades estruturais. O *Transformer*, base de sistemas de tradução e assistentes virtuais (MARTÍNEZ et al., 2025), também é responsável pela capacidade de modelos de linguagem memorizarem dados sensíveis durante o treinamento (CARLINI et al., 2021). Essa sobreposição entre eficiência e risco revela um dilema fundamental: as mesmas propriedades que permitem gerar texto coerente e fluido são aquelas que podem expor informações privadas e comprometer a

confidencialidade dos usuários. O desafio, portanto, não está apenas na evolução da tecnologia, mas na definição de limites éticos e técnicos que garantam o uso seguro dessas ferramentas.

De modo semelhante, a integração entre modelos de linguagem e sistemas de visão computacional ampliou o potencial de criação de conteúdo multimodal, mas também deu origem a uma nova categoria de ameaça: os *deepfakes*. Esses conteúdos, gerados por redes adversárias generativas (*GAN*s), reproduzem expressões faciais, vozes e gestos com um grau de realismo capaz de enganar até mesmo sistemas automatizados de verificação. O impacto disso vai além da manipulação de imagens; afeta diretamente a autenticidade das mensagens e a confiança social nos meios digitais. Nesse cenário, a visão computacional discriminativa, quando usada para detectar falsificações, e não para produzi-las, torna-se uma contramedida essencial. Ou seja, ferramentas originalmente criadas para classificação e reconhecimento visual passam a ser reutilizadas como barreiras contra o uso indevido da IA generativa.

A desinformação multimodal agrava esse panorama ao combinar texto e imagem para criar narrativas falsas com alto poder de persuasão. Shen et al. (2024) destacam que a mesma integração entre modelos linguísticos e visuais que permite identificar inconsistências também pode ser explorada para gerar desinformação. Esse fenômeno demonstra que as fronteiras entre detecção e criação são tênues: os sistemas multimodais podem tanto combater quanto ampliar o problema. Essa simetria evidencia a urgência de desenvolver mecanismos de controle que diferenciem uso legítimo de manipulação intencional, especialmente em contextos políticos e institucionais.

Diante desse cenário de riscos, as contramedidas tecnológicas se consolidam como um campo de pesquisa em rápido amadurecimento. Diversos estudos apontam que a defesa contra abusos da IA generativa depende de uma combinação de abordagens técnicas, organizacionais e éticas. No âmbito técnico, algoritmos de detecção baseados em *CNN*s e *LSTM* têm demonstrado resultados promissores na identificação de *deepfakes*, analisando padrões sutis de inconsistência em quadros de vídeo ou variações temporais em expressões faciais (SINGH et al., 2024). Esses modelos, de natureza discriminativa, atuam como filtros automatizados que reconhecem sinais residuais de manipulação, mesmo quando o conteúdo aparenta autenticidade.

Complementarmente, métodos de detecção multimodal expandem essa capacidade ao analisar simultaneamente texto e imagem. Ao correlacionar legendas, descrições e contextos visuais, esses sistemas conseguem identificar incongruências

semânticas que indicam manipulação (SHEN et al., 2024). Essa integração tem se mostrado especialmente útil no combate à desinformação online, onde a coerência entre texto e mídia visual é frequentemente artificial.

Além das técnicas de detecção, destacam-se as estratégias de mitigação preventiva, que buscam proteger os dados antes e durante o treinamento de modelos generativos. A privacidade diferencial, por exemplo, introduz ruído estatístico nas informações originais para impedir que dados pessoais sejam recuperados ou reconhecidos nos resultados de um modelo (FERETZAKIS et al., 2024). O aprendizado federado distribui o processo de treinamento entre diferentes dispositivos, evitando o armazenamento centralizado de dados sensíveis e reduzindo o risco de vazamento. Já o *watermarking* digital, técnica descrita por Kuditipudi et al. (2023), insere marcas invisíveis nos conteúdos gerados por IA, permitindo rastrear sua origem mesmo após edições. Em conjunto, essas práticas não apenas reduzem o potencial de danos, mas também estabelecem um padrão de responsabilidade no desenvolvimento e uso de IA generativa.

Contudo, as contramedidas ainda enfrentam limitações técnicas e conceituais. Nenhum dos métodos existentes é totalmente confiável diante da velocidade com que os modelos generativos evoluem. A eficácia dos detectores baseados em *CNN* e *LSTM* tende a cair conforme novos modelos são treinados para imitar os sinais que essas redes procuram identificar. Da mesma forma, o *watermarking* pode ser removido por técnicas de pós-processamento ou compressão de mídia. Além disso, há desafios éticos: a introdução de ruído para preservar a privacidade pode comprometer a precisão dos modelos; o aprendizado federado, embora proteja dados locais, exige infraestrutura complexa e comunicação contínua entre servidores; e a marcação de conteúdo automatizado levanta questões sobre vigilância e controle informacional.

Essas tensões indicam que a segurança na era da IA generativa não pode ser reduzida apenas a soluções técnicas. É necessário um ecossistema de governança que envolva políticas regulatórias, padrões de transparência e educação digital. As contramedidas tecnológicas devem ser complementadas por práticas institucionais que garantam accountability, rastreabilidade e ética na criação de conteúdo artificial. Assim, o debate sobre segurança da informação transcende o domínio computacional e passa a integrar questões sociais e normativas.

Em síntese, os resultados da revisão apontam para uma realidade ambivalente: a IA generativa representa simultaneamente uma força de inovação e um vetor de

vulnerabilidade. A mesma capacidade de criar conteúdo autêntico e inclusivo pode ser explorada para desinformação, fraude e exposição de dados. As contramedidas emergem como resposta técnica necessária, mas ainda insuficiente. A consolidação de uma comunicação digital segura dependerá de uma integração equilibrada entre avanços tecnológicos, responsabilidade ética e regulação adequada, um desafio que continuará a definir o rumo da segurança da informação nos próximos anos.

## CONCLUSÃO

Conclui-se que a IA generativa, centrada em PLN e Visão Computacional, apresenta dupla natureza na segurança da comunicação. Por um lado, democratiza o acesso à informação (através de tradução instantânea, síntese de fala, ferramentas de acessibilidade); por outro, introduz novos vetores de ameaça, tais como *deepfakes* sofisticados e violações de privacidade via vazamento de dados treinados. A análise documental indicou que, sem controles apropriados, os benefícios comunicacionais podem ser ofuscados pelos perigos de desinformação, fraudes e perda de confidencialidade.

Os achados deste estudo realçam a necessidade de políticas e técnicas de mitigação: por exemplo, adotar práticas de segurança "*privacy by design*" em sistemas de IA, regulamentar a geração de mídia sintética e desenvolver algoritmos que detectem conteúdos gerados. Cabe aos gestores de tecnologia e legisladores equilibrar o fomento à inovação com salvaguardas éticas. A criação de *frameworks* legais (semelhantes ao *AI Act* europeu) e a conscientização pública sobre *deepfakes* são caminhos urgentes. Em síntese, a IA generativa pode aprimorar a comunicação, mas somente se for empregada de forma consciente e supervisionada. Futuras pesquisas devem investigar protocolos de validação de fontes e métodos criptográficos avançados para proteger a integridade da informação em canais digitais.

## REFERÊNCIAS


ALALI, A.; THEODORAKOPOULOS, G. Partial Fake Speech Attacks in the Real World Using Deepfake Audio. **Journal of Cybersecurity and Privacy**, v. 5, n. 1, 2024. Disponível em: https://doi.org/10.3390/jcp5010006. Acesso em: 22 maio 2025.

CARLINI, N.; TRAMÈR, F.; WALLACE, E.; JAGIELSKI, M.; HERBERT-VOSS, A.; LEE, K.; ROBERTS, A.; BROWN, T.; SONG, D.; ERLINGLESSON, Ú.; OPREA, A.; RAFFEL, C. Extracting Training Data from Large Language Models. **30th USENIX Security Symposium (USENIX Security 21)**, 2021. Proceedings..., p. 2633–2650. Disponível em: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. Acesso em: 22 maio 2025.

DFRLAB – Digital Forensic Research Lab. AI tools usage for disinformation in the war in Ukraine. DFRLab, 9 jul. 2024. Disponível em: https://dfrlab.org/2024/07/09/ai-tools-usage-for-disinformation-in-the-war-in-ukraine/. Acesso em: 23 maio.2025.

FERETZAKIS, G. et al. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. **Information**, v. 15, n. 11, p. 697, 2024.

KAZIM, M. Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning**. Information**, v. 16, n. 4, p. 270, 2025.

KUDITIPUDI, R.; THICKSTUN, J.; HASHIMOTO, T.; LIANG, P. Robust Distortion-free Watermarks for Language Models. **arXiv preprint**, 28 jul. 2023 (última revisão em 6 jun. 2024). Disponível em: https://arxiv.org/abs/2307.15593. Acesso em: 23 maio 2025.

MARTÍNEZ, J. R. et al. Generative Artificial Intelligence and Machine Translators in Spanish Translation of Early Vulnerability Cybersecurity Alerts. **Applied Sciences**, v. 15, n. 8, p. 4090, 2025.

SHEN, X. et al. Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. **Frontiers in Computer Science**, v. 6, p. 1473457, 2024.

SINGH, L. H. et al. Advancements in Detecting Deepfakes: AI Algorithms and Future Prospects – A Review. **Discover Internet of Things**, 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention Is All You Need. In: **Advances in Neural Information Processing Systems**, v. 30, 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 17 set. 2025.