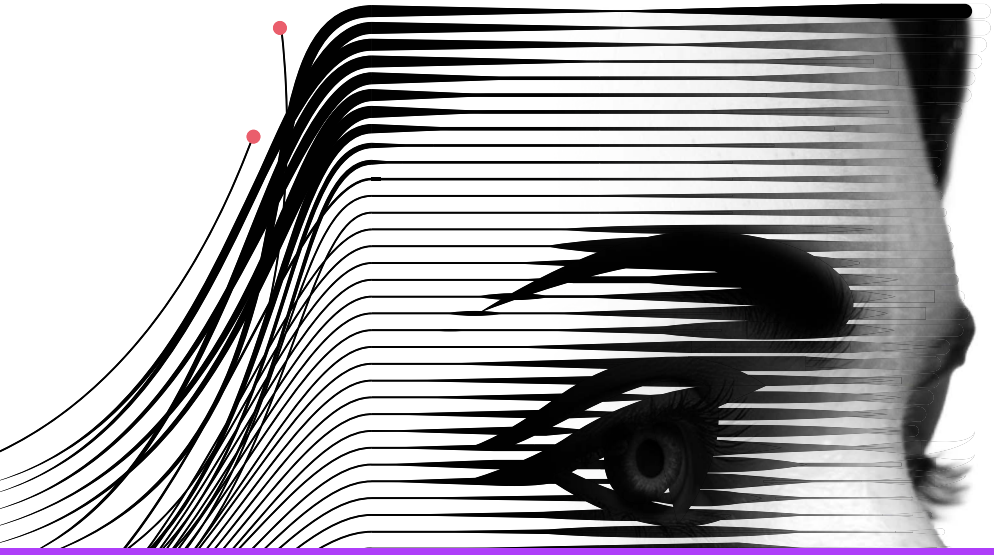
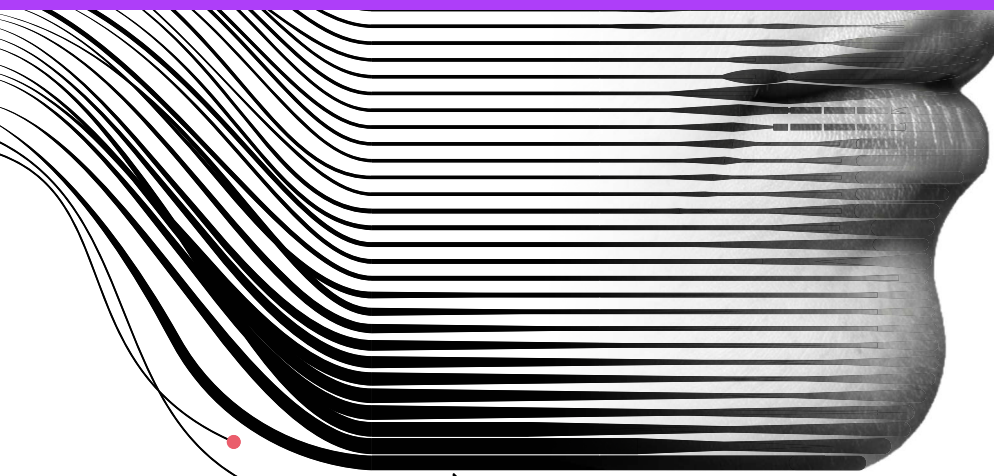


‹ radar tecnológico ›

deepfakes



ANPD

Agência Nacional de Proteção de Dados

‹ radar tecnológico ›

n° 6

deepfakes

Gabriela Campos Alkmin
Ingrid David Alves de Carvalho
Roseane Salvio
Roberta Fernandes Mendiondo Nunes
Mariana Luisa da Costa Lage
Ticianne de Gois Ribeiro Darin

ANPD
Brasília, DF
2026

ANPD
Agência Nacional de Proteção de Dados

Diretor-Presidente
Waldemar Gonçalves Ortunho Junior

Diretores
Iagê Zendron Miola
Lorena Giuberti Coutinho
Miriam Wimmer

Coordenação e administração do projeto
Superintendência de Inovação Tecnológica (SITEC)
Ticianne de Gois Ribeiro Darin
Angela Halen Claro Franco
Lucas Costa dos Anjos

Equipe de elaboração
Gabriela Campos Alkmin
Ingrid David Alves de Carvalho
Roseane Salvio
Roberta Fernandes Mendiondo Nunes
Mariana Luisa da Costa Lage
Ticianne de Gois Ribeiro Darin

Revisão e edição
José Pedro Morais de Araujo

Projeto gráfico / capa
André Scofano Maia Porto

Editoração eletrônica
André Scofano Maia Porto
Bruna Assi Hernandes

Produção de elementos gráficos
Ewerton Luiz Costadelle

Dados Internacionais de Catalogação na Publicação (CIP)

A2656d

Deepfakes/ Agência Nacional de Proteção de Dados. – Brasília, DF: ANPD, 2026.

163 p.il.: Color. (Radar tecnológico; n. 6)

ISBN: 978-65-82658-05-1

Inclui bibliografia.

1. Deepfakes. 2. Inteligência artificial. 3. Proteção de dados pessoais. I. Agência Nacional de Proteção de Dados - ANPD. II. Título.

CDU 004.62:004.8

Ficha Catalográfica elaborada pela bibliotecária
Angela Halen Claro Franco, CRB1- S023

Como referenciar esta publicação:

ANPD. **Deepfakes**. Brasília, DF: ANPD, 2026. (Radar Tecnológico, n. 6). Disponível em: incluir *link* do documento. Acesso em: incluir data de acesso ao documento.

ANPD
SCN, Qd. 6, Conj. A,
Ed. Venâncio 3000, Bl. A, 9º andar
Brasília, DF · Brasil · 70716-900
t. (61) 2025-8101
www.gov.br/anpd

‹ sobre a série ›

A série "Radar Tecnológico" é uma produção periódica da ANPD que objetiva realizar abordagens concisas de tecnologias emergentes que vão impactar ou já estejam impactando o cenário nacional e internacional da proteção de dados.

Sem a intenção de esgotar as temáticas ou firmar posicionamentos institucionais, o propósito da série é agregar informações relevantes ao debate da proteção de dados no País, com textos estruturados de forma didática e acessível ao público em geral.

Para cada tema, são abordados os conceitos principais, as potencialidades e as perspectivas de futuro sempre com ênfase na proteção de dados no contexto brasileiro. ▪

‹ lista de siglas ›

ANPD – Agência Nacional de Proteção de Dados
API – Application Programming Interface
BNCC – Base Nacional Comum Curricular
C2PA – Coalition for Content Provenance and Authenticity
CDC – Código de Defesa do Consumidor
CNN – Rede Neural Convolucional
CPF – Cadastro de Pessoas Físicas
CSIM – Similaridade de Cosseno de Incorporação Facial
DCGAN – Deep Convolutional Generative Adversarial Network
EBM – Modelo Baseado em Energia
EBEM – Estratégia Brasileira de Educação Midiática
ECA – Estatuto da Criança e do Adolescente
ECA Digital – Estatuto da Criança e do Adolescente Digital
ELBO – Limite Inferior de Evidência
ENISA – European Union Agency for Cybersecurity
EUA – Estados Unidos da América
FID – Fréchet Inception Distance
GAN – Rede Adversarial Generativa
GPS – Global Positioning System
GPU – Unidade de Processamento Gráfico
IA – Inteligência Artificial
IAG – Inteligência Artificial Generativa
IS – Inception Score
KL – Kullback Leibler (Divergência de Kullback Leibler)
LFW – Labeled Faces in the Wild
LGPD – Lei Geral de Proteção de Dados
LLM – Modelos de Linguagem de Grande Escala
LSE-C – Lip Sync Error - Confidence
LSE-D – Lip Sync Error Error -Distance
MEC – Ministério da Educação
MPF – Ministério Público Federal
NeRF – Campo de Radiância Neural
NLP – Natural Language Processing
ONU – Organização das Nações Unidas
PE – Estado de Pernambuco

PII – Informações Pessoalmente Identificáveis
PLN – Processamento de Linguagem Natural
PSNR – Peak Signal-to-Noise Ratio
RIPD – Relatório de Impacto à Proteção de Dados Pessoais
RNA – Rede Neural Artificial
RNN – Rede Neural Recorrente
Senacon – Secretaria Nacional do Consumidor
SSIM – Structural Similarity Index
STF – Supremo Tribunal Federal
TSE – Tribunal Superior Eleitoral
TTS – Texto para Fala (Text-to-Speech)
Unicef – Fundo das Nações Unidas para a Infância
VAE – Autocodificador Variacional
VAEBM – Autocodificador Variacional Baseado em Energia
VC – Conversão de Voz
ViT – Vision Transformers
VPN – Virtual Private Network
XAI – Inteligência Artificial Explicável

‹ sumário ›

‹ 1 introdução ›

Você já assistiu a um vídeo pensando que se tratava de uma situação real e te contaram, em seguida, que o conteúdo na verdade havia sido gerado por inteligência artificial? Ou já recebeu um áudio no WhatsApp com a voz de uma pessoa que você conhece apenas para descobrir, depois, que não partiu dela? Ou já teve acesso a um vídeo íntimo repetidamente compartilhado nas redes sociais envolvendo uma pessoa famosa que afirma que o conteúdo é falso? Ou, ainda, já se deparou com um áudio viral de um candidato à eleição dizendo algo que, segundo ele, foi manipulado digitalmente? Situações como essas têm se tornado cada vez mais comuns com os chamados *deepfakes*, mídias criadas ou manipuladas artificialmente que, devido à elevada aparência de autenticidade, parecem reais para o público geral.

A criação e a edição de conteúdos falsos, sejam analógicos ou digitais, não são novidade. O uso de montagens fotográficas e manipulação de imagens já no início do século XX, especialmente em contextos políticos e propagandísticos (King, 1997), é um exemplo disso. Porém, nos últimos anos, impulsionadas pelo avanço da inteligência artificial[1] e pelo maior acesso a ferramentas capazes de gerar conteúdo sintético, essas atividades ganharam uma nova abrangência. Devido às técnicas utilizadas para sua formulação, bem como por sua verossimilhança, os *deepfakes* são usados crescentemente para fraudes, golpes, roubos de identidade, violações de dados, disseminação em massa de informações falsas e produção de conteúdo pornográfico – que representava 98% de todo o volume de *deepfakes* em circulação até 2023 (Security Hero, 2023).[2] Em razão desses usos e do potencial de amplificar ataques de engenharia social, campanhas de desinformação e outras formas de manipulação digital, os *deepfakes* passaram a figurar entre as principais preocupações de organismos internacionais e relatórios globais de segurança cibernética, como o *Global Cybersecurity Outlook 2025*, do World Economic Forum, o *ENISA Threat Landscape*, de 2025, e

1 *A inteligência artificial é a área do conhecimento humano que estuda o desenvolvimento de sistemas capazes de melhorar seu próprio desempenho a partir dos dados que recebe e gerar saídas tais como predição, conteúdo, recomendações ou decisões que podem influenciar ambientes físico ou virtual (ANPD,2024).*

2 *Em 2023, a organização Security Hero analisou 95.820 vídeos de deepfake, 85 canais dedicados a essas mídias em diferentes plataformas e 100 websites vinculados ao seu ecossistema. O relatório aponta que 98% dos deepfakes em circulação eram de conteúdo pornográfico e que 99% das vítimas eram mulheres. Embora a Security Hero não seja uma instituição acadêmica ou especializada em pesquisa científica, seu relatório tem sido amplamente utilizado por pesquisadores e reguladores internacionais como referência para o mapeamento dos usos dessas mídias, motivo pelo qual é aqui mencionado.*

o relatório da Europol sobre os desafios dos *deepfakes* para a aplicação da lei, de 2024.

Inicialmente restritos a ambientes acadêmicos ou a aplicações experimentais, esses recursos passaram a circular amplamente nas redes sociais, muitas vezes na forma de vídeos ou áudios manipulados de figuras públicas. A popularização do termo e da prática nas redes sociais partiu de um usuário anônimo da plataforma Reddit, que utilizou o termo *deepfake* para explicar o uso de técnicas de aprendizado profundo (*deep learning*) em um código que compartilhou para criar conteúdo falso (*fake*). Esse código permitia inserir a imagem de celebridades em vídeos de conteúdo adulto, sem o devido consentimento. A partir daí, o interesse dos usuários da mesma plataforma em produzir esse tipo de material levou à produção em massa de conteúdos falsos.

Os primeiros alvos dos *deepfakes* incluíram atores (por exemplo, Emma Watson e Scarlett Johansson), cantores (por exemplo, Katy Perry) e políticos (por exemplo, os presidentes dos EUA Barack Obama e Donald Trump), cujos rostos foram sobrepostos, sem sua permissão, aos de outras pessoas (Kietzmann *et al.*; 2020, p.136). Esse fenômeno contribuiu, por um lado, para que o público em geral tomasse conhecimento dessa tecnologia e, por outro, para intensificar as preocupações com a desinformação, a manipulação da opinião pública e os danos à reputação de indivíduos.

Diante desse cenário, é imprescindível promover a reflexão crítica e a conscientização da sociedade sobre os *deepfakes*, especialmente no contexto da segurança digital, do dever de cuidado e da integridade da informação. Compreender os *deepfakes* e conhecer os riscos associados a eles ajuda a desenvolver o senso crítico no consumo de conteúdos digitais, colaborando para o letramento digital e para a promoção de reflexões éticas sobre o uso de tecnologias. Além disso, no panorama da proteção de dados pessoais, esse entendimento torna-se urgente, já que a manipulação

de rostos e vozes envolve diretamente o tratamento de dados pessoais sensíveis e exige uma perspectiva regulatória atenta à garantia dos direitos fundamentais. Dessa forma, disseminar esse conhecimento ajuda a equilibrar inovação e responsabilidade, protegendo a sociedade contra usos maliciosos – principalmente em relação a grupos vulneráveis, como mulheres e crianças.

Este Radar Tecnológico tem por objetivo, portanto, apresentar e fundamentar os principais conceitos relacionados aos *deepfakes*, bem como identificar potenciais riscos impostos por essas ferramentas à privacidade, à proteção de dados e à criação de ambientes digitais seguros. Para mapear os debates, tecnologias e necessidades regulatórias em torno desse tema, foram pesquisados artigos, documentos técnicos e relatórios publicados entre os anos de 2017 e 2026, período que abrange desde o primeiro uso do termo até o atual momento de consolidação dessas tecnologias. Além desta introdução, o estudo apresenta a caracterização técnica de *deepfakes* e conceitos relacionados, seguida da análise de casos e exemplos do contexto brasileiro para reflexão e, por fim, discute as perspectivas futuras de aplicação e fiscalização de *deepfakes*.

Além disso, este estudo possui caráter técnico-informativo e prospectivo, voltado à sistematização de conceitos, riscos e tendências relacionados a *deepfakes* sob a perspectiva da proteção de dados pessoais. O documento não constitui orientação normativa, decisão administrativa ou manifestação conclusiva da Agência Nacional de Proteção de Dados (ANPD) sobre casos concretos. A atuação da Agência no tema observa os limites de suas competências legais e incidirá, especialmente, quando houver tratamento de dados pessoais, nos termos da Lei Geral de Proteção de Dados (LGPD), sem prejuízo da cooperação institucional com outros órgãos e autoridades competentes.

‹ 2 conceitos principais ›

A manipulação de mídias visuais e auditivas é uma prática documentada na evolução dos meios de comunicação que vai desde fotomontagens analógicas no século XIX até técnicas digitais de aperfeiçoamento de imagem e voz em tempo real, como filtros de redes sociais e *software* que corrige a entonação da voz (*autotune*). O surgimento dos *deepfakes*, por sua vez, marca um ponto de virada tanto na escala quanto na acessibilidade e no realismo da criação de conteúdos adulterados. Essa tecnologia se difundiu a partir dos avanços recentes em inteligência artificial e aprendizado de máquina, que disponibilizaram procedimentos automatizados para gerar imagens, vídeos e áudios de alta fidelidade, dificultando a detecção de falsificações por observadores humanos (Kietzmann *et al.*, 2020, p. 135). Como exemplos desses avanços, podemos citar o próprio amadurecimento da inteligência artificial generativa[3] e o desenvolvimento de tecnologias que permitem a criação de conteúdos sintéticos[4] altamente realistas, como autocodificadores (*autoencoders*), Redes Adversariais Generativas, do inglês *Generative Adversarial Networks* (GANs) e modelos de difusão (*diffusion models*), sobre os quais trataremos adiante.

A ampliação do acesso a ferramentas de geração de imagens e vídeos sintéticos reduziu as barreiras técnicas desse processo, permitindo que pessoas sem formação especializada produzissem esse tipo de conteúdo rapidamente (Irfan *et al.*, 2025, p. 1928). Contudo, o emprego indiscriminado dessa tecnologia pode gerar impactos substanciais à sociedade, introduzindo riscos relacionados à falsificação de identidades, ameaças à segurança, ampla disseminação de conteúdos enganosos e invasão de privacidade (Pei *et al.*, 2026, p. 273). Por outro lado, as tecnologias que possibilitam a criação de *deepfakes* apresentam potencial para serem aplicadas em setores como o do entretenimento, da produção cinematográfica e da construção de imagens humanas digitais (Pei *et al.*, 2026, p. 273).

3 *A inteligência artificial (IA) generativa é uma subárea do aprendizado profundo. Os modelos generativos descrevem como um conjunto de dados é gerado em termos de um modelo probabilístico, sendo capazes de criar dados por meio de amostragem (ANPD, 2024, p. 7).*

4 *Conteúdo sintético é o conjunto de informações ou dados gerados artificialmente por sistemas computacionais, em contraste direto com os dados reais, que são oriundos da realidade física ou de interações humanas diretas (ANPD, 2024).*

Diante desse cenário, torna-se fundamental estruturar conceitualmente esse ecossistema. Esta seção tem por objetivo descrever e organizar os fundamentos dos diversos tipos de *deepfakes* e suas ramificações, a fim de diminuir ambiguidades teóricas e classificatórias. Para isso, iniciaremos abordando o conceito de mídias sintéticas, e partiremos, então, para a definição de *deepfake*, apresentando suas classificações quanto a aplicações e tipos de uso. Também discutiremos o que não é *deepfake*, diferenciando as adulterações algorítmicas de manipulações audiovisuais mais simples ou edições básicas de contexto. Essa distinção técnica nos permitirá mapear com clareza as diferentes modalidades de falsificação de alta complexidade, com especial atenção à clonagem de voz.

A delimitação dessas fronteiras estruturais servirá para diferenciar o uso não malicioso de mídias sintéticas das adulterações enganosas e ilícitas. Discutiremos, na sequência, os limites que, quando desrespeitados, tornam abusivo o uso de *deepfakes* e as maneiras pelas quais a compreensão dessas definições técnicas podem impactar sua regulação. Explicitaremos, assim, que a clareza conceitual pode contribuir com a forma como o direito e as instituições públicas devem estruturar normas de proteção, responsabilização e governança no ciberespaço.

2.1 Conceitos guarda-chuva: conteúdo e mídias sintéticas

Assim, "**conteúdo sintético**" é um termo técnico amplo e se refere a qualquer conteúdo gerado artificialmente ou manipulado, ou seja, que não provém diretamente de um evento ou de um dado real, mas de um processo de produção artificial. Esses conteúdos podem ser criados utilizando diversas ferramentas e técnicas, dependendo do propósito. Neste documento, focamos em conteúdos sintéticos gerados por meios computacionais, já que esse é o universo relevante

para o fenômeno de *deepfakes*. Especificamente no contexto de *deepfakes*, esses conteúdos são obtidos com técnicas de IA generativa, independentemente de serem dados estruturados, como tabelas e registros, ou conteúdos de mídia, como imagens, áudios, textos e vídeos (ANPD, 2024, p. 11). Já o termo "**mídia sintética**" é a aplicação desses conteúdos em formatos audiovisuais e comunicativos, como imagens, vídeos, áudios e outros materiais destinados à comunicação. *Deepfakes* são, portanto, uma subcategoria de mídias sintéticas (CNTI, 2025).

Em 2021, o Parlamento Europeu realizou um estudo e, a partir dele, publicou o relatório intitulado *Tackling deepfakes in European policy* (Enfrentando *deepfakes* nas políticas europeias, em tradução livre), cujo foco está em vídeos *deepfake*, clonagem de voz e síntese de texto, além de animações 3D (European Parliament, 2021, p. 5). De acordo com o relatório, a evolução das capacidades de criação e manipulação de mídias decorreu de três desenvolvimentos principais: a criação de algoritmos capazes de mapear automaticamente características faciais; a ampla disponibilidade de dados audiovisuais na internet; e o aumento correspondente das capacidades de perícia forense digital como resposta à maior necessidade de detecção de falsificações. Esse cenário permitiu o surgimento de modelos de inteligência artificial que utilizam um ciclo de aprendizado contínuo para criar simulações realistas e aprender a corrigir falhas com base em técnicas desenvolvidas para detecções forenses (European Parliament, 2021, p. 5).

Criar mídias falsificadas altamente realistas exigia, tradicionalmente, grandes volumes de dados e elevado tempo de processamento. No entanto, avanços nos modelos generativos e nas redes discriminativas – que atuam na validação dos dados, distinguido registros reais de conteúdos sintéticos – têm ampliado substancialmente a capacidade de geração de imagens e vídeos, permitindo produzir conteúdos cada vez mais convincentes e semelhantes a registros autênticos (Pei *et al.*,

2026, p. 3, 5). Apesar desse progresso tecnológico, análises de especialistas indicam que, na prática, os riscos mais imediatos decorrem frequentemente de conteúdos descontextualizados ou de formas mais simples de manipulação, apresentadas de modo a sustentar narrativas enganosas (CNTI, 2025). Nesse contexto, expomos, na próxima seção, as diferentes formas de se definir *deepfakes*, bem como seus principais tipos.

2.2 *Deepfakes*

2.2.1 Definição

O termo "*deepfake*", de forma simplificada, serve para descrever áudios, fotos e vídeos modificados ou artificialmente criados por meio da inteligência artificial (CNTI, 2025; Twomey *et al.*, 2024, p. 1). Porém, o que de fato caracteriza *deepfakes* e os distingue dentro da categoria mais ampla de mídias sintéticas é que o resultado se trata de uma simulação hiper-realista da identidade e da ação humana (European Union, 2024; Witness, 2022).

Existem diferentes tipos de *deepfakes*. A clonagem de voz, também conhecida como *deepfake* de áudio, utiliza inteligência artificial para criar imitações altamente realistas, combinando a reprodução de tom de voz com a síntese de texto para imitar o vocabulário e o estilo da pessoa, aumentando a credibilidade do clone (European Parliament, 2021, p. 6). Já a síntese de texto é aplicada na criação de *deepfakes* utilizando técnicas de Processamento de Linguagem Natural (PLN), do inglês *Natural Language Processing* (NLP), para reproduzir o estilo de comunicação de uma pessoa a partir da análise de grandes volumes de textos associados ao indivíduo (alvo), como transcrições, em que os sistemas identificam padrões linguísticos, nuances expressivas e intenções. Com base nesses dados, a inteligência artificial aprende a imitar a forma característica de se expressar de uma pessoa para gerar novos conteúdos (European Parliament, 2021, p. 6).

Deepfakes podem surgir tanto de processos de geração sintética via **comandos** *(prompts)*, em que não há correspondência direta com dados existentes, quanto de processos de **manipulação de dados reais** (Meding; Sorge, 2025, p. 5), como a substituição de faces, a alteração de expressões ou a modificação de contextos. Para evidenciar a extensão dessas adulterações em registros reais, a Organização Witness exemplificou diversas formas de manipulação de conteúdo, como: a inclusão ou remoção de objetos em vídeos; a alteração do cenário, como condições climáticas; a simulação de movimentos faciais ou corporais; a sincronização artificial de fala com movimentos labiais; a geração e a modulação de vozes para imitar indivíduos específicos; e a criação de imagens realistas de pessoas ou eventos inexistentes, inclusive a partir de descrições textuais (Witness, 2022).

Segundo a organização Witness (2022), a substituição de rostos em vídeos é a forma mais comum de *deepfakes*. Essa visão é complementada por Altuncu, Franqueira e Li (2024, p. 2), que estendem a classificação para além do suporte em vídeo, mencionando também imagens estáticas de rostos falsos, falsificações de fala. Os autores destacam ainda os vídeos híbridos, que combinam diferentes modalidades de falsificação, integrando imagens manipuladas e falas sintéticas (áudio) em um único arquivo.

Embora a palavra *deepfake* seja muito popular, a comunidade científica ainda debate qual é o limite exato entre o que é e o que não é um *deepfake* (Altuncu; Franqueira; Li, 2024, p. 2; CNTI, 2025). Em um levantamento feito com 826 artigos científicos publicados, pesquisadores da Universidade de Maryland identificaram três dimensões que ajudam a definir *deepfakes* de uma maneira mais abrangente, dependendo da fonte da identidade, do propósito e da extensão da alteração, como ilustra o diagrama da Figura 1 (Liu; Padmanabhan; Viswanathan, 2025, p. 4). Segundo a perspectiva desses autores, *deepfakes* podem ser criados de três formas: Transferência de Original para Alvo (A → B), em que uma

identidade substitui completamente outra (por exemplo, troca de rostos); Modificação do Original (A → A'), em que os atributos são alterados, preservando-se a identidade central (por exemplo, progressão etária, mudança de expressão); e Síntese Completa (ø → B), que envolve a criação de identidades completamente artificiais sem referentes no mundo real (por exemplo, atores sintéticos).

Além disso, Liu, Padmanabhan e Viswanathan (2025, p. 5) diferenciam *deepfakes* a partir de outras duas dimensões. A primeira delas é a "intenção", que distingue o propósito por trás da criação desse tipo de conteúdo. A intenção pode ter aplicações enganosas destinadas a induzir as pessoas ao erro quanto à autenticidade (por exemplo, desinformação, fraude) ou usos não enganosos criados para fins legítimos, com transparência quanto à natureza artificial (por exemplo, entretenimento, pesquisa, educação). A segunda dimensão é a "granularidade da manipulação", que caracteriza a extensão da alteração feita. Elas podem ser alterações holísticas, ou seja, envolvendo mudanças abrangentes que afetam toda a identidade ou cena (por exemplo, substituição completa da identidade); ou podem ser modificações mais direcionadas, envolvendo alteração seletiva de certos atributos ou de características específicas (por exemplo, alterar apenas a cor dos olhos ou a expressão facial).

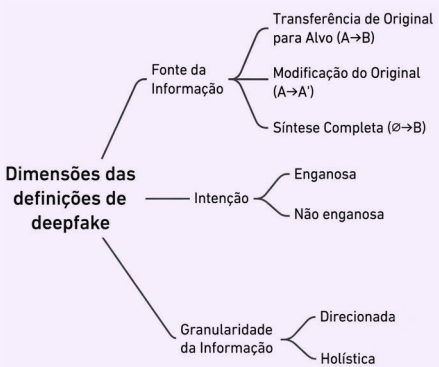


Figura 1 *Três principais dimensões das definições de deepfakes*

Fonte: Elaboração própria a partir da conceituação de Liu, Padmanabhan e Viswanathan (2025).

O cruzamento dessas três variáveis gera diferentes definições possíveis na literatura acadêmica. O Quadro 1 resume como os pesquisadores articulam esses critérios para classificar e caracterizar *deepfakes*.

Quadro 1 *Exemplos de diferentes definições de deepfake por dimensões*

Fonte	Intenção	Granularidade	Aplicações	Definição resumida
A → B	Enganosa	Holística	Desinformação e roubo de identidade (Nguyen *et al.* 2019)	Tecnologia orientada por IA projetada para enganar, na qual identidades completas são transferidas de uma fonte original para um alvo usando manipulação holística, com o objetivo de induzir indivíduos a acreditar em sua autenticidade.
A → B	Enganosa	Direcionada	Ataques de manipulação direcionada (Prezja *et al.* 2022)	Conteúdo gerado por IA no qual atributos-alvo específicos de alguma fonte são manipulados para criar mídia enganosa, com foco em alterações precisas de características específicas, como voz, tom facial, traços faciais ou expressões.
A → B	Não enganosa	Holística	Pesquisa e entretenimento (Eberl *et al.* 2022)	Tecnologia que transfere identidades completas de fontes originais para alvos com intenção não enganosa, usada para fins legítimos, incluindo aplicações acadêmicas, de pesquisa, entretenimento e educacionais.
A → A′	Enganosa	Direcionada	Falsificação em nível de atributo (Khochare *et al.* 2021)	Tecnologia que gera conteúdo sintético modificando atributos específicos dos dados originais, projetada para enganar ao mirar características específicas, como características biológicas faciais.

Fonte	Intenção	Granularidade	Aplicações	Definição resumida
A → A′	Não enganosa	Holística	Aprimoramento de pesquisa (Thambawita *et al.* 2021)	Tecnologia que modifica dados originais para gerar conteúdo sintético sem intenção enganosa, com foco na criação de alternativas realistas para áreas médicas, educacionais ou de pesquisa.
A → A′	Não enganosa	Direcionada	Aprimoramento de conteúdo (Herring *et al.* 2022)	Tecnologia que cria conteúdo sintético modificando atributos específicos dos dados originais para fins legítimos, como filtros de beleza e entretenimento.
Ø → B	Enganosa	Holística	Criação de identidade sintética (Xie *et al.* 2024)	Tecnologia que cria identidades inteiramente sintéticas sem depender de correspondente no mundo real, projetada para enganar espectadores, levando-os a acreditar que o conteúdo fabricado representa indivíduos ou eventos autênticos.
Ø → B	Não enganosa	Holística	Conteúdo sintético legítimo (Kaate *et al.* 2024a)	Tecnologia que cria conteúdo sintético inteiramente novo com intenção transparente e não enganosa, projetada para aplicações legítimas de entretenimento, pesquisa ou criação.

Fonte: Adaptação própria a partir de Liu, Padmanabhan e Viswanathan (2025, p. 7–8, tradução nossa).

Em termos técnicos, as diferentes definições de *deepfake* compartilham pontos fundamentais. O levantamento realizado por Liu, Padmanabhan e Viswanathan (2025, p. 39) mostra que todos os artigos científicos analisados caracterizam *deepfake* pela modificação ou criação de conteúdo sintético por meio de inteligência artificial. O objetivo comum está na

alteração de mídias para gerar representações alternativas de indivíduos, vozes ou eventos.

Porém, não há convergência entre as definições sobre a origem da identidade, os objetivos do criador e a extensão da modificação. Essa multiplicidade de formas de enxergar e definir *deepfakes* traz diversos desafios regulatórios, como discutiremos na **seção 6**.

A definição que melhor se alinha com a perspectiva adotada neste documento é a primeira do quadro acima (Nguyen *et al.* (2019)), **que foca em simular ou substituir a identidade de pessoas reais** e, potencialmente, gerar um engano coletivo devido à alta verossimilhança técnica. No entanto, a escolha desta abordagem não implica que a alteração de um atributo pessoal (A → A'), ou mesmo a síntese facial de uma pessoa fictícia (Ø → B), não sejam, tecnicamente, *deepfakes*.

Já que a urgência regulatória, jurídica e social está concentrada nas técnicas que reutilizam, extraem e manipulam dados biométricos de indivíduos reais com aplicação enganosa, este documento também adota, de forma complementar, as perspectivas de Khochare (2021), voltadas à **modificação maliciosa de atributos específicos**, e de Xie *et al.* (2024), que se refere à **síntese facial**. Afinal, a primeira é baseada em dados capturados de pessoas reais e possui alto potencial de simular pessoas autênticas, com aplicação enganosa; já a segunda, embora não corresponda a nenhuma pessoa específica, possui alto potencial de simular eventos e fatos como se fossem autênticos, contribuindo para o mesmo tipo de engano coletivo.

Assim, considerando o escopo de atuação da ANPD, no decorrer deste documento consideramos que *deepfakes* envolvem tanto a transferência holística de identidade (A → B) e a manipulação de atributos de um indivíduo (A → A'), quanto a síntese facial de um personagem (Ø → B), utilizadas com intenção de lesar o público. Casos que abordam cada uma dessas perspectivas serão discutidos nas seções 4, 5 e 6.

2.2.2 Classificação de uso

A literatura classifica *deepfakes* em diferentes categorias, com base no tipo de alteração ou geração que o programa de computador realiza na mídia digital (Irfan *et al.*, 2025, p. 1930). Isso significa organizar essas tecnologias a partir do objetivo prático da manipulação, ou seja, da compreensão de qual elemento visual o algoritmo foi programado para alterar ou fabricar. Dentro dessa divisão, as tecnologias de *deepfake* podem ser agrupadas conforme a atuação das mídias sintéticas, nas seguintes categorias de operação: troca de rostos (*face swapping*), reencenação facial (*face reenactment*), manipulação de atributos (face-attribute manipulation) e síntese facial (*face synthesis*).

Na categoria de troca de rostos, o sistema substitui as feições de uma pessoa pela face de outra, preservando o corpo e o cenário da gravação original (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 4). Na reencenação facial não há a troca do indivíduo, mas a falsificação de suas ações e atitudes. Na manipulação ou edição de atributos faciais, a ferramenta edita características físicas pontuais, como envelhecer a pele, alterar o gênero ou modificar o cabelo (Irfan *et al.*, 2025, p. 1931; Pei *et al.*, 2026, p. 5). Por fim, na geração ou síntese facial, a ferramenta realiza um processo de geração que dispensa o uso de qualquer imagem de base, podendo tanto fabricar uma fisionomia e identidade inéditas, de seres humanos fictícios, quanto criar representações fotorrealistas de pessoas que existem no mundo real (Irfan *et al.*, 2025, p. 1930).

Nesta seção, são descritos as características e o funcionamento de cada categoria, bem como exemplos de riscos associados a aplicações destinadas a enganar pessoas em diferentes contextos. Algumas aplicações de mídias sintéticas possuem finalidades lícitas e socialmente benéficas, abordadas na **seção 2.4**, enquanto exemplos mais detalhados de usos maliciosos dessas tecnologias serão discutidos ao longo das **seções 4** e **5**. As bases tecnológicas que viabilizam

os diferentes tipos de *deepfakes* apresentados a seguir serão explicadas em detalhes na **seção 3**.

Troca de rosto (Face swapping)

A troca de rostos é a categoria de manipulação de *deepfake* mais conhecida e fácil de ser visualizada na prática. O sistema extrai o rosto de uma pessoa, chamada de fonte, e o transfere para o corpo de outra pessoa, chamada de alvo, em uma fotografia ou vídeo. Nesse processo, o cenário, as roupas e a movimentação do corpo original permanecem intactos, ocorrendo apenas a substituição da identidade facial de quem aparece na tela (Pei *et al.*, 2026, p. 4). A troca de rosto pode ser ilustrada pela Figura a seguir:

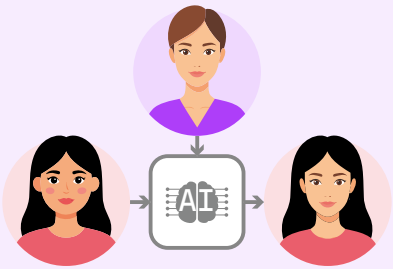


Figura 2 *Esquema ilustrativo da troca de rosto*

Fonte: Elaboração própria.

Nesse caso, o grau de realismo gerado pelo algoritmo é tão alto que a nova face se adapta à iluminação da cena e copia as atitudes originais, piscando e sorrindo em sincronia com o vídeo de destino. Para exemplificar, Kietzmann *et al.* (2020, p. 137) citam o famoso caso do vídeo em que o rosto do ator e comediante Jim Carrey foi sobreposto ao corpo da atriz Alison Brie durante uma entrevista no programa de televisão Late Night with Seth Meyers. Qualquer pessoa que assistisse ao vídeo modificado poderia ter a ilusão de que era Jim Carrey quem estava ali dando a entrevista, pois o algoritmo mapeou e adaptou os traços do ator aos gestos da atriz com alta

precisão. Embora possua aplicações legítimas na indústria do audiovisual e do entretenimento, a troca de rosto é uma tecnologia bastante utilizada para práticas criminosas e abusivas.

No mercado publicitário, a manipulação não autorizada facilita fraudes cibernéticas. O rosto e a voz de celebridades, jornalistas e apresentadores conhecidos têm sido clonados por algoritmos e inseridos em anúncios falsos nas redes sociais para vender produtos enganosos ou promover golpes (Fukuha; Archegas; Fürst, 2026, p. 187). Nessas situações, o ator ou a figura pública não apenas perde o controle de sua identidade digital, mas corre sérios riscos de ter a sua credibilidade e reputação associadas a crimes (Fukuha; Archegas; Fürst, 2026, p. 187).

Entre as diversas práticas graves e nocivas viabilizadas pelo uso da técnica de troca de rosto, destaca-se a criação de pornografia não consensual, em razão da gravidade dos danos que produz às vítimas e de sua crescente incidência. Nesse tipo de conduta, que é considerada crime no Brasil[5], os falsificadores extraem o rosto de atrizes famosas, ou mesmo de mulheres comuns, e o inserem no corpo de atrizes de filmes adultos (Kietzmann *et al.*, 2020, p. 142; Kirchengast, 2020, p. 2). O resultado é uma mídia enganosa que viola severamente a imagem e a intimidade da vítima e que o público comum pode acreditar ser real.

Reencenação facial (Face reenactment)

Na reencenação facial, de forma simplificada, o sistema inicia o processo observando o vídeo da pessoa fonte (quem faz a ação). Em seguida, ele copia os gestos que essa pessoa faz, como virar a cabeça, piscar os olhos e mexer a boca. Isso cria uma espécie de molde das expressões que serão usadas na edição (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). Depois, o programa analisa a imagem da pessoa-alvo (quem sofre a manipulação) e separa os traços físicos que a identificam das

5 *Desde 2018, o art. 218-C do Código Penal brasileiro prevê como crime as práticas de "Oferecer, trocar, disponibilizar, transmitir, vender ou expor à venda, distribuir, publicar ou divulgar, por qualquer meio - inclusive por meio de comunicação de massa ou sistema de informática ou telemática -, fotografia, vídeo ou outro registro audiovisual que contenha cena de estupro ou de estupro de vulnerável ou que faça apologia ou induza a sua prática, ou, sem o consentimento da vítima, cena de sexo, nudez ou pornografia" (Brasil, 1940).*

expressões que ela demonstra na foto original. Isso garante que o rosto da vítima (pessoa-alvo) não fique deformado e mantenha suas características únicas ao receber os novos movimentos.

Por fim, o sistema aplica os gestos copiados da pessoa fonte diretamente no rosto da pessoa-alvo. Assim, a imagem da vítima passa a reproduzir as ações do outro vídeo. Para que o resultado pareça verdadeiro, o sistema revisa e corrige a imagem várias vezes. Ele elimina os defeitos visuais até que os movimentos pareçam naturais e o vídeo gerado pareça autêntico. Um esquema que busca representar a reencenação facial é exposto na Figura a seguir:

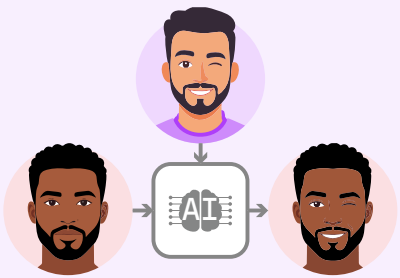


Figura 3 *Esquema ilustrativo da reencenação facial*

Fonte: Elaboração própria.

A reencenação facial possui aplicações lícitas e muito positivas, especialmente nas áreas da educação, da saúde, do entretenimento e do cinema (Fukuha; Archegas; Fürst, 2026, p. 177), mas há também diversos riscos associados ao uso dessa tecnologia. Por exemplo, na área da saúde, a capacidade de fabricar vídeos hiper-realistas é frequentemente usada para aplicar golpes e espalhar desinformação médica. Falsificadores clonam o rosto e as atitudes de médicos renomados, jornalistas ou celebridades e os colocam em anúncios falsos recomendando tratamentos perigosos ou vendendo produtos "milagrosos" para emagrecimento que não têm qualquer comprovação científica (Fukuha; Archegas; Fürst, 2026, p. 187).

No campo político, essa tecnologia pode ser utilizada para espalhar desinformação, aplicar golpes e manipular a opinião pública (Fukuha; Archegas; Fürst, 2026, p. 181). Como o sistema consegue criar discursos falsos com alto nível de realismo, ele pode ser usado para simular declarações ou condutas falsas atribuídas a agentes públicos, líderes políticos, jornalistas ou outras figuras de destaque, o que pode influenciar processos eleitorais, comprometer a reputação de indivíduos e enfraquecer a confiança nas instituições democráticas (Fukuha; Archegas; Fürst, 2026, p. 182; Kietzmann *et al.*, 2020, p. 141).

Em um vídeo criado em 2018 pelo comediante norte-americano Jordan Peele, a técnica de reencenação facial é aplicada para transferir os seus próprios movimentos labiais e a sua voz para o rosto do ex-presidente dos Estados Unidos, Barack Obama, fazendo parecer que Obama estava ofendendo gravemente um adversário político (Irfan *et al.*, 2025, p. 1928; Fukuha; Archegas; Fürst, 2026, p. 182). Embora esse vídeo tenha sido feito justamente para alertar o público sobre os riscos das notícias falsas, ele demonstrou a facilidade de enganar pessoas por meio de conteúdos sintéticos (Kietzmann *et al.*, 2020, p. 141).

Manipulação de atributos (Face-attribute manipulation)

Na manipulação de atributos, o sistema altera apenas detalhes específicos da imagem ou do vídeo, como a idade, a cor do cabelo ou as expressões, sem substituir o rosto da pessoa (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). Para realizar essa edição, o algoritmo mapeia a face e separa a identidade do indivíduo da característica exata que será modificada, o que garante que o restante da aparência continue intacta. Após aplicar a mudança nesse ponto isolado, o algoritmo revisa o material para eliminar pequenos defeitos visuais e corrigir falhas de continuidade, assegurando que o novo detalhe acompanhe os movimentos de forma fluida, e o

resultado pareça totalmente natural e autêntico (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). A manipulação de atributos é esquematizada na Figura 4.

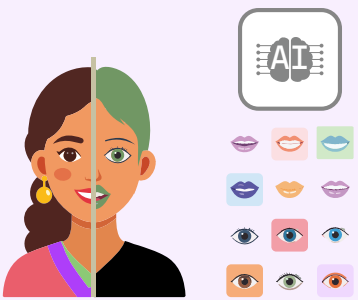


Figura 4 *Esquema ilustrativo da manipulação de atributos*

Fonte: Elaboração própria.

Síntese facial (Face-synthesis)

A síntese facial é o processo de criação de imagens ou vídeos hiper-realistas de rostos humanos que, na verdade, não existem no mundo físico. Em lugar de alterar a fotografia de uma pessoa real, a tecnologia atua construindo uma aparência inédita, gerando a figura de um indivíduo que parece autêntico, mas que é inteiramente fabricado pela tecnologia (Irfan *et al.*, 2025, p. 1930; Lazzarini, 2025, p. 12).

Para isso, o algoritmo começa analisando uma quantidade massiva de fotos de pessoas reais disponíveis na internet. Ao ser treinado a partir dessas imagens, ele aprende todas as características que formam a face humana, como a distância entre os olhos, as proporções do nariz e a forma como a luz reflete na pele. O algoritmo não faz uma montagem no sentido de "recortar e colar pedaços" de fotos existentes, em vez disso ele entende o padrão visual de como um rosto deve se parecer e o produz sem necessariamente reproduzir nenhuma das imagens com as quais foi treinado (Lazzarini, 2025, p. 10). A síntese facial é ilustrada esquematicamente na Figura 5.

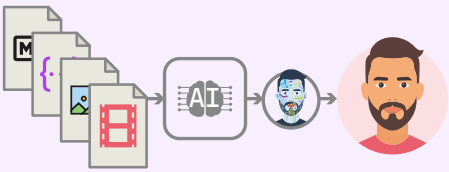


Figura 5 *Esquema ilustrativo da síntese facial*

Fonte: Elaboração própria.

Clonagem de voz (Voice cloning)

A manipulação acústica, popularmente conhecida como clonagem de voz, abrange sistemas de texto para fala (*text-to-speech* – TTS) e conversão de voz (*voice conversion* – VC), e representa a vertente sonora das mídias sintéticas e *deepfakes*. Diferentemente dos formatos visuais, que processam geometrias faciais, o foco dessa tecnologia está em isolar e transferir as características biométricas da fala de um indivíduo-alvo (Dehghani; Saberi, 2025, p. 24–26).

Segundo Dehghani e Saberi (2025, p. 24–25), a literatura divide esses sistemas entre métodos de treinamento paralelos – que exigem enunciados idênticos gravados pelas duas partes – e não paralelos, que utilizam coleções diversas e desalinhadas de áudio, conferindo maior flexibilidade prática à clonagem.

O processo começa com a coleta de gravações da voz da pessoa que se deseja reproduzir. Essas amostras servem como referência para que o sistema aprenda as características da fala do indivíduo. A quantidade de áudio disponível influencia diretamente a qualidade do resultado, já que amostras muito curtas tendem a dificultar a reprodução fiel da voz, enquanto o acréscimo de grandes volumes de gravações gera ganhos cada vez menores após determinado ponto (Dehghani; Saberi, 2025, p. 26).

Em seguida, o algoritmo examina as gravações para identificar os elementos que tornam aquela voz única. Entre esses aspectos estão características como a forma de falar, o ritmo da fala, a entonação e outras particularidades que ajudam a diferenciar uma pessoa das demais. Essa etapa é importante para que a voz sintetizada mantenha o máximo possível de semelhança com a voz original (Dehghani; Saberi, 2025, p. 26).

Após aprender os padrões característicos da voz, o sistema passa a gerar novos áudios que reproduzem essas características. Isso pode ocorrer de duas formas principais. Na primeira, o usuário fornece um texto escrito, que é transformado em fala utilizando a voz da pessoa-alvo. Na segunda, o sistema recebe uma gravação feita por outra pessoa e altera suas características vocais para que o resultado soe como se tivesse sido pronunciado pelo indivíduo cuja voz foi aprendida, mantendo a mensagem originalmente falada (Dehghani; Saberi, 2025, p. 24).

2.2.3 *Cheapfakes* e *shallowfakes*

Enganar as pessoas na internet não exige apenas tecnologias avançadas, inteligência artificial ou computadores superpotentes. Embora se comente muito sobre os *deepfakes*, o perigo mais frequente do dia a dia pode estar em fraudes muito mais simples (Elliott, 2023). Nesse contexto é que surgem, na literatura, os conceitos de *cheapfakes* (falsificações baratas) e *shallowfakes* (falsificações rasas), os quais são bastante diferentes dos *deepfakes*.

Essa diferença começa na origem dos termos e se estende à tecnologia empregada em cada tipo de manipulação. O termo "*deep*" refere-se ao uso de técnicas de aprendizado profundo (*deep learning*), enquanto "*shallow*" (raso) e "*cheap*" (barato) indicam que a edição foi feita de forma superficial e de baixo custo. Assim, enquanto os *deepfakes* são constru-

ídos por sistemas de inteligência artificial capazes de gerar falsificações altamente convincentes (Altuncu, Franqueira e Li, 2024), os *cheapfakes* e *shallowfakes* são produzidos por humanos, por meio de *softwares* de edição convencionais e gratuitos, usados para cortar, acelerar ou alterar o contexto de um vídeo (Brady, 2020, p. 2; Stoll, 2020), sem o uso de técnicas de aprendizado profundo (Çetin, 2022; Weikmann; Egelhofer; Lecheler, 2025, p. 2; Weikmann; Lecheler, 2022, p. 3).

Para facilitar a compreensão dessas diferenças, é útil observar o ecossistema das manipulações digitais como uma linha contínua: de um lado, técnicas altamente complexas criadas por sistemas de inteligência artificial; do outro, edições simples que podem ser feitas por qualquer usuário no celular. Como ilustrado por Brady (2020), essa divisão fica clara quando analisamos três pontos principais: o nível de sofisticação, a tecnologia usada e a facilidade de detecção (Figura 6)

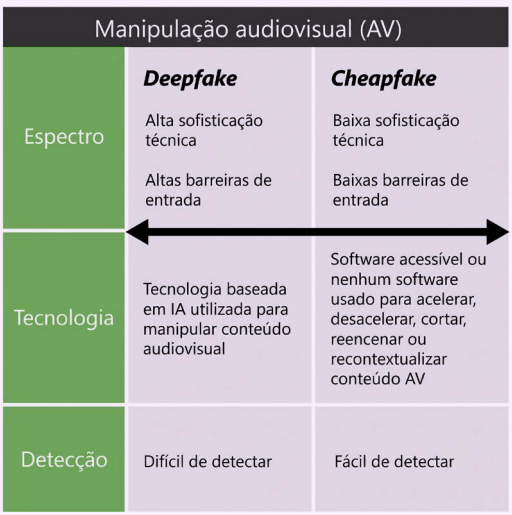


Figura 6 *Manipulação audiovisual analisada a partir do nível de sofisticação, tecnologia utilizada e facilidade de detecção*

Fonte: Elaboração própria a partir de Brady (2020, p. 3, tradução nossa).

Apesar disso, especialistas conseguem detectar esses materiais com facilidade; ainda assim, eles enganam milhares de pessoas devido à rápida velocidade de disseminação.

Parte da literatura trata *cheapfakes* e *shallowfakes* como sinônimos, dada a baixa sofisticação técnica de ambos. Contudo, Twomey *et al.* (2024, p. 7) propõem distinções mais claras ao considerar que a divisão entre *deepfake* e *shallowfake* baseia-se no nível tecnológico empregado: as ferramentas baseadas em inteligência artificial pertenceriam apenas ao primeiro grupo, enquanto mídias do segundo grupo derivariam de republicações em novos contextos ou de alterações estruturais básicas no material original, como cortes e sobreposição de faixas sonoras.

Os mesmos autores propõem um mapeamento gradual para os *cheapfakes*, sem divisões fechadas. Dessa forma, o espectro *cheapfake-deepfake* organiza as técnicas em uma linha contínua definida por custo, barreiras de acesso e sofisticação. Os *deepfakes* ocupam o topo da escala, com modificações digitais completas, enquanto os *cheapfakes* e *shallowfakes* ocupam a base, englobando desde a recontextualização de arquivos autênticos até edições manuais, como técnicas de rotoscopia para substituição facial e alteração na velocidade de reprodução. Assim, o mapeamento proposto indica que o **perigo informacional imediato está na simplicidade e na recontextualização**, não apenas na sofisticação algorítmica.

Essa leitura é reforçada por Weikmann, Egelhofer e Lecheler (2025, p. 1020), para quem a desinformação visual simples dos *cheapfakes* gera impacto psicológico e comportamental persistente, muitas vezes superando *deepfakes* em efeitos de desinformação — evidência de que o risco cognitivo desses conteúdos não depende da sofisticação técnica empregada. Essa perspectiva é complementada por Brady (2020, p. 3), que argumenta que o impacto social dessas adulterações de baixo custo é potencializado por canais de comunicação diretos entre os usuários. Isso porque a disseminação de informações de uma pessoa para outra por aplicativos de mensagens privadas oculta o rastreamento da desinformação. Esse processo pode reforçar vieses cognitivos que validam o conteúdo manipulado antes mesmo de qualquer checagem técnica, especialmente em contextos de polarização política.

A esse respeito, Çetin (2022) ressalta que ferramentas cotidianas e filtros de redes sociais exemplificam o caráter acessível dessas adulterações rudimentares. Ainda assim, subestimar a simplicidade operacional dessas mídias representa um risco crítico no ecossistema de segurança cibernética. Sob a ótica prática da autenticação corporativa, Çetin (2022) alerta que se os mecanismos de segurança de uma organização forem frágeis, até mesmo essas técnicas simplificadas são capazes de burlar com facilidade os testes de vivacidade (*liveness tests*) em procedimentos de verificação de identidade remota.

Entender claramente os nomes e as diferenças que dividem e interligam os *deepfakes*, *cheapfakes* e *shallowfakes* é indispensável para criar defesas eficazes no cenário tecnológico contemporâneo. A ameaça à integridade informacional e à segurança digital não provém exclusivamente de algoritmos sofisticados de IA.

2.3 O que não é *deepfake*

Para delimitar com precisão o escopo do debate sobre mídias sintéticas, é fundamental estabelecer uma linha clara entre o que constitui um *deepfake* e o que são apenas ferramentas comuns de edição. Em primeiro lugar, o *deepfake* não se confunde com edições assistivas, rotineiras ou ajustes triviais de imagem e som. Intervenções cotidianas como correção de cor, cortes de áudio, remoção de ruídos de fundo ou a aplicação de filtros comuns de iluminação não entram nessa categoria. A razão para essa exclusão é que **essas alterações não modificam a semântica, o contexto ou o sentido original da mensagem**.

No âmbito europeu, essa diferenciação encontra amparo legal direto no artigo 50 do *AI Act* (European Union, 2024). O texto regulatório europeu estabelece obrigações rígidas de transparência e rotulagem apenas para sistemas de

inteligência artificial que geram ou manipulam conteúdos capazes de enganar o público. Contudo, a legislação protege explicitamente a liberdade criativa e a rotina digital, ao prever que essas regras de rotulagem não se aplicam a ferramentas de IA utilizadas puramente para edição rotineira ou que não alterem o significado real do conteúdo (European Union, 2024), como correções técnicas ou filtros assistivos.

Em segundo lugar, os limites para definir o que é ou não um *deepfake* ainda são confusos, o que gera ambiguidade conceitual e instabilidade terminológica em torno do termo. Estudos interdisciplinares apontam que essa indefinição está associada à multiplicidade de enquadramentos provenientes de áreas como computação, direito e comunicação, o que contribui para a existência de interpretações heterogêneas do fenômeno (Twomey *et al.*, 2024).

Em resposta a essa limitação conceitual, parte da literatura propõe distinguir os *deepfakes* como um subconjunto da categoria mais ampla de mídias sintéticas, a qual engloba diferentes formas de conteúdos gerados ou manipulados por inteligência artificial, inclusive aqueles que não se baseiam em indivíduos reais (CNTI, 2025; Twomey *et al.*, 2024; Roe *et al.* 2025). Paralelamente, investigações empíricas demonstram que o uso do termo *deepfake* tende a evocar percepções associadas a risco e desinformação, ao passo que expressões alternativas, como "mídias sintéticas", podem favorecer leituras menos negativas e mais alinhadas a aplicações legítimas da tecnologia (Rauchfleisch; Vogler; De Seta, 2025).

Nesse contexto, Ding (2020) analisou o relatório do *Tencent Research Institute*, que propõe o uso de terminologias mais abrangentes e semanticamente neutras, a exemplo de síntese profunda (*deep synthesis*), em substituição ou como alternativa ao termo *deepfake*, considerado reducionista e potencialmente estigmatizante. Contudo, tal proposta não constitui padrão consolidado na literatura científica internacional, na qual permanece predominante o uso dos termos *deepfake*

e mídia sintética, ainda que reconhecidas suas limitações conceituais.

Por fim, o que diferencia uma mídia sintética permitida de uma criada com objetivos nocivos não é apenas a técnica de aprendizado profundo utilizada, mas sim a combinação entre a sua intenção e a granularidade da manipulação, como visto na **seção 2.2**. Conforme aponta o CNTI (2025), as legislações modernas buscam conciliar os usos não enganosos, puramente artísticos ou utilitários com a proibição de práticas criminosas.

O foco das políticas eficazes está em isolar e punir os usos prejudiciais destinados a lesar o público. Isso abrange desde o engano coletivo por alta verossimilhança na transferência holística de identidade, até a modificação maliciosa de atributos específicos e a síntese facial de personagens para fins de fraude ou pornografia não consensual. Assim, se o conteúdo gerado por IA é transparente e não possui o objetivo ou o efeito de induzir o público ao erro, ele se enquadra na categoria de expressão criativa ou utilitária, afastando-se do escopo restritivo e punitivo aplicado aos *deepfakes* com aplicação enganosa. Essas especificidades regulatórias serão retomadas na **seção 6.1**.

2.4 Uso não malicioso de mídias sintéticas

No contexto de *deepfake*, o uso não malicioso refere-se à criação e à aplicação de mídias sintéticas hiper-realistas (sejam vídeos, áudios ou imagens gerados por inteligência artificial) sem a intenção de enganar, fraudar, difamar ou causar danos a terceiros. Assim, essa tecnologia traz potencial de usos legítimos e inovação em diversos setores, incluindo intervenções artísticas, produção de filmes, geração de conteúdo de entretenimento e construção de modelos humanos digitais (Pei *et al.*, 2026, p. 28), como descrito a seguir. É importante destacar, porém, que esses usos nem sempre são isentos de riscos ou de preocupações éticas.

Entre os usos legítimos, destaca-se, no comércio, o emprego dessas ferramentas na criação de manequins virtuais para os clientes testarem roupas ou cosméticos. Já na indústria audiovisual, as aplicações são diversas: a tecnologia possibilita fazer a dublagem visual de modo automático, ajustando os movimentos da boca do ator para que sincronizem perfeitamente com o áudio do dublador em um idioma estrangeiro, bem como permite que os produtores corrijam palavras ditas de forma errada em uma gravação sem precisarem chamar os atores de volta ao estúdio para refazerem a cena, dispensando a necessidade de regravar cenas originais e permitindo às produtoras economizar recursos e ampliar o trabalho criativo (Kietzmann *et al.*, 2020, p. 141). A troca de rosto, por exemplo, permite que os produtores coloquem o rosto de um ator principal sobre o corpo de um dublê durante a gravação de cenas de ação muito perigosas, gerando um resultado visual altamente realista e imperceptível para os espectadores (Kietzmann *et al.*, 2020, p. 142).

No entanto, o uso dos *deepfakes* na indústria criativa também pode causar debates, pois mesmo nos usos permitidos existem dúvidas éticas sobre a autenticidade, a originalidade e a autoria das obras (Santos Junior, 2025). Mudar um roteiro ou copiar o estilo de um artista, por exemplo, esbarra nas regras de direitos autorais (Kietzmann *et al.*, 2020, p. 145). Além disso, alterar a atuação original levanta questões sobre a falta de consentimento dos atores no uso da sua imagem (Kietzmann *et al.*, 2020, p. 142). Exemplo disso é o caso envolvendo a atriz Scarlett Johansson, que teve atributos de sua voz replicados por um sistema de inteligência artificial da empresa OpenAI sem o seu consentimento (Lazzarini, 2025, p. 7). Por causa desses conflitos legais e artísticos, a economia de recursos feita pelas produtoras recebe críticas.

Há, ainda, a delicada e polêmica questão da "ressurreição digital", que ocorre quando a inteligência artificial é usada para recriar o rosto e a voz de artistas que já faleceram (Lazzarini, 2025, p. 23). Isso já aconteceu em grandes produ-

ções de Hollywood, como a recriação do ator Peter Cushing na franquia *Star Wars* (Fukuha; Archegas; Fürst, 2026, p. 175; Kirchengast, 2020, p. 2), e na publicidade brasileira, com a recriação da cantora Elis Regina em um comercial de televisão (Lazzarini, 2025, p. 23). A fabricação desses "*deepfakes* póstumos" levanta discussões sobre os limites da tecnologia e o respeito à memória de pessoas falecidas. O debate jurídico atual foca em definir quem tem o poder de autorizar essa recriação comercial. Discute-se, por exemplo, se os herdeiros e familiares possuem esse direito, ou se uma legislação deveria exigir que o próprio artista tenha deixado uma permissão prévia e expressa, ainda em vida, para que o computador possa fabricar novas imagens suas no futuro (Lazzarini, 2025, p. 44).

Essa mesma capacidade de reanimar imagens do passado, quando pautada pela ética e pela transparência, também começa a ser explorada em propostas artísticas e culturais de preservação histórica. O artista Guilherme Bretas, por exemplo, utilizou essa tecnologia para unir inteligência artificial e arquivos de memória, dando movimento a retratos de pessoas negras e originárias do século XIX. Ao animar rostos guardados em museus e projetar essas imagens em espaços públicos, o artista buscou romper o silêncio e o controle herdados da época colonial (Santos Junior, 2025).

Em outro exemplo, o Museu da Memória Negra em IA (@museudamemorianegraemia) utiliza a tecnologia artificial para recriar imagens históricas e produzir imagens de personagens que, no tempo em que viveram, não foram retratados em fotos ou pinturas. Nesses projetos, o uso do algoritmo abriu uma nova possibilidade narrativa, uma ressignificação, permitindo vislumbrar vidas e histórias que os registros oficiais apagaram ou nunca incluíram, bem como contribuiu como ferramenta didática para a não reprodução de estereótipos racistas.

Outra aplicação relacionada à preservação da memória é a restauração de arquivos visuais, na qual a inteligência artificial

consegue recompor partes que faltam de um rosto, melhorar fotos com baixa qualidade e até colorir fotografias em preto e branco, a fim de preservar filmes antigos (Pei *et al.*, 2026, p. 273; Santos Junior, 2025).

No campo educacional, essa lógica de dar vida a arquivos e memórias ganha aplicações pedagógicas concretas. A reencenação facial pode funcionar como uma ferramenta para tornar o ensino mais acessível e inclusivo, pois possibilita a criação de vídeos educativos sob medida para alunos com dislexia, deficiência auditiva ou outras dificuldades de aprendizado, ajustando o estilo de comunicação e sincronizando perfeitamente os movimentos da boca com a leitura de legendas, ou com a interpretação em língua de sinais (Fukuha; Archegas; Fürst, 2026, p. 176–177). A tecnologia permite, ainda, "dar vida" a personagens do passado: o algoritmo consegue criar vídeos interativos em que figuras históricas como cientistas, escritores e líderes políticos contam suas próprias histórias e ensinam conceitos aos alunos de forma moderna e acessível. Um exemplo marcante desse uso ocorreu no Museu do Holocausto de Illinois, nos Estados Unidos, que utilizou inteligência artificial para criar hologramas de sobreviventes reais da tragédia, permitindo que os visitantes do museu interagissem e "conversassem" diretamente com eles (Fukuha; Archegas; Fürst, 2026, p. 176).

Essas iniciativas ilustram como a criação de mídias sintéticas pode servir de instrumento técnico para trazer reflexões e reparar lacunas da história, indicando um caminho que pode ser adotado em contextos semelhantes, desde que o uso da tecnologia seja pautado pela ética e pela transparência na informação (Assis Junior; Hessel, 2021, p. 81). O perigo, no entanto, está na distorção dos fatos: assim como a tecnologia pode "reviver" uma figura histórica para ensinar algo correto, ela pode ser usada para falsificar a história, criar discursos que nunca existiram e manipular a opinião dos estudantes e da sociedade. Por esse motivo, a literatura alerta que, para essas inovações serem benéficas na saúde e no ensino, elas

precisam ser aplicadas com total transparência, avisando o público de que se trata de uma inteligência artificial, com o consentimento das pessoas retratadas e sob supervisão rigorosa, evitando que a linha entre a educação personalizada e a manipulação maliciosa seja ultrapassada (Fukuha; Archegas; Fürst, 2026, p. 177–178).

Como recurso de promoção da acessibilidade, tecnologias de clonagem tem potencial de restaurar e criar vozes para indivíduos que perderam sua capacidade de falar (Dehghani; Saberi, 2025, p. 2). Na prática clínica, podem ainda se desdobrar em dispositivos assistivos para pacientes acometidos por condições severas, como Esclerose Lateral Amiotrófica ou intervenções laríngeas, que buscam clonar suas vozes antes do agravo clínico, para preservar sua identidade.

Paralelamente, na indústria cinematográfica e de entretenimento, as aplicações validadas pelos autores incluem o desenvolvimento de avatares digitais, a atualização de trechos de filmes sem necessidade de refilmagem e a dublagem realista de produções estrangeiras (Dehghani; Saberi, 2025, p. 2) – mercado que se expande para a tradução automatizada capaz de preservar o timbre original do ator e de restaurar figuras históricas. Ademais, a partir da síntese de voz, o setor de serviços estende essas fronteiras para a customização de assistentes virtuais de atendimento, ferramentas de navegação por satélite (GPS), audiolivros e assistentes residenciais a partir de identidades vocais corporativas ou personalizadas.

Já o setor de publicidade e cultura pode usar a geração de mídias para compor músicas e produzir peças publicitárias com novos recursos sonoros e visuais. Embora não seja necessariamente enganoso, o uso do *deepfake* nessas áreas reacende, entre profissionais, as preocupações já mencionadas sobre a autoria, a originalidade e a autenticidade das obras, sobretudo quando se copia o estilo ou se usa a imagem de figuras públicas sem o seu consentimento, prática que pode violar as regras de direitos autorais (Santos Junior,

2025; Kietzmann *et al.*, 2020, p. 142, 145). Além dos conflitos artísticos e de imagem, existe o perigo de a tecnologia reproduzir preconceitos e desigualdades de gênero, de classe e de raça, na medida em que os algoritmos aprendem e copiam os vieses que já existem nos grandes bancos de dados (Santos Junior, 2025).

As mídias sintéticas podem, ainda, ter aplicações que trazem benefícios à área da saúde. Na comunicação em saúde pública, a reencenação facial pode servir para quebrar as barreiras de idioma. Um exemplo mundialmente conhecido foi a campanha de combate à malária chamada *Malaria Must Die*, na qual um sistema de reencenação facial alterou os movimentos labiais e faciais do ex-jogador David Beckham para que ele parecesse falar, fluentemente, em nove idiomas diferentes, ajudando a espalhar uma mensagem médica vital e de conscientização para o mundo todo (Fukuha; Archegas; Fürst, 2026, p. 177). Outra aplicação possível é a proteção da privacidade de pacientes, alterando a identidade de uma pessoa em um vídeo de diagnóstico médico, de modo que equipes compartilhem essas mídias sem revelar quem é o paciente (Pei *et al.*, 2026, p. 11). Por outro lado, essa técnica não descarta eventuais perdas de precisão nas imagens, que podem ter seu uso inviabilizado para determinados objetivos de pesquisa ou para aplicações que requeiram imagens fidedignas.

A tecnologia também produz inovações para o mercado de livros e de videogames. Na produção de audiolivros, por exemplo, os sistemas mudam a voz da narração para interpretar diferentes papéis. Eles geram vozes mais jovens ou mais velhas, masculinas ou femininas, com sotaques diferentes para cada personagem da história (Kietzmann *et al.*, 2020, p. 141).

Já nos jogos digitais, a ferramenta de alteração de rostos permite que os jogadores coloquem as suas próprias faces nos personagens. Contudo, considerando a ampla adesão aos jogos digitais por crianças e adolescentes, essa hiperpersona-

lização gera riscos significativos associados à vulnerabilidade desse público, que pode apresentar maior dificuldade em distinguir os conteúdos sintéticos da realidade, o que os deixa mais expostos a violações de privacidade, ao *cyberbullying* e a práticas de aliciamento, quando ofensores se aproveitam de identidades fabricadas e ferramentas generativas para a manipulação infantil no ambiente virtual (Negreiro, 2025, p. 2).

Essas transformações, por meio de ferramentas de *deepfake*, são realizadas com aprendizado de máquina para automatizar o processo. Dessa forma, o jogador precisa fornecer apenas algumas fotos ou *selfies* comuns, e com elas o algoritmo faz a troca de rostos ou a transição contínua (*face-morphing*) (Kietzmann *et al.*, 2020, p. 137). A grande diferença em relação às tecnologias anteriores está na qualidade e na facilidade de realizar a alteração. Anteriormente, criar efeitos semelhantes à realidade em personagens de jogos exigia equipamentos caros e profissionais treinados. Atualmente, as tecnologias de *deepfake* permitem que o jogador faça isso em poucos segundos, sem precisar de um estúdio de escaneamento 3D (Kietzmann *et al.*, 2020, p. 137).

2.5 Papel das definições técnicas na proteção de dados

A definição e a compreensão compartilhada de conceitos técnicos são indispensáveis para orientar respostas regulatórias compatíveis com os desafios impostos pelos *deepfakes*. Nos Estados Unidos, dois projetos de lei foram propostos nesse âmbito, mas, apesar de suas contribuições para o debate conceitual dessas ferramentas, eles não chegaram a ser sancionados. O primeiro deles, o projeto de lei *Malicious Deep Fake Prohibition Act* (Lei de Proibição de *Deepfakes* Maliciosos), de 2018, propôs regras rigorosas contra as adulterações digitais. A norma definia o *deepfake* como um registro audiovisual criado ou alterado de forma ilusória, cujo objetivo é parecer autêntico para um observador comum. Dessa forma, a mani-

pulação buscaria imitar de modo exato a fala ou a conduta real de um indivíduo (United States of America, 2018, p. 3). Esse projeto focava em punir o uso malicioso da tecnologia e proibia a criação e a distribuição de arquivos sintéticos que tivessem a intenção de facilitar crimes ou atos ilícitos.

Essa iniciativa ilustra uma dificuldade conceitual no debate regulatório global, que é a ausência de uma definição técnica única e estável para as mídias sintéticas. A rápida e contínua evolução das arquiteturas de inteligência artificial generativa torna definições estáticas elaboradas pelo legislador rapidamente obsoletas diante de novos modelos matemáticos de geração de conteúdo (Fukuha; Fürst; Archegas, 2026, p. 188–189).

No ano seguinte, em 2019, o projeto *DEEP FAKES Accountability Act* (Lei de Responsabilização sobre *Deepfakes*) trouxe novas propostas de estratégias de regulação. Além de punir abusos, a proposta exigia regras rigorosas de transparência para frear a desinformação (United States of America, 2019, p. 2). A norma obrigaria a inserção de marcas d'água digitais e avisos verbais e escritos nas mídias manipuladas que informassem ao público de modo claro que o conteúdo de vídeo, áudio ou imagem havia sofrido alterações técnicas.

O texto do projeto apresentava uma definição técnica central, descrevendo o *deepfake* como qualquer representação tecnológica que fosse substancialmente derivada da fala ou da conduta original de uma pessoa. Isso mostrava que a regra focava na adulteração feita por programas avançados de computador, e não na imitação humana física ou vocal. O ilícito ocorreria quando o sistema fizesse parecer que a vítima fez algo que, na verdade, ela não fez (United States of America, 2019, p. 15).

Ao contrastar as duas propostas, podemos observar uma nítida mudança na matriz conceitual adotada pelos legisladores, o que ilustra a volatilidade do debate regulatório em

torno das mídias sintéticas. O projeto de 2018 apoiava-se em um modelo de definição funcional e cognitiva, classificando *deepfakes* essencialmente pelo seu efeito psicológico de iludir um "observador razoável" quanto à autenticidade do arquivo (United States of America, 2018, p. 2). Em contrapartida, o projeto de 2019 migrou para um modelo de definição técnico-operacional, separando a adulteração algorítmica da mera imitação física ou verbal humana ao estipular que a falsificação audiovisual deve ser "substancialmente dependente de meios técnicos" (United States of America, 2019, p. 15–16).

Essa mudança de posicionamento já apontava para um desafio prático: o foco regulatório migrou do resultado perceptivo (a capacidade do arquivo de enganar) para o processo computacional de geração da mídia sintética.

A aprovação dos projetos não ocorreu em razão de impasses políticos, da complexidade do processo legislativo e de fortes pressões de grupos de interesse (Fukuha; Fürst; Archegas, 2026). Nesse cenário, os critérios que definem as aplicações como inofensivas ou lícitas estão intrinsecamente ligados ao direito à liberdade de informação, à liberdade criativa de desenvolvedores individuais e à otimização de processos mercadológicos em empresas de tecnologia e em indústrias criativas. Nessas instâncias funcionais, a redução de custos e o ganho de eficiênciasão usados como justificativa para a adoção e a defesa da inteligência artificial generativa (Hachkevych, 2025, p. 42), perspectiva distinta daquela a ser pensada pelo Estado na proteção dos direitos das pessoas, especialmente de grupos vulneráveis, como crianças e adolescentes.

A análise para diferenciar as mídias sintéticas inofensivas daquelas consideradas perigosas costuma ser estruturada a partir do conflito de interesses entre os desenvolvedores, os indivíduos e o Estado, ao invés da sua definição puramente técnica. Ainda assim, a compreensão técnica dessa mecânica de geração massiva de mídias sintéticas pode impactar diretamente o alcance da regulação (Fukuha; Fürst; Archegas,

2026, p. 172–173). Isso significa que a regulação estatal deve incorporar definições técnicas precisas, que não se limitem a proibir a arquitetura computacional em si, mas foquem na finalidade de uso e no dano provocado pela mídia sintética (Fukuha; Fürst; Archegas, 2026; Kirchengast, 2020, p. 222), como será retomado na **seção 6.1**.

Nesse cenário, observamos que a forma como o legislador traduz tecnicamente a ferramenta e define o *deepfake* na norma pode ditar as consequências práticas da lei, pois conceitos estáticos, ou excessivamente amplos, podem gerar defasagem estrutural e criminalização de usos benéficos (Fukuha; Fürst; Archegas, 2026, p. 188–189; Kirchengast, 2020, p. 415). Por outro lado, o esforço por definições equilibradas permitiria ao Estado cumprir sua missão dupla de assegurar a confiança na autenticidade das mídias e de resguardar os direitos fundamentais no ambiente digital (Fukuha; Fürst; Archegas, 2026, p. 168). Assim, analisar as definições de *deepfake* consideradas nas propostas de leis de outros países pode contribuir para o debate regulatório no Brasil.

‹ 3 como funcionam os *deepfakes* ›

Agora que já foram estabelecidas as fronteiras e distinções conceituais das mídias sintéticas, apresentaremos, nesta seção, uma visão geral sobre a arquitetura técnica dos *deepfakes*. O estudo técnico dos *deepfakes* se divide em dois processos: um que cria mídias sintéticas e outro que tenta identificá-las, denominados, respectivamente, de *pipeline*[6] de geração e de pipeline de detecção (Irfan *et al.*, 2025, p. 1928). O primeiro trata dos algoritmos dedicados à geração das mídias sintéticas e o segundo dos modelos projetados para a sua detecção.

6 *Um* pipeline *é um fluxo de trabalho para projetar, criar e implementar soluções tecnológicas, Ele pode representar o caminho sequencial, composto por várias etapas passo a passo, que os dados percorrem dentro de um sistema para realizar uma tarefa completa, do início ao fim.*

O ***pipeline* de geração** é uma operação matemática de transformação condicional. Isso significa que dado um vídeo ou imagem alvo e uma condição de controle (como a identidade facial de outra pessoa), a rede produz uma nova mídia que preserva o contexto original, mas altera os atributos desejados. Para atingir esse objetivo, os sistemas generativos utilizam técnicas avançadas de aprendizado profundo, como os autocodificadores, os modelos de difusão e as Redes Adversariais Generativas (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 273). As três arquiteturas diferem-se na estratégia matemática: autocodificadores usam compressão e reconstrução, modelos de difusão usam reversão de corrupção progressiva e Redes Adversariais Generativas usam competição entre redes. Cada uma produz resultados distintos em termos de controle, qualidade e custo computacional.

Embora essas abordagens possam ser utilizadas em vários contextos, no caso de *deepfakes*, os algoritmos são treinados massivamente para extrair características latentes do rosto humano e reconstruí-las de maneira automatizada, alterando identidades, expressões ou movimentos com alta fidelidade (Irfan *et al.*, 2025, p. 1930; Kietzmann *et al.*, 2020, p. 144). Dessa forma, os modelos aprendem a representar o rosto humano por meio de vetores numéricos compactos, os quais capturam padrões como estrutura óssea, textura de pele e dinâmica muscular, e que podem ser modificados e decodificados para gerar novos rostos. Esses modelos serão explicados a seguir.

Por outro lado, o ***pipeline* de detecção** atua como um mecanismo de defesa e controle forense, sendo modelado tecnicamente como uma operação matemática de classificação em nível de imagem ou de *pixel*. Isso significa que a detecção funciona como uma classificação binária: a rede decide, imagem a imagem, ou *pixel* a *pixel*, se o conteúdo é real ou sintético. A granularidade (imagem *versus pixel*) determina se o modelo apenas sinaliza a manipulação ou se também localiza onde ela ocorre (Pei *et al.*, 2026, p. 4). Esse macroprocesso emprega

redes neurais de classificação, a exemplo das Redes Neurais Convolucionais, do inglês *Convolutional Neural Network* (CNNs), e das Redes Neurais Recorrentes, do inglês *Recurrent Neural Networks* (RNNs), para extrair automaticamente características discriminativas e identificar artefatos, inconsistências espaço-temporais ou traços deixados pelos algoritmos de síntese (Irfan *et al.*, 2025, p. 1932). As CNNs analisam padrões espaciais em imagens individuais (como bordas irregulares, texturas inconsistentes), enquanto as RNNs analisam sequências temporais em vídeos (como movimentos faciais que violam a continuidade natural). Ambas buscam os rastros que os algoritmos de síntese, inevitavelmente, deixam.

Assim, no processo de engenharia de inteligência artificial, os geradores e os detectores de *deepfake* encontram-se engajados em uma constante competição, na qual os sistemas de detecção evoluem para expor as adulterações, enquanto os métodos de geração se aprimoram continuamente para contornar esses mecanismos de segurança e produzir conteúdos sintéticos cada vez mais indetectáveis (Irfan *et al.*, 2025, p. 1934).

Para sistematizar a compreensão da engenharia por trás da criação de mídias sintéticas, podemos distinguir a **infraestrutura** computacional subjacente, os **sistemas operacionais** aplicados e a **lógica matemática** que rege o aprendizado da máquina. Podemos apresentar essa distribuição conceitual da seguinte forma:

+ 1. **Infraestruturas:** Referem-se às arquiteturas computacionais de base e às estruturas topológicas que fornecem o alicerce para o processamento de alto volume de dados. As redes neurais constituem a arquitetura lógica que processa os dados. O suporte físico são as unidades de processamento gráfico, ou *Graphic Processing Units* (GPUs), necessárias pela escala dos cálculos envolvidos. O aprendizado profundo fornece a base sobre a qual todos os modelos são construídos: as redes neurais profundas aprendem

representações hierárquicas, desde as características mais simples (bordas, cores) às mais complexas (expressão facial, identidade), empilhando múltiplas camadas de transformações matemáticas.

+ 2. **Modelos Matemáticos:** Correspondem às funções, equações estatísticas, distribuições e problemas de otimização que realizam o treinamento interno dos modelos generativos. Trata-se da lógica matemática estrita que permite aos algoritmos ajustarem seus parâmetros para gerar mídias sintéticas realistas.

+ 3. **Modelos Generativos:** Os modelos representam os arranjos arquitetônicos, infraestruturas ou sistemas específicos construídos sobre as infraestruturas de Redes Neurais Artificiais (RNAs) para executar a tarefa direcionada de síntese ou manipulação de conteúdo. Eles definem a estratégia operacional de como as redes interagem para criar o arquivo final. Os principais modelos utilizados na atualidade são:

    - **Redes Adversariais Generativas (GANs):** Treinam dois modelos simultaneamente com objetivos opostos: o gerador produz imagens sintéticas tentando enganar o discriminador; o discriminador classifica imagens como reais ou falsas, tentando não ser enganado. Esse processo de treinamento atinge o seu objetivo quando o gerador aprende a produzir imagens com a mesma variedade e com características visuais estatisticamente muito semelhantes àquelas dos rostos verdadeiros.
    - **Autocodificadores:** Comprimem uma imagem em um vetor numérico compacto e depois a reconstrói a partir dele. Em *deepfakes* de troca de faces, dois autocodificadores compartilham o mesmo codificador, mas têm decodificadores distintos, um para cada identidade, permitindo que o vetor de uma pessoa seja decodificado com as feições de outra.

- **Modelos de Difusão:** Criam imagens realistas por meio de um processo gradual de limpeza, no qual o computador aprende a remover, passo a passo, um ruído inicial semelhante a uma tela borrada até formar uma figura nova e perfeitamente nítida (Pei *et al.*, 2026, p. 6).
- **Campos de Radiância Neural (NeRFs):** Funcionam como construtores virtuais que analisam imagens comuns para aprender as formas geométricas e a iluminação de um rosto ou cenário e usam essas informações para reconstruir figuras tridimensionais (3D) altamente realistas (Pei *et al.*, 2026, p. 6).

Ao longo desta seção, detalharemos esses modelos e estruturas que compõem a engenharia da inteligência artificial, os quais permitem o funcionamento das ferramentas de *deepfake*.

3.1 Bases para os algoritmos que produzem *deepfakes*

Para compreender como funcionam as ferramentas de *deepfakes*, precisamos olhar para um universo tecnológico mais amplo, que é o campo da inteligência artificial. Dentro desse universo, existe uma subárea fundamental chamada de aprendizado de máquina (*machine learning*). Nela, os computadores deixam de apenas seguir ordens e regras rígidas de programação e passam a aprender a partir da análise de um grande volume de dados. Dentro do aprendizado de máquina, encontramos as Redes Neurais Artificiais e o seu nível de operação mais avançado, o aprendizado profundo (Fukuha; Fürst; Archegas, 2026, p. 170–171).

Essa infraestrutura mais profunda e sofisticada viabiliza a Inteligência Artificial Generativa (IAG), uma vertente tecnológica cuja característica principal é a capacidade de aprender padrões matemáticos complexos para gerar e sintetizar conteúdos textuais, sonoros e imagéticos inéditos (Alves Junior, 2025, p. 10–11).

Cada um desses conceitos e, em especial, seus papéis na produção de *deepfakes*, são detalhados a seguir.

3.1.1 Aprendizado de Máquina ou *Machine Learning*

Um programa de computador utiliza uma técnica de aprendizado de máquina quando aprende a partir da experiência com determinadas tarefas e o seu desempenho nessas tarefas melhora à medida que acumula dados (Mitchell, 1997, p. 2). Isso ocorre por meio de ciclos repetidos de treino e teste do modelo de IA, nos quais o sistema analisa grandes volumes de dados históricos para identificar padrões e ajustar seus parâmetros internos, aprendendo a lógica do problema a partir da experiência acumulada (ANPD, 2024, p. 7).

Algoritmos tradicionais de aprendizado de máquina fazem isso a partir de entradas de **dados estruturados** e tabulares (por exemplo: idade, quantidade de compras, tempo de uso, categoria de produto) e utiliza uma matriz de atributos para mapear matematicamente a relação entre as características dos dados e o que elas representam, aprendendo uma função que toma decisões ou faz predições diretamente sobre essas variáveis. Essa abordagem se baseia no princípio de que, ao acumularem experiência com esses dados históricos, os sistemas podem melhorar seu desempenho de forma autônoma (ANPD, 2024). Nesse contexto, é preciso que engenheiros humanos escolham quais dados e detalhes o sistema deve considerar para tomar uma decisão (AWS, 2023). Portanto, diferentemente dos modelos mais modernos de aprendizado profundo, os modelos clássicos dependem dessas "características artesanais" (*handcrafted features*) desenvolvidas por especialistas para classificar informações (Dehghani; Saberi, 2026, p. 29)

É assim que funcionam, por exemplo, os filtros de e-mail que barram mensagens de *spam* ou as recomendações personali-

zadas de conteúdo da Netflix. Nesses casos, o sistema analisa dados históricos estruturados que os usuários de alguma forma forneceram, como palavras específicas de um texto ou o gênero dos filmes assistidos, para tentar prever o seu próximo passo (Goodfellow; Bengio; Courville, 2016, p. 3).

Portanto, as técnicas clássicas de aprendizado de máquina são valorizadas por sua simplicidade, eficiência e interpretabilidade. Uma vez que suas decisões são baseadas em regras e atributos claramente definidos por humanos, elas apresentam bom desempenho em conjuntos de dados menores, exigem menos recursos computacionais e são amplamente utilizadas em aplicações como modelagem financeira e detecção de fraudes (Rahnama, 2025).

3.1.2 Aprendizado de Máquina Profundo ou *Deep Learning*

O aprendizado de máquina evoluiu de uma abordagem predominantemente discriminativa para o desenvolvimento de modelos generativos sofisticados. A **abordagem discriminativa** é focada em aprender funções de decisão para classificar dados em grupos conhecidos ou prever probabilidades com base em atributos, muitas vezes, selecionados de forma manual (ANPD, 2024, p. 10–11). Por isso, o processamento de imagens e vídeos com redes discriminativas era focado principalmente na rotulagem e na identificação, como a detecção de rostos (Pei *et al.*, 2026, p. 15).

Já os **modelos generativos** consolidaram tarefas como troca de rostos, reencenação facial, geração de rostos falantes e edição de atributos faciais (Pei *et al.*, 2026, p. 2). Isso porque, diferentemente dos modelos discriminativos, que aprendem a classificar o que existe, os modelos generativos aprendem a distribuição de probabilidade implícita nos dados de treinamento, o que lhes permite sintetizar amostras estatisticamente plausíveis e inéditas (ANPD, 2024, p. 11). Ao extrair automaticamente representações de alto nível diretamente

dessas informações não estruturadas, o sistema torna-se capaz de sintetizar conteúdos inéditos e altamente realistas. E, como muitos conteúdos no ambiente digital são compostos por dados não estruturados, como imagens, áudios e vídeos brutos, isso impõe limites aos algoritmos tradicionais.

É justamente nesse aspecto que o aprendizado de máquina profundo se torna indispensável. Enquanto o aprendizado de máquina clássico oferece a base estatística para o reconhecimento de padrões, o aprendizado profundo utiliza múltiplas camadas para extrair conhecimento a partir de **dados brutos** (informações que ainda não passaram por nenhum tipo de processamento, filtragem ou organização), transformando-os incrementalmente para obter características de alto nível (ANPD, 2024, p. 7) e gerando conteúdos sintéticos altamente realistas.

Assim, o aprendizado profundo é um subconjunto do aprendizado de máquina orientado por **Redes Neurais Artificiais (RNA)**, que são modelos matemáticos e computacionais criados com o objetivo de imitar, de forma simplificada, o funcionamento do cérebro humano para resolver problemas complexos. Conforme Haykin (2007, p. 186), a estrutura fundamental dessas redes é o neurônio artificial, que atua como uma unidade de processamento capaz de aprender e armazenar informações. O autor explica que, na prática, cada neurônio recebe sinais de entrada, multiplica esses dados numéricos por "pesos", os quais indicam a importância de cada informação, e gera uma resposta de saída. Antes de transmitir esse resultado para a camada seguinte, o neurônio aplica uma função de ativação, que introduz não-linearidade ao sistema e é o que permite à rede aprender padrões complexos que vão além de simples relações proporcionais entre variáveis.

Essa abordagem permite que o sistema, à medida que constrói, também aprenda conceitos complexos a partir de representações mais simples organizadas em múltiplos níveis (Goodfellow; Bengio; Courville, 2016, p. 1, 8). A arquitetura de uma RNA é dividida em blocos interconectados, em que

os sinais de entrada (estímulos) chegam primeiro à "camada de entrada" e, em seguida, são repassados para as chamadas "camadas ocultas", representadas na imagem como primeira e segunda camada oculta, locais onde a maior parte dos cálculos e do aprendizado do sistema de fato acontece.

Após passar por todo esse processamento interno, os dados alcançam a "camada de saída", que tem a função de fornecer o resultado ou a resposta final (Haykin, 2007, p. 186–188). Essa arquitetura é estruturada para funcionar como uma linha de montagem de dados, em que os dados brutos vão sendo transformados passo a passo por meio das múltiplas etapas de cálculo. Cada nova camada da rede obtém e refina características de nível mais alto a partir da camada anterior, de forma autônoma e com uma menor necessidade de ajustes humanos (ANPD, 2024, p. 7; Irfan *et al.*, 2025, p. 1932). A Figura 7, a seguir, ilustra a estrutura de uma RNA.

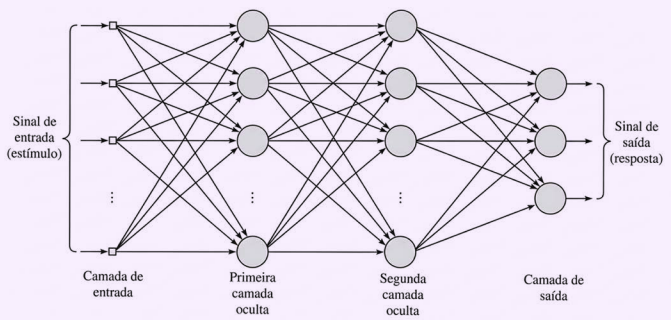


Figura 7 *Exemplo de RNA do tipo Perceptron multicamadas*

Fonte: Haykin (2007, p. 186).

É importante destacar que o aprendizado profundo abrange diferentes tipos de redes estruturadas em múltiplas camadas, e não apenas uma única arquitetura (Irfan *et al.*, 2025, p. 1932). Contudo, para a geração de mídias sintéticas como os *deepfakes*, o aprendizado profundo costuma utilizar uma arquitetura de rede neural específica, conhecida como Rede Neural Convolucional (CNN), que frequentemente trabalha em conjunto com outras estruturas profundas, como os

autocodificadores e as Redes Adversariais Generativas (Fukuha; Fürst; Archegas, 2026, p. 171). Esses tipos de redes são diferenciados adiante.

Nas redes neurais, as representações das informações processadas são organizadas em um **espaço latente** (ou espaço de representação), uma estrutura matemática que captura as características estatísticas principais dos dados (Pei *et al.*, 2026). É o espaço latente que permite ao algoritmo de aprendizado profundo extrair e organizar os padrões dos dados com mínima intervenção humana (Goodfellow; Bengio; Courville, 2016, p. 4; Tiu, 2020). Assim, ele comprime informações complexas e detalhadas dos dados não estruturados em um formato matemático simplificado, mantendo apenas o que é essencial para o modelo (Tiu, 2020). Isso permite a manipulação e a síntese de novas informações.

Como uma simplificação didática para exemplificar este conceito. considere um modelo generativo de aprendizado profundo treinado com imagens de móveis recebe como entrada uma imagem de uma cadeira (que é um dado bruto, não estruturado). Quando essa foto passa pelas camadas da rede neural, o modelo descarta informações redundantes ou irrelevantes para seu objetivo, nesse caso, o fundo ou a iluminação. Porém, o sistema guarda propriedades essenciais, como representações abstratas da estrutura visual contida na imagem. No mundo real, tais representações equivalem a padrões visuais, como a proporção entre altura e largura, a densidade de linhas verticais e a simetria das formas. Se representássemos essas características em um gráfico, os vetores de cadeiras ficariam agrupados por similaridade estatística e afastados dos vetores de mesas (Tiu, 2020).

Assim, esse processo converte dados brutos em vetores compactos que organizam os padrões fundamentais do conjunto original, os quais, no exemplo, seriam móveis (Goodfellow; Bengio; Courville, 2016; Tiu, 2020). Ao combinar as coordenadas de uma cadeira e de uma mesa, o sistema calcula uma

representação visual intermediária inédita (Tiu, 2020). Essa capacidade de realizar operações matemáticas na estrutura permite gerar novas imagens e serve como base técnica para a síntese de mídias de alta fidelidade. Na prática, essa combinação viabilizou algoritmos que fundem ou recriam rostos com alta precisão. A criação de *deepfakes* apoia-se em três abordagens principais baseadas em aprendizado profundo: modelos generativos; arquiteturas de redes neurais; e os métodos de representação de cena do tipo Campo de Radiância Neural, do inglês *Neural Radiance Field* (NeRF). Cada tecnologia apresenta características operacionais específicas no contexto da criação de mídias sintéticas, conforme veremos adiante.

3.2 Principais abordagens técnicas para a geração de *deepfakes*

A geração algorítmica de *deepfakes* apoia-se em diferentes modelos, cada qual com vantagens e desvantagens técnicas. As GANs, por exemplo, oferecem a vantagem de produzir imagens de alta fidelidade visual por meio de uma competição contínua entre uma rede geradora e uma discriminadora (EPRS, 2025, p. 1; Pei *et al.*, 2026, p. 273). Contudo, apresentam desvantagens notáveis, como instabilidade durante o treinamento e suscetibilidade à criação de artefatos indesejados, revelando dificuldades substanciais para lidar com oclusões severas ou grandes variações de pose facial (Pei *et al.*, 2026, p. 11–12).

Por sua vez, os Modelos de Difusão ganharam a vanguarda do fotorrealismo por sua superioridade na qualidade e resolução das imagens, tratando a síntese como um processo iterativo de remoção gradual de ruído para revelar imagens e vídeos íntegros (Pei *et al.*, 2026, p. 273). Porém, apresentam desvantagens que residem na alta complexidade computacional e no extenso tempo de processamento exigido pelas suas diversas etapas algorítmicas (Pei *et al.*, 2026, p. 14).

Por outro lado, a integração de Modelos de Linguagem de Grande Escala, do inglês *Large Language Models* (LLMs), e arquiteturas multimodais no processamento de vídeos traz a vantagem de aliar comandos textuais rígidos à geração visual e de atuar de forma forense, traduzindo as inconsistências de um arquivo falso diretamente para explicações analíticas em linguagem natural (Qian; Xia, 2025, p. 11–12). Os modelos multimodais são arquiteturas avançadas de inteligência artificial projetadas para processar, integrar e aprender relações estatísticas inerentes entre diferentes tipos de modalidades de dados simultaneamente, englobando texto, imagens, áudio e vídeo. O uso desses modelos, entretanto, esbarra em desvantagens como o elevado custo computacional estrutural e a dificuldade técnica de garantir um controle refinado sobre a intensidade de expressões emocionais contínuas nos rostos simulados em vídeo (Pei *et al.*, 2026, p. 16).

Essas diferentes abordagens são explicadas em maiores detalhes a seguir.

3.2.1 Redes Adversariais Generativas (GANs)

As Redes Adversariais Generativas (do inglês *Generative Adversarial Networks* - GANs) são um subcampo revolucionário do aprendizado profundo. Introduzidas por Ian Goodfellow em 2014, elas se consolidaram como uma das arquiteturas mais influentes na criação de mídias sintéticas hiper-realistas, os *deepfakes* (Patel *et al.*, 2023, p. 143301). Diferente de abordagens generativas baseadas em reconstrução, como os autocodificadores tradicionais – que operam por meio do mapeamento determinístico e da compressão de amostras individuais (Dehghani; Saberi, 2025, p. 9; Patel *et al.*, 2023, p. 143301) –, as GANs geram dados inteiramente novos, imitando fielmente a distribuição de dados reais nos quais foram treinadas (Patel *et al.*, 2023, p. 143297, 143301- 143302).

Esse refinamento matemático e a automação do processo foram os grandes precursores para a disseminação massiva de mídias sintéticas altamente realistas na internet, pois permitiram que usuários sem conhecimentos técnicos profundos gerassem falsificações visuais e sonoras ultraconvincentes em larga escala, utilizando *softwares* com interfaces simplificadas que "escondem" a complexidade matemática dessas tecnologias por trás de botões intuitivos (Dehghani; Saberi, 2025, p. 5; Patel *et al.*, 2023, p. 143297).

O funcionamento técnico das GANs é semelhante a um jogo não-cooperativo de rivalidade estratégica (Goodfellow *et al.*, 2014, p. 1, 3). Esse ecossistema é composto por duas Redes Neurais Artificiais concorrentes que são treinadas simultaneamente: o Gerador (G) e o Discriminador (D) (Dehghani; Saberi, 2025, p. 12; Patel *et al.*, 2023, p. 143297), em que:

+ **Gerador (G):** Tem por única função criar conteúdos sintéticos artificiais (sejam imagens, vídeos ou áudios) (Patel *et al.*, 2023, p. 143301). Ele recebe como entrada um vetor de ruído aleatório e tenta transformá-lo em uma amostra de mídia visual ou sonora plausível (Dehghani; Saberi, 2025, p. 12; Patel *et al.*, 2023).

+ **Discriminador (D):** Atua como um classificador binário. Ele recebe, de forma mista, tanto as amostras reais pertencentes ao banco de dados original quanto as mídias sintéticas criadas pelo Gerador, estimando a probabilidade de a amostra ser legítima ou falsa (Patel *et al.*, 2023, p. 143298, 143301).

A Figura 8 a seguir ilustra o funcionamento técnico das GANs.

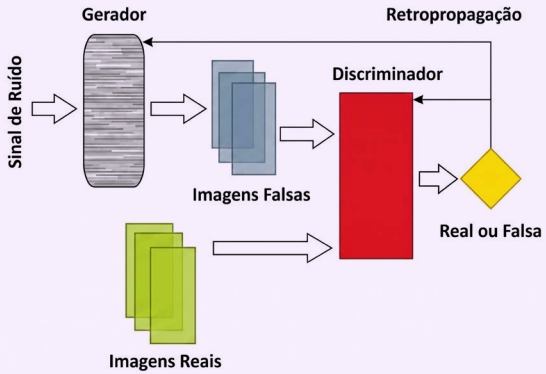


Figura 8
Funcionamento técnico das GANs

Fonte: Dehghani e Saberi (2025, p. 12, tradução nossa).

O dinamismo do treinamento ocorre por meio do ajuste contínuo de funções de perda baseadas no formato de um problema "*min-max*"[7]. O processo funciona da seguinte forma: enquanto o Discriminador é treinado para maximizar a sua probabilidade de acerto, o Gerador é atualizado para maximizar os erros do Discriminador, tentando enganá-lo. À medida que os ciclos de processamento avançam, o Gerador aprende a mapear texturas ocultas, contrastes e padrões estatísticos complexos. O treinamento atinge o equilíbrio ideal quando o Gerador produz amostras tão realistas que o Discriminador atinge uma taxa de acerto de aproximadamente 50% (Patel *et al.*, 2023, p. 143297- 143299, 143301-143302).

Essa arquitetura básica das GANs evoluiu para contornar limitações iniciais, como imagens borradas ou de baixa resolução (Patel *et al.*, 2023, p. 143302). Versões avançadas moldam o cenário atual de mídias sintéticas. Entre elas estão DCGAN (*Deep Convolutional* GAN), StyleGAN, CycleGAN e StarGAN.

Com seu papel fundamental no avanço do realismo visual, as GANs estabeleceram uma dinâmica de coevolução entre geradores e detectores de *deepfakes* no ecossistema digital (Dehghani; Saberi, 2025, p. 3–5). À medida que os modelos generativos mitigavam falhas iniciais por meio de sucessivas

7 *Segundo Patel et al. (2023), durante o treinamento de Redes Adversariais Generativas (GANs), o gerador tenta enganar o discriminador ao criar dados sintéticos realistas, enquanto o discriminador tenta maximizar a detecção dessas falsificações. Esse processo de otimização "min-max" promove um ajuste contínuo, aprimorando a capacidade do modelo de gerar conteúdos indistinguíveis dos originais.*

iterações sobre os dados – como as inconsistências nos padrões de piscada de olhos descritas por Patel *et al.* (2023, p. 143308) –, os sistemas forenses tradicionais baseados em análise do domínio espacial tornaram-se obsoletos. Esse cenário forçou a transição para arquiteturas de detecção híbridas e multimodais baseadas em aprendizado profundo (Heidari *et al.*, 2024). Tais defesas buscam identificar anomalias complexas, como as dissonâncias e incompatibilidades funcionais entre áudio, vídeo e movimentos labiais (Patel *et al.*, 2023, p. 143311–143312, 143317).

A detecção dessas mídias constitui um desafio crítico devido à falha dos modelos existentes ao enfrentar novas técnicas de geração e à vulnerabilidade dos sistemas diante de conteúdos compactados ou de baixa qualidade (Heidari *et al.*, 2024; Patel *et al.*, 2023, p. 143314–143315). Diante disso, o futuro do combate aos *deepfakes* depende de algoritmos de detecção cada vez mais sofisticados, além do desenvolvimento de ferramentas acessíveis e fáceis de usar que possam ser disponibilizadas ao público em geral (Dehghani; Saberi, 2025).

Embora as GANs se destaquem pela alta fidelidade visual de suas amostras, o caráter competitivo de seu treinamento frequentemente resulta em instabilidades matemáticas e problemas como o colapso de modo (Bond-Taylor *et al.* 2022). Em termos práticos, em vez de aprender a modelar a diversidade total da distribuição original (por exemplo, gerando rostos com diferentes gêneros, etnias e expressões), o modelo entra em um ciclo vicioso em que ignora a maior parte dos modos (agrupamentos) de dados e foca em apenas um ou poucos formatos específicos (Bond-Taylor *et al.*, 2022). Essa perda de diversidade compromete a utilidade da rede para a síntese realista e a manipulação ampla de mídias, o que motivou a comunidade científica a buscar funções de perda baseadas na distância de Wasserstein ou a migrar para alternativas probabilísticas como autocodificadores variacionais, do inglês *Variational Autoencoders* (VAEs), e os modelos de difusão (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

3.2.2 Autocodificadores Variacionais (VAEs)

Entre as arquiteturas generativas pioneiras da aprendizagem profunda destacam-se os Autocodificadores Variacionais (VAEs) (Dehghani; Saberi, 2025; Bond-Taylor *et al.*, 2022). Um VAE é um modelo de inteligência artificial generativa projetado para compactar dados e criar versões sintéticas baseadas neles, amplamente utilizado no processamento de imagens e edição de características faciais (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025). Além disso, essa arquitetura estende-se à geração de conteúdos sintéticos de áudio e música (Bond-Taylor *et al.*, 2022), bem como à clonagem e síntese de voz humana (Dehghani; Saberi, 2025).

Diferente dos modelos baseados em energia, do inglês *Energy-Based Models* (EBMs), que dependiam de simulações estatísticas lentas para gerar conteúdo, os VAEs revolucionaram o campo ao integrar a inferência variacional diretamente ao treinamento de redes neurais. Isso permitiu gerar novas amostras com alta eficiência e rapidez, em apenas uma passagem pela rede (Bond-Taylor *et al.*, 2022).

Para entender um VAE, imagine que você quer ensinar um computador a desenhar rostos humanos. O VAE faz isso trabalhando como uma dupla de redes neurais integradas: o Codificador (*Encoder*), que é a rede de inferência, e o Decodificador (*Decoder*), que é a rede generativa.

+ **Codificador:** É a rede neural responsável por mapear as imagens de alta dimensão em um espaço latente de menor dimensão, convertendo os atributos visuais em distribuições de probabilidade contínuas (geralmente definidas por médias e desvios-padrão) (Bond-Taylor *et al.*, 2022). No exemplo apresentado anteriormente, em vez de tentar decorar cada *pixel* de uma foto, o Codificador extrai as características fundamentais do rosto, como a cor dos olhos, a largura do nariz ou a presença de um sorriso, e as organiza estatisticamente nesse espaço resumido.

Vale notar que essa correspondência direta entre uma característica e uma dimensão isolada do espaço latente é uma simplificação didática; em um VAE convencional, essas características tendem a ficar emaranhadas nas dimensões latentes, e não isoladas uma a uma.

+ **Decodificador:** É a rede neural que faz o caminho inverso. gerara partir de uma amostra extraída desse espaço latente estruturado, o Decodificador reconstrói e gera novas amostras de dados que compartilham das mesmas propriedades matemáticas da distribuição original (Bond-Taylor *et al.* 2022). Assim, ele utiliza a lista de características abstratas gerada no espaço latente como um "manual de instruções" para desenhar uma face inédita, garantindo que o resultado realmente se pareça com um rosto humano real.

A Figura 9 ilustra a arquitetura e operação dos VAEs nas fases de treinamento de testes.

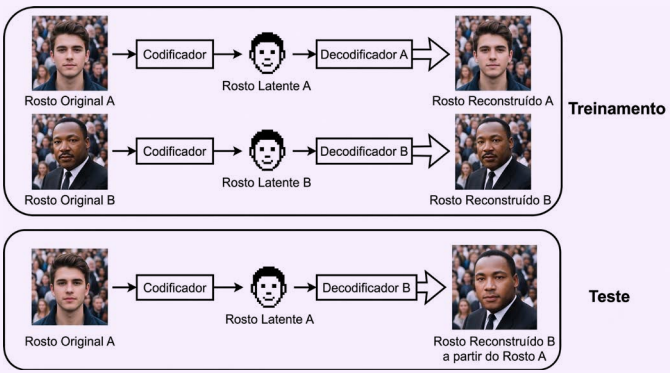


Figura 9 *Arquitetura e operação de VAEs nas fases de treinamento e testes*

Fonte: Adaptado de Dehghani e Saberi (2025, p. 31, tradução nossa).

A principal evolução conceitual que diferencia os VAEs dos autocodificadores tradicionais (cujo propósito se restringe à compressão e reconstrução determinística de dados) reside na modelagem explícita da distribuição de probabilidade do espaço latente, viabilizando o processo de amostragem e geração de novos dados (Bond-Taylor *et al.*, 2022; Dehghani;

Saberi, 2025). Em um autocodificador convencional, cada amostra de entrada é mapeada em um ponto fixo e determinístico do espaço latente. Essa ausência de regularização estatística resulta em um espaço descontínuo e esparso. Consequentemente, ao decodificar uma coordenada aleatória situada em uma região não mapeada durante o treinamento, o decodificador falha em estruturar a informação, produzindo apenas ruído visual desconexo.

Os VAEs superam essa limitação aplicando um mecanismo de regularização que garante a densidade, a suavidade e a continuidade do espaço latente (Dehghani; Saberi, 2025). Sob essa arquitetura, torna-se possível navegar pelas coordenadas latentes de forma linear para realizar interpolações suaves, como a transição gradual entre as características de diferentes faces, ou amostrar qualquer ponto aleatório para gerar um dado sintético inédito e visualmente realista (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

O processo de otimização de um VAE fundamenta-se no equilíbrio de uma métrica conhecida como Limite Inferior de Evidência, do inglês *Evidence Lower Bound* (ELBO) (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025). Por permitir um controle refinado sobre o seu espaço latente, o VAE tornou-se uma ferramenta para a edição precisa de mídias (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025), atuando como um dos modelos por trás de tecnologias de troca de rostos. Nesses sistemas, o codificador isola características, como expressões faciais de um indivíduo, para que o decodificador consiga projetá-las na identidade de outro. Além disso, essa flexibilidade matemática permite que a arquitetura se destaque tanto na síntese e clonagem de voz quanto na reconstrução e restauração de imagens corrompidas (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

Apesar de apresentarem velocidade de amostragem e estabilidade de treinamento superiores às das GANs, os VAEs convencionais enfrentam uma limitação histórica: a tendência

de gerar imagens levemente desfocadas ou borradas (Bond-Taylor *et al.*, 2022). Para mitigar esse problema, a área evoluiu para o desenvolvimento de VAEs hierárquicas e de arquiteturas híbridas que combinam a regularização dos VAEs com a nitidez de outras técnicas de modelagem (Bond-Taylor *et al.*, 2022).

Os autocodificadores variacionais baseados em energia, do inglês *Variational Autoencoder Energy-Based Models* (VAEBMs), por exemplo, são sistemas híbridos que integram a velocidade de geração dos VAEs ao refinamento de alta fidelidade visual dos EBM, superando o aspecto borrado das abordagens tradicionais (Bond-Taylor *et al.*, 2022). Ao acoplar ambas as estruturas, o VAE delimita a organização estrutural global do dado no espaço latente, enquanto o componente EBM atua diretamente no espaço de dados para corrigir imperfeições e resgatar detalhes de alta frequência (Bond-Taylor *et al.*, 2022).

Embora as limitações clássicas de nitidez visual tenham impulsionado o desenvolvimento de variantes hierárquicas e sistemas híbridos complexos como os VAEBMs, a busca por uma fidelidade ainda maior e por processos de geração mais robustos levou a comunidade científica a explorar novos caminhos (Bond-Taylor *et al.*, 2022). Diante desse cenário, a próxima seção abordará os modelos de difusão (*diffusion models*), uma classe de arquiteturas generativas que reformulou o estado da arte em síntese de mídia ao abordar a geração de dados sob uma perspectiva iterativa de remoção de ruído.

3.2.3 Modelos de difusão

Os modelos de difusão representam a evolução mais recente e impactante no campo da inteligência artificial generativa, superando as limitações históricas de seus predecessores (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025). Enquanto as GANs sofrem com o colapso de modo (gerando imagens repetitivas e sem variedade) e os VAEs produzem imagens de aspecto

borrado, essas novas arquiteturas alcançaram estabilidade matemática e alta fidelidade visual (Bond-Taylor *et al.*, 2022; Pei *et al.*, 2026). Em vez de tentarem gerar um dado complexo de uma única vez, os modelos de difusão quebram essa tarefa em um processo iterativo, inspirado em princípios da física estatística e da termodinâmica (Bond-Taylor *et al.*, 2022).

O mecanismo se estrutura em duas etapas (Bond-Taylor *et al.*, 2022):

+ Processo direto (*Forward process*): adiciona ruído gaussiano progressivamente a uma imagem real, segundo um cronograma predefinido, até convertê-la em ruído puro (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025).

+ Processo reverso (*Reverse process*): uma rede neural é treinada para estimar o ruído inserido em cada etapa e revertê-lo gradualmente, reconstruindo a imagem passo a passo (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025).

Assim, a geração parte de um tensor de ruído puro, lapidado passo a passo até se tornar um dado sintético realista (Bond-Taylor *et al.*, 2022; Pei *et al.*, 2026) — um processo mais estável, que evita o problema de colapso de modo das GANs (Liu *et al.*, 2025; Pei *et al.*, 2026).

Diferente dos VAEs, que comprimem os dados em um espaço latente reduzido (gerando perda de nitidez), os modelos de difusão preservam a dimensão completa da imagem original, resultando em texturas mais realistas (Liu *et al.*, 2025) — consolidando-se como o alicerce atual da síntese e edição hiper-realista de conteúdo (Liu *et al.*, 2025; Pei *et al.*, 2026).

Esse mecanismo funciona como um sistema condicionado: a rede recebe uma imagem facial alvo e informações adicionais (texto, áudio ou movimento) para gerar a mídia customizada, aplicando-se às quatro categorias de manipulação já introduzidas (**seção 2.2**). Na troca de rosto, os modelos DiffSwap

e DiffFace reduzem recortes artificiais, mas ainda falham sob oclusão severa ou má iluminação (Pei *et al.*, 2026, p. 10–12); na reencenação facial, o DiffusionAct permite maior controle sobre o resultado, embora métodos tradicionais ainda sejam mais robustos em movimentos amplos (Pei *et al.*, 2026, p. 12–14); na geração de rosto falante, DAE-Talker e VASA-1 dispensam marcações manuais no rosto, ao custo de alto processamento computacional (Pei *et al.*, 2026, p. 15–16); e na edição de atributos, o DiffusionRig preserva identidade e detalhes finos, enfrentando o desafio de evitar alterações indesejadas em outras partes do rosto (Pei *et al.*, 2026, p. 17–18).

Para avaliar essas mídias, a literatura adota métricas automatizadas voltadas a dois pilares: fidelidade visual e diversidade amostral (Liu *et al.*, 2025). As principais são:

+ **FID:** compara estatisticamente as imagens reais e sintéticas, quanto menor o valor, maior a fidelidade (Liu *et al.*, 2025; Pei *et al.*, 2026).

+ **IS:** mede nitidez e diversidade a partir da confiança de um classificador, quanto maior, melhor (Liu *et al.*, 2025).

+ **SSIM:** avalia a semelhança estrutural (luminosidade, contraste, estrutura) em escala de 0 a 1, de forma mais próxima da percepção humana que a comparação *pixel* a *pixel* (Liu *et al.*, 2025; Pei *et al.*, 2026).

+ **PSNR:** compara a imagem sintetizada com a original *pixel* a *pixel*, quanto maior o valor, mais próxima do original (Liu *et al.*, 2025; Pei *et al.*, 2026).

Além dessas métricas globais, a literatura especializada em manipulação facial usa critérios específicos: o CSIM, que mede preservação de identidade via reconhecimento facial, e as métricas LSE-C e LSE-D, que validam a sincronização labial em rostos falantes. Segundo Pei *et al.* (2026), a convergência

desses indicadores garante não só fidelidade visual, mas também consistência temporal e integridade de identidade.

Apesar dos avanços, a transição dos modelos de difusão do laboratório para sistemas práticos de larga escala ainda impõe desafios críticos de robustez (Liu *et al.*, 2025, p. 24–25; Pei *et al.*, 2026, p. 2, 5, 26).

3.2.4 Redes Neurais Discriminativas e Redes Neurais Convolucionais

Uma **Rede Neural Convolucional** (CNN) é uma arquitetura especializada de aprendizado de máquina profundo projetada para processar dados que possuem uma topologia "em grelha", como os *pixel*s que compõem uma imagem.

Tecnicamente, é uma rede do tipo alimentação direta ou *feedforward* (Dehghani; Saberi, 2025). Isto significa que é um modelo no qual os dados transitam em uma única direção, partindo da camada de entrada e atravessando as camadas ocultas até atingirem a camada de saída, sem que a informação retorne para camadas anteriores durante o processamento da inferência. Dessa forma, para um conjunto fixo de dados de entrada, o sistema produzirá exatamente a mesma saída em todas as execuções.

Diferente de modelos como os LLMs, que frequentemente operam de forma probabilística, amostrando o próximo *token*, as CNNs são arquiteturas fundamentalmente determinísticas. Embora, na prática possam ocorrer pequenas variações numéricas, devido a particularidades de *hardware* ou *software*, sua natureza determinística as torna uma escolha comum para aplicações que exigem alta consistência e reprodutibilidade, como é o caso de muitas tarefas forenses.

Ela é composta por múltiplas camadas, sendo as principais: camadas convolucionais, camadas de agrupamento (*pooling*)

e camadas totalmente conectadas. Cada uma dessas camadas serve para realizar a extração e a transformação de características de forma hierárquica e incremental. As camadas convolucionais utilizam filtros digitais (núcleos) que deslizam sobre a imagem para detectar padrões como bordas e texturas; as camadas de agrupamento realizam a subamostragem, reduzindo a dimensão espacial dos dados para diminuir a carga computacional e manter apenas as informações mais relevantes; e as camadas totalmente conectadas recebem os atributos processados para realizar a classificação final ou a predição.

O funcionamento central de uma CNN é baseado na operação matemática de convolução, onde filtros digitais deslizam sobre os dados de entrada para extrair padrões de forma hierárquica. Devido aos mecanismos de compartilhamento de pesos e conectividade local, as CNNs possuem a capacidade de detectar características (como o formato de um olho) independentemente de sua posição na imagem (Dehghani; Saberi, 2025). CNNs são utilizadas para identificar automaticamente pontos específicos na face, como cantos dos olhos e boca, permitindo criar uma malha digital para a reencenação facial (Pei *et al.*, 2026). Modelos como o MesoNet e o Xception utilizam arquiteturas convolucionais para realizar a classificação binária de vídeos, distinguindo se um rosto é autêntico ou sintético com base em anomalias de textura.

Já as **Redes Neurais Discriminativas** representam uma categoria de modelos cujo objetivo é aprender a relação matemática entre as características de um dado de entrada e a categoria a que ele pertence (ANPD, 2024). Elas focam em aprender uma função de decisão que permite classificar informações em grupos conhecidos ou prever a probabilidade de um dado pertencer a uma classe específica. O modelo é treinado para separar o espaço de dados em regiões distintas. Em tarefas de classificação, ele categoriza os dados em classes previamente conhecidas, enquanto em tarefas de regressão, ele prediz valores numéricos contínuos (ANPD, 2024).

Isso significa que o sistema funciona mapeando as propriedades estatísticas das entradas diretamente para um rótulo ou valor esperado, atuando como um mecanismo de verificação que decide se uma amostra atende aos critérios técnicos de uma determinada categoria com base no que foi aprendido durante o treinamento. Assim, em uma GAN o componente discriminativo atua como um validador técnico que avalia a autenticidade das amostras geradas, tentando identificar falhas estatísticas que as diferenciem dos dados reais (Pei *et al.*, 2026). Modelos discriminativos são treinados para identificar materiais maliciosos ou inconsistências fisiológicas, como a ausência de piscadas de olhos ou falhas na sincronização labial, que indicam que o conteúdo foi manipulado (ANPD, 2024).

3.2.5 Campo de Radiância Neural (NeRF)

O Campo de Radiância Neural (NeRF) é uma tecnologia de representação de cenas tridimensionais baseada em redes neurais que permite reconstruir objetos e ambientes complexos a partir de imagens bidimensionais (Pei *et al.*, 2026). No contexto de mídias sintéticas e *deepfakes*, essa abordagem é utilizada para garantir que o conteúdo gerado possua profundidade e consistência espacial.

O NeRF utiliza redes neurais para representar uma cena 3D de forma implícita. Diferente de modelos tradicionais que usam pontos ou malhas geométricas, o NeRF mapeia coordenadas espaciais (e o ângulo de visão) diretamente para valores de densidade de volume e cor (Pei *et al.*, 2026). Isso ocorre por meio do Campo Neural Implícito, que é a estrutura de dados operados pela renderização volumétrica, processo por meio do qual esses dados são extraídos para serem transformados em imagem.

Diferente de formatos tradicionais como o JPG (que guarda *pixel*s em uma grade 2D), o NeRF armazena a cena dentro

dos pesos de uma rede neural, a qual funciona como uma função. Ao receber uma coordenada tridimensional e um ângulo de visão, a rede calcula instantaneamente a cor e a densidade daquele ponto específico. A densidade se refere à opacidade. Uma densidade alta indica que o ponto faz parte de um objeto sólido (como a pele de um rosto), enquanto uma densidade zero indica espaço vazio. Por ser uma função contínua, o sistema não sofre com "*pixel*s estourados" ao dar *zoom*, permitindo uma reconstrução de alta fidelidade em altíssima resolução dentro do espaço modelado.

Em seguida, por meio da renderização volumétrica, representações tridimensionais complexas são projetadas em uma imagem bidimensional, determinando o valor visual de cada *pixel* na tela do espectador. Para realizar essa projeção, utiliza-se o método de marcha de raios (*ray marching*), no qual "raios virtuais" são emitidos a partir da perspectiva da câmera e atravessam o volume da cena modelada. Durante o trajeto de cada raio, o sistema executa a amostragem e consulta, selecionando múltiplos pontos espaciais para extrair informações de cor e densidade de volume diretamente dos pesos da rede neural que compõe o campo neural implícito. O estágio final é a integração numérica, uma operação matemática que soma as contribuições de luz de todas as amostras coletadas ao longo do raio. Esse cálculo considera a emissão de luz de cada ponto e o quanto essa luminosidade é reduzida pelos pontos que estão situados à frente na trajetória do raio, resultando no valor cromático final de um único *pixel*.

Em *deepfakes*, essa integração é o que permite a consistência de múltiplos ângulos (*multi-view*). Como a identidade da face está guardada de forma volumétrica (no campo implícito), o sistema pode renderizar o rosto de qualquer perspectiva nova simplesmente mudando a trajetória dos raios no processo de renderização, garantindo que a iluminação e a geometria do rosto falso acompanhem o movimento da cabeça de forma realista (Dehghani; Saberi, 2025). Diversas arquiteturas utilizam o NeRF para tarefas específicas de geração de conteúdo

sintético, por exemplo: AD-NeRF (*Audio Driven NeRF*) HiDe-NeRF FENeRF e FaceDNeRF.

3.3 Identificação de *deepfakes*

Diferentemente das adulterações manuais, a sofisticação do conteúdo classificado como *deepfake* está em seu processo automatizado de produção (Altuncu; Franqueira; Li, 2024). Os *softwares* conseguem mapear feições anatômicas com precisão e extrair o vocabulário e as nuances emocionais da fala de um alvo para gerar discursos inéditos, diminuindo as barreiras técnicas de acesso a esse tipo de manipulação (Witness, 2022). Exatamente por isso, é desafiador detectar *deepfakes* utilizando ferramentas computacionais. A maioria dos métodos de detecção de *deepfakes* ou de manipulação facial baseados em aprendizado profundo não é confiável e não explica a razão por trás do resultado da detecção (Akhtar, 2023). Isso se deve principalmente ao fato de as técnicas de aprendizado profundo serem, por natureza, uma "caixa-preta"[8].

Mesmo com o desafio de funcionarem como caixas-pretas, onde não se vê o processo interno de decisão, os atuais detectores de *deepfakes* ou de manipulação facial fornecem um rótulo, uma porcentagem de confiança ou uma pontuação de probabilidade de falsificação (Akhtar, 2023). Novas tecnologias baseadas em *Transformers* mostram boa capacidade técnica para identificar *deepfakes* de forma automática. Por exemplo, o sistema *Vision Transformers* (ViT) alcança 94% de precisão com um tempo de resposta de apenas 26,7 milissegundos. Além disso, os modelos híbridos funcionam bem mesmo com vídeos compactados, que o estado comum dos arquivos nas redes sociais (Khalid *et al.*, 2025; Charpe; Khokale; Dheeraj, 2025). Essas tecnologias lidam bem com os problemas mais frequentes nesse contexto: a perda de qualidade dos arquivos, novos tipos de adulteração (que o sistema não viu no treinamento) e falhas visuais ou de movimento

8 *Considera-se que técnicas de aprendizado profundo funcionam como uma "caixa preta" porque seus processos internos de decisão nem sempre são facilmente interpretáveis. Embora seja possível saber quais dados foram usados como entrada e qual resultado foi gerado, a forma como o modelo combinou milhares ou milhões de parâmetros para chegar a esse resultado pode ser difícil de explicar de modo transparente.*

que ocorrem ao longo do vídeo (Saga *et al.*, 2025; Charpe; Khokale; Dheeraj, 2025).

Por outro lado, essas pesquisas costumam medir o tempo de processamento do sistema de forma isolada, em laboratório. Elas não consideram o tempo total que leva desde o envio do vídeo pelo usuário até a decisão final de moderação, que é o que de fato causaria a remoção automática do *deepfake* no mundo real (Charpe; Khokale; Dheeraj, 2025; Raza; Ding, 2022). Assim, aplicar essa tecnologia em contextos reais para detecção e remoção ainda é um ponto crítico, porque: os modelos de laboratório perdem precisão quando enfrentam métodos de falsificação inéditos ou dados reais de baixa qualidade; e agentes mal-intencionados podem camuflar as falhas visuais em momentos-chave do vídeo para enganar a detecção (Rani *et al.*, 2025; Lu *et al.*, 2025; Wang *et al.*, 2022).

Também falta padronização para avaliar as explicações que os próprios detectores dão sobre como identificaram a fraude (Qian *et al.*, 2025). Para resolver isso, existem propostas promissoras, como transformar as notas brutas do sistema em cálculos de probabilidade que comparam as chances de o arquivo ser real ou falso (Guo; Li; Tang, 2026). Outra saída é o uso de *hashing*, identificadores únicos e históricos de processamento para rastrear as provas (Bhondave *et al.*, 2026; Williams *et al.*, 2025). Isso significa que a área avançou mais em descobrir a falsificação do que em provar, de forma repetível e válida na Justiça, como o sistema chegou àquela conclusão (Kraetzer *et al.*, 2022; Qian *et al.*, 2025).

Esses detectores só ganham confiança e validade jurídica quando funcionam como ferramentas de apoio para peritos humanos, permitindo verificação e contestação, mas não como detectores completamente automáticos (Siegel *et al.*, 2024; Sandoval *et al.*, 2024). A literatura científica ainda aponta para outras soluções em desenvolvimento, como marcas d'água, *blockchain* e rastreamento de origem, mas elas exigem testes mais amplos, certificação oficial dos países

e adaptação às leis locais antes de se tornarem ferramentas consolidadas (Bhondave *et al.*, 2026; Kraetzer *et al.*, 2022).

3.4 Limitações técnicas, incertezas e vieses

Apesar dos recentes avanços tecnológicos, a identificação de *deepfakes* ainda esbarra em diversas limitações técnicas e incertezas que dificultam a precisão dos resultados. Na prática, as ferramentas de análise atuais apresentam dificuldades para processar mídias com baixa resolução ou com durações muito variadas, o que frequentemente gera avaliações inconsistentes, como apontar falsos positivos em conteúdos originais e falsos negativos em vídeos que de fato foram manipulados (Moraes; Batista, 2025).

Além das falhas técnicas de detecção, os sistemas de verificação enfrentam um problemarelacionado aos vieses algorítmicos. Grande parte dos sistemas é treinado com bancos de dados altamente homogêneos, ou seja, que possuem uma variedade muito restrita de etnias, de gêneros e de idades. Essa falta de diversidade nas amostras de treinamento permite que a inteligência artificial aprenda os defeitos e as características dos rostos de forma limitada, o que gera diferenças consideráveis de desempenho ao analisar grupos demográficos distintos, e compromete a confiabilidade dessas ferramentas em situações do mundo real (Prabhu *et al.*, 2026, p. 6).

Como as barreiras tecnológicas atuais ainda falham com frequência devido a variações de resolução e vieses dos algoritmos, os conteúdos falsos conseguem contornar a verificação e se disseminar, causando danos profundos que comprometem desde a credibilidade das fontes de mídia até a autenticidade de provas no sistema de justiça (Moraes; Batista, 2025, p. 3; Prabhu *et al.*, 2026, p. 1). A seguir, discutiremos mais detalhadamente algumas das limitações técnicas e sobre os possíveis vieses algorítmicos.

Tecnicamente, as limitações dos sistemas de detecção ocorrem devido a ineficiências nas regras matemáticas que processam os dados no sistema, o que leva a avaliações inconsistentes em cenários reais.

Algumas técnicas de identificação de *deepfakes* são classificadas sob o guarda-chuva dos métodos tradicionais forenses e da análise de sinais biológicos (Kaushal *et al.*, 2025, p. 5; Prabhu *et al.*, 2026, p. 6). Métodos tradicionais forenses são métodos baseados em sinais e características manuais dos arquivos. Já a análise de sinais biológicos é uma técnica em que o sistema investiga os sinais naturais de vida no vídeo e, para isso, rastreia detalhes físicos e de comportamento, como o ritmo das piscadas, a pulsação do sangue e as pequenas expressões do rosto. Diferentemente das arquiteturas modernas de aprendizado profundo que descobrem padrões automaticamente, essas ferramentas operam a partir de classificadores estatísticos, métricas físicas e testes matematicamente pré-definidos. Por dependerem de sinais muito explícitos ou de defeitos específicos de *software*, esses modelos enfrentam obsolescência frente à capacidade das redes neurais modernas de regularizar e apagar tais imperfeições (Kaushal *et al.*, 2025, p. 5; Prabhu *et al.*, 2026, p. 6 7).

Já em sistemas baseados em Redes Neurais Convolucionais, os principais problemas técnicos são o "superajuste" (*overfitting*) e a dependência excessiva dos bancos de dados. Em lugar de aprender a identificar a essência da manipulação visual, o algoritmo frequentemente memoriza defeitos específicos do seu treinamento, como o nível de compressão do vídeo, o padrão de iluminação, ou até o fundo da imagem. Devido a esse viés, quando o sistema precisa analisar um *deepfake* criado por um método novo que ele não conhece, a sua precisão cai drasticamente (Prabhu *et al.*, 2026, p. 8).

Outra técnica de identificação é a análise da frequência das imagens, que, em lugar de focar nas cores, nas formas e nos pontos da imagem que nossos olhos conseguem ver,

examina a foto como se ela fosse um conjunto de ondas e de vibrações. Na vida real, as fotografias autênticas capturadas por câmeras possuem um fluxo natural, suave e contínuo nessas ondas (Prabhu *et al.*, 2026, p. 6).

Por outro lado, quando uma inteligência artificial tenta fabricar um rosto, ela não consegue imitar perfeitamente essa harmonia. Isso ocorre porque a forma como o sistema de IA constrói a mídia sintética deixa "cicatrizes" matemáticas no arquivo, invisíveis aos nossos olhos. Como o sistema monta a imagem por etapas e aplica filtros artificiais, ela acaba criando repetições e picos de ondas anormais que não existem na realidade. Assim, os programas de detecção procuram ativamente por essas "impressões digitais" deixadas pelo sistema para descobrir se a imagem é falsa (Prabhu *et al.*, 2026, p. 6).

A grande vulnerabilidade dessa abordagem é que essas cicatrizes digitais são muito frágeis e fáceis de disfarçar. Se o sistema falsificador aplicar um pequeno efeito na imagem essas cicatrizes podem desaparecer. Algumas formas de fazer isso são desfocar a imagem, aplicar um borrão, melhorar a resolução, ou postar o vídeo em uma rede social, o que comprime o arquivo. Com esses rastros apagados pela edição, o sistema de verificação não consegue mais encontrar as falhas no *deepfake* e acaba sendo enganado, o que torna a análise ineficaz se for usada sozinha (Prabhu *et al.*, 2026, p. 6–7).

Nesse cenário, há modelos de inteligência artificial mais avançados, que tentam analisar o contexto global do rosto e da imagem, como a arquitetura *Transformers*, que é uma arquitetura de redes neurais construída com base em um recurso chamado de autoatenção, ou mecanismo de atenção, que ensina o modelo a analisar toda a informação ao mesmo tempo e a focar nos detalhes essenciais (Prabhu *et al.*, 2026, p. 8).

Esses modelos de verificação baseados em Transformers dividem a imagem em vários blocos pequenos e utilizam o mecanismo matemático chamado "autoatenção" para conec-

tar e comparar todas essas partes ao mesmo tempo. Com isso, o sistema consegue avaliar o relacionamento entre áreas distantes da foto, verificando, por exemplo, se a iluminação do rosto faz sentido com a sombra do fundo, ou se a proporção geométrica entre os olhos e o queixo é natural. Dessa forma, o programa procura por erros no conjunto da obra, focando na coerência da cena para descobrir a falsificação (Prabhu *et al.*, 2026, p. 8–9).

A vulnerabilidade dessa abordagem, não está na fragilidade das pistas digitais, mas sim no seu alto custo operacional e estrutural. Como o sistema precisa cruzar as informações de todos os blocos da imagem, simultaneamente, para entender o contexto global, ele exige um poder de processamento computacional extremamente elevado. Essa exigência técnica torna o programa muito pesado e lento, o que inviabiliza o seu uso para analisar vídeos de em tempo real, ou a sua aplicação em dispositivos comuns e celulares do dia a dia, limitando o seu uso a computadores mais potentes e laboratórios especializados (Prabhu *et al.*, 2026, p. 9).

Outras abordagens de detecção também possuem vulnerabilidades técnicas claras. Somadas às falhas de arquitetura da máquina, as ferramentas comerciais de verificação esbarram em limitações operacionais diretas ligadas às características físicas do arquivo.

Testes práticos mostram que os algoritmos perdem a capacidade de análise ao processar vídeos de baixa resolução ou de longa duração. Nesses casos, os sistemas geram falsos negativos. Isso significa que a ferramenta aponta como real um conteúdo que é falso. Esse erro ocorre porque o tempo prolongado do vídeo introduz um excesso de variabilidade nos dados faciais analisados, o que compromete a identificação da fraude (Moraes; Batista, 2025, p. 3, 6). Por outro lado, vídeos muito curtos não fornecem amostras visuais suficientes para o sistema trabalhar, o que frequentemente causa resultados incertos, conflitos entre os próprios módulos

de avaliação da ferramenta e falsos positivos em conteúdos perfeitamente autênticos (Moraes; Batista, 2025, p. 4–6).

Além dos aspectos técnicos, relacionados aos algoritmos, o problema dos vieses não está somente associado à falta de diversidade. Outra questão importante na construção de bancos de dados de treinamento dos algoritmos é o foco excessivo em imagens de celebridades. Quando o sistema de detecção, treinado majoritariamente com rostos de pessoas famosas, tenta analisar vídeos de indivíduos comuns ou com características que fogem do padrão inserido na máquina, ele frequentemente falha e apresenta resultados imprecisos (Kaushal *et al.*, 2025, p. 6).

Esse cenário altamente homogêneo cria o que os pesquisadores chamam de "correlações espúrias", ou seja, falsas associações. Como a inteligência artificial aprende a partir de exemplos muito limitados, ela pode buscar atalhos e associar elementos que não têm relação direta com a falsificação facial em si (Prabhu *et al.*, 2026, p. 6). Por exemplo, o algoritmo pode atrelar um determinado tipo de iluminação ou a cor do cenário de um vídeo do banco de dados ao conceito de "falso", passando a considerar que esses detalhes irrelevantes de fundo são a verdadeira prova de que a mídia foi manipulada (Prabhu *et al.*, 2026, p. 6).

Outro viés estrutural ocorre quando o algoritmo fica "viciado" no próprio programa que gerou as imagens usadas no seu treinamento (Prabhu *et al.*, 2026, p. 6). Em lugar de aprender a procurar os rastros reais e fundamentais de uma manipulação, o sistema acaba apenas memorizando a "impressão digital" daquele *software* específico, como se ele decorasse as falhas e os defeitos visuais exclusivos de um único gerador (Prabhu *et al.*, 2026, p. 6).

Nesse cenário, a detecção de *deepfakes* enfrenta uma corrida constante contra a evolução das técnicas de manipulação (Prabhu *et al.*, 2026, p. 11). As limitações técnicas e os vieses

de dados mostram que ainda não existe um sistema infalível. Para superar essas barreiras, os pesquisadores indicam que os futuros sistemas precisam ser treinados com exemplos mais diversos e funcionar de forma rápida, em tempo real (Kaushal *et al.*, 2025, p. 6).

Também é necessário investir na criação de modelos multimodais. Essas são ferramentas mais completas, que analisam o áudio, a imagem e os movimentos da pessoa ao mesmo tempo. Outro avanço essencial é o uso da Inteligência Artificial Explicável (XAI). Esse recurso detalha exatamente como a máquina chegou àquela conclusão, o que traz mais transparência e confiança para os resultados analisados (Kaushal *et al.*, 2025, p. 6; Prabhu *et al.*, 2026, p. 11).

Por fim, a tecnologia não atua sozinha. Para garantir uma análise realmente segura e eficiente, é crucial combinar o uso dessas ferramentas automatizadas com a avaliação crítica humana (Moraes; Batista, 2025, p. 6).

‹ 4 *deepfake* e proteção de dados pessoais ›

Esta seção se dedica a compreender como a emergência e a disseminação de *deepfakes* impõem riscos de violação às normas que regem a proteção de dados pessoais. Esses riscos estão vinculados a três finalidades básicas de sua produção, que são: (i) o treinamento de modelos, quando os dados são utilizados como insumo para o aprendizado do algoritmo; (ii) a síntese ou manipulação da mídia visual ou auditiva, como conteúdo gerado em si; e (iii) a disseminação no ambiente digital. Para todos esses casos, o ordenamento jurídico brasileiro já oferece salvaguardas e permite que se identifiquem

os riscos à proteção de dados relacionados à produção e à circulação de *deepfakes*.

Assim, a proteção de dados pessoais corre riscos no treinamento de modelos de inteligência artificial para a geração de *deepfakes*. Isso ocorre, por exemplo, quando esse tratamento é realizado sem hipótese legal adequada, sem transparência, para usos secundários ou sem salvaguardas proporcionais, especialmente quando envolve dados sensíveis, como dados biométricos ou relativos à vida sexual. De igual modo, entra também nesse grupo a utilização de dados de crianças e adolescentes sem observância de seu melhor interesse e dos requisitos aplicáveis à hipótese legal utilizada – quando a base for o consentimento, este deve ser específico, em destaque e dado por ao menos um dos pais ou responsável legal.

Em ambos os casos, o tratamento de dados para treinamento de *deepfakes* implica limitações na autodeterminação informativa[9] dos titulares e restrições à sua privacidade, liberdade e livre desenvolvimento da personalidade. Esses são direitos protegidos por normas constitucionais, pela Lei nº 13.709 de 2018 (Lei Geral de Proteção de Dados – LGPD), pela Lei nº 15.211 de 2025 (ECA Digital) e por outras legislações, como os Decretos nº 12.975 e 12.976 de 20 de maio de 2026.

Quanto à finalidade de efetiva geração de mídias sintéticas ou manipuladas, quando o processo resulta em um *deepfake*, os riscos impostos à proteção de dados se ampliam. A produção de *deepfakes* com imagens ou vídeos aparentemente autênticos de alguém sem sua autorização pode trazer riscos associados ao livre desenvolvimento da personalidade e ao direito de imagem do titular. Isso porque a imagem de uma pessoa se conecta a características que a representam na vida social, como sua honra, reputação ou dignidade. Uma ameaça à imagem pública de alguém, nesse sentido, pode lhe trazer riscos, mesmo nos casos em que o tema da mídia não seja ilícito em si mesmo, mas que ameace sua imagem pública, por exemplo.

9 *A autodeterminação informativa é o princípio segundo o qual uma pessoa deve ter o poder de controlar as informações sobre si mesma e, portanto, sobre seus dados pessoais. Ele garante que o indivíduo tenha autonomia para decidir quando, como, por quem e para qual finalidade seus dados podem ser tratados. Na Lei Geral de Proteção de Dados, a autodeterminação informativa aparece como um dos fundamentos da proteção de dados pessoais, prevista em seu art. 2º, II (Brasil, 2018).*

Esse fenômeno se torna particularmente preocupante quando é produzido com a finalidade de disseminar conteúdo falso, difamatório e/ou humilhante no ambiente digital, como nos casos da síntese e da manipulação de mídias pornográficas não consentidas. Dados da organização Security Hero revelam que, em 2023, 98% das *deepfakes* envolviam pornografia e, em 99% desses materiais, as vítimas eram mulheres (Security Hero, 2023). Tais casos ameaçam a dignidade das vítimas, podendo implicar-lhes danos à imagem, à honra e à intimidade, e constituem forma de violência digital contra a mulher, prevista expressamente no Decreto nº 12.976/2026[10], que busca garantir a proteção de mulheres na internet.

A ampla circulação de mídias pornográficas, por sua vez, aumenta o risco de exposição de crianças e adolescentes a conteúdos impróprios, inadequados ou proibidos pelo ECA Digital. Conforme prevê a lei de 2025, fornecedores de produtos ou serviços digitais, provedores de aplicações, *sites* e demais agentes alcançados pela norma devem adotar medidas compatíveis para proteger crianças e adolescentes e assegurar que acessem conteúdos adequados à sua faixa etária, respeitada sua autonomia progressiva[11]. Nesse contexto, o elevado volume de mídias pornográficas de *deepfakes* favorece o desrespeito às normas protetoras do público infantojuvenil no ambiente digital.

Diante desse cenário, as próximas subseções são dedicadas a compreender e detalhar os riscos impostos por *deepfakes* à proteção de dados pessoais. Para tanto, analisam-se as categorias de dados tratadas, os métodos de coleta e processamento e a relação entre essas práticas e os princípios da Lei Geral de Proteção de Dados (LGPD). Ao final, são discutidos riscos sociais decorrentes de usos indevidos de *deepfakes*, viabilizados por seu funcionamento técnico, que extrapolam o ecossistema da proteção de dados pessoais e alcançam outros direitos.

10 *O Decreto reconhece a seriedade do risco ora descrito e proíbe que provedores de aplicações de internet gerem ou modifiquem conteúdo íntimo de terceiros, determinando que sejam implementadas salvaguardas técnicas e procedimentais para identificar e bloquear solicitações desse tipo nos provedores baseados em inteligência artificial (Brasil, 2026).*

11 *O tema da aferição de idade, que se relaciona à autonomia progressiva, foi discutido pelo Radar Tecnológico n. 5, elaborado pela ANPD e disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar-tecnologico-5-mecanismos-de-afericao-de-idade.pdf.*

4.1 Identificação dos dados pessoais tratados

Quando falamos em *deepfakes*, o primeiro passo para uma análise adequada sob a perspectiva da proteção de dados é entender ou, de forma mais precisa, reconhecer e qualificar, quais dados pessoais estão sendo utilizados ao longo desse processo. Essa identificação nem sempre é imediata ou evidente. Diferentemente de tratamentos tradicionais, os *deepfakes* operam por camadas, envolvendo dados fornecidos, observados, derivados e inferidos.

De acordo com Mussalo, Gain e Hotti (2018, p. 56–57), essas classificações podem ser elaboradas da seguinte forma: os dados fornecidos (*provided data*) são aqueles entregues conscientemente e de forma ativa pelos indivíduos, representando a categoria onde o titular possui o mais alto nível de percepção sobre a coleta; os dados observados (*observed data*) são registrados de maneira automática a partir das ações ou do ambiente do usuário (como câmeras, cliques ou sensores de internet); os dados derivados (*derived data*) são novos elementos criados mecanicamente a partir de outros dados, utilizando operações lógicas, matemáticas ou aritméticas relativamente simples; e, por fim, os dados inferidos (*inferred data*) que consistem na produção de perfis, categorizações ou tendências prováveis resultantes de processos analíticos complexos que buscam correlações entre diversas bases de dados. Os autores destacam que, à medida que os dados avançam de "fornecidos" para "inferidos", o nível de consciência do indivíduo a respeito dos seus próprios dados pode diminuir consideravelmente, e que podem ser altamente transformados.

Por isso, os *deepfakes* envolvem o processamento de múltiplas categorias de dados pessoais. Isso evidencia o papel central dos dados no funcionamento da tecnologia, os quais podem ser agrupados conforme o Quadro 2, a seguir:

Quadro 2 *Categorias de dados pessoais*

Categoria de dado	Descrição	Exemplos	Implicações regulatórias e riscos (LGPD)
Dado pessoal	Informação relacionada a pessoa natural identificada ou identificável	Nome, CPF, e-mail, RG etc.	O tratamento só pode ser realizado se houver uma base legal (ex: consentimento, cumprimento de obrigação legal, execução de contrato). Em caso de incidentes de segurança ou tratamento irregular, o controlador/operador é obrigado a reparar danos patrimoniais, morais, individuais ou coletivos
Dado pessoal sensível	Dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, além de dados referentes à saúde ou à vida sexual e dados genéticos ou biométricos, sempre que estiverem vinculados a uma pessoa natural	Dados genéticos ou biométricos vinculados a pessoa natural, incluindo padrões faciais, vocais, gestuais ou comportamentais quando tratados para identificação, autenticação ou singularização da pessoa.	Para o tratamento de dados, exige hipótese legal prevista no art. 11 da LGPD. O controlador deve utilizar apenas os dados estritamente necessários para a finalidade pretendida, implementar medidas de segurança adequadas aos riscos e garantir transparência sobre o tratamento realizado. O consentimento, quando utilizado, deve ser específico e destacado para uma finalidade determinada.
Dado anonimizado	Dado relativo a um titular que não possa mais ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento	Dados utilizados em bases para estudos estatísticos ou de saúde pública geridos por órgãos de pesquisa, onde a identificação do indivíduo foi eliminada ou mantida separadamente em ambiente controlado.	Não são considerados dados pessoais para os fins da LGPD. O principal risco é a reversão da anonimização: se a anonimização puder ser revertida por meios próprios ou esforços razoáveis, o dado poderá voltar a ser tratado como dado pessoal, com incidência das obrigações da LGPD.

Categoria de dado	Descrição	Exemplos	Implicações regulatórias e riscos (LGPD)
Dados de crianças e adolescentes	Informações relativas a crianças e adolescentes, cujo tratamento deve observar, de forma prioritária, o melhor interesse, a transparência adequada à idade, a minimização e salvaguardas reforçadas.	Informações coletadas para participação em jogos, aplicações de internet e outras atividades on-line	O tratamento de dados de crianças exige consentimento específico e em destaque dado por pelo menos um dos pais ou responsável legal, quando essa for a hipótese legal utilizada. É proibido condicionar a participação da criança em jogos ou aplicações ao fornecimento de mais informações do que as estritamente necessárias. As informações sobre o tratamento devem ser acessíveis e adequadas ao entendimento da criança

Fonte: Elaboração própria com base em Lei Geral de Proteção de Dados (LGPD).

Um aspecto frequentemente negligenciado é que o *deepfake* consegue **reaproveitar e recombinar conteúdos** que estejam publicamente disponíveis para criar simulações altamente realistas. Nesse viés, considerando que distintas jurisdições têm se dedicado à elaboração de normas que regem a proteção de dados pessoais, sobretudo os dados sensíveis, como proceder quando a própria pessoa titular do dado publica um vídeo seu em sua rede social? Quando, conscientemente, ela decide preencher formulários, cadastrar biometrias, compartilhar fotos, emitir opiniões? Ainda que haja proteção para conteúdos públicos, a questão central é se o novo uso, criado pelo *deepfake*, é compatível com a finalidade original e com a expectativa legítima da pessoa titular.

Pesquisas recentes demonstram que bases de treinamento obtidas por raspagem de dados[12] frequentemente contêm dados pessoais identificáveis, mesmo após tentativas de

12 *Raspagem de dado ou* Web scraping *é a técnica de extrair automaticamente dados de sites da internet, transformando conteúdos não estruturados em informações organizadas e planilhadas para análise. Também conhecido como "raspagem de dados da web", esse processo permite que, em vez de copiar dados manualmente, um programa realize a coleta de forma rápida e sistemática.*

anonimização. Em um estudo forense de um conjunto de dados popular de larga escala, Hong *et al.* (2026) notaram uma presença considerável de informações pessoalmente identificáveis (PII) apesar dos esforços de sanitização das plataformas, corroborando a preocupação de que bases obtidas por raspagem podem conter dados pessoais identificáveis ou reidentificáveis, inclusive dados sensíveis. Aplicando essa lógica ao contexto brasileiro, a disponibilização pública de um conteúdo não afasta, por si só, a incidência da LGPD, devendo o novo tratamento observar finalidade, boa-fé, interesse público, direitos dos titulares e hipótese legal adequada. Como complementa Lazzarini (2025, p. 10), o treinamento desses sistemas de inteligência artificial generativa depende da ingestão massiva de dados coletados na internet, os quais são incorporados aos modelos de treinamento muitas vezes sem a transparência adequada ou o consentimento dos titulares.

O caso do *DataComp CommonPool*,[13] apresentado por Hong. *et al.* (2026), por exemplo, ganhou repercussão mundial após uma auditoria acadêmica revelar que um dos maiores conjuntos de dados abertos usados para treinar modelos de inteligência artificial (com cerca de 12,8 bilhões de pares imagem-texto) contém uma quantidade massiva de dados pessoais, inclusive dados sensíveis, coletados por raspagem de dados ao longo de anos. Apesar de ter analisado uma fração mínima da base (cerca de 0,1%), a investigação encontrou milhares de exemplos de documentos pessoais como passaportes, cartões de crédito, certidões e carteiras de habilitação, além de imagens com rostos nitidamente identificáveis.

O aspecto mais preocupante relatado nesse artigo foi a identificação de documentos ligados à vida civil e profissional, como currículos e candidaturas a emprego, incluindo informações sobre raça, deficiência, endereço, entre outras. A partir dessa amostra, os pesquisadores estimaram que o conjunto completo pode conter centenas de milhões de registros desse tipo de conteúdo.

13 *DataComp CommonPool (https://www.datacomp.ai/dcclip/) é um conjunto grande dados de 12,8 bilhões de pares de imagem-texto projetado para pesquisa de IA multimodal.*

No contexto brasileiro, o debate sobre *deepfakes* e a identificação dos dados pessoais tratados se conecta diretamente com a LGPD, especialmente no que se refere ao tratamento de dados acessíveis publicamente e à observância dos princípios da finalidade e da transparência, bem como do dever geral de boa-fé que orienta o tratamento de dados pessoais na LGPD. A constatação de que dados sensíveis, como raça, saúde ou informações que possam reforçar vulnerabilidades, podem ser capturados e reutilizados em larga escala evidencia os limites da interpretação de dados pessoais acessíveis publicamente prevista no art. 7º, de dados sensíveis, prevista no art. 11 da LGPD, e tensiona práticas baseadas em raspagem indiscriminada de dados. Trata-se de um risco já amplamente debatido no país, com reflexos concretos em fenômenos como discriminação algorítmica e uso indevido de dados em contextos fraudulentos, tema que se conecta diretamente ao eixo central desta edição do Radar Tecnológico.

O episódio do *DataComp CommonPool* reforça a urgência de orientações mais específicas por parte dos reguladores e dos próprios provedores de tecnologia, bem como a necessidade de atualização e revisão dos marcos normativos aplicáveis aos conjuntos de dados de treinamento de IA. O caso demonstra que não se trata de um "vazamento" tradicional ou pontual, mas de um processo estrutural de coleta, reutilização e redistribuição de dados pessoais em escala industrial, frequentemente naturalizado e alheio ao conhecimento ou controle dos titulares.

Essa dinâmica revela, portanto, que o problema está na própria lógica estrutural de produção de sistemas de IA e *deepfakes*, que normaliza a apropriação massiva de informações pessoais sob o rótulo de "dados públicos". Nesse cenário, a identificação dos dados pessoais não se limita ao dado original, mas abrange também as formas como ele é transformado, combinado e reapresentado em novos contextos digitais, exigindo uma leitura ampliada do ciclo de vida dos dados. Isso é particularmente relevante porque a geração de *deepfakes*

depende de modelos treinados com grandes volumes de dados reais, capazes de extrair padrões biométricos e comportamentais em larga escala para seu contínuo aperfeiçoamento.

4.2 Como os dados pessoais são obtidos, processados e compartilhados?

Como acompanhamos até aqui, a coleta dos dados pessoais é uma etapa fundamental e ocorre por vias que são, muitas vezes, combinadas para alimentar os sistemas. Esses dados são indispensáveis para o treinamento de modelos de inteligência artificial, como as GANs e os modelos de difusão. Por sua vez, tais tecnologias são responsáveis por reproduzir rostos, vozes e gestos de maneira sintética com um impressionante nível de realismo.

A raspagem de dados é uma das formas mais comuns de coleta, por envolver a extração automatizada de conteúdos disponíveis publicamente em redes sociais por materiais em vídeos e fotos, notícias e entrevistas públicas, *sites* e bancos de imagens. Neste processo, já são coletadas imagens faciais em diferentes ângulos, áudios da voz das pessoas, expressões, movimentos e padrões de fala.

Outra forma de obtenção de dados é através de bases de dados estruturadas, como conjuntos de dados, a exemplo do que ocorreu no caso da *DataComp CommonPool*, em que há o uso de conjuntos de dados já organizados para treinamento. Outros exemplos de conjuntos de dados apresentados por Pei *et al.* (2026) são: ***Labeled Faces in the Wild (LFW)*** – conjunto de dados de geração com foco em imagem; **VoxCeleb** – conjunto de dados de vídeo e áudio voltados para fala humana com mais de 100 mil amostras; e **FaceForensics++** – conjunto de dados para detecção de falsificações. Frequentemente, esses conjuntos de dados incluem dados pessoais identificáveis ou potencialmente reidentificáveis, inclusive dados biométricos, o que gera muitas controvérsias sobre o consentimento dos titulares.

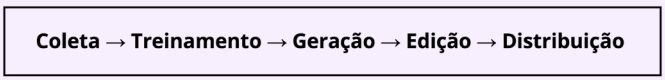
Uma quarta forma de coleta que podemos considerar aqui é a captura direta e intencional de dados por meio de aplicativos e plataformas que pedem *selfies*, gravações e/ou verificação biométrica. Nesse caso, pode haver o consentimento explícito do titular ou mesmo o uso indevido, com propósitos ilegítimos. Tal hipótese nos faz refletir sobre uma quinta possibilidade de coleta de dados, realizada a partir de conteúdo vazado ou obtido irregularmente, como, por exemplo, a partir de hackeamento de contas de usuários.

No contexto de *deepfakes*, é importante observar que os dados não são apenas coletados, mas também extraídos, inferidos, combinados e reutilizados em novos contextos. No que se refere ao processamento, os dados pessoais passam por etapas técnicas que incluem, além do armazenamento, a limpeza e a organização dos dados, a rotulagem (*labeling*) para identificar características específicas, a extração de atributos biométricos, e o treinamento iterativo dos modelos de aprendizado de máquina, como vimos na **seção 2** deste Radar Tecnológico. Ainda, o processamento pode envolver técnicas de ampliação artificial de dados (*data augmentation*) e a transformação desses dados em representações matemáticas abstratas, o que dificulta a identificação direta, mas não elimina o risco de reidentificação.

Após coletados e processados, com os modelos devidamente treinados, os dados podem ser distribuídos, ou seja, compartilhados. O compartilhamento de dados e de modelos pode ocorrer em diferentes níveis do ecossistema digital, incluindo o intercâmbio entre empresas, pesquisadores e plataformas tecnológicas, e envolve tanto o compartilhamento de conjuntos de dados quanto de modelos pré-treinados, APIs e infraestruturas de IA, muitas vezes em escala transnacional. Esse compartilhamento pode ter inúmeras finalidades, por exemplo acadêmicas, comerciais e/ou políticas, de um lado, mas também ilegítimas ou ilícitas, por outro. Ou seja, os dados podem ser reutilizados em contextos distintos daqueles originalmente previstos, o que dificulta a rastreabilidade,

a responsabilização e o controle pelos titulares. A disseminação final pode ser viral e descentralizada a partir de perfis em redes sociais, aplicativos de assinaturas mensais e *sites* fraudulentos, por exemplo. Há inúmeros outros riscos que serão apresentados adiante, bem como sua conexão com os princípios da LGPD.

Figura 10 *Fluxo de dados na formulação e na disseminação de deepfakes*



Fonte: Elaboração própria.

Em suma, como se esquematiza na Figura 10, o fluxo de obtenção, processamento, manipulação, arquivamento e compartilhamento dos dados possui esse sentido, em que os dados pessoais percorrem um ciclo contínuo e cumulativo de tratamento, no qual cada etapa potencializa os riscos à privacidade, especialmente em razão da possibilidade de recombinação e uso secundário não previsto.

4.3 Conexão com os princípios da LGPD

As últimas seções demonstraram como os riscos associados ao ecossistema de *deepfakes* atravessam todo o ciclo do tratamento de dados pessoais, abrangendo a coleta, o processamento e o compartilhamento desses dados para treinamento de modelos e geração de *deepfakes*. Esses riscos têm direta conexão com os princípios estruturantes da LGPD, que oferecem parâmetros normativos para a avaliação da legitimidade e da adequação dessas práticas.

O primeiro elemento que chama a atenção na leitura de *deepfakes* pelas lentes da LGPD é o fato de essas ferramentas dependerem, em grande medida, da coleta massiva de dados

pessoais, frequentemente obtidos por meio de técnicas de raspagem de dados em larga escala, discutidas acima. Desse processo decorrem diferentes aspectos de contrariedade aos princípios da mencionada lei.

Como exposto anteriormente, o treinamento com elevados volumes de dados pode envolver dados de identificação direta, como imagens e registros de voz, e mesmo dados sensíveis, como informações relacionadas a raça, saúde ou orientação sexual e às informações biométricas de uma pessoa identificada. Na maior parte dos casos, esses dados são coletados sem o conhecimento ou expectativa legítima dos titulares, de modo a contrariar o princípio da **transparência**, segundo o qual os titulares devem ser informados de maneira clara e acessível sobre as finalidades e as formas de tratamento de seus dados.

Esse contexto de coletas difusas e automatizadas também envolve, com frequência, a reutilização dos dados para tratamentos distintos daqueles originalmente previstos. Este é o caso de treinamentos de modelos feitos a partir de dados compartilhados para fins sociais, profissionais ou informativos, por exemplo. Esse tipo de prática contraria o princípio da **finalidade**, que determina que o tratamento de dados pessoais deve ocorrer para propósitos legítimos, específicos e previamente informados ao titular. Um tratamento secundário que não esteja claro nem seja informado ao titular constitui ferimento a esse princípio.

Além disso, em cenários de coleta automatizada, como ocorre no uso de raspagem de dados para treinamento de modelos, os titulares muitas vezes não sabem que seus dados foram coletados e, por consequência, não conseguem acessar informações sobre como eles estão sendo utilizados, por quanto tempo ou por quais agentes. Isso compromete, portanto, o princípio do **livre acesso** previsto na LGPD.

Para além das questões relativas aos usos secundários de dados pessoais envolvidos em práticas de raspagem de dados, a agregação de grandes bases de dados para treinamento de modelos aumenta o risco de vazamentos, acessos não autorizados e reutilizações indevidas – problemas que afrontam o princípio da **segurança** expresso na LGPD. Adicionalmente, os usos *deepfakes* para comprometimento de sistemas de autenticação biométrica e verificação de idade, por exemplo, ampliam os riscos à segurança dos dados pessoais de ecossistemas amplos.

A rápida evolução técnica de *deepfakes*, também discutida ao longo deste Radar Tecnológico, evidencia, por vezes, a insuficiência de medidas preventivas adotadas para a proteção de dados pessoais. Assim, a ausência de salvaguardas eficazes a essas ferramentas favorece a ocorrência de fraudes, extorsões, violência digital e outros riscos que decorrem da inobservância do princípio da **prevenção** na geração dessas mídias.

Em sentido similar, o uso de *deepfakes* para fins indevidos ou ilícitos, como nos casos de fraudes baseadas em clonagem de voz, imagem e técnicas multimodais e de pornografia não consensual, que serão discutidas adiante, constituem violações ao princípio da **não-discriminação** da LGPD. Isso porque tais ferramentas viabilizam essas práticas abusivas, que impactam desproporcionalmente grupos socialmente vulneráveis – pessoas idosas, adolescentes e mulheres, por exemplo. Nota-se, assim, como a aplicação mal-intencionada de certas tecnologias pode aprofundar desigualdades já existentes na sociedade.

Finalmente, a natureza distribuída e transnacional do ecossistema de *deepfakes* dificulta a identificação de responsáveis pelo tratamento indevido de dados pessoais e pelos usos ilegais dessas mídias. Compromete-se, assim, a demonstração de conformidade e de adoção de medidas eficazes de governança preconizadas pelo princípio da **responsabilização e prestação de contas** da LGPD.

Convém salientar que, diante dos riscos emergentes das práticas comumente associadas à geração de *deepfakes* e de seu potencial de violação de princípios da LGPD, o uso do consentimento como base legitimadora do tratamento de dados pessoais precisa ser questionado. Nos termos da LGPD, o consentimento constitui "manifestação livre, informada e inequívoca pela qual o titular concorda com o tratamento de seus dados pessoais para uma finalidade determinada" (Brasil, 2018). Isso se mostra inviável em operações de coleta massiva e indeterminada de dados disponíveis virtualmente, sobretudo quando estes são reutilizados para finalidades distintas daquelas originalmente previstas.

Nesse contexto, a substituição do consentimento pelo legítimo interesse também deve ser encarada com ressalvas, na medida em que o legítimo interesse não é hipótese legal prevista no art. 11 da LGPD para o tratamento de dados pessoais sensíveis. Assim, quando imagens, voz ou padrões comportamentais forem tratados como dados biométricos vinculados à pessoa natural, deverá ser identificada a adequada hipótese legal do mencionado artigo. Considerando-se que imagens faciais e registros de voz são elementos centrais para o treinamento e a geração de *deepfakes*, sua utilização em larga escala com base no legítimo interesse inspira preocupações quanto à proporcionalidade, à necessidade e ao impacto desse tratamento para direitos fundamentais dos titulares, como discutido ao longo desta seção.

Por fim, o tratamento de dados pessoais, incluindo dados sensíveis, sob a justificativa de que eles são "manifestamente públicos", por terem sido, muitas vezes, compartilhados virtualmente pelos próprios titulares, constitui leitura inadequada do art. 7º, § 4º da LGPD, que dispensaria o consentimento nesses casos[14].

Dados disponibilizados em contextos comunicacionais específicos, como interações em redes sociais, envio de documentos para contratações profissionais e compartilhamento

14 *"É dispensada a exigência do consentimento previsto no caput deste artigo para os dados tornados manifestamente públicos pelo titular, resguardados os direitos do titular e os princípios previstos nesta Lei" (Brasil, 2018).*

de informações pessoais em ambientes virtuais não devem ser utilizados para finalidades imprevistas, desvinculadas dos contextos em que essas informações foram expostas. Não se pode tomar como automática a publicidade desses dados pessoais, nem considerar tal publicidade como apta a afastar as salvaguardas da LGPD e de outros dispositivos legais do ordenamento jurídico nacional nesses contextos.

É importante destacar, por conseguinte, que o tratamento de dados pessoais, e particularmente dos dados sensíveis, deve sempre respeitar os contornos legais que autorizam sua realização. Embora os *deepfakes* sejam ferramentas tecnológicas recentes, sua geração não se dá em um vácuo normativo. Especialmente a partir de seus princípios, a LGPD já confere balizas para cada etapa de sua elaboração, garantindo que o tratamento de dados pessoais envolvidos nesse processo não implique violação aos direitos dos titulares.

Em aplicações de maior risco, especialmente quando envolverem dados biométricos, crianças e adolescentes, autenticação, perfilamento, geração de conteúdo íntimo ou circulação massiva de mídia sintética, recomenda-se documentação reforçada da governança, avaliação de riscos, medidas de segurança, retenção limitada, controles de acesso, mecanismos de atendimento aos direitos dos titulares e, quando cabível, elaboração de Relatório de Impacto à Proteção de Dados Pessoais (RIPD).

4.4 Riscos específicos conectados ao funcionamento técnico

Para além dos riscos relacionados aos tipos de dados tratados para a geração de *deepfakes* e do modo como esses dados são coletados e processados, essas mídias sintéticas ou manipuladas também implicam riscos sociais variados decorrentes de suas aplicações, que extrapolam o âmbito da proteção de dados pessoais. Como descrito ao longo deste Radar

Tecnológico, o funcionamento técnico de *deepfakes* envolve, com frequência, o treinamento de modelos de inteligência artificial com grande volume de dados, como imagens, vídeos e registros de voz de indivíduos reais. Com isso, a reprodução, a combinação e a simulação de características biométricas tornam-se cada vez mais realistas aos olhos humanos – e, também, a mecanismos tecnológicos de segurança baseados nesses dados pessoais.

Nesse contexto, uma das possibilidades de aplicação do *deepfakes* que tem ocasionado elevados riscos sociais é a de **contornar mecanismos de verificação de identidade.** Os *deepfakes* multimodais em tempo real – isto é, aqueles que combinam imagem, vídeo e áudio e simulam movimento – podem comprometer mecanismos de autenticação baseados em biometria (Altuncu; Franqueira; Li, 2024, p. 2).

Sistemas de reconhecimento facial e da chamada detecção de vivacidade (*liveness detection*), por exemplo, buscam verificar se o dado biométrico capturado provém de uma pessoa natural presente no momento da autenticação e não de uma reprodução estática ou sintética. Por esse motivo, são muito utilizados por serviços financeiros e bancários, para desbloqueio de dispositivos e aplicações, por serviços públicos digitais de identidade e por plataformas de comércio virtual e serviços digitais de maior risco.

Embora ofereçam elevados níveis de segurança, mesmo esses sistemas podem ser vulneráveis a *deepfakes* avançados, capazes de reproduzir sinais de vivacidade em tempo real (Altuncu; Franqueira; Li, 2024, p. 2; Pei *et al.*, 2026, p. 9). Nesses casos, o engano possibilitado por *deepfakes* pode ocasionar importantes danos sociais, por exemplo, ao permitir o acesso indevido a contas bancárias ou a outros serviços financeiros; abrir margem ao sequestro de identidade ou possibilitar o roubo de dados pessoais sensíveis (como é o caso dos próprios dados biométricos).

Essas mesmas funcionalidades permitem, ademais, a burla de **mecanismos de aferição de idade**. Como mencionado anteriormente, o ECA Digital determinou a adoção de mecanismos de aferição de idade por fornecedores de produtos ou serviços de tecnologias direcionados ou provavelmente acessados pelo público infantojuvenil. Uma relevante ferramenta técnica que possibilita essa aferição é, precisamente, a biometria facial. Nesse sentido, *deepfakes* que permitem contornar mecanismos de identidade são também capazes de ultrapassar barreiras etárias no ambiente virtual, funcionando como um dos mecanismos acessíveis para crianças e adolescentes burlarem barreiras destinadas a protegê-las (Colman, 2026; Lucca; Braga, 2026). Isso torna a experiência on-line desse público mais exposta a conteúdos impróprios, inadequados ou proibidos para sua faixa etária.

Os riscos emergentes do funcionamento de *deepfakes* não envolvem apenas burlas a mecanismos técnicos de segurança, mas também implicam possíveis usos para esquemas de engenharia social. Nesse contexto, é preocupante a utilização crescente dessas ferramentas para a prática de **fraudes financeiras corporativas ultrarrealistas**. Em tais situações, a sofisticação técnica de *deepfakes* é usada para simular videoconferências ou falsificar a voz de altos executivos de empresas, de modo que agentes mal-intencionados solicitem transferências bancárias em seu nome a funcionários ou parceiros econômicos dessas empresas, por exemplo (Fukuha; Furst; Archegas, 2026, p. 185–186).

Outros casos de engenharia social que lançam mão de *deepfakes* envolvem a **formulação de propagandas publicitárias falsas** contendo vídeos manipulados – e não autorizados – de celebridades nacionais e internacionais (Fukuha; Furst; Archegas, 2026, p. 186–187). Nesses exemplos, a ferramenta é utilizada para simular a participação de pessoas conhecidas pelo público em anúncios virtuais. Nas peças publicitárias, vídeos aparentemente autênticos de celebridades oferecendo produtos ou serviços reforçam a credibilidade dos anúncios,

que demandam, em geral, a inserção de dados pessoais ou o pagamento de valores falsamente indicados, como frete ou taxa, pelo consumidor (Romagna, 2025).

Ainda nos casos de engenharia social, destaca-se o crescente uso de *deepfakes* para a aplicação de **golpes contra pessoas idosas mediante a clonagem de voz** de familiares ou conhecidos. Por vezes a clonagem de voz é feita a partir de poucos segundos de áudio da pessoa clonada, coletados em redes sociais ou mensagens de áudio, e resulta em vozes sintéticas altamente convincentes. Esses casos exploram a precisão técnica proporcionada por *deepfakes*, as lacunas de letramento digital de populações idosas e o forte apelo emocional das mensagens, tornando extorsões e estelionatos direcionados a esse público cada vez mais comuns no Brasil (Fukuha; Furst; Archegas, 2026, p. 211; Lazzarini, 2025, p. 13).

Finalmente, riscos com alto potencial de dano relativos aos *deepfakes* se vinculam aos usos dessas ferramentas para a **geração de pornografia não consensual** e para a **exploração sexual infantil** (Kshetri; Voas, 2026, p. 105; Meding; Sorge, 2025, p. 158).

O funcionamento técnico de *deepfakes*, discutido neste Radar, permite a criação de imagens e vídeos falsos, totalmente sintéticos, sem qualquer participação ou conhecimento da pessoa retratada. Os modelos são treinados com extensas bases de imagens e vídeos, capazes de "inserir" o rosto de uma pessoa específica em conteúdo sexualizado, ou de "suprimir as roupas" de uma pessoa a partir de uma foto em que ela não está nua – o chamado *deepnude*. A mídia resultante desse processo pode ser compartilhada com outras pessoas, distribuída em larga escala ou mesmo utilizada para **extorsão sexual** – chantagem na qual criminosos ameaçam as vítimas de divulgar conteúdos íntimos manipulados com sua imagem caso não recebam determinada quantia (Fukuha; Furst; Archegas, 2026, p. 184).

A pornografia está na origem desse tipo de mídia sintetizada ou manipulada e altamente convincente, e segue representando a maioria do volume de *deepfakes* circulantes atualmente. O uso dessas ferramentas para a produção de conteúdos sexuais é facilitado por condicionantes técnicos, como o fato de aplicações de geração de *deepfakes* oferecerem, de modo cada vez mais acessível, recursos para a criação de conteúdo pornográfico: segundo dados de 2023, uma em cada três ferramentas relacionadas a essas mídias permitiam a síntese de pornografia (Security Hero, 2023).

O resultado da facilidade técnica para a usos discriminatórios é mensurável pela literatura especializada. Uma pesquisa que coletou e analisou 105 aplicações disponíveis na *Google Play Store* em dezembro de 2024 constatou que 89,5% das aplicações de *deepfakes* representa um perigo para as mulheres, isto é, "normaliza a objetificação, perpetua estereótipos de gênero e/ou potencialmente explorada para gerar conteúdo sexual dos corpos das mulheres" (Cuzcano-Chavez; González-Véliz, 2025, p. 53, tradução nossa).

O levantamento propõe, ainda, que os perigos às mulheres extrapolam a elaboração de *deepfakes* pornográficos e envolvem opções de *design* que favorecem usos sexistas dessas ferramentas. Exemplos disso são funcionalidades que permitem a simulação de beijos e abraços (além de cenas íntimas) entre o usuário e um terceiro e a disponibilidade de predefinições (*templates*) explicitamente desenhadas para a erotização feminina, que exibem mulheres com atributos corporais exagerados, decotes e roupas íntimas e homens em vestimentas completamente cobertas. Esses casos sugerem, portanto, que o funcionamento técnico de aplicações de *deepfakes* podem induzir usos maliciosos dessa tecnologia, com alto potencial danoso para mulheres. A Figura 11 ilustra parte desses achados:

Figura 11 *Capturas de tela de aplicações de deepfake*

Fonte: Cuzcano-Chavez; González-Véliz, 2025, p. 55.

Dessa maneira, constata-se que o uso de *deepfakes* para a geração de mídias íntimas sem autorização é uma espécie de abuso sexual baseado em imagem, uma faceta da violência de gênero facilitada pelo uso de tecnologias (Brigham; Wei; Kohno; Redmiles, 2024, p. 374). Embora se origine no ambiente virtual, a criação de *deepfakes* pornográficos não consentidos gera impactos concretos na saúde e na vida social de suas vítimas, ocasionando-lhes, entre outros males, estresse pós-traumático, ansiedade, depressão, isolamento, baixa autoestima, problemas de confiança e sentimento de violação corporal (Brigham; Wei; Kohno; Redmiles, 2024, p. 374).

A produção de *deepfakes* sexuais não consentidos é, portanto, um problema de ordem tanto quantitativa quanto qualitativa para a garantia de ambientes digitais seguros. Diante desse cenário, como adiantado na introdução a esta seção, o Decreto nº 12.976, de 20 de maio de 2026, proibiu a geração e a modificação de conteúdo íntimo de terceiros mediante uso de inteligência artificial por provedores de aplicações de internet.

Para além de mídias sexuais relativas a pessoas adultas, há ainda casos em que a geração de materiais sintéticos é feita para exploração sexual infantil (UNICEF, 2026). Nessas hipóteses, os conteúdos podem ou não corresponder a vítimas reais e identificáveis, mas mesmo nos casos em que a mídia é totalmente sintetizada, sua produção pode se basear na coleta e

no processamento de dados reais de crianças e adolescentes (imagens, rostos, características físicas, por exemplo). Uma vez gerados, esses conteúdos podem ser utilizados em contextos de aliciamento, extorsão, *cyberbullying* ou abuso infantil (Kshetri; Voas, 2026, p. 105), contrariando a proteção integral de crianças e adolescentes preconizada pela Constituição Federal, pelo Estatuto da Criança e do Adolescente (ECA) e pelo ECA Digital.

Em resumo, observa-se que, de distintas maneiras, os usos de *deepfakes* descritos ao longo das últimas seções apresentam conflitos com o arcabouço normativo brasileiro e riscos sociais que se relacionam à proteção de dados e extrapolam seu escopo. Na próxima seção, discutiremos como essas ferramentas têm sido observadas no contexto brasileiro.

‹ 5 *deepfake* no contexto brasileiro ›

Abordamos até aqui o que é um *deepfake*, quais são os seus conceitos centrais, aplicações, riscos, relação com a LGPD e limitações técnicas. Pode-se afirmar, até o momento, que a existência de *deepfakes* representa um dos exemplos mais avançados da convergência entre inovação tecnológica e disseminação de informações em larga escala no ambiente digital contemporâneo.

No contexto brasileiro, esse avanço tecnológico já se traduz em ocorrências concretas e mensuráveis. Observa-se o aumento do uso de conteúdos sintéticos em fraudes digitais, desinformação e manipulação de identidade, especialmente em ambientes de alta circulação de informações, como redes sociais e aplicativos de mensagens. Esse cenário evidencia

que o *deepfake* deixou de ser uma preocupação futurística para se tornar um problema de impacto econômico e social atual, relevante e urgente.

A partir do Decreto nº 12.975/2026, de 20 de maio de 2026, observa-se a consolidação de um arcabouço de responsabilização, governança e moderação de conteúdo que passa a abranger, de forma indireta, esse tipo de material no ambiente digital. O decreto altera a regulamentação do Marco Civil da Internet (Lei n° 12.965/2014) e estabelece deveres ampliados para provedores de aplicações. Além disso, ele incorpora a mudança de entendimento do Supremo Tribunal Federal (STF) sobre o art. 19 dessa lei, promovendo a transição de um modelo predominantemente reativo, ou seja, de uma resposta a um descumprimento de ordem judicial, por exemplo, para um modelo orientado ao dever de cuidado e à gestão de riscos.

O Decreto nº 12.976/2026, por sua vez, trata expressamente da geração e da modificação de conteúdo íntimo por inteligência artificial ou recursos tecnológicos semelhantes. Nesse decreto, que estabelece diretrizes para a proteção de mulheres na internet, provedores de aplicação de internet são proibidos de gerar tais conteúdos e devem implementar salvaguardas técnicas para bloquear solicitações nesse sentido.

Esse movimento regulatório não ocorre de forma isolada. Pelo contrário, é acompanhado pela atuação de diferentes instituições no país. Destacam-se iniciativas do Tribunal Superior Eleitoral (TSE), que tem disciplinado o uso de inteligência artificial em campanhas eleitorais; da própria ANPD, com orientações voltadas ao tratamento de dados pessoais em contextos de alto risco; dentre outras, que veremos nas seções e subseções a seguir.

Lazzarini (2025, p. 7) destaca que a produção de imagens artificiais gera uma pressão sobre as estruturas tradicionais do Direito Civil no Brasil. Como essas novas mídias são fabricadas

a partir de análises de dados e não da captura de momentos concretos, elas exigem uma reavaliação do direito de imagem, historicamente ancorado apenas em registros reais de fotografias e filmagens.

Para além da dimensão conceitual, esse tensionamento também se manifesta no plano prático, especialmente no que diz respeito à produção de prova e à verificação da autenticidade de conteúdos digitais. A dificuldade de distinguir registros reais de conteúdos sintéticos impacta diretamente a atuação do Poder Judiciário, das autoridades administrativas e das próprias plataformas, exigindo o desenvolvimento de ferramentas técnicas de detecção, bem como critérios jurídicos adaptados a esse novo contexto.

O Decreto nº 12.975/2026 estabelece deveres de cuidado, governança e responsabilização de plataformas digitais, inserindo determinados usos abusivos de *deepfakes* no conjunto de riscos regulatórios sistêmicos[15] associados à atuação de provedores, especialmente quando relacionados a fraudes, anúncios ilícitos, redes artificiais de distribuição ou violações a direitos. Esse contexto exige prevenção, detecção e resposta ativa pelos provedores, em consonância com a proteção de dados pessoais e direitos da personalidade. Em outras palavras, práticas frequentemente associadas ao uso de *deepfakes*, como fraudes digitais, violência contra mulheres, manipulação de identidade, estelionato, dentre outras, passam a ser enquadradas como categorias de risco cuja mitigação demanda remoção. Isso também requer implementação de mecanismos estruturais de prevenção e moderação por parte das plataformas.

Nesse contexto, apresentamos nesta quinta seção casos ilustrativos, a fim de subsidiar e aprofundar a reflexão sobre o tema. Discutiremos casos como o uso de *deepfakes* pornográficos contra meninas e mulheres; em contextos de racismo; em contexto eleitoral; na violência política de gênero; na aplicação de golpes financeiros e em propagandas enganosas.

15 *A compreensão de risco sistêmico se refere a ameaças aos direitos fundamentais geradas pela arquitetura ou pelo funcionamento de serviços ofertados por provedores de aplicações on-line, com impactos que ultrapassam a esfera individual.*

5.1 *Deepfakes* pornográficos contra meninas e mulheres

A utilização de *deepfakes* para a produção de conteúdo pornográfico não consentido tem se consolidado como uma das manifestações mais graves e difundidas do uso abusivo de inteligência artificial. Trata-se de um fenômeno que incide de forma desproporcional sobre meninas e mulheres, reproduzindo padrões estruturais de violência de gênero no ambiente digital (McGlynn; Toparlak, 2025). Evidências empíricas recentes indicam que as vítimas identificadas em levantamentos nacionais são, em sua totalidade ou em ampla maioria, do sexo feminino, incluindo crianças e adolescentes (Safernet Brasil, 2025).

O principal *locus* de incidência desses casos tem sido o ambiente educacional. Levantamento conduzido pela SaferNet Brasil identificou, entre 2023 e 2025, ao menos 16 casos documentados em instituições de ensino de 10 unidades federativas, envolvendo 72 vítimas e 57 agressores (Safernet Brasil, 2025). Os episódios caracterizam-se pela criação, por meio de ferramentas de IA generativa, de imagens de nudez ou de conteúdo sexual em que rostos reais – frequentemente extraídos de redes sociais – são sobrepostos a corpos nus, sem qualquer consentimento das pessoas retratadas.

Casos concretos ilustram a dinâmica desses ilícitos. Em diferentes estados, estudantes foram investigados por produzir e compartilhar *nudes* falsos de colegas e até de professoras, utilizando aplicações de inteligência artificial acessíveis ao público. Em uma escola no Rio de Janeiro, por exemplo, adolescentes criaram e disseminaram imagens íntimas falsas de colegas, o que levou à abertura de uma investigação pela Delegacia de Proteção à Criança e ao Adolescente (Neto, 2023). Em outro caso, no Mato Grosso, quatro alunos foram responsabilizados por produzir montagens envolvendo mais de 30 vítimas, incluindo estudantes e uma professora, com ampla circulação em grupos digitais (G1 MT, 2024). Esses

episódios revelam não apenas a banalização da prática, mas também sua inserção em dinâmicas de *bullying*, de humilhação pública e de violência simbólica.

Além da dimensão quantitativa, destaca-se também a dimensão qualitativa desses danos. *Deepfakes* pornográficos constituem uma forma particularmente intensa de violação da dignidade humana, pois combinam elementos de exposição sexual, falsidade material e perda de controle sobre a própria imagem. Estudos acadêmicos brasileiros apontam que esse tipo de prática deve ser compreendida como uma manifestação contemporânea de violência de gênero, na medida em que instrumentaliza o corpo feminino como objeto de humilhação e de controle social (Valente, 2023; Rodrigues, 2025). Ademais, a natureza sintética do conteúdo não reduz seus efeitos — pelo contrário, dificulta sua contestação social e jurídica, prolongando os danos reputacionais e psicológicos às vítimas (Rodrigues, 2025).

Outro aspecto crítico é a escala e velocidade de disseminação desses conteúdos. Dados da SaferNet Brasil (2025) indicam que *link*s relacionados à circulação de *deepfakes* sexuais e de material de abuso associado vêm sendo identificados desde 2023, circulando em redes sociais, aplicativos de mensagens e até em fóruns na *dark web*. A atuação frequentemente envolve grupos organizados que utilizam mecanismos automatizados (como *bots*) para ampliar a distribuição, o que contribui para a persistência e replicabilidade do dano.

Caso Real

Processo de fiscalização da ANPD sobre deepfakes pornográficos envolvendo mulheres e crianças/adolescentes

Segundo a Nota Técnica nº 1/2026/FIS/CGF/ANPD, a ANPD recebeu, no dia 14 de janeiro de 2026, uma

Representação e Petição apresentadas pela deputada federal Érika Hilton, denunciando a irregularidade do sistema de inteligência artificial Grok em face da LGPD.

Em 20 de janeiro de 2026, a ANPD, o Ministério Público Federal (MPF) e a Secretaria Nacional do Consumidor (Senacon), do Ministério da Justiça e Segurança Pública, expediram Recomendação Conjunta à empresa X Brasil Internet Ltda., responsável pela plataforma X (antigo Twitter). A iniciativa decorre de denúncias e evidências sobre o uso da ferramenta de inteligência artificial Grok, integrada à plataforma, para a criação de conteúdos sintéticos potencialmente ilícitos, especialmente *deepfakes* de caráter sexual envolvendo mulheres e crianças e adolescentes.

Nela, as instituições destacam que o Grok passou a permitir a edição automatizada de imagens de terceiros sem verificação de consentimento ou de finalidade legítima, o que viabiliza a produção de conteúdos falsos altamente realistas. Há registros de uso da ferramenta para gerar imagens sexualizadas e pornográficas sem a autorização das pessoas retratadas, incluindo menores de idade, com ampla disseminação pública na plataforma. Tal cenário potencializa danos à dignidade, à imagem, à privacidade e à integridade física e psicológica das vítimas. O documento fundamenta-se em diversos marcos normativos nacionais e internacionais. Entre eles, destacam-se a Constituição Federal, o Estatuto da Criança e do Adolescente (ECA), a Lei Geral de Proteção de Dados (LGPD), o Marco Civil da Internet e o Código de Defesa do Consumidor (CDC). Também são citadas normas mais recentes, como o ECA Digital (Lei nº 15.211/2025), que reforça a obrigação de adotar medidas preventivas desde a concepção dos serviços digitais. O texto ressalta que a proteção de crianças e adolescentes deve observar o princípio da proteção integral e a prioridade absoluta, e que a criação de *deepfakes* sexualizados pode configurar, inclusive, crimes previstos na legislação penal.

Em relação às mulheres, o uso da ferramenta para gerar conteúdo sexual não consensual é entendido como uma possível forma de violência psicológica de gênero, tipificada no Código Penal, além de violar compromissos internacionais assumidos pelo Brasil, como a Convenção de Belém do Pará. O documento também problematiza a responsabilidade da plataforma, argumentando que, ao oferecer e estruturar a ferramenta de IA, a empresa não atua como mero intermediário e pode ser considerada coautora dos conteúdos gerados.

As instituições ressaltam que, mesmo à luz do Marco Civil da Internet, há um dever de cuidado reforçado por parte dos provedores, especialmente após decisões recentes do Supremo Tribunal Federal que relativizam a imunidade prevista no art. 19. Além disso, evidenciam que a conduta da empresa pode violar normas consumeristas ao ofertar serviço potencialmente perigoso e inconsistente com suas próprias políticas internas, caracterizando, inclusive, possível publicidade enganosa.

Diante desse contexto, a recomendação estabeleceu uma série de medidas obrigatórias à empresa, incluindo: (i) implementação imediata de mecanismos para impedir a geração de conteúdos sexualizados sem consentimento, sobretudo envolvendo menores; (ii) criação de procedimentos eficazes para identificação e remoção desses conteúdos já existentes na plataforma; (iii) aplicação rigorosa das políticas internas da empresa, com suspensão de contas infratoras; (iv) disponibilização de canais acessíveis para denúncias e exercício de direitos pelos titulares de dados; e (v) elaboração de um Relatório de Impacto à Proteção de Dados (RIPD) detalhado sobre as atividades do Grok.

Por fim, no dia 11 de fevereiro de 2026, as instituições envolvidas na Recomendação Conjunta avaliaram como insuficientes as providências informadas, e a ANPD

expediu medida preventiva com fundamento nos arts. 32, III, e 35 do Regulamento de Fiscalização, de modo a impedir que a ferramenta de inteligência artificial Grok gere conteúdos que representem crianças e adolescentes ou pessoas maiores de idade identificadas ou identificáveis em contextos sexualizados ou erotizados, sem autorização. O objetivo central da medida era prevenir violações de direitos fundamentais e assegurar a proteção efetiva de usuários brasileiros diante dos riscos associados ao uso de inteligência artificial generativa na plataforma.

Os casos brasileiros evidenciam que os *deepfakes* pornográficos constituem uma forma emergente de violência digital, com forte recorte de gênero e marcada pela facilidade tecnológica, pela vulnerabilidade de adolescentes e pela insuficiência de mecanismos de prevenção e de resposta. A recorrência desses episódios, especialmente em ambientes escolares, sugere que o fenômeno deve ser tratado como uma tendência estruturante do uso indevido de inteligência artificial, com implicações diretas para políticas públicas, regulação e proteção de dados pessoais.

5.2 *Deepfakes* e racismo

É importante destacar que os impactos dos *deepfakes* não se restringem à dimensão sexual, uma vez que essas tecnologias também têm sido empregadas para reproduzir e amplificar outras formas históricas de discriminação e violência, incluindo o racismo. Nesse contexto, a manipulação audiovisual a partir da inteligência artificial passa a integrar repertórios já conhecidos de estigmatização, ridicularização e desinformação direcionados a grupos racializados.
Quem nunca recebeu pelo celular ou se deparou nas redes sociais com vídeos aparentemente autênticos em que figuras públicas negras são retratadas de forma caricatural, animali-

zada ou ofensiva? Casos envolvendo conteúdos manipulados atribuídos à ex-primeira-dama dos Estados Unidos Michelle Obama, à atriz e duquesa de Sussex Meghan Markle ou ao jogador brasileiro Vinícius Júnior ilustram como a tecnologia de *deepfake* pode ser utilizada para associar indivíduos negros a comportamentos jocosos, discursos inadequados ou situações vexatórias que jamais ocorreram. Embora tais conteúdos sejam frequentemente apresentados como piadas ou entretenimento, eles reproduzem padrões históricos de representação racista e reforçam estereótipos socialmente prejudiciais.

Ao longo da história, a raça tem sido utilizada como um marcador de desigualdade, opressão e violência, operando não como um fato biológico ou genético, mas como uma ficção ideológica e um dispositivo sociopolítico historicamente fabricado para codificar a divisão humana, criar hierarquias e autorizar a exclusão (Mbembe, 2014, p. 71). Entre as diversas formas de manifestação do racismo, destaca-se o uso do humor como instrumento de desumanização e inferiorização de pessoas negras. Se antes essa prática encontrava espaço em programas de televisão, espetáculos humorísticos e outras produções midiáticas tradicionais, atualmente ela se manifesta também por meio de memes, *gifs*, vídeos curtos e publicações virais nas plataformas digitais. Em muitos casos, pessoas negras e suas características físicas, culturais ou comportamentais são reduzidas a representações ofensivas, pejorativas, caricaturais e até hipersexualizadas.

Como observa Dantas (2025, p. 30), essa dinâmica não é propriamente nova, mas foi consideravelmente transformada pelas tecnologias digitais. Segundo a autora, "a diferença é que, agora, esse tipo de humor está a um clique de distância, viraliza em minutos e ganha camadas tecnológicas cada vez mais sofisticadas" (Dantas, 2025, p. 30). Os *deepfakes* representam justamente uma dessas novas camadas tecnológicas, permitindo que conteúdos discriminatórios sejam produzidos com maior realismo, capacidade de engajamento e potencial de disseminação.

Nesse sentido, os *deepfakes* podem ser compreendidos como mais uma ferramenta a serviço do chamado racismo recreativo. Conforme destaca Dantas (2025, p. 30), "o racismo recreativo, termo cunhado por Adilson Moreira para descrever a prática de desumanizar pessoas negras sob o pretexto de fazer humor, não é uma novidade. Ele apenas ganhou novos cenários e ferramentas" (Dantas, 2025, p. 30). A tecnologia, portanto, remodela as formas de execução do preconceito e oferece novos meios para sua reprodução, circulação e amplificação em larga escala.

Esse fenômeno também se conecta a uma prática cada vez mais comum nas plataformas digitais: o chamado *rage bait*. O termo:

> *Surgiu no discurso informal da internet (Scott-Railton, 2022) para descrever a prática manipulativa de provocar indignação e reações agressivas com o objetivo de aumentar o tráfego de um perfil, o número de seguidores, o posicionamento nos algoritmos das plataformas e, em alguns casos, a receita financeira obtida online (Cover, 2025, p. 3).*

Diferentemente de manifestações espontâneas de opinião, o *rage bait* é planejado estrategicamente para gerar reações emocionais intensas e maximizar o engajamento.

Uma das formas mais comuns dessa estratégia consiste justamente na publicação de conteúdos racistas ou hostis contra grupos minorizados[16]. Conforme observa Cover (2025, p. 3), o *rage bait* opera por meio da provocação deliberada da raiva e da indignação, dinâmica que pode ser potencializada pelo uso da inteligência artificial na produção automatizada de conteúdos. Nesses casos, conteúdos discriminatórios são utilizados como gatilhos emocionais para transformar preconceitos sociais em mecanismos de engajamento digital, ampliando sua circulação por meio das reações que provocam.

Para corroborar ainda mais com essa reflexão, um estudo conduzido por Kolluri *et al.* (2025, p. 8) sobre a dinâmica

16 *"Minorizado" é um termo que tem sido utilizado na literatura contemporânea para designar coletividades que, embora nem sempre constituam minorias numéricas, ocupam posições de desvantagem social, política ou econômica em razão de processos históricos de exclusão e sub-representação. Diferentemente da noção de "minoria", que pode sugerir um critério quantitativo, a expressão "minorizado" destaca as relações de poder que produzem e mantêm situações de vulnerabilidade e marginalização social.*

de disseminação do discurso de ódio na rede social *Parler* identificou que conteúdos direcionados contra pessoas negras, juntamente com conteúdos antissemitas, apresentaram os maiores níveis de propagação de mensagens de ódio na plataforma. Os resultados demonstraram que comentários publicados em postagens de conteúdo antinegro tinham 227 vezes mais probabilidade de conter discurso de ódio racial quando comparados a outras interações analisadas, conforme Kolluri *et al.* (2025, p. 8). Esses dados sugerem que conteúdos racistas voltados à população negra possuem maior capacidade de atrair engajamento e estimular novas manifestações discriminatórias, contribuindo para a rápida amplificação da hostilidade e da violência simbólica nos ambientes digitais. Dessa forma, o estudo apresenta como o racismo on-line se difunde com facilidade nas redes sociais e tende a gerar ciclos de reforço e intensificação do discurso de ódio.

Embora a literatura acadêmica ainda apresente relativamente poucos estudos que examinem especificamente a relação entre *deepfakes*, *rage bait* e discursos racistas, as evidências disponíveis indicam uma convergência importante entre esses fenômenos. Como apontam Godulla *et al.* (2021, p. 76, tradução nossa), "o uso malicioso de *deepfakes* está fortemente associado a campanhas automatizadas de propaganda e desinformação, bem como à disseminação de informações falsas e teorias da conspiração por meio das redes sociais". Dessa forma, tecnologias sintéticas, quando usadas de maneira maliciosa, contribuem para ampliar narrativas discriminatórias já existentes, conferindo-lhes aparência de autenticidade e maior capacidade de persuasão.

Além da instrumentalização de preconceitos raciais, essas mesmas estratégias podem ser mobilizadas em disputas políticas e ideológicas. Vídeos manipulados por inteligência artificial podem atribuir declarações falsas a candidatas e candidatos, lideranças políticas ou figuras públicas, a fim de ampliar a polarização, gerar tensões ideológicas e explorar debates e conflitos entre os usuários.

5.3 *Deepfakes* no contexto eleitoral

O uso de *deepfakes* representa um dos principais desafios contemporâneos para a integridade do debate público e para a confiança nas instituições democráticas no contexto eleitoral. A possibilidade de produzir vídeos ou áudios falsos envolvendo candidatas e candidatos, autoridades, agentes ou lideranças políticas, amplia o potencial de campanhas de desinformação altamente persuasivas, capazes de influenciar percepções, reforçar polarizações e comprometer a tomada de decisão do eleitorado, especialmente em períodos de intensa circulação de informações.

Além da falsificação de declarações ou acontecimentos, a inteligência artificial, por meio de *deepfakes*, tem sido utilizada para criar personagens inteiramente sintéticos que simulam cidadãos comuns, influenciadores ou comentaristas políticos. Dois exemplos que ganharam grande destaque são os perfis "Dona Maria" e "Sô Zé da Feira" (Figuras 12 e 13), personagens gerados integralmente por IA que aparentam ser pessoas reais. Carregando características de pessoas de mais de 60 anos e com históricos criados de vida simples, os personagens usam linguagem coloquial para comentar acontecimentos políticos e compartilhar opiniões sobre temas de interesse público. Construídos a partir de elementos facilmente reconhecíveis do cotidiano brasileiro (como a figura da aposentada, da dona de casa ou do trabalhador da feira), esses avatares exploram características de pessoas comuns para despertar identificação, proximidade e confiança junto ao público.

Figura 12 *Personagem "Dona Maria", criada por inteligência artificial*

Fonte: Reprodução da internet.

A estratégia traz luz a uma mudança importante na forma como a desinformação pode ser produzida e disseminada. Em vez de recorrer apenas à manipulação de imagens de figuras públicas, conteúdos sintéticos passam a construir falsos porta-vozes da sociedade, criando a impressão de que determinadas opiniões representam manifestações espontâneas de cidadãos comuns. Quando esses personagens não informam de forma objetiva e transparente que foram gerados por inteligência artificial, muitas pessoas passam a acreditar que estão assistindo a depoimentos autênticos, o que aumenta o poder de convencimento das mensagens e dificulta a identificação de que se trata de um conteúdo artificial. Esse mecanismo interfere diretamente no exercício da cidadania, pois pode influenciar a formação de opiniões e decisões políticas a partir de uma percepção artificialmente construída sobre aquilo que a sociedade supostamente pensa ou defende.

Łabuz e Nehring (2024) destacam a preocupação com o uso e com os avanços da inteligência artificial generativa no cenário eleitoral, e do quanto isso pode minar a confiança nos processos democráticos. Ainda que a ideia central desses autores não seja exatamente sobre *deepfake*, é possível estender a preocupação para esse contexto.

Esse cenário torna o enfrentamento à desinformação mais complexo. Diferentemente das *fakenews* tradicionais, os *deepfakes* e demais conteúdos sintéticos exploram recursos audiovisuais que dificultam sua contestação mesmo quando posteriormente desmentidos. Em um ambiente digital marcado pela velocidade do compartilhamento e pelo consumo fragmentado de informação, conteúdos manipulados podem alcançar milhões de pessoas antes que sejam identificados, corrigidos ou removidos.

Figura 13 *Personagem "Sô Zé da Feira", criado por inteligência artificial*

Fonte: Reprodução da internet.

O impacto dessa estratégia é potencializado pela falta de transparência e pela falta de controle. Um levantamento do Observatório IA nas Eleições[17] identificou que 61% dos perfis políticos criados por inteligência artificial nas redes sociais não informam aos usuários que são produzidos por IA, apesar das regras estabelecidas pela Justiça Eleitoral para a identificação desse tipo de conteúdo. O estudo também verificou que mais da metade das publicações analisadas continha algum tipo de desinformação, revelando que esses personagens sintéticos ocultam sua natureza artificial e são frequentemente utilizados para impulsionar narrativas enganosas.

Diante desse contexto, o debate sobre *deepfakes* nas eleições exige um equilíbrio entre a proteção da liberdade de expressão, a promoção da inovação tecnológica e, sobretudo, a preservação da integridade do processo democrático. O principal desafio, nesse caso, está na garantia de transparência quanto ao uso de inteligência artificial na produção de conteúdo, para que os cidadãos possam exercer seu direito ao voto de maneira consciente, crítica e informada.

No Brasil, esse risco não tem passado despercebido pelo TSE, que vem adotando uma postura proativa para enfrentar esse desafio e se adaptar ao novo contexto digital. Em resoluções recentes sobre propaganda eleitoral e combate à desinformação, o Tribunal passou a abordar explicitamente conteúdos gerados ou manipulados por inteligência artificial, reconhe-

17 *O Observatório IA nas Eleições é um portal de referência para pesquisadores e especialistas sobre uso de IA generativa para manipulação de informações que possam interferir nos processos eleitorais e na integridade da democracia.*

cendo seus impactos no processo eleitoral. Para além do aspecto jurídico, o TSE também tem investido na cooperação com plataformas digitais e na realização de campanhas de conscientização voltadas à sociedade. A atuação normativa do Tribunal será detalhada na seção 6.1, que trata da abordagem legal e regulatória dos *deepfakes*.

5.4 *Deepfakes* como ferramenta de violência política de gênero

O uso de *deepfakes* em contextos políticos tem introduzido uma nova camada de complexidade à já conhecida violência política de gênero. Se historicamente mulheres em posições de visibilidade pública enfrentam ataques à sua honra, reputação e credibilidade, a introdução de tecnologias de manipulação audiovisual amplia significativamente o alcance, a sofisticação e o potencial de dano dessas agressões (Azevedo; 2025, p. 85). A inovação aqui reside na capacidade que essas ferramentas têm de produzir conteúdos altamente realistas, em larga escala e com rápida disseminação.

Segundo a Lei nº 14.192/2021, a violência política de gênero pode ser caracterizada como toda ação, conduta ou omissão com a finalidade de impedir, obstaculizar ou restringir os direitos políticos da mulher, seja ela candidata ou eleita (Brasil, 2021). No que diz respeito ao espaço virtual, podemos citar como exemplos a divulgação de fotos ou vídeos inverídicos durante a campanha eleitoral ou o mandato eletivo e difamação da candidata, atribuindo a ela fato que seja ofensivo. Estudos recentes indicam que os *deepfakes* constituem uma extensão direta das dinâmicas de violência de gênero no ambiente digital, sendo empregados estrategicamente para deslegitimar a participação política feminina e silenciar vozes dissidentes. Relatório da ONU Mulheres aponta que ataques coordenados envolvendo inteligência artificial são frequentemente direcionados a mulheres em posições públicas, com o objetivo de minar sua credibilidade profissional e reputação

pessoal (ONU News, 2026). Nesse contexto, a manipulação de imagens e vozes é tanto um mecanismo de desinformação quanto um instrumento de coerção simbólica e de exclusão política (Azevedo; 2025, p. 106).

A literatura e os dados empíricos identificam três padrões principais no uso de *deepfakes* contra mulheres na política: i) sexualização e pornografia não consentidas, com a criação de imagens íntimas falsas para gerar humilhação pública; ii) simulação de falas ou comportamentos, por meio de áudios e vídeos manipulados, a fim de provocar escândalos artificiais; e iii) campanhas de desinformação coordenadas que utilizam conteúdos sintéticos para influenciar as percepções do eleitorado.

Esses padrões frequentemente se combinam, reforçando estereótipos de gênero, explorando vulnerabilidades específicas associadas à presença feminina na esfera pública e corroborando a narrativa de que espaços de poder público "não deveriam" ser ocupados por mulheres. A própria ONU destaca que tais práticas produzem efeitos concretos de retração democrática: cerca de 41% das mulheres em posições públicas relatam autocensura para evitar novos ataques, enquanto uma parcela significativa reduz sua atuação profissional ou política (ONU Mulheres, 2026).

No Brasil, embora ainda não haja sistematização jurisprudencial específica sobre *deepfakes* na violência política de gênero, episódios concretos evidenciam o fenômeno. Um caso relevante ocorreu durante o processo eleitoral municipal de 2024, quando a candidata à prefeitura de Ipojuca (PE), Adilma Lacerda, foi alvo de um vídeo manipulado em que sua voz foi artificialmente alterada para simular ofensas a eleitores (Souza, 2024). O conteúdo, amplamente disseminado em aplicativos de mensagens, atribuía à candidata declarações depreciativas, com claro potencial de prejudicar sua imagem pública e comprometer sua campanha. A investigação policial identificou o uso de tecnologia de manipulação de áudio,

caracterizando o episódio como uso ilícito de inteligência artificial para fins eleitorais.

Casos como esse demonstram que a violência política mediada por *deepfakes* não se limita à disseminação de desinformação, mas também incorpora elementos de violência simbólica, sexualização e humilhação pública, afetando diretamente a participação feminina na política. Outro exemplo relevante disso envolve a circulação de imagens íntimas falsas de mulheres ocupantes de cargos públicos. Em 2024, a Prefeita de Bauru (SP) se deparou com a veiculação de imagem *deepfake* com seu rosto sobre o corpo de uma mulher nua, o que fez com que ela procurasse as autoridades policiais para denunciar o uso indevido de sua imagem (Piassi, 2024). Em outra situação, a deputada Tabata Amaral foi associada a duas imagens manipuladas de uma criadora de conteúdo adulto, fazendo com que a até então deputada e candidata à prefeitura de São Paulo tomasse medidas legais contra os autores das postagens (Folha de S. Paulo, 2024).

Nesse cenário, a atuação institucional, incluindo a da ANPD, revela-se essencial para a mitigação dessas práticas. A articulação entre proteção de dados, regulação eleitoral e responsabilização jurídica tende a se tornar cada vez mais necessária diante do caráter transversal dessas violações.

5.5 *Deepfakes* para aplicação de golpes financeiros

Nos últimos anos, o uso de *deepfakes* passou a integrar um tipo mais sofisticado de fraude digital no Brasil, especialmente em golpes financeiros. Diferentemente dos tradicionais esquemas de *phishing*[18] ou de engenharia social baseados em texto e envio de *link*s, por exemplo, esses golpes, embora baseados em dinâmicas de fraude já conhecidas, adquirem nova sofisticação com o uso de inteligência artificial, especialmente em esquemas que exploram con-

18 Phishing *é um tipo de golpe digital em que criminosos tentam enganar pessoas para obter informações sensíveis, como senhas, dados bancários ou números de documentos, geralmente se passando por instituições confiáveis (como bancos, órgãos públicos ou empresas) por meio de e mails, mensagens ou* sites *falsos (ANPD, 2024).*

fiança, urgência e autoridade, adicionando imagem e voz. Essas situações são muito bem planejadas para construir uma aparência de legitimidade bem convincente. Há registros recentes de vídeos manipulados circulando em redes sociais e aplicativos de mensagens, nos quais supostos especialistas ou pessoas públicas, políticos e até mesmo celebridades recomendam investimentos falsos. Em alguns casos, até mesmo a identidade visual de empresas conhecidas é replicada para aumentar a credibilidade da fraude.

A seguir, apresentaremos três casos brasileiros de golpes financeiros que ganharam bastante repercussão no país nos últimos dois anos:

O caso brasileiro envolvendo o uso de *deepfake* com a imagem de Brad Pitt ganhou projeção a partir de um episódio ocorrido no Rio Grande do Sul, no fim de 2025, quando uma mulher afirmou manter um relacionamento virtual com o ator e chegou a aguardar sua chegada em um aeroporto local (Henrique, 2026). Segundo relatos, ela dizia ter realizado diversas videochamadas com o suposto Brad Pitt e acreditava que o encontro presencial culminaria em um relacionamento formal. A convicção da vítima chamou a atenção das autoridades, especialmente porque ela descartava a hipótese de fraude com base na experiência visual proporcionada pelas chamadas, por conta das tecnologias de manipulação que aconteciam em tempo real.

Nos desdobramentos legais, o caso passou a ser analisado pela Polícia Civil, que buscou verificar se houve efetivamente estelionato sentimental (modalidade de fraude em que há exploração emocional para obtenção de vantagem econômica) ou se a situação envolvia outros ilícitos, como possível comunicação falsa de crime. Posteriormente, a mulher alterou sua versão, alegando se tratar de uma "brincadeira", o que gerou dúvidas quanto à eventual existência de prejuízo financeiro ocultado ou à veracidade inicial dos fatos. Diante dessas inconsistências, as autoridades mantiveram a investigação

aberta para apurar tanto a eventual atuação de terceiros (golpistas) quanto a conduta da própria envolvida, à luz do Código Penal brasileiro.

A mesma estratégia de golpe foi usada em outros países. Na França, por exemplo, uma mulher perdeu mais de € 830 mil (equivalente a pouco mais de R$ 5 milhões) (Mulher..., 2025). Por conta disso, o próprio Brad Pitt acionou um porta-voz para se pronunciar e lamentar o caso, alertando os fãs sobre quadrilhas on-line que utilizavam sua imagem sem consentimento.

Esse episódio se insere em um contexto mais amplo de crescimento acelerado de fraudes com *deepfake* no Brasil, especialmente após a popularização de ferramentas de IA generativa, conforme temos visto aqui. De acordo com o *Identity Fraud Report* 2025–2026, da empresa de verificação de identidade *Sumsub*, o Brasil concentrou cerca de 39% dos casos de *deepfake* detectados na América Latina em 2025, além de registrar um aumento de 126% nos ataques envolvendo *deepfake* e identidades sintéticas entre 2024 e 2025. Esses dados reforçam a expansão quantitativa dessas práticas e a sofisticação qualitativa, marcada também pelo uso de narrativas afetivas e pela simulação de interações "reais", o que intensifica riscos jurídicos, desafios regulatórios e vulnerabilidades no contexto da proteção de dados e da segurança digital.

Outro caso de *deepfake* envolvendo golpe financeiro é o vídeo fraudulento que utilizava a imagem e a fala de Pedro Moreira Salles, copresidente do Conselho de Administração do Banco Itaú (Itaú, 2025), amplamente divulgado no Facebook em 2025. O conteúdo atribuía declarações falsas ao executivo com o objetivo de convencer usuários a aderirem a um suposto investimento "milagroso". Para reforçar a credibilidade, o golpe direcionava as vítimas a um *link* que imitava o site de um veículo de imprensa, criando uma aparência de legitimidade. O Itaú esclareceu que o executivo não tinha qualquer relação com a publicação e que o material

apresentava sinais típicos de fraude, como ausência de fontes confiáveis e promessas irreais de retorno financeiro. A Figura 14 ilustra o conteúdo que foi compartilhado à época, a fim de trazer maior compreensão ao caso.

Figura 14 *Imagem ilustrando o caso de golpe financeiro com deepfake de representante de um grande banco*

Fonte: Elaboração própria, desenvolvida de forma interativa com a ferramenta Gemini Flash 3.5.

Do ponto de vista jurídico, embora o caso específico seja tratado sobretudo como alerta preventivo (sem detalhamento público de processos judiciais decorrentes), ele se enquadra em categorias já consolidadas no direito brasileiro que veremos mais detalhadamente à frente. Já no plano regulatório mais amplo, o episódio ilustra um problema crescente que tem motivado respostas institucionais e discussões legislativas para obrigar a identificação de conteúdos sintéticos (como projetos de lei específicos sobre *deepfakes*), bem como o aumento da pressão sobre plataformas digitais para moderar anúncios fraudulentos e reduzir incentivos econômicos à desinformação.

O terceiro caso de golpe financeiro envolve a Serasa Experian (Domingos, 2024), empresa brasileira especializada em análise de dados para decisões de crédito. Na prática, ela opera como um grande banco de dados que reúne o histórico

financeiro dos consumidores a partir de movimentações vinculadas ao CPF, auxiliando empresas na avaliação do risco de inadimplência antes da concessão de empréstimos, financiamentos ou crediário.

De acordo com dados de maio de 2026 do Mapa da Inadimplência e Negociação de Dívidas do Serasa (Mapa..., 2026), o Brasil possui cerca de 83,5 milhões de pessoas com o nome negativado, o que corresponde a aproximadamente metade da população adulta do país. Além disso, estima-se que 57 milhões de consumidores possuam dívidas e não saibam, pois não acompanham regularmente seu CPF. Diante desse cenário de alto nível de endividamento e o desejo de regularizar a situação financeira, esse público se torna especialmente vulnerável a golpes que exploram o nome da empresa.

Nos últimos dois anos, circularam, amplamente, conteúdos falsos que prometiam supostas indenizações por um vazamento de dados, bem como ofertas enganosas de quitação de dívidas mediante pagamento de taxas reduzidas ou mesmo repasse de crédito. A gravidade do caso foi ampliada pelo uso de *deepfakes*, com a criação de vídeos falsos de figuras públicas, como o então senador Sérgio Moro (Amado; Campos, 2024), que posteriormente veio a público esclarecer a falsidade do conteúdo. Os materiais também reproduziam elementos visuais de telejornais, para conferir maior credibilidade à fraude.

Os vídeos afirmavam que haveria o pagamento de indenizações de até R$ 30 mil para pessoas supostamente afetadas por terem dados expostos, como números de CPF, nomes completos, endereços, telefones, e até informações salariais, benefícios sociais e dados do imposto de renda. Ao final, o usuário era incentivado a clicar em um *link* para checar uma lista e verificar se teria direito ao benefício. Para dar aparência de legitimidade, os golpistas utilizaram uma reportagem real de 2021 sobre o vazamento de dados de aproximadamente 223 milhões de pessoas, à época associado a base da Serasa Experian. No

entanto, até o momento desta publicação não houve condenação da empresa ao pagamento de qualquer indenização.

Em 2026, o referido megavazamento de dados ocorrido em 2021 voltou à tona e ganhou novos desdobramentos no campo jurídico. O caso passou a ser objeto de uma ação coletiva na justiça inglesa contra o grupo Experian, ao qual a Serasa está vinculada. A iniciativa foi adotada porque a controladora da empresa tem sede no Reino Unido, o que permite esse tipo de demanda. No Brasil, o tema também segue em discussão judicial: o Ministério Público Federal (MPF) atua como coautor de uma ação civil pública que busca a responsabilização e uma eventual indenização pelos danos causados. Todavia, até o momento ainda não há comprovação técnica definitiva sobre a origem do vazamento, nem decisão judicial que tenha determinado o pagamento de indenizações aos consumidores.

Outro formato de golpe que vem se intensificando envolve a clonagem de voz ou imagem de pessoas que realmente fazem parte do convívio da vítima, como o chefe, a diretora da empresa em que ela trabalha, seu familiar ou alguém próximo, para induzir a realização de pagamentos, transferências bancárias ou compartilhamento de dados sensíveis. Geralmente essas solicitações vêm atreladas a um senso de urgência, em que o contexto de pressão e imediatismo se sobrepõe ao discernimento e à prudência, fazendo com que a vítima não tenha tempo de refletir criticamente sobre o que está sendo solicitado e se torne mais suscetível a "cair" no golpe.

Nesse contexto, o *deepfake* atua como um catalisador de fraudes já conhecidas, elevando o nível de aprimoramento e exigindo novos mecanismos de verificação, tanto técnicos quanto comportamentais, para reduzir riscos. Em um cenário em que já há uma dificuldade de distinguir o que é autêntico do que é sintético, tanto em razão da sofisticação das tecnologias aplicadas à criação de *deepfakes* quanto da redução drástica dos sinais tradicionais de alerta que o contexto de urgência

ou de suposta ajuda familiar provoca, o risco para indivíduos e instituições é consideravelmente ampliado, especialmente em ambientes nos quais decisões rápidas são valorizadas.

5.6 *Deepfakes* para propaganda enganosa

O uso de *deepfakes* tem se expandido rapidamente, também, no campo da publicidade digital, configurando uma nova modalidade de propaganda enganosa altamente sofisticada, baseada na simulação realista de identidades, vozes e reputações. Diferentemente de formas tradicionais de fraude, os conteúdos gerados por inteligência artificial permitem a criação de anúncios visualmente plausíveis e emocionalmente persuasivos, o que dificulta a distinção entre publicidade legítima e conteúdo fraudulento.

No Brasil, a disseminação desses anúncios tem ocorrido principalmente em plataformas digitais, como redes sociais e aplicativos de mensagens, nas quais conteúdos patrocinados são direcionados com base no perfil do usuário. Nesse contexto, *deepfakes* são utilizados para simular o endosso de produtos e serviços por figuras públicas – médicos, jornalistas, artistas – cuja credibilidade atua como vetor de confiança para o consumidor.

É possível identificar um padrão de operação nesses casos. Em geral, há três etapas: i) criação de conteúdo sintético utilizando a imagem e a voz de figuras públicas; ii) impulsionamento do conteúdo nas redes sociais; e iii) redirecionamento das vítimas para ambientes fraudulentos, como *sites* clonados ou plataformas falsas, onde são solicitados pagamentos ou dados pessoais.

Assim, diferentes técnicas de *marketing* digital e de manipulação psicológica são associadas ao intuito de gerar confiança e induzir o comportamento do consumidor. Como exemplo disso, podemos citar produtos com benefícios imediatos, descontos, brindes ou curas milagrosas.

No Brasil, há múltiplos registros de uso de *deepfakes* em propaganda enganosa envolvendo celebridades e especialistas. Um dos casos mais relevantes envolveu uma organização criminosa que utilizou *deepfakes* de celebridades como Gisele Bündchen, Angélica, Juliette e Sabrina Sato para divulgar produtos inexistentes. No caso de Gisele Bündchen, por exemplo, o grupo criminoso vendia um "kit antirrugas grátis" cujo pagamento do frete era feito via PIX, mas o produto não era entregue. A investigação policial apontou uma movimentação de mais de R$ 20 milhões da quadrilha, o que resultou na prisão de seus integrantes (Souza, Figueiredo, 2025).

Outro vetor recorrente é o uso de imagens de profissionais de saúde em campanhas fraudulentas. O médico Drauzio Varella, por exemplo, teve sua imagem e voz repetidamente manipuladas para simular recomendações de medicamentos e tratamentos sem qualquer respaldo científico (Fantástico, 2025). Esses conteúdos, muitas vezes apresentados como entrevistas ou reportagens jornalísticas, exploram a autoridade do especialista para induzir consumidores a adquirir produtos potencialmente nocivos, como ilustrado pela figura a seguir.

Figura 15
Representação fictícia de deepfake de propaganda enganosa (à esquerda), utilizando indevidamente a imagem publicada por um profissional de saúde (à direita)

Fonte: Elaboração própria a partir da ferramenta Gemini Flash 3.5.

A imagem apresenta dois vídeos, em que o da esquerda (identificado como a conta @dr.mauro_corella_tv) ilustra um exemplo de *deepfake*. Nessa representação fraudulenta, a imagem do profissional de saúde é utilizada sem autorização para promover falsas promessas de um "suplemento milagroso", um tipo de conteúdo perigoso do ponto de vista sanitário e enganoso sob a perspectiva da publicidade. O painel da direita, por outro lado, representa o que seria o vídeo autêntico do mesmo profissional (da conta oficial e legítima), abordando um tema de saúde pública legítimo, que poderia ser utilizado como base para a criação da mídia manipulada.

A principal característica desses golpes é a escala massiva, associada a perdas individuais reduzidas. Em muitos casos, cada vítima perde valores relativamente baixos, o que reduz a probabilidade de denúncia e de investigação, permitindo que os criminosos operem com elevada lucratividade agregada (Romagna, 2025).

Além do impacto econômico direto, esses esquemas produzem efeitos relevantes em outras dimensões, como: i) risco à saúde pública, especialmente quando envolve produtos médicos ou terapias não comprovadas; ii) comprometimento da credibilidade de profissionais e instituições; iii) violação de direitos de personalidade, em especial o uso indevido de imagem e voz; e iv) exposição de dados pessoais, uma vez que as vítimas frequentemente fornecem informações em *sites* fraudulentos.

Portanto, o uso de *deepfakes* em propaganda enganosa representa uma convergência entre desinformação, fraude econômica e o uso indevido de dados pessoais. A inovação tecnológica introduzida pela inteligência artificial amplia significativamente a eficácia desses golpes, ao permitir a simulação convincente de figuras de autoridade e de confiança.

‹ 6 perspectivas de futuro ›

A evolução de tecnologias para a síntese e a manipulação de mídias visuais e auditivas, que constituem o ecossistema de *deepfakes*, inspira preocupações contemporâneas e projeta desafios de ordem técnica, regulatória e social para o futuro do ambiente digital. Para enfrentar esses desafios, é fundamental observar as soluções regulatórias emergentes especificamente ligadas aos *deepfakes* no Brasil e em outras jurisdições, verificando os conceitos de *deepfake* mobilizados por diferentes ordenamentos jurídicos e o escopo de limitações ou vedações relativas à geração e à disseminação desses conteúdos em cada país.

Adicionalmente, em razão de sua complexidade e crescente sofisticação, a governança dessas ferramentas demanda abordagens que transcendam o campo estritamente normativo e incorporem a promoção de competências informacionais e midiáticas entre usuários e o desenvolvimento de mecanismos técnicos de mitigação de riscos. Essas estratégias se tornam particularmente relevantes diante das dificuldades discutidas ao longo deste Radar Tecnológico, seja com relação ao controle de conteúdos sintéticos já criados e circulantes, seja com relação à natureza distributiva e transnacional de infraestruturas digitais.

Nesse contexto, esta seção discute, de maneira exploratória, perspectivas de futuro relacionadas aos *deepfakes*, as quais estão organizadas em três eixos: i) o tratamento legal e regulatório dessas tecnologias no Brasil e em outros países e jurisdições; ii) a relevância do letramento digital como elemento de resiliência social perante os danos ligados a usos indevidos de *deepfakes*; e iii) o desenvolvimento de ferramentas aptas a mitigar os riscos associados a esse ecossistema.

A análise a seguir não pretende esgotar essas discussões, mas oferecer um panorama de questões emergentes que tendem a orientar o debate regulatório e informacional dos próximos anos. Passemos a elas.

6.1 Perspectiva legal e regulatória

As seções anteriores discutiram os riscos decorrentes da geração e da disseminação de *deepfakes* para a proteção de dados pessoais e os danos que essas ferramentas podem ocasionar quando utilizadas para a geração de pornografia não consentida, para a aplicação de fraudes e golpes financeiros, para a disseminação de distorções factuais no contexto eleitoral e para a promoção da violência política de gênero, por exemplo. Por tudo isso, mostra-se evidente que o uso indevido dessas tecnologias ameaça uma série de direitos fundamentais, como o direito à privacidade, à imagem, à honra e à liberdade, além da dignidade da pessoa humana.

Diante desse cenário, autoridades ao redor do mundo têm buscado conter esses e outros riscos por meio da regulação específica de *deepfakes*. Esse esforço normativo, contudo, enfrenta importantes desafios, que são reconhecidos por reguladores de diferentes jurisdições. Alguns desses desafios são resumidos na Figura 16 a seguir e discutidos ao longo desta seção.

Figura 16 *Desafios regulatórios relacionados aos deepfakes*

Fonte: Elaboração própria.

A **ambiguidade conceitual** dos *deepfakes* impacta diretamente o desenho das normas que buscam regulá-los. De maneira simplificada, é possível distinguir duas formas de conceber *deepfakes*, às quais correspondem dois tipos de regulação.

De um lado, compreender *deepfakes* como ferramentas tecnológicas específicas implica constituir regulações de **caráter técnico**. Exemplo disso é o já citado projeto *DEEP FAKES Accountability Act*, apresentado em 2019 nos Estados Unidos, que concebia o *deepfake* como qualquer representação tecnológica substancialmente derivada da fala ou da conduta de uma pessoa. O foco dessa norma estava na adulteração de características humanas por meio de programas de computador, isto é, na aplicação de ferramentas tecnológicas avançadas para simular traços de determinada pessoa, situação diversa do que ocorre em uma imitação humana não mediada por tecnologias, por exemplo.

De outro lado, interpretar *deepfakes* a partir dos possíveis danos que eles podem ocasionar conduz à elaboração de regulações **baseadas no risco**. Essas abordagens deixam de lado os meios pelos quais os *deepfakes* são gerados e se atentam às finalidades para as quais eles são produzidos.

Essa perspectiva tem sido a preferida nas leis sobre *deepfakes* atualmente em vigor ao redor do mundo, como se observará adiante, porque ela permite enfrentar, ao mesmo tempo, dois importantes desafios de governança ligados a esse ecossistema. Em primeiro lugar, temos a **defasagem regulatória**, quer dizer, a lacuna naturalmente existente entre a demora exigida pelo processo legislativo e o acelerado avanço tecnológico por que passam essas ferramentas.

Em segundo lugar, temos as barreiras impostas pela **opacidade tecnológica**, que dificultam a identificação exata das tecnologias utilizadas para a geração de *deepfakes* e podem impedir a aplicabilidade de legislações destinadas apenas a

certos tipos de tecnologias – e não a outros. Nesse contexto, normas baseadas nas (más) aplicações de *deepfakes* deslocam a centralidade regulatória das tecnologias utilizadas para os riscos sociais a elas associados, tornando-se mais perenes e eficientes para mitigar tais ameaças a direitos.

No âmbito normativo centrado nos riscos, legislações internacionais têm seguido três grandes padrões normativos em relação aos *deepfakes*: i) determinações ligadas à **transparência**, que obrigam a sinalização de conteúdo sintético; ii) proibições relativas ao **contexto eleitoral**, que impõem vedações de usos de *deepfakes* em situações consideradas particularmente arriscadas do ponto de vista político; e iii) **criminalizações** de abusos, que preveem penas para a produção ou a disseminação de *deepfakes* com conteúdo sexual/pornográfico.

Exemplo do primeiro padrão, focado em **transparência**, é encontrado no *AI Act*, da União Europeia, que afirma, em seu art. 50, item 1, que "os fornecedores de sistemas de IA [...] que gerem conteúdos sintéticos de áudio, imagem, vídeo ou texto devem assegurar que os resultados [...] sejam marcados e detectáveis como tendo sido gerados ou manipulados artificialmente" (European Union, 2024, tradução nossa).

De modo complementar, o mesmo artigo determina, em seu item 3, que "os responsáveis pela utilização de um sistema de IA que gere ou manipule conteúdos de imagem, áudio ou vídeo constitutivos de *deepfake* devem divulgar que o conteúdo foi gerado ou manipulado artificialmente [...]" (European Union, 2024, tradução nossa).

Nos dispositivos europeus, observamos a centralidade conferida à informação de manipulação ou completa síntese de mídia artificial, que responde às preocupações ligadas ao potencial engano que *deepfakes* produzem na sociedade, em razão de seu alto nível de verossimilhança. De acordo com as normas anteriores, qualquer conteúdo de *deepfake*, independentemente de seu contexto de uso e circulação, deve vir

acompanhado de sinalização expressa sobre sua constituição – marcas d'água, textos persistentes na tela, avisos prévios à reprodução –, seja pelo fornecedor de sistema de IA, seja pelo responsável por sua geração.

Já a segunda abordagem normativa, marcada por vedações ligadas ao **contexto eleitoral**, tem sido observada em diferentes países do mundo. Exemplo disso é a Coreia do Sul, que prevê no artigo 82-8 de seu *Public Official Election Act* que "ninguém pode produzir, editar, distribuir, exibir ou publicar vídeos *deepfake* para fins de campanha eleitoral, contados de 90 dias antes até o próprio dia da eleição" (South Korea, 2024, tradução nossa).

No Brasil, a Justiça Eleitoral tem assumido protagonismo na regulação de *deepfakes*. Isso porque a disseminação de *deepfakes* enganosos ou inverídicos apresenta elevados riscos para a saúde democrática e para a adequada condução de campanhas eleitorais no país. Diante da sensibilidade da matéria, desde 2024 o TSE vem abordando expressamente hipóteses de limitação e vedação de uso dessas ferramentas no contexto eleitoral.

Assim, o Brasil previu, na Resolução nº 23.732/2024 do TSE – que altera a redação da Resolução nº 23.610/2019 do mesmo tribunal –, a proibição do uso, tanto para o favorecimento quanto para o prejuízo de candidaturas, de "conteúdo sintético em formato de áudio, vídeo ou combinação de ambos, que tenha sido gerado ou manipulado digitalmente, ainda que mediante autorização, para criar, substituir ou alterar imagem ou voz de pessoa viva, falecida ou fictícia (*deep fake*)" (Brasil, 2019). Observamos, portanto, a explícita nomeação de "*deepfakes*" para se referir a ferramentas de síntese e manipulação digital de mídias por parte da Justiça Eleitoral.

A mesma Resolução determina que "é vedada a utilização, na propaganda eleitoral, [...] de conteúdo fabricado ou manipulado para difundir fatos notoriamente inverídicos ou descontextua-

lizados com potencial para causar danos ao equilíbrio do pleito ou à integridade do processo eleitoral" (Brasil, 2019). Nesse caso, a legislação nacional se preocupou com os riscos associados aos *deepfakes* na disseminação de desinformação, que se tornam especialmente graves em período eleitoral, na medida em que tais conteúdos podem influenciar o voto dos cidadãos.

Vale ressaltar que os padrões normativos propostos nem sempre aparecem de forma isolada na prática. Proibições eleitorais como as expostas acima frequentemente são associadas a determinações relacionadas à transparência, por exemplo. Esse é o caso de outro dispositivo da mesma Resolução, que exige a sinalização de fabricação ou manipulação de conteúdo em propagandas eleitorais:

> *a utilização na propaganda eleitoral, em qualquer modalidade, de conteúdo sintético multimídia gerado por meio de inteligência artificial ou tecnologia equivalente para criar, substituir, omitir, mesclar ou alterar a velocidade ou sobrepor imagens ou sons, impõe ao responsável pela propaganda o dever de informar, de modo explícito, destacado e acessível, que o conteúdo foi fabricado ou manipulado e qual tecnologia foi utilizada (Brasil, 2019).*

Como se lê do trecho anterior, a legislação eleitoral brasileira admite o uso de inteligência artificial para a realização de alterações de conteúdos multimídia que se enquadram no conceito de *cheapfakes* ou *shallowfakes* discutidos neste Radar Tecnológico. Entretanto, mesmo nas hipóteses que envolvem mídias com alterações mais facilmente identificáveis pelo usuário comum de internet, exige-se a informação explícita de alteração de conteúdo, para garantir a ciência dos cidadãos a seu respeito.

Em 2026, o TSE publicou nova Resolução sobre propaganda eleitoral (Resolução nº 23.755/2026), na qual incluiu como regra vedação ainda mais rigorosa relativa a conteúdo manipulado ou sintetizado por inteligência artificial:

> *Ficam vedadas a publicação e a republicação, ainda que gratuitas, bem como o impulsionamento pago de novos conteúdos sintéticos produzidos ou alterados por inteligência artificial ou por tecnologias equivalentes que utilizem imagem, voz ou manifestação de candidata ou candidato ou de pessoa pública, mesmo que rotulados e em conformidade com as demais exigências deste artigo, no período compreendido entre as 72 (setenta e duas) horas que antecedem e as 24 (vinte e quatro) horas que sucedem o término do pleito (Brasil, 2019).*

Dessa maneira, a nova regra brasileira reconhece a criticidade do período imediatamente anterior e posterior ao pleito eleitoral, e proíbe qualquer circulação de *deepfakes* nessa janela temporal. Mais uma vez, busca-se evitar que mídias sintéticas ou manipuladas altamente verossímeis sejam disseminadas e, assim, conduzam eleitores a votarem a partir de informações enganosas.

O caso dos Estados Unidos também é ilustrativo dessa espécie de regulação. Marcado por um federalismo forte e descentralizado, o panorama de governança e de regulação de *deepfakes* estadunidense se desenvolveu de maneira fragmentada e heterogênea (Fukuha; Fürst; Archegas, 2026, p. 201–202). Na ausência de uma lei nacional unificada e abrangente regulando os *deepfakes*, diferentes estados criaram suas próprias normas para coibir usos indevidos dessas ferramentas.

Assim, entre 2019 e 2025, 28 dos 50 estados estadunidenses aprovaram leis para regular o uso de mídias sintéticas ou manipuladas em campanhas eleitorais. Estados como a Califórnia e o Texas despontaram na vanguarda dessas ações, aprovando pacotes de leis bastante rigorosos (Fukuha; Fürst; Archegas, 2026, p. 202; Kshetri; Voas, 2026). De modo geral, as regras aprovadas obrigavam provedores de internet e responsáveis por campanhas eleitorais a usarem ferramentas para detectar e aplicar rótulos claros nos arquivos artificiais (Kshetri; Voas, 2026), demonstrando, novamente, a associação

de padrões regulatórios relativos à transparência e à vedação de usos eleitorais indevidos de *deepfakes*.

Finalmente, quanto às abordagens centradas no terceiro uso, isto é, na **criminalização**, diferentes países do mundo têm se mobilizado para a penalização de *deepfakes* em casos de elevado dano social, geralmente relativos a ofensas sexuais.

A criminalização de condutas inspira preocupações particulares no contexto de *deepfakes*. Sendo ela uma ferramenta que pode tanto ser utilizada para finalidades indevidas quanto para objetivos lícitos, sua regulação deve ser sempre capaz de **equilibrar direitos fundamentais**. Nesse sentido, para a preservação da liberdade de pensamento e do incentivo à inovação tecnológica, ordenamentos jurídicos ao redor do mundo têm evitado penalizar atos preparatórios mediante uso de *deepfakes*, ou ainda a mera intenção de cometer ilicitudes por meio dessas mídias. Isso porque a antecipação de punições pode aumentar de maneira expressiva o risco de supercriminalização estatal e impor ameaças ao livre exercício da liberdade de expressão (Kirchengast, 2020, p. 6; 10).

Por outro lado, a ausência de previsão criminal expressa relacionada aos *deepfakes* pode impedir que condutas extremamente danosas sejam puníveis no âmbito penal. Na Coreia do Sul, por exemplo, antes das reformas de 2020 e 2024, a ausência de leis claras contra essas ferramentas levou à absolvição de um homem que distribuiu imagens íntimas falsas (Kshetri; Voas, 2026, p. 6). Um problema parecido ocorreu no México, onde a Justiça absolveu um homem que criou imagens sexuais falsas de mais de mil mulheres. O agressor ficou impune da fraude digital, porque não existia uma regra específica para condená-lo (Kshetri; Voas, 2026, p. 2).

Nesse contexto, ordenamentos jurídicos ao redor do globo têm sido alterados para incluir esse tipo específico de violência de gênero facilitada pelo uso de tecnologias em suas legislações penais. A Coreia do Sul é um exemplo claro desse

processo. De acordo com dados da Security Hero, 53% dos *deepfakes* pornográficos circulantes até 2023 retratava vítimas sul-coreanas, sendo que 8 das 10 pessoas que mais constavam em *deepfakes* pornográficos naquele período eram cantoras da Coreia do Sul (Security Hero, 2023).

Diante desse cenário, o país ampliou, em 2020, sua legislação criminal sobre crimes sexuais em ambientes digitais, prevendo no art. 14 do *Act on Special Cases Concerning the Punishment, etc. of Sexual Crimes* pena de prisão para a gravação de vídeo ou fotografia com teor sexual sem o consentimento da pessoa retratada, bem como para distribuição, venda, divulgação, posse e visualização desse material. Em 2024, o art. 14-2 da mesma legislação passou ainda a incluir proibições expressas relativas a edição, síntese, processamento e publicação de fotos, vídeos ou áudios que incluam o rosto, a voz ou o corpo de pessoas específicas com teor sexual (South Korea, 2025). A atuação legal buscou, portanto, punir de maneira direta a geração e a circulação de *deepfakes* de natureza pornográfica.

No Reino Unido, a criminalização de *deepfakes* pornográficos ocorreu gradualmente, por meio de adaptações de leis em vigor no país. Inicialmente, o *Online Safety Act* 2023 alterou o *Sexual Offences Act* 2003 para criminalizar o compartilhamento de imagens íntimas sem consentimento, incluindo imagens sintetizadas artificialmente (United Kingdom, 2003). Em seguida, o *Data (Use and Access) Act* 2025 previu a criminalização da criação e da solicitação de imagens íntimas sintetizadas sem consentimento (United Kingdom, 2025). Nas legislações britânicas, as proibições recaem sobre "imagens simuladas" ou "imagens que parecem ser reais" (*purported images*), o que revela a preocupação do país com a verossimilhança dessas mídias, e não tanto com o modo como são geradas.

No mesmo sentido, a França incluiu em seu Código Penal dispositivo direcionado à proibição de *deepfakes* pornográficos focando em sua aparência de autenticidade. O art. 226-8 da

norma francesa prevê punição de prisão para a disseminação de "montagem realizada com as falas ou a imagem de uma pessoa sem seu consentimento, caso não seja evidente que se trata de uma montagem ou caso isso não seja expressamente indicado" (France, 2026, tradução nossa). A mesma punição é aplicada à disseminação de conteúdo visual ou sonoro "gerado por processamento algorítmico e que represente a imagem ou as falas de uma pessoa, sem seu consentimento, caso não seja evidente que se trata de uma montagem ou caso isso não seja expressamente indicado" (France, 2026, tradução nossa).

Em 2024, foi a vez de a Austrália aprovar emenda ao seu Código Criminal explicitamente direcionada a material sexual de *deepfake*. A lei australiana prevê pena de prisão para quem se utiliza de serviço de comunicação para transmitir material de terceiros que retrate ou aparente retratar "a outra pessoa envolvida em uma pose ou atividade sexual; [...] ou [...] os órgãos sexuais ou a região anal da outra pessoa; ou [...] se a outra pessoa é mulher – seus seios" (Austrália, 2024, tradução nossa), sempre que o material tenha sido transmitido sem o conhecimento ou o consentimento da pessoa retratada. A emenda prevê ser irrelevante se o material é transmitido sem alterações ou se ele foi criado ou alterado mediante uso de tecnologia. Embora não conceitue o que é um *deepfake*, uma nota a esse artigo da emenda prevê que os materiais sobre os quais ele trata incluem:

> *imagens, vídeos ou áudios que representem uma pessoa e que tenham sido editados ou integralmente criados por meio de tecnologia digital (incluindo inteligência artificial), gerando uma representação realista, porém falsa, da pessoa. Exemplos desse tipo de material são os "deepfakes" (Austrália, 2024, tradução nossa).*

Após observar a multiplicação de leis sobre *deepfakes* em seus estados, os Estados Unidos também previram a criminalização da disseminação intencional de representações

visuais íntimas não consentidas por meio de seu *Take it Down Act,* de 2025. No texto da legislação estadunidense, de âmbito nacional, o *deepfake* é compreendido como "falsificação digital" (*digital forgery*), como se lê adiante:

> *O termo 'falsificação digital' significa qualquer representação visual íntima de um indivíduo identificável criada por meio do uso de software, aprendizado de máquina, inteligência artificial ou quaisquer outros meios tecnológicos ou computacionais, inclusive por meio da adaptação, modificação, manipulação ou alteração de uma representação visual autêntica, que, quando considerada em seu conjunto por uma pessoa razoável, seja indistinguível de uma representação visual autêntica do indivíduo (United States of America, 2025, p. 2–3, tradução nossa).*

A lei estadunidense prevê pena de prisão para a publicação de *deepfake* de pessoas adultas sem seu consentimento, para a publicação de *deepfake* de menores de idade e para a ameaça de divulgação de *deepfake* íntimo com o objetivo de intimidar, coagir, extorquir ou causar sofrimento na vítima da mídia. Além de inaugurar a criminalização de práticas relativas ao abuso dessas ferramentas, a lei também impõe às plataformas virtuais a obrigação de remover, após notificação, o conteúdo não consensual.

No Brasil, ainda não há um tipo penal específico para tratar de usos abusivos de *deepfake*, tampouco para o enquadramento de *deepfakes* pornográficos, como os vistos anteriormente. A prática pode ser enquadrada no art. 218-C do Código Penal, que trata da divulgação de cenas de estupro e estupro de vulnerável ou de sexo, nudez ou pornografia sem consentimento. O dispositivo foi incluído ao Código em 2018, e teve sua pena aumentada em 2025 (Brasil, 1940). Também em 2025, a Lei nº 15.123 alterou o art. 147-B do Código Penal, que trata da violência psicológica contra a mulher. A partir do novo texto, há aumento de pena quando a violência é praticada mediante o uso de inteligência artificial ou de outro recurso tecnológico capaz de alterar a imagem ou o som da vítima (Brasil, 1940),

o que demonstra maior atenção aos riscos especificamente atrelados aos *deepfakes*. Além disso, quando a mídia sexual envolve o público infantojuvenil, o ECA Digital previu, na esfera civil (e não criminal), que fornecedores de produtos ou serviços de tecnologia da informação deverão remover e comunicar conteúdos de aparente exploração ou abuso sexual de crianças e adolescentes (Brasil, 2025).

***

As legislações anteriormente descritas oferecem algumas pistas para a compreensão da perspectiva legal e regulatória de *deepfakes* no Brasil e no mundo, no presente e no futuro próximo. Em primeiro lugar, observa-se que, diante da heterogeneidade conceitual que marca o ecossistema de *deepfakes*, a postura regulatória internacional tem sido a de proibir ou limitar **usos indevidos** dessas ferramentas, com base nos riscos que elas impõem a direitos fundamentais. Assim, a maior parte das legislações aprovadas sobre a matéria ao redor do globo tem se direcionado a aplicações particularmente críticas de *deepfakes*, como ocorre em contextos eleitorais e em relação a conteúdos sexuais/pornográficos.

Em segundo lugar, a constatação de heterogeneidade técnica de *deepfakes* também tem conduzido a regulações centradas na **verossimilhança** dessas mídias, que se tornam cada vez mais convincentes e indistinguíveis de conteúdos autênticos. Dessa maneira, mais do que identificar o modo como os *deepfakes* são produzidos, autoridades judiciais e administrativas de distintos países têm buscado conter os danos que emergem do poder de engano dessas mídias sintéticas ou manipuladas.

Em terceiro lugar, a postura regulatória internacional tem priorizado a regulação **de temas com alto potencial de dano** a direitos dos cidadãos e às democracias contemporâneas em detrimento de vedações genéricas de *deepfakes* na abordagem normativa dessas ferramentas. Isso se nota pela ausência

de proibições totais à geração de *deepfakes* e pela atenção dedicada a essas mídias em legislações relativas a eleições e pornografia não consensual, por exemplo, que se multiplicam em distintos países.

Alguns desafios regulatórios, contudo, permanecem prementes no enfrentamento de riscos associados aos *deepfakes*. Um importante gargalo na governança dessas ferramentas é a dificuldade de **identificar e responsabilizar** atores envolvidos no processo que vai do treinamento à disseminação de *deepfakes* (Kshetri; Voas, 2026, p. 110). Nesse contexto, estratégias para a garantia de anonimato de autores dessas mídias (como o uso de VPNs ou de redes descentralizadas) impõem entraves à individualização de condutas ilícitas. Além disso, a própria natureza transfronteiriça da internet, marcada pelo livre trânsito dos dados e por controles normativos heterogêneos de país para país, obstaculiza a efetividade de regulações já vigentes nessas nações.

Diante desse cenário, abordagens normativas aptas a enfrentar os desafios dos *deepfakes* parecem ser aquelas que seguem modelos multifacetados e integrados de regulação. Isso significa, por um lado, restringir a penalização aos casos de graves transgressões, como extorsão e abuso de imagens íntimas (Kirchengast, 2020, p. 11). Por outro lado, incorporar mecanismos civis e administrativos de responsabilização é uma medida importante para abranger usos abusivos menos gravosos, mas ainda indevidos, dessas ferramentas.

No Brasil, esse movimento tem envolvido, por exemplo, a previsão de mecanismos de responsabilização de outros agentes, além dos autores de *deepfakes*, como provedores de aplicações de internet. Nesse sentido, o Decreto nº 12.976/2026 previu a vedação da geração e da modificação "de conteúdo íntimo de terceiro mediante uso de inteligência artificial ou de qualquer outro recurso tecnológico que altere imagem ou som da vítima", aplicável a provedores de aplicações de internet (Brasil, 2026b).

O mesmo Decreto, que estabelece diretrizes para a proteção de mulheres no ambiente digital, determinou que "os provedores de aplicações de internet baseados em funcionalidades de inteligência artificial ou recurso tecnológico equivalente deverão implementar salvaguardas técnicas e procedimentais para identificar e bloquear solicitações de geração" desses conteúdos (Brasil, 2026b). Nesses casos, proíbe-se, então, a síntese e a manipulação de conteúdo íntimo de terceiros, obrigando que provedores de aplicações capazes de auxiliar tais práticas impeçam tecnicamente essas solicitações. Na perspectiva da LGPD, essas salvaguardas são relevantes quando a geração ou modificação de conteúdo íntimo envolver tratamento de imagem, voz, padrões biométricos ou outros atributos pessoais da vítima, especialmente diante dos princípios da prevenção, segurança, necessidade, não discriminação e responsabilização.

Por sua vez, o Decreto nº 12.975/2026, que regulamenta o Marco Civil da Internet, prevê que a responsabilidade do provedor de aplicações de internet é presumida quando houver veiculação de conteúdo ilícito em anúncios ou impulsionamentos pagos. Além disso, essa responsabilidade é presumida caso o conteúdo em questão seja disseminado por meio de redes artificiais de distribuição de conteúdos, independentemente de notificação (Brasil, 2026a). Dessa maneira, as aplicações onde circulam anúncios pagos de *deepfakes* expressamente orientados à promoção de golpes e fraudes serão consideradas formalmente responsáveis por sua disseminação.

Nos dois casos, nota-se o deslocamento de um modelo normativo centrado na responsabilização individual, como ocorre em diferentes regulações citadas anteriormente, para um modelo que inclui obrigações expressamente direcionadas a provedores de aplicações de internet. Procedimento semelhante foi adotado pelo ECA Digital, que vedou a esses agentes a monetização e o impulsionamento de conteúdos que erotizem crianças e adolescentes e obrigou provedores a

indisponibilizar e reportar conteúdos de aparente exploração e abuso sexual (Brasil, 2025). Outras legislações internacionais centradas em *deepfakes* também adotaram essa perspectiva de responsabilização de provedores, como o *AI Act* europeu e o *Take it Down Act* dos Estados Unidos.

O movimento global de regulação de *deepfakes*, tratado brevemente nesta seção, é parte de um esforço mais amplo da governança da inteligência artificial, que também se observa em distintos países e jurisdições do mundo. Embora algumas previsões legais referentes ao uso de IA se apliquem aos *deepfakes*, as particularidades dessas ferramentas de mídia altamente convincentes demandam tratamento legal também específico para conter seus riscos. Em consonância com essa percepção, o Relatório do Grupo de Trabalho sobre *Deepfakes* da Secretaria de Políticas Digitais da Presidência da República propôs, recentemente, que a redação do Projeto de Lei nº 2338/2023, que dispõe sobre o uso da inteligência artificial e tramita no Congresso Nacional, inclua dispositivos especificamente atinentes ao fenômeno de *deepfakes* (Brasil, 2026c)

As propostas do Grupo de Trabalho envolvem, entre outros elementos: a inclusão ou o aperfeiçoamento do conceito de "conteúdo sintético" previsto na proposta legislativa; a ampliação dos direitos de pessoas ou grupos afetados por sistemas de IA em casos de uso não autorizado de imagem, voz ou outros atributos pessoais; a criação de obrigações de transparência e rastreabilidade para conteúdos sintéticos e a exigência de avaliação específica de riscos ligados a *deepfakes* por parte de desenvolvedores de sistemas de IA generativa. No que diz respeito aos *deepfakes* pornográficos, o Relatório propõe múltiplos requisitos para a geração de conteúdo sintético retratando pessoa natural em nudez ou ato sexual, incluindo-se identificação do usuário responsável pela geração e consentimento expresso da pessoa retratada (Brasil, 2026c).

O esforço do Relatório se coaduna com as propostas descritas ao longo desta seção: em lugar de simplesmente proibir *deepfakes*, recomenda-se aos desenvolvedores de IA generativa a imposição de obrigações de transparência, rastreabilidade e avaliação e mitigação de riscos, além da proteção especial contra abusos de natureza sexual. Embora se trate de recomendação não vinculante, o documento aponta para propostas em debate no processo legislativo, cuja redação e alcance devem ser verificados no momento da publicação.

O fortalecimento de medidas que coíbam usos indevidos e altamente danosos de *deepfakes* passa, portanto, por uma abordagem legal e regulatória multidimensional. Isso pode incluir criminalizações e vedações pontuais; mecanismos de reparação civil e fiscalizações em âmbito administrativo que garantam o adequado cumprimento de leis e das diretrizes de privacidade e proteção de dados de provedores de aplicações de internet. Nesse contexto, é fundamental encarar o ecossistema dessas mídias em sua complexidade, responsabilizando autores, mas também envolvendo empresas na desafiadora tarefa de constituir ambientes digitais mais seguros.

Além disso, essas medidas podem ser complementadas por condutas que partam de cidadãos usuários de aplicações de internet e dos próprios provedores de aplicações. Do ponto de vista de cidadãos, é imprescindível consolidar competências informacionais e midiáticas na população, de modo a desenvolver no público geral uma capacidade aumentada de identificação de conteúdos enganosos e fraudulentos. Do ponto de vista das aplicações, é necessário que os importantes avanços técnicos que tornam as próprias *deepfakes* possíveis sejam também utilizados para sua detecção, para o bloqueio de solicitações de geração de conteúdo ilícito e para a restrição da circulação de mídias com alto potencial de dano. Nas próximas subseções, trataremos desses aspectos.

6.2 Competência informacional e midiática

Diante dos riscos e dos desafios regulatórios decorrentes dos *deepfakes*, é fundamental envolver usuários na mitigação de possíveis riscos associados ao mau uso dessas ferramentas. Trata-se de um modo de encarar o ecossistema de *deepfakes* a partir da resiliência social e da prevenção de danos, que busca complementar – jamais substituir – os mecanismos normativos de controle de geração e disseminação dessas mídias, abordados anteriormente. Nesse contexto, o desenvolvimento de competências informacionais e midiáticas no público geral é uma importante medida de prevenção centrada nos indivíduos.

A competência informacional refere-se ao desenvolvimento de capacidades de **avaliação crítica de conteúdos e informações**, contemplando também informações produzidas e disseminadas por tecnologias digitais Essas capacidades implicam uma postura reflexiva diante de mídias digitais que, em vez de serem presumidas como autênticas, passam a ser analisadas com atenção, levando-se em conta aspectos como a fonte da informação, o reconhecimento de sinais de manipulação e o exame do contexto de circulação dos conteúdos.

Esse letramento implica a redução dos danos da desinformação política, que são acelerados por ferramentas como os *deepfakes*, mas também a diminuição da vulnerabilidade de usuários a fraudes e golpes realizados por clonagem de voz, validações falsas de identidade ou chamadas de vídeo sintéticas, por exemplo. Dessa forma, a conscientização sobre a existência e as possibilidades técnicas de *deepfakes* contribui para reduzir a efetividade de ataques de engenharia social, que se aproveitam do desconhecimento da população a respeito dessas ferramentas. Além disso, o letramento ainda tem o efeito de ampliar, em organizações de diferentes naturezas, a inclusão de procedimentos de autenticação complementar de identidade para o reforço da segurança em sistemas técnicos que são crescentemente utilizados na vida cotidiana dos brasileiros.

Os benefícios da competência informacional são experimentados de maneira transversal pela população, mas são particularmente relevantes para a **proteção de grupos vulneráveis**. Como exposto neste Radar Tecnológico, os riscos associados aos *deepfakes* são aprofundados para algumas populações, como é o caso de crianças e adolescentes, pessoas idosas e mulheres. A promoção de competências midiáticas para esses grupos pode incluir, por exemplo, a alfabetização digital nas escolas, a capacitação de cuidadores de pessoas idosas, a implementação de programas de educação digital voltados à comunidade, com módulos específicos para pessoas idosas, e a conscientização, em diferentes arenas, sobre a existência e os danos de violências de gênero facilitadas pelo uso de tecnologias.

Nesse mesmo sentido, a população geral precisa compreender, ainda que de maneira simplificada, como funciona a inteligência artificial, o que é possível fazer com esses sistemas, o que são conteúdos sintéticos e quais são as possíveis manipulações de mídia disponíveis atualmente. Embora esse seja um desafio considerável em um país tão desigual quanto o Brasil, é preciso reforçar que escolas, universidades, bibliotecas, associações de bairro, empresas de tecnologia e mídia e outras instituições públicas e privadas têm papéis complementares e fundamentais no propósito de ampliação do letramento digital da população brasileira.

Essa dimensão crítica do letramento digital se conecta, ainda, a uma noção mais ampla de autonomia dos usuários diante das tecnologias. Como observa Darin (2024), sistemas digitais são capazes de moldar comportamentos de forma positiva ou negativa, de modo que o exercício consciente da experiência digital, isto é, compreender como as aplicações operam, que interesses orientam seu *design* e como elas afetam decisões cotidianas, é uma condição para que as pessoas preservem sua capacidade de autodeterminação em ambientes hiperconectados. No contexto dos *deepfakes*, essa perspectiva reforça que a competência informacional não se limita à verificação pontual de conteúdos suspeitos, mas envolve uma postura

permanente de reflexividade sobre as próprias mediações tecnológicas por meio das quais informações são produzidas, recomendadas e consumidas.

Nesse contexto, exemplo de iniciativa de letramento digital é encontrado na Estratégia Brasileira de Educação Midiática (EBEM), que busca desenvolver competências informacionais e midiáticas ao promover a formação crítica de cidadãos diante da desinformação e da manipulação algorítmica. A Estratégia, formulada no âmbito da Semana Brasileira de Educação Midiática, objetiva formar indivíduos capazes de interpretar, produzir e compartilhar informações de forma ética e responsável no ambiente digital. Entre os eixos estruturantes da EBEM, encontra-se a inteligência artificial, que procura fortalecer capacidades de identificação e compreensão de conteúdos sintéticos, a fim de reduzir seu potencial para a desinformação e a violação de direitos (Brasil, 2023).

A Educação Digital e Midiática é parte da atual Base Nacional Comum Curricular (BNCC) do Brasil, e visa a incentivar o uso efetivo de tecnologias digitais nas atividades de ensino e aprendizagem e promover o desenvolvimento da literacia digital para adequado aproveitamento dessas ferramentas. Para além da EBEM, outras iniciativas também procuram fortalecer as competências informacionais de distintos segmentos da população brasileira, como é o caso da Disciplina de Cidadania Digital do SaferNet Brasil, que oferece planos de aulas e materiais para educadores implementarem a BNCC de educação digital (Disciplina..., 2026). O programa Cidadão Digital, também do SaferNet, é outro esforço nesse sentido, que busca fomentar diálogos com adolescentes e jovens sobre segurança on-line, combate à desinformação e respeito nas redes (Cidadania..., 2026). Além disso, o Ministério da Educação (MEC) e a Secretaria de Comunicação Social da Presidência da República desenvolvem diversas ações de educação digital por meio da Coleção Brasileira de Educação Digital e Midiática e da plataforma AVAMEC. O debate sobre o desenvolvimento de competências informacionais não é

novo. Ele se torna, porém, um imperativo para a promoção de ambientes digitais seguros.

Apesar da importância do desenvolvimento de competências informacionais e midiáticas, é preciso ressaltar que essas são medidas complementares às práticas regulatórias que limitam usos indevidos e proíbem abusos relacionados aos *deepfakes*. A eficácia do letramento digital frente à crescente verossimilhança de conteúdos sintéticos e à assimetria de poder existente entre empresas desenvolvedoras de sistemas de IA e usuários individuais é bastante limitada. Dessa maneira, a promoção do letramento digital deve ser compreendida como componente de uma estratégia ampla de governança do ecossistema de *deepfakes*, associada a práticas legais e a ferramentas técnicas de mitigação de riscos, como se discutirá a seguir.

6.3 Ferramentas de mitigação e possíveis soluções

Os avanços da inteligência artificial generativa impulsionaram uma evolução acelerada dos *deepfakes*, tornando-os cada vez mais sofisticados e difíceis de distinguir de conteúdos autênticos. Nesse contexto, torna-se necessário o desenvolvimento contínuo de mecanismos capazes de identificar manipulações digitais e de conferir maior transparência à origem e ao histórico dos conteúdos. A mitigação dos riscos associados a essa tecnologia depende, assim, tanto do aperfeiçoamento das ferramentas de detecção quanto da implementação de mecanismos de autenticação e rastreabilidade capazes de fortalecer a confiança no ambiente digital.

Assim como os modelos de aprendizado de máquina e aprendizado profundo empregados na geração de conteúdos sintéticos, as ferramentas de identificação utilizam algoritmos treinados para reconhecer padrões e inconsistências imperceptíveis ao olho humano. Entre os principais indícios

analisados estão alterações na iluminação, sincronização labial, padrões biométricos faciais, metadados, artefatos digitais e outras características que podem revelar a manipulação artificial de imagens, áudios e vídeos.

Nesse contexto, diferentes soluções vêm sendo desenvolvidas para aumentar a confiabilidade dos processos de verificação, destacando-se:

+ **Detecção automatizada**, baseada em algoritmos de IA treinados para identificar padrões característicos de conteúdos manipulados;

+ **Assinaturas criptografadas**, que permitem verificar a integridade do conteúdo desde sua origem, assegurando que o arquivo não tenha sofrido alterações após sua criação;

+ **Marcas d'água digitais *(watermarking)***, que consistem na inserção de sinais visíveis ou invisíveis que identificam conteúdos produzidos ou modificados por sistemas de inteligência artificial, facilitando sua posterior identificação;

+ **Padrões de rastreabilidade, como *Coalition for Content Provenance and Authenticity* (C2PA),** que estabelecem mecanismos padronizados para registrar informações sobre origem, autoria, histórico de edição e cadeia de produção de conteúdos digitais, promovendo maior transparência e verificabilidade.

Esses mecanismos demonstram uma mudança de paradigma na mitigação dos riscos associados aos *deepfakes*. Além da identificação posterior de manipulações, busca-se documentar a cadeia de produção dos conteúdos digitais desde sua origem, favorecendo processos de autenticação, auditoria e responsabilização.

No último ano, consolidaram-se no mercado diversas ferramentas on-line usadas para a identificação de manipulações

em vídeos e mídias digitais. Três delas: *DeepWare Scanner, Attestiv* e *TrueMedia* foram analisadas comparativamente do ponto de vista da usabilidade e do ponto de vista técnico. O resultado desta análise foi publicado em um artigo que versa sobre os desafios e limitações de *deepfakes* (Moraes; Batista, 2025). Os resultados evidenciaram, à época, a necessidade de investigação humana complementar, pois nenhuma identificou de maneira certeira quais eram os casos de *deepfakes* e outras falsificações. Ou seja, todas deram respostas incertas tanto para produtos sintéticos quanto para materiais originais, de resolução mais baixa ou produzidos em ambiente muito controlado, como por exemplo um episódio da antiga série de TV Chaves ou um trecho de uma exibição do Jornal Nacional. De acordo com os autores,

> *[...] as ferramentas analisadas têm suas forças e limitações, sendo necessárias melhorias contínuas para aumentar a precisão e confiabilidade dos resultados. A variabilidade nos desempenhos evidencia que, apesar do progresso tecnológico, a detecção de deepfakes ainda enfrenta desafios. Portanto, a combinação dessas ferramentas com a avaliação humana é crucial para um processo de verificação mais robusto e eficiente (Moraes; Batista, 2025).*

Os resultados desse estudo apresentam que, embora os avanços tecnológicos tenham ampliado consideravelmente a capacidade de identificação de conteúdos manipulados, as ferramentas atualmente disponíveis ainda apresentam limitações. A detecção de *deepfakes* permanece sujeita a margens de erro e a diferentes níveis de confiança, especialmente diante de conteúdos de baixa resolução, materiais históricos ou registros produzidos em condições específicas de captura. Nesse contexto, a verificação da autenticidade demanda uma abordagem integrada, que combine recursos tecnológicos, análise contextual e supervisão humana.

Paralelamente aos mecanismos de detecção, cresce a importância das iniciativas voltadas à transparência e à governança

da inteligência artificial. A identificação da origem dos conteúdos, o registro de seu histórico de modificações e a indicação de que determinado material foi produzido ou alterado por sistemas de IA representam medidas capazes de reduzir assimetrias informacionais e fortalecer a confiabilidade do ecossistema digital. Ao mesmo tempo, instrumentos de rastreabilidade e auditoria contribuem para demonstrar conformidade com princípios de *accountability* e com os requisitos de governança previstos nos modelos regulatórios mais recentes.

Essas iniciativas dialogam com uma agenda mais ampla de responsabilidade no *design* de sistemas interativos. Nesse sentido, Darin (2024) destaca que o desenvolvimento tecnológico não deve orientar-se exclusivamente por métricas de eficiência, engajamento e retenção, cabendo a desenvolvedores refletir sobre o impacto ético de suas criações e adotar escolhas de *design* que respeitem os limites humanos e o bem-estar dos usuários. Transposta ao ecossistema de *deepfakes*, essa perspectiva sugere que salvaguardas técnicas como a marcação de conteúdo sintético, o bloqueio de solicitações ilícitas de geração e os mecanismos de rastreabilidade não deveriam ser apenas obrigações regulatórias, mas expressões de um compromisso de projeto (*by design*) com a integridade informacional e com a proteção das pessoas retratadas.

A partir desse diagnóstico, impõe-se uma análise jurídica mais crítica, como a desenvolvida por Moraes e Batista (2025) e, que questiona a premissa amplamente adotada pela indústria de que dados "publicamente disponíveis" podem ser livremente reutilizados no treinamento de sistemas de IA. Tal entendimento se mostra incompatível com os marcos contemporâneos de proteção de dados, pois desconsidera princípios fundamentais como finalidade, adequação, necessidade e transparência. No contexto dos *deepfakes*, esses princípios assumem especial importância, exigindo avaliar se o uso dos dados atende a propósitos legítimos ou se favorece práticas de manipulação e engano, bem como

se há excessos no tratamento e na ausência de informação adequada aos titulares.

Por fim, uma vez incorporados ao treinamento desses sistemas, os dados passam a gerar riscos persistentes, como memorização, reidentificação e propagação em múltiplos sistemas downstream, o que dificulta a governança e compromete a efetiva tutela dos direitos dos titulares. Assim, os *deepfakes* podem representar um desafio tecnológico e um problema regulatório estrutural, demandando a reinterpretação de categorias tradicionais da proteção de dados diante de um ambiente marcado pela recombinação massiva de informações e por impactos sociais potencialmente representativos.

Desse modo, a mitigação dos riscos associados aos *deepfakes* exige uma abordagem multidimensional que articule soluções tecnológicas, mecanismos de transparência, padrões de rastreabilidade, estruturas regulatórias compatíveis com a velocidade da inovação e, principalmente, supervisão humana qualificada. Faz-se necessário construir um ambiente digital pautado pela autenticidade da informação, pela proteção de direitos fundamentais e pela responsabilização dos diferentes agentes envolvidos na produção e circulação de conteúdos sintéticos.

‹ 7 considerações finais ›

Os *deepfakes* representam uma das manifestações mais emblemáticas da atual convergência entre inteligência artificial, dados pessoais e comunicação digital. Embora a manipulação de imagens, vídeos e áudios não seja um fenômeno novo, os avanços recentes das técnicas de aprendizado profundo ampliaram consideravelmente a capacidade de produzir conteúdos sintéticos altamente realistas, acessíveis e escaláveis. Como demonstrado ao longo deste número da série Radar Tecnológico, essas tecnologias podem ter potencial para aplicações legítimas em áreas como entretenimento, educação, saúde, preservação histórica e acessibilidade, mas também introduzem riscos relevantes à privacidade, à proteção de dados pessoais, à segurança digital e à integridade informacional.

A análise realizada evidencia que os *deepfakes* não constituem apenas um desafio tecnológico. Sua relevância para o cenário de proteção de dados decorre, sobretudo, do fato de que essas aplicações dependem intensamente da coleta, do tratamento e da reprodução de elementos capazes de identificar ou singularizar indivíduos, especialmente dados biométricos como imagem facial, voz, gestos e padrões comportamentais. Nesse contexto, os *deepfakes* desafiam conceitos tradicionais de autenticidade, consentimento, identidade e responsabilização, exigindo novas reflexões sobre os mecanismos de proteção aplicáveis aos ambientes digitais.

Os exemplos analisados de usos indevidos de *deepfakes* demonstram que os impactos negativos dessas ferramentas se distribuem de forma desigual entre diferentes grupos sociais. Mulheres e meninas permanecem desproporcionalmente afetadas pela pornografia sintética não consensual e por outras formas de violência digital. No campo político, a manipulação audiovisual tem sido empregada para desinformar eleitores, comprometer reputações e amplificar dinâmicas de violência política de gênero. Já no ambiente

econômico, anúncios fraudulentos e golpes financeiros baseados na clonagem de voz e de imagem exploram as relações de confiança para induzir consumidores e usuários ao erro. *Deepfakes* são, ainda, ferramentas que produzem e reproduzem práticas racistas. Essas múltiplas manifestações revelam que os danos causados pelos *deepfakes* extrapolam a esfera individual, alcançando dimensões coletivas e institucionais.

Ao mesmo tempo, este Radar Tecnológico demonstrou que a evolução das técnicas de geração e manipulação de mídias sintéticas ocorre em ritmo acelerado e frequentemente superior à capacidade de adaptação dos mecanismos tradicionais de detecção e resposta. Ferramentas de identificação automatizada, marcas d'água digitais, assinaturas criptográficas e iniciativas de rastreabilidade de conteúdo represen*tam* avanços importantes, mas ainda enfrentam limitações técnicas e operacionais importantes em sua implementação em contextos reais, como plataformas de mídias sociais. Em muitos casos, a análise humana especializada permanece, até o momento, indispensável para avaliar a autenticidade de conteúdos potencialmente manipulados.

Sob a perspectiva regulatória, observam-se diferentes estratégias internacionais para enfrentar o fenômeno. Algumas iniciativas priorizam obrigações de transparência e rotulagem de conteúdo sintético; outras se concentram na proibição de usos específicos, como na interferência eleitoral; outras, ainda, têm criminalizado abusos ligados à pornografia não consensual; enquanto modelos mais recentes procuram combinar abordagens baseadas em risco, governança tecnológica e responsabilização dos diversos atores envolvidos. Apesar das diferenças, emerge uma convergência internacional em torno da necessidade de ampliar a transparência, proteger os indivíduos contra usos abusivos e fortalecer mecanismos de supervisão, prestação de contas e responsabilização.

Por fim, a principal conclusão deste Radar Tecnológico é que os *deepfakes* devem ser compreendidos como parte de um

processo mais amplo de transformação dos mecanismos de produção, circulação e validação da informação na sociedade contemporânea. O enfrentamento de seus riscos dependerá de soluções técnicas, de instrumentos jurídicos específicos, e, principalmente, do fortalecimento do letramento digital, da educação midiática, da transparência algorítmica e da construção de uma cultura de uso responsável da inteligência artificial. Em um contexto em que evidências visuais e sonoras deixam de constituir, por si sós, garantias de autenticidade, a preservação da confiança nos ambientes digitais se torna um dos principais desafios para a proteção de direitos e a manutenção de sociedades informadas e democráticas.

‹ referências ›

AKHTAR, Zahid. *Deepfakes* generation and detection: a short survey. **Journal of Imaging**, v. 9, n. 1, 2023. DOI: 10.3390/jimaging9010018. Disponível em: https://doi.org/10.3390/jimaging9010018. Acesso em: 15 jun. 2026.

ALTUNCU, Enes; FRANQUEIRA, Virginia N. L.; LI, Shujun. *Deepfake*: definitions, performance metrics and standards, datasets, and a meta-review. **Frontiers in Big Data**, [*S. l.*], v. 7, p. 1–23, 2024. doi: 10.3389/fdata.2024.1400024

AMADO, Guilherme; CAMPOS, João Pedroso de. Moro avisa de golpe de inteligência artificial: "Não vendo crédito". **Metrópoles**, 7 jun. 2024. Disponível em: https://www.metropoles.com/colunas/guilherme-amado/moro-avisa-de-golpe-de-inteligencia-artificial-nao-vendo-credito. Acesso em: 21 jul. 2026.

ANPD. **Inteligência artificial generativa**. Brasília, DF: ANPD, 2024. (Radar Tecnológico, n. 3). Publicação digital. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar_tecnologico_ia_generativa_anpd.pdf. Acesso em: 19 jun. 2026.

ASSIS JUNIOR, Fabio de Paula; HESSEL, Ana Maria Di Grado. Entre ver e crer: *deepfake* e criação para arte e entretenimento. **TECCOGS – Revista Digital de Tecnologias Cognitivas**, n. 23, p. 79–89, jan./jun. 2021. Disponível em: https://revistas.pucsp.br/index.php/teccogs/article/view/55980/37928. Acesso em: 4 jul. 2025.

AWS. **Qual é a diferença entre machine learning e deep learning?** Seattle: Amazon Web Services, 2023. Disponível em: https://aws.amazon.com/compare/the-difference-between-machine-learning-and-deep-learning/. Acesso em: 10 jun. 2026.

AZEVEDO, Ingrid Borges de. *Deepfakes* **contra candidatas nas eleições:** o uso da inteligência artificial na perpetuação da violência política de gênero. 2025. Dissertação (Mestrado em Direito, Estado e Constituição) – Faculdade de Direito, Universidade de Brasília, Brasília, 2025. Disponível em: https://repositorio.unb.br/bitstream/10482/54125/1/IngridBorgesDeAzevedo_DISSERT.pd. Acesso em: 6 jul. 2026.

BHONDAVE, A. *et al. Deepfake* Forensics Tool: AI-Driven Multi-Modal Media Authentication with Forensic Evidence Reporting. **International Journal of Scientific Research in Engineering and Management**, v. 10, n. 3, mar. 2026. DOI: 10.55041/ijsrem57455.

BOND-TAYLOR, Sam; LEACH, Adam; LONG, Yang; WILLCOCKS, Chris G. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 44, n. 11, p. 7327–7347, nov. 2022. Disponível em: https://doi.org/10.1109/TPAMI.2021.3116668. Acesso em: 6 jul. 2026.

BRADY, Madeline. *Deepfakes*: a new desinformation threat? Technical report, **Democracy Reporting International**, 31 jul. 2020. Disponível em: https://democracyreporting.s3.eu-central-1.amazonaws.com/images/20842020-09-01-DRI-*deepfake*-publication-no-1.pdf. Acesso em: 15 jun. 2026.

BRASIL. **Decreto-Lei n. 2.848**, de 7 de dezembro de 1940. Código Penal. Rio de Janeiro: Presidência da República, 1940. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: 06 jul. 2026. ~

BRASIL. **Lei n. 8.069, de 13 de julho de 1990**. Dispõe sobre o Estatuto da Criança e do Adolescente e dá outras providências. Brásilia, DF: Presidência da República, 1990.

BRASIL. **Lei n. 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Congresso Nacional, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 06 jul. 2026.

BRASIL. **Resolução nº 23.610, de 18 de dezembro de 2019**. Dispõe sobre a propaganda eleitoral (Redação dada pela Resolução nº 23.732/2024). Brasília: Presidência da República, 2019. Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2019/resolucao-no-23-610-de-18-de-dezembro-de-2019. Acesso em: 06 jul. 2026.

BRASIL. **Lei nº 14.192, de 4 de agosto de 2021**. Estabelece normas para prevenir, reprimir e combater a violência política contra a mulher; e altera a Lei nº 4.737, de 15 de julho de 1965 (Código Eleitoral), a Lei nº 9.096, de 19 de setembro de 1995 (Lei dos Partidos Políticos), e a Lei nº 9.504, de 30 de setembro de 1997 (Lei das Eleições), para dispor sobre os crimes de divulgação de fato ou vídeo com conteúdo inverídico no período de campanha eleitoral, para criminalizar a violência política contra a mulher e para assegurar a participação de mulheres em debates eleitorais proporcionalmente ao número de candidatas às eleições proporcionais. Diário Oficial da União: seção 1, Brasília, DF, 5 ago. 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14192.htm. Acesso em: 6 jul. 2026.

BRASIL. **Lei n. 15.211, de 17 de setembro de 2025**. Dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (Estatuto Digital da Criança e do Adolescente). Brasília: Congresso Nacional, 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Acesso em: 06 jul. 2026.

BRASIL. **Decreto nº 12.975, de 20 de maio de 2026**. Altera o Decreto nº 8.771, de 11 de maio de 2016, que regulamenta a Lei nº 12.965, de 23 de abril de 2014. Brasília: Presidência da República, 2026a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12975.htm. Acesso em: 06 jul. 2026.

BRASIL. **Decreto n. 12.976, de 20 de maio de 2026**. Estabelece diretrizes para a proteção de mulheres na internet e para o enfrentamento da violência contra mulheres em ambiente digital. Presidência da República: Brasília, 2026b. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12976.htm. Acesso em: 11 jun. 2026.

BRASIL. Secretaria de Políticas Digitais. Secretaria de Comunicação Social da Presidência da República. **Relatório Grupo de Trabalho** *Deepfakes*. Brasília, 2026c. Disponível em: https://www.gov.br/secom/pt-br/arquivos/2606_gt_relatorio-*deepfake*-v4-web.pdf/. Acesso em: 06 jul. 2026.

BRASIL. Secretaria de Comunicação Social. **Estratégia Brasileira de Educação Midiática**. Brasília, 2023. Disponível em: https://www.gov.br/secom/pt-br/arquivos/2023_secom-spdigi_estrategia-brasileira-de-educacao-midiatica.pdf. Acesso em: 07 jul. 2026.

BRIGHAM, Natalie Grace ; WEI, Miranda ; KOHNO, Tadayoshi ; REDMILES, Elissa W. Violation of my body: Perceptions of AI-generated non-consensual (intimate) imagery. *In*: SYMPOSIUM ON USABLE PRIVACY AND SECURITY, 20., 2024, Philadelphia. **Proceedings [...]**. Philadelphia: USENIX, 2024. p. 373-392. Disponível em: https://arxiv.org/abs/2406.05520. Acesso em: 6 jul. 2025.

CNTI. **Synthetic media &** ***deepfakes*****.** CNTI, 6 out. 2025. Disponível em: https://cnti.org/issue-primers/synthetic-media-*deepfakes*/. Acesso em: 1 jun. 2026.

ÇETIN, Reycan. Differences Between *Cheapfake* and *Deepfake* and the Threats They Pose in Identity Fraud. **Techsign**, 2022. Disponível em: https://www.techsign.com.tr/en-blog-posts/differences-between-*cheapfake*-and-*deepfake*-and-the-threats-they-pose-in-identity-fraud. Acesso em: 15 jun. 2026.

CHARPE, Aditya; KHOKALE, Rahul; GHAGHRE, Dheeraj. Real-Time *Deepfake* Detection: A Systematic Review of Generative Adversarial Networks (GANs) and Generative Transformer Networks (GTNs). **International Journal for Research in Applied Science and Engineering Technology**, v. 13, n. 5, May 2025. Disponível em: https://doi.org/10.22214/ijraset.2025.71021 Acesso em: 6 jul. 2026.

CIDADANIA Digital para adolescentes e jovens. SaferNet Brasil, cursos online. Salvador, 2026. Disponível em: https://ead.safernet.org.br/cidadaodigital/. Acesso em: 06 jul. 2026.

COLMAN, Ben. **Real-time** *deepfakes* **are rewriting the rules of child safety**. World Economic Forum, Opinion, 27 abr. 2026. Disponível em: https://www.weforum.org/stories/all/ai-voice-*deepfakes*-child-safety/. Acesso em: 06 jul. 2026.

COVER, Rob. AI generation of *rage bait*: Implications for digital harms. **New Media & Society**, 16 dez. 2025. https://doi.org/10.1177/14614448251400675.

CUZCANO-CHAVEZ, Ximena; GONZÁLEZ-Véliz, Constanza. Análisis del abuso de aplicaciones *deepfake* y su impacto en violencia digital de género em América Latina y el Caribe. **Veritas**, Valparaíso, n. 61, p. 32–67, 2025. Disponível em: https://www.scielo.cl/pdf/veritas/n61/0718-9273-veritas-61-32.pdf. Acesso em: 6 jul. 2026.

DANTAS, Glenda. Racismo recreativo em tempos de *deepfake*. *In*: SOUZA, Gustavo; SILVA, Tarcizio (org.). **Enfrentando** *Deepfakes*. Desvelar: Brasília, 2025. [Livro eletrônico]. Disponível em: desvelar.org/enfrentando-*deepfakes*. Acesso em: 6 jul. 2026.

DARIN, Ticianne. **Manual da autonomia digital:** bem-estar na experiência do usuário. Fortaleza: Editora UFC, 2024. (Coleção Flecha)

DEHGHANI, Arash.; SABERI, Hossein. Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools. **arXiv**, 7 jan. 2025. https://doi.org/10.48550/arXiv.2501.06227. Disponível em: https://arxiv.org/abs/2501.06227v1. Acesso em: 6 jul. 2026.

DING, Jeffrey. **CHINAI #104**: Tencent 2020 AI white paper. ChinAI, 27 jul. 2020. Disponível em: https://chinai.substack.com/p/chinai-104-tencent-2020-ai-white. Acesso em: 12 jun. 2026.

DISCIPLINA de Cidadania Digital. **SaferNet Brasil**, Salvador, 2026. Disponível em: https://cidadaniadigital.org.br/. Acesso em: 6 jul. 2026.

DOMINGOS, Roney. É #FAKE vídeo que diz que governo federal condenou Serasa a pagar indenização de R$ 30 mil. **G1**, 30 mar. 2024. Disponível em: https://g1.globo.com/fato-ou-fake/noticia/2024/03/30/e-fake-video-que-diz-que-governo-federal-condenou-serasa-a-pagar-indenizacao-de-r-30-mil.ghtml. Acesso em: 21 jul. 2026.

EBERL, A.; KÜHN, J.; WOLBRING, T. Using *deepfakes* for experiments in the social sciences: A pilot study. **Frontiers in Sociology**, v. 7, p. 907199, 2022.

ELLIOTT, Vittoria. ¿Qué son los *cheapfakes* y por qué serán más graves que los *deepfakes* en la escena política? **WIRED**, 19 dic. 2023. Disponível em: https://es.wired.com/articulos/*cheapfakes*-y-por-que-seran-mas-graves-que-*deepfakes*-en-escena-politica. Acesso em: 23 jun. 2026.

EUROPEAN PARLIAMENT. European Parliamentary Research Service (EPRS). **Tackling** *deepfakes* **in European policy**. Luxembourg: Publications Office of the European Union, 2021. Disponível em: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf. Acesso em: 1 jun. 2026.

EUROPEAN UNION. *AI Act*. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. Estrasburgo: Parlamento Europeu, 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Acesso em: 6 jul. 2026.

FANTÁSTICO. Drauzio Varella processa Meta por anúncios falsos com sua imagem: "Existe curso online para ensinar falsários a imitar minha voz". **G1**, 22 dez. 2025. Disponível em: https://g1.globo.com/fantastico/noticia/2025/12/22/drauzio-varella-processa-meta-por-anuncios-falsos-com-sua-imagem-existe-curso-online-para-ensinar-falsarios-a-imitar-minha-voz.ghtml. Acesso em: 7 jul. 2026.

FOLHA DE S. PAULO. Tabata Amaral é alvo de campanha de desinformação que a compara a criadora de conteúdo adulto. **Folha de S. Paulo**, set. 2024. Disponível em: https://www1.folha.uol.com.br/poder/2024/09/tabata-amaral-e-alvo-de-campanha-de-desinformacao-que-a-compara-a-criadora-de-conteudo-adulto.shtml. Acesso em: 7 jul. 2026.

FRANCE. **Code pénal.** Paris: Parlement, 2026. Disponível em: https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006070719/LEGISCTA000006165310/?anchor=LEGIARTI000049571542&__cf_chl_f_tk=gEJLxZw8A2.kRORiNeqG47sEN2k9gFe0y1NhBrXPD5A-1783347616-1.0.1.1-Upn7J0pkckwyJRIAZTdR8GFkW410hIskuTiWcySyFYA#LEGIARTI000049571542. Acesso em: 6 jul. 2026.

FUKUHA, Ana Carolina; ARCHEGAS, João Victor; FÜRST, Maria Eduarda. *Deepfake* como desafio jurídico e social: análise jurídico comparada em perspectiva internacional. **Revista Direito FAE**, Curitiba, v. 18, n. 1, p. 166-226, 2026. Disponível em: https://revistadedireito.fae.edu/direito/article/view/179. Acesso em: 06 jul. 2026.

GODULLA, Alexander.; HOFFMANN, Christian Pieter; BITTON, Daniel Bendahan. Dealing with *deepfakes* – an interdisciplinary examination of the state of research and implications for communication studies. **Studies in Communication and Media**, v. 10, n. 1, p. 72–96, 2021. Disponível em: https://doi.org/10.5771/2192-4007-2021-1-72. Acesso em: 6 jul. 2026.

GOLPE no Facebook usa IA e imagem falsa de Pedro Moreira Salles. **Blog Itaú #ÉFake**, 4 jun. 2025. Disponível em: https://blog.itau.com.br/efake/e-fake-golpe-no-facebook-usa-ia-e-imagem-falsa-de-pedro-moreira-salles. Acesso em: 21 jul. 2026.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep Learning. Cambridge, MA: MIT Press, 2016. Disponível em: https://www.deeplearningbook.org/. Acesso em: 11 jun. 2026. GOODFELLOW, Ian J.; POUGET-ABADIE, Jean; MIRZA, Mehdi; BING, Xu; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative adversarial nets. *In*: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 27., 2014, Montreal. **Anais eletrônicos** [...]. Montreal: Neural Information Processing Systems Foundation, 2014. p. 2672-2680. Disponível em: https://papers.nips.cc/paper/5423-generative-adversarial-nets. Acesso em: 29 jun. 2026.

GUO, T.; LI, J.; TANG, Y. A score based likelihood ratio framework for *deepfake* image identification in forensic science. **Scientific Reports**, v. 16, 2026. Disponível em: https://www.nature.com/articles/s41598-026-42176-w. Acesso em: 6 jul. 2026. DOI: 10.1038/s41598-026-42176-w.

G1 MT. Alunos são expulsos após usar inteligência artificial para criar *nudes* falsos de professora e colegas em escola particular de Cuiabá. **G1 Mato Grosso**, 25 set. 2024. Disponível em: https://g1.globo.com/mt/mato-grosso/noticia/2024/09/25/alunos-sao-expulsos-apos-usar-inteligencia-artificial-para-criar-*nudes*-falsos-de-professora-e-colegas-em-escola-particular-de-cuiaba.ghtml. Acesso em: 7 jul. 2026.

HACHKEVYCH, Andrii. The impact of generative artificial intelligence on the legal systems of contemporary states. **Law and Innovative Society**, v. 1, n. 24, p. 37-46, 2025. Disponível em: https://apir.org.ua/lais/en/article/view/467. Acesso em: 6 jul. 2026.

HAYKIN, Simão. **Redes Neurais: Princípios e Prática** . 2. ed. Porto Alegre: Bookman, 2007. *E-book*. pág.185. ISBN 9788577800865. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788577800865/. Acesso em: 06 jun. 2026.

HEIDARI, Arash; NAVIMIPOUR, Nima Jafari; DAG, Hasan; UNAL, Mehmet. *Deepfake* detection using deep learning methods: A systematic and comprehensive review. **WIREs Data Mining and Knowledge Discovery**, v. 14, n. 2, e1520, 2024.

HENRIQUE, Everton. Mulher acreditava estar "noiva" de Brad Pitt e descobre que caiu em golpe. **Terra**, 7 jan. 2026. Disponível em: https://www.terra.com.br/diversao/gente/mulher-acreditava-estar-noiva-de-brad-pitt-e-descobre-que-caiu-em-golpe,314ff7b946892c73e5fc85a6dc1d7d52zzodh96o.html. Acesso em: 21 jul. 2026.

HERRING, Susan. C.; DEDEMA, Meredith; RODRIGUEZ, Enrique.; YANG, Leo. Gender and culture differences in perception of deceptive video filter use. *In*: INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION, 2022. **Proceedings [...]**. Springer, 2022. p. 52-72.

HONG, Rachel; HUTSON, Jevan; AGNEW, William; HUDA, Immad; KOHNO, Tadayoshi; MORGENSTERN, Jamie. A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset. *In*: ACM SYMPOSIUM ON COMPUTER SCIENCE AND LAW, 5., 2026. **Proceedings [...]**. March 2026, p. 1-14.

IRFAN; ARORA; SANDOTRA; RAZA. On Machine Learning and Deep Learning based *Deepfake* Generation and Detection. **Procedia Computer Science**, [*S. l.*], v. 259, p. 1927-1936, 2025. DOI: https://doi.org/10.1016/j.procs.2025.04.148. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050925012505. Acesso em: 06jul. 2026.

ITAÚ. Golpe no Facebook usa IA e imagem falsa de Pedro Moreira Salles. **Blog Itaú #ÉFake**, 4 jun. 2025. Disponível em: https://blog.itau.com.br/efake/e-fake-golpe-no-facebook-usa-ia-e-imagem-falsa-de-pedro-moreira-salles. Acesso em: 21 jul. 2026.

KAATE, Ilkka.; SALMINEN, Joni.; JUNG, Soon.-Gyo; RIZUN, Nina; REVINA, Aleksandra; JANSEN, Bernard J. Getting Emotional Enough: Analyzing Emotional Diversity in *Deepfake* Avatars. *In*: NORDIC CONFERENCE ON HUMAN-COMPUTER INTERACTION, 13., 2024, Uppsala. **Proceedings [...]**. Uppsala, Sweden, 2024. p. 1–12. Disponível em: https://doi.org/10.1145/3679318.3685398. Acesso em: 6 jul. 2026.

KARRAS, Tero; LAINE, Samuli; AILA, Timo. A style-based generator architecture for generative adversarial networks. *In*: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2019, Long Beach. **Proceedings [...]**. Long Beach: IEEE, 2019. p. 4396-4405. Disponível em: https://doi.org/10.1109/CVPR.2019.00453. Acesso em: 29 jun. 2026.

KAUSHAL, Anukriti; YADAV, Uttam Kumar; PARIHAR, Yogesh; TOMAR, Vatan. Study On *Deepfake* Detection Techniques: Challenges, Limitations and Future Directions. *In*: INTERNATIONAL CONFERENCE ON DATA INTELLIGENCE AND COGNITIVE INFORMATICS (ICDICI), 6., 2025, Tirunelveli. **Proceedings [...]**. Tirunelveli, India, 2025. Disponíel em: https://ieeexplore.ieee.org/document/11135143. Acesso em: 6 jul. 2026.

KHALID, Omer; SAAD, Fatima; SAFDAR, Aimen; JAMIL, Akhtar; HAMEED, Alaa Ali; MUÑOZ, Carlos Quiterio Gómez. Benchmarking Vision Transformers Against Traditional Architectures: a Systematic Analysis for Real-Time *Deepfake* Detection. *In*: IEEE INTERNATIONAL CONFERENCE ON COMPUTING AND MACHINE INTELLIGENCE (ICMI), 4., 2025. **Proceedings [...]**. IEEE, 2025. p. 01-07. Disponível em: https://ieeexplore.ieee.org/document/11141063/authors. Acesso em: 6 jul. 2026.

KHOCHARE, Javani.; JOSHI, Chaitali; YENARKAR, Bakul; SURATKAR, Shraddha; KAZI, Faruk A deep learning framework for audio *deepfake* detection. **Arabian Journal for Science and Engineering**, v. 47, n. 3, p. 3447-3458, 2022. Disponível em: https://*link*.springer.com/article/10.1007/s13369-021-06297-w. Acesso em: 6 jul. 2026.

KIETZMANN, Jan; LEE, Linda W.; MCCARTHY, Ian P.; KIETZMANN, Tim C. *Deepfakes*: Trick or treat? **Business Horizons**, v. 63, n. 2, p. 135–146, 2020. Disponível em: https://doi.org/10.1016/j.bushor.2019.11.006. Acesso em: 6 jul. 2026.

KING, David. **The Commissar Vanishes**: The Falsification of Photographs and Art in Stalin's Russia. New York: Metropolitan Books; Henry Holt and Company, 1997. Disponível em: https://archive.nytimes.com/www.nytimes.com/books/first/k/king-commissar.html. Acesso em: 15 jun. 2026.

KIRCHENGAST, Tyrone. *Deepfakes* and image manipulation: criminalisation and control. **Information & Communications Technology Law**, v. 29, n. 3, p. 308-323, 2020. Disponível em: https://doi.org/10.1080/13600834.2020.1794615. Acesso em: 6 jul. 2026.

KOLLURI, Akaash.; MURTHY, Dhiraj; VINTON, Kami. Quantifying the spread of racist content on fringe social media: A case study of Parler. **Big Data & Society**, v. 12, n. 2, p. 1–19, abr./jun. 2025. Disponível em: https://doi.org/10.1177/20539517251321752. Acesso em: 6 jul. 2026.

KRAETZER, C. *et al.* Process-Driven Modelling of Media Forensic Investigations — Considerations on the Example of *Deepfake* Detection. **Sensors**, v. 22, n. 9, 2022. Disponível em: https://www.mdpi.com/1424-8220/22/9/3137. Acesso em: 6 jul. 2026. DOI: 10.3390/s22093137.

KSHETRI, Nir; VOAS, Jeffrey. The *deepfake* governance gap. **Computer**, v. 59, n. 3, p. 105-112, Mar. 2026. Disponível em: https://www.computer.org/csdl/magazine/co/2026/03/11404325/2egupWkE1kA. Acesso em: 6 jul. 2026.

ŁABUZ, Mateusz; NEHRING, Christopher. *On the way to deepfake democracy? Deepfakes in election campaigns in* 2023. **European Political Science**, v. 23, p. 454–473, 2024. Disponível em: https://*link*.springer.com/article/10.1057/s41304-024-00482-9. Acesso em: 7 jul. 2026.

LAZZARINI, Bruno Justo. **Proteção dos direitos de imagem na produção de conteúdos por Inteligência Artificial Generativa:** *deepfakes*, avatares digitais e os limites do marco jurídico brasileiro. 2025. Trabalho de Conclusão de Curso (Direito) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2025. Disponível em: https://repositorio.pucsp.br/jspui/handle/handle/45762. Acesso em: 6 jul. 2026.

LIU, Baoping; LIU, Bo; ZHU, Tianqing; DING, Ming. A Review of *Deepfake* and Its Detection: From Generative Adversarial Networks to Diffusion Models. **International Journal of Intelligent Systems**, v. 2025, n. 1, p. 1-32, 2025. Disponível em: https://onlinelibrary.wiley.com/doi/10.1155/int/9987535. Acesso em: 6 jul. 2026.

LIU, Yizhi; PADMANABHAN, Balaji; VISWANATHAN, Siva. What exactly is a *deepfake*? **arXiv**, arXiv:2510.22128, out. 2025. Disponível em: https://ui.adsabs.harvard.edu/abs/2025arXiv251022128L. Acesso em: 15 jun. 2026.

LU, Dingyu; LIU, Zihou; ZHANG, Dongming; ZHANG, Jing; JIN, Guoqing. *et al.* Spatial-temporal transformer network for protecting person-of-interest from deepfaking. **Multimedia Systems**, v. 31, 2025. Disponível em: https://*link*.springer.com/article/10.1007/s00530-024-01655-8. Acesso em: 6 jul. 2026. DOI: 10.1007/s00530-024-01655-8.

LUCCA, Bruno; BRAGA, Tomás. Jovens compartilham tutoriais para burlar ECA Digital; site pornô ensina a fraudar bloqueio. **Folha de S. Paulo**, São Paulo, Internet, 10 abr. 2026. Disponível em: https://www1.folha.uol.com.br/cotidiano/2026/04/jovens-compartilham-tutorias-para-burlar-eca-digital-site-porno-ensina-a-fraudar-bloqueio.shtml. Acesso em: 06 jul. 2026.

MAPA da inadimplência e renegociação de dívidas no Brasil. **Serasa**, 2026. Disponível em https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/. Acesso em: 15 jun. 2026.

MBEMBE, Achille. **Crítica da razão negra**. Tradução de Marta Lança. 1. ed. Lisboa: Antígona, 2014. Acesso em: 30 de jun. 2026.

MCGLYN, Clare; TOPARLAK, Rüya Tuna. The 'new voyeurism': criminalizing the creation of '*deepfake* porn'. **Journal of Law and Society**, v. 52, n. 2, p. 204-228, 2025. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/jols.12527. Acesso em: 6 jul. 2026.

MEDING, Kristof; SORGE, Christoph. What constitutes a Deep Fake? The blurry line between legitimate processing and manipulation under the EU *AI Act*. *In*: SYMPOSIUM ON COMPUTER SCIENCE AND LAW (CSLAW), 2025, New York. **Proceedings [...]**. New York: Association for Computing Machinery, p. 152–159, 2025. Disponível em: https://doi.org/10.1145/3709025.3712218. Acesso em 6 jul. 2026.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997.

MORAES, Felipe Soares; BATISTA, Mariana Sousa. Desafios e limitações na detecção de *deepfakes*: uma análise comparativa de ferramentas online. **Revista Panorama-Revista de Comunicação Social**, Goiânia, Brasil, v. 15, n. 1, 2025. Disponível em: https://seer.pucgoias.edu.br/index.php/panorama/article/view/15148. Acesso em: 6 jul. 2026. DOI: 10.18224/pan.v15i1.15148.

MULHER perde mais de R$ 5 milhões ao cair em golpe de namoro com falso Brad Pitt. **Terra**, 14 jan. 2025. Disponível em: https://www.terra.com.br/noticias/mundo/mulher-perde-mais-de-r-5-milhoes-ao-cair-em-golpe-de-namoro-com-falso-brad-pitt,7e5080db558ee2a21b871d71aeba0d35z9fsuyyh.html. Acesso em: 21 jul. 2026.

MUSSALO, Petteri; GAIN, Ulla; HOTTI, Virpi. Types of Data Clarify Senses of Data Processing Purpose in Health Care. In: MANTAS, J. *et al.* (Ed.). Decision Support Systems and Education. [*S. l.*]: IOS Press, 2018. p. 55-59.

NETO, Vital. Alunos de colégio tradicional do Rio usam IA para criar imagens íntimas de meninas; polícia investiga. **CNN Brasil**, Rio de Janeiro, 2 nov. 2023. Disponível em: https://www.cnnbrasil.com.br/nacional/alunos-de-colegio-tradicional-do-rio-usam-ia-para-criar-imagens-intimas-de-meninas-policia-investiga/. Acesso em: 7 jul. 2026.

NGUYEN, Than Thi; NGUYEN, Quoc Viet Hung; NUGYEN, Dung Tien; NGUYEN, Duc Thanh; HUHYNH-THE, Thien; NAHAVANDI, Saeid; NGUYEN, Thanh Tam; PHAM, Quoc-Viet; NGUYEN, Cuong M. Deep learning for *deepfakes* creation and detection. **arXiv**, arXiv:1909.11573, v. 1, n. 2, p. 2, 2019. Disponível em: https://arxiv.org/abs/1909.11573. Acesso em: 6 jul. 2026.

ONU MULHERES BRASIL. Novo relatório da ONU Mulheres destaca autocensura e avanço da violência online contra mulheres jornalistas. **ONU Mulheres Brasil**, 30 abr. 2026. Disponível em: https://brasil.unwomen.org/pt-br/stories/noticia/2026/05/relatorio-ponto-de-virada. Acesso em: 20 jul. 2026.

ONU NEWS. ONU Mulheres aponta riscos da Inteligência Artificial para igualdade de gênero. **ONU News**, 22 jun. 2026. Disponível em: https://news.un.org/pt/story/2026/06/1853511. Acesso em: 7 jul. 2026.

PATEL, Yogesh; TANWAR, Sudeep; GUPTA, Rajesh; BHATTACHARYA, Pronaya; DAVIDSON, Innocent Ewean; NYAMEKO, Royi; ALUVALA, Srinivas; VIMAL, Vrince. *Deepfake* Generation and Detection: Case Study and Challenges. **IEEE Access**, v. 11, p. 143296-143323, 2023. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10354308. Acesso em: 6 jul. 2026.

PIASSI, Paulo. Prefeita de Bauru registra boletim de ocorrência contra veiculação de *deepfake* com seu rosto sobre corpo nu. **G1**, Bauru e Marília, 19 set. 2024. Disponível em: https://g1.globo.com/sp/bauru-marilia/eleicoes/2024/noticia/2024/09/19/prefeita-de-bauru-registra-boletim-de-ocorrencia-contra-veiculacao-de-deep-fakes-com-seu-rosto-sobre-corpo-nu.ghtml. Acesso em: 7 jul. 2026.

PEI, Gan; ZHANG, Jiangning; HU, Menghan; ZHANG, Zhenyu; WANG, Chengjie; WU, Yunsheng; ZHAI, Guangtao; YANG, Jian; TAO, Dacheng. *Deepfake* **generation and detection:** a benchmark and survey. ACM Computing Surveys, v. 58, p. 1–41, 2026. Disponível em: https://doi.org/10.1145/3801962. Acesso em: 6 jul. 2026.

PRABHU, Omkar; NAIK, Sanketh S.; BK, Prarthana; PRABHU, Srikanth; CHADAGA, Krishnaraj. Countering synthetic realities: Advances, challenges, and future directions in *deepfake* image detection. **Systems and Soft Computing**, v. 8, June 2026. Disponível em: https://doi.org/10.1016/j.sasc.2026.200476. Acesso em: 6 jul. 2026.

PREZJA, Fabi.; PALONEVA, Juha.; PÖLÖNEN, Ilkka; NIINIMÄKI, Esko; ÄYRÄMÖ, Sami. *Deepfake* knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. **Scientific Reports**, v. 12, 2022. Disponível em https://www.nature.com/articles/s41598-022-23081-4#citeas. Acesso em: 6 jul. 2026.

QIAN, Hanwey; XIA, Lingling; GE, Ruihao; FAN, Yiming; WANG, Qun; JING, Zhengjun. From Black Boxes to Glass Boxes: Explainable AI for Trustworthy *Deepfake* Forensics. **Cryptography**, v. 9, n. 61, p. 1-19, 2025. Diponível em: https://doi.org/10.3390/cryptography9040061. Acesso em: 6 jul. 2026.

RAHNAMA, Hossein. AI Glossary/Dictionnary. **MIT Media Lab**, 8 jan. 2025. Disponível em: https://www.media.mit.edu/tools/ai-glossary-dictionary/. Acesso em: 4 jun. 2024.

RANI, Geeta; KOTHEKAR, Atharv; PHILIP, Shawn George; DHAKA, Vijaypal Singh; ZUMPANO, Ester; VOCATURO, Eugenio. Lightweight and hybrid transformer-based solution for quick and reliable *deepfake* detection. **Frontiers in Big Data**, v. 8, 2025. Disponível em: https://doi.org/10.3389/fdata.2025.1521653. Acesso em: 6 jul. 2026.

RAUCHFLEISCH, Adrian; VOGLER, Daniel; DE SETA, Gabriele. *Deepfakes* or synthetic media? The effect of euphemisms for labeling technology on risk and benefit perceptions. **Social Media + Society**, v. 11, n. 3, July 2025. Disponível em: https://doi.org/10.1177/20563051251350975. Acesso em: 15 jun. 2026.

RAZA, Shaina; DING, Chen. Fake news detection based on news content and social contexts: a transformer-based approach. **International Journal of Data Science and Analytics**, v. 13, p. 335-362, 2022. Disponível em: https://doi.org/10.1007/s41060-021-00302-z. Acesso em: 6 jul. 2026.

RODRIGUES, Me. Paulo Gustavo. *Deepfakes* pornográficas não-consensuais: a busca por um modelo de criminalização. **Revista Brasileira de Ciências Criminais**, São Paulo, v. 199, n. 199, p. 277–311, 2023. DOI: 10.5281/zenodo.8380977. Disponível em: https://publicacoes.ibccrim.org.br/index.php/RBCCRIM/article/view/267. Acesso em: 7 jul. 2026.

ROE, Jasper; PERKINS, Mike; SOMORAY, Klaire; MILLER, Dan; FURZE, Leon. To *deepfake* or not to *deepfake*: higher education stakeholders' perceptions and intentions towards synthetic media. **arXiv**, v. 1, 2025. Disponível em: https://arxiv.org/pdf/2502.18066. Acesso em: 12 jun. 2026.

ROMAGNA, Duda. Promessa de produto grátis e site falso para pagar frete: como funcionava golpe com *deepfake* de famosos. **G1**, Rio Grande do Sul, 1 out. 2025. Disponível em: https://g1.globo.com/rs/rio-grande-do-sul/noticia/2025/10/01/promessa-de-produto-gratis-e-site-falso-para-pagar-frete-como-funcionava-golpe-com-*deepfake*-de-famosos.ghtml. Acesso em: 11 jun. 2026.

SOUZA, Beto; FIGUEIREDO, Carolina. Grupo usava *deepfake* de Gisele Bündchen para aplicar golpes milionários. **CNN Brasil**, São Paulo, 1 out. 2025. Disponível em: https://www.cnnbrasil.com.br/nacional/sul/rs/golpe-usava-*deepfake*-de-gisele-bundchen-para-aplicar-golpes-milionarios/. Acesso em: 7 jul. 2026.

SAFERNET BRASIL. Mapeamento da SaferNet identifica *deepfakes* sexuais em escolas em 10 dos 27 estados brasileiros. **SaferNet Brasil**, 6 out. 2025. Disponível em: https://new.safernet.org.br/content/mapeamento-da-safernet-identifica-*deepfakes*-sexuais-em-escolas-em-10-dos-27-estados. Acesso em: 6 jul. 2026.

SAGA, Amna; A, Lili N.; KHALID, Fatimah; SANI, Nor Fazlida Mohd; ABDALLA, Hussna E. M.; SYAHPUTRA, Zulfahmi; WIJAYA, Rian Farta. Detecting Low-Quality *Deepfake* Videos Using 3D Residual Vision Transformer. **International Journal of Advanced Computer Science and Applications**, v. 16, n. 12, 2025. Disponível em: https://dx.doi.org/10.14569/IJACSA.2025.0161260. Acesso em: 6 jul. 2026.

SANDOVAL, Maria-Paz; VAU, Maria de Almeida; SOLAAS, John; RODRIGUES, Luano. Threat of *deepfakes* to the criminal justice system: a systematic review. **Crime Science**, v. 13, 2024. Disponível em: https://doi.org/10.1186/s40163-024-00239-1. Acesso em: 6 jul. 2026.

SANTOS JUNIOR, Francisco Alves dos. Inteligência Artificial, memória e reimaginação histórica. *In*: REUNIÃO DE ANTROPOLOGIA DO MERCOSUL, 15., 2025, Salvador. **Anais** [...]. Salvador: UFBA, 2025. Disponível em: https://www.ram2025.sinteseeventos.com.br/arquivo/downloadpublic?q=eyJwYXJhbXMiOiJ7XCJJRF9BUlFVSVZPXCI6XCIyOTEyXCJ9IiwiaCI6ImM1MmFmZTIwZTU0OWI4NTEyYmI5M2YxOWQ3MTUxNjJmIn0%3D. Acesso em: 6 jul. 2026.

SECURITY HERO. 2023 State of *Deepfakes:* realities, Threats, and Impact. **Security Hero**, 2023. Disponível em: https://www.securityhero.io/state-of-*deepfakes*/#key-findings. Acesso em: 11 jun. 2026.

SIEGEL, Dennis; KRAETZER, Christian; SEIDLITZ, Stefan; DITTMANN, Jana. Media Forensic Considerations of the Usage of Artificial Intelligence Using the Example of *Deepfake* Detection. **Journal of Imaging**, v. 10, n. 2, 2024. Disponível em: https://doi.org/10.3390/jimaging10020046. Acesso em: 6 jul. 2026.

SOUTH KOREA. **90 Days Until National Assembly Elections:** Election Campaigns Using *Deepfake* Restricted. Seul, Relatório de imprensa, 02 jun. 2024. Disponível em: https://www.nec.go.kr/site/eng/ex/bbs/View.do?bcIdx=226657&cbIdx=1270. Acesso em: 6 jul. 2026.

SOUTH KOREA. **Act on Special Cases Concerning the Punishment of Sexual Crimes**. Seul: Assembleia Nacional, 2025. Disponível em: https://law.go.kr/LSW/lsInfoP.do?lsiSeq=277347&chrClsCd=010203&urlMode=engLsInfoR&viewCls=engLsInfoR#0000. Acesso em: 6 jul. 2026.

SOUZA, Rafael. **TRE manda tirar do ar e proíbe compartilhamento de vídeo em que candidata a prefeita de Ipojuca supostamente xinga eleitores. G1 Pernambuco**, 16 set. 2024. Disponível em: https://g1.globo.com/pe/pernambuco/eleicoes/2024/noticia/2024/09/16/tre-manda-tirar-do-ar-e-proibe-compartilhamento-de-video-em-que-candidata-a-prefeita-de-ipojuca-supostamente-xinga-eleitores.ghtml. Acesso em: 7 jul. 2026.

STOLL, Ashley. *Shallowfakes* and their potential for fake news. **Washington Journal of Law, Technology & Arts**, 13 jan. 2020. Disponível em: https://wjlta.com/2020/01/13/*shallowfakes*-and-their-potential-for-fake-news/. Acesso em: 12 jun. 2026.

THAMBAWITA, Vajira.; ISAKSEN, Jonas L.; HICKS, Steven A.; GHOUSE, Jonas; AHLBERG, Gustav; LINNEBERG, Allan; GRARUP, Niels; ELLERVIK, Christina; OLESEN, Morten Salling; HANSEN, Torben; GRAFF, Claus; HOLSTEIN-RATHLOU, Niels-Henrik; STRÜMKE, Inga; HAMMER, Hugo L.; MALECKAR, MARY M.; HALVORSEN, Pal; RIEGLER, Michael A.; KANTERS, Jorgen K. *Deepfake* electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. **Scientific Reports**, v. 11, 2021. Disponível em: https://doi.org/10.1038/s41598-021-01295-2. Acesso em: 6 jul. 2026.

TIU, Ekin. Understanding Latent Space in Machine Learning. **Medium**, 4 fev. 2020. Disponível em: https://medium.com/data-science/understanding-latent-space-in-machine-learning-de5a7c687d8d. Acesso em: 22 jun. 2026.

TWOMEY, John; CHING, Didier; AYLETT, Matthew Peter; QUAYLE, Michael; LINEHAN, Conor; MURPHY, Gillian. What is so deep about *deepfakes*? A multi-disciplinary thematic analysis of academic narratives about *deepfake* technology. **IEEE Transactions on Technology and Society**, preprint, 2024. Disponível em: https://www.researchgate.net/publication/385916734. Acesso em: 15 jun. 2026.

UN WOMEN. **Reports to police of online violence against women journalists double since 2020, with one in four experiencing related anxiety and/or depression**. New York: UN Women, 30 abr. 2026.

Disponível em: https://www.unwomen.org/en/news-stories/press-release/2026/04/reports-to-police-of-online-violence-against-women-journalists-double-since-2020-with-one-in-four-experiencing-related-anxiety-andor-depression. Acesso em: 7 jul. 2026.

UNICEF. Artificial Intelligence and Child Sexual Abuse and Exploitation. **Unicef**, February 2026. Disponível em: https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_2.pdf. Acesso em: 24 jun. 2026.

UNITED KINGDOM. **Sexual Offences Act 2003**. Londres: Congresso Nacional, 2003. Disponível em: https://www.legislation.gov.uk/ukpga/2003/42/part/1/crossheading/other-offences. Acesso em: 06 jul. 2026.

UNITED KINGDOM. **Data (Use and Access) Act 2025**. Londres: Congresso Nacional, 2025. Disponível em: https://www.legislation.gov.uk/ukpga/2025/18/section/138. Acesso em: 06 jul. 2026.

UNITED STATES OF AMERICA. S. 3805: **Malicious *Deep Fake* Prohibition Act of 2018. To amend title 18, United States Code, to prohibit certain fraudulent audiovisual records, and for other purposes**. Washington, DC: Senado dos Estados Unidos, 21 dez. 2018. Disponível em: https://www.govinfo.gov/app/details/BILLS-115s3805is. Acesso em: Acesso em: 10 jun. 2026.

UNITED STATES OF AMERICA. H.R. 3230: ***DEEP FAKES* Accountability Act. To combat the spread of disinformation through restrictions on deep-fake video alteration technology**. Washington, DC: Câmara dos Representantes dos Estados Unidos, 12 jun. 2019. Disponível em: https://www.govinfo.gov/app/details/BILLS-116hr3230ih. Acesso em: Acesso em: 10 jun. 2026.

UNITED STATES OF AMERICA. *Take it Down Act*. An act to require covered platforms to remove nonconsensual intimate visual depictions, and for other purposes. Washington: Congresso Nacional, 2025. Disponível em: https://www.congress.gov/119/bills/s146/BILLS-119s146es.pdf. Acesso em: 06 jul. 2026.

VALENTE, Mariana. **Misoginia na internet:** uma década de disputas por direitos. São Paulo: Fósforo, 2023.

WANG, Tianyi; CHENG, Harry; CHOW, Kam Pui; NIE, Liqiang. Deep Convolutional Pooling Transformer for *Deepfake* Detection. **ACM Transactions on Multimedia Computing, Communications and Applications**, v. 19, n. 6 p. 1-20, 2022. Disponível em: https://doi.org/10.1145/3588574. Acesso em: 6 jul. 2026.

WEIKMANN, Teresa; EGELHOFER, Jana Laura; LECHELER, Sophie. Beyond Credibility: The Effects of Different Forms of Visual Disinformation. **Journalism & Mass Communication Quarterly**, v. 102, n. 4, 2025. Disponível em: https://journals.sagepub.com/doi/10.1177/10776990251357299. Acesso em: 15 jun. 2026.

WEIKMANN, Teresa; LECHELER, Sophie. Visual disinformation in a digital age: A literature synthesis and research agenda. **New Media & Society**, v. 25, n. 12, dez. 2022. Disponível em: https://journals.sagepub.com/doi/10.1177/14614448221141648. Acesso em: 15 jun. 2026.

WILLIAMS, Emily; RODGERS, Jason; CHANDLER-CRNIGOJ, Sebastian; JONES, Karl; ROBINSON, Colin. Pandora's Black or White Box: Are AI Tools Undermining Evidence? *In*: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND TECHNOLOGIES (CompSysTech), 2025, Ruse. **Proceedings [...]**. Ruse, Bulgária, 2025. p. 1–6. Disponível em: https://ieeexplore.ieee.org/document/11137140. Acesso em: 6 jul. 2026 .

WITNESS. Backgrounder: *deepfakes* in 2022. **Witness**, Mar. 2022. Disponível em: https://lab.witness.org/backgrounder-*deepfakes*-in-2022/. Acesso em: 1 jun. 2026.

ANPD
Agência Nacional de Proteção de Dados
t. (61) 2025-8101
www.gov.br/anpd

www.gov.br/anpd

ANPD

Agência Nacional de
Proteção de Dados