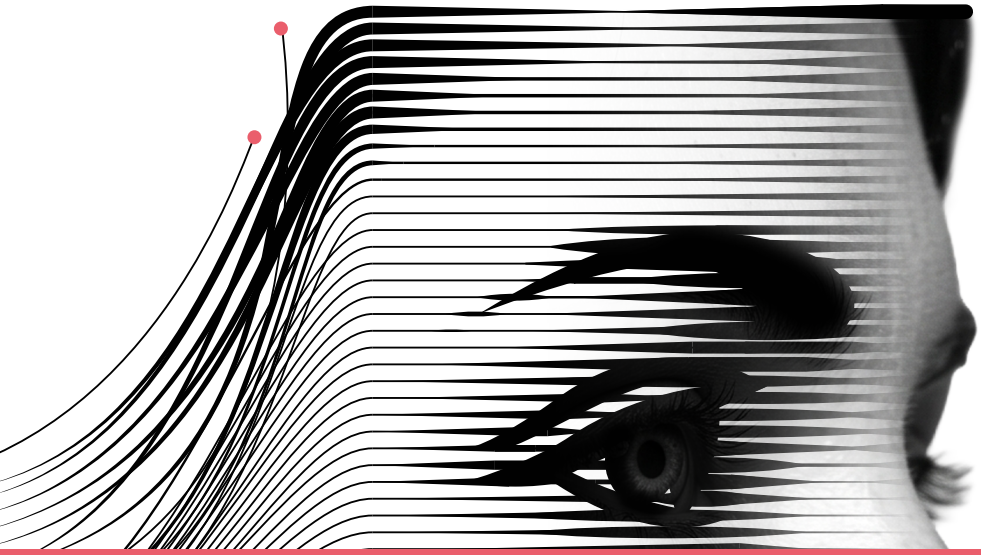
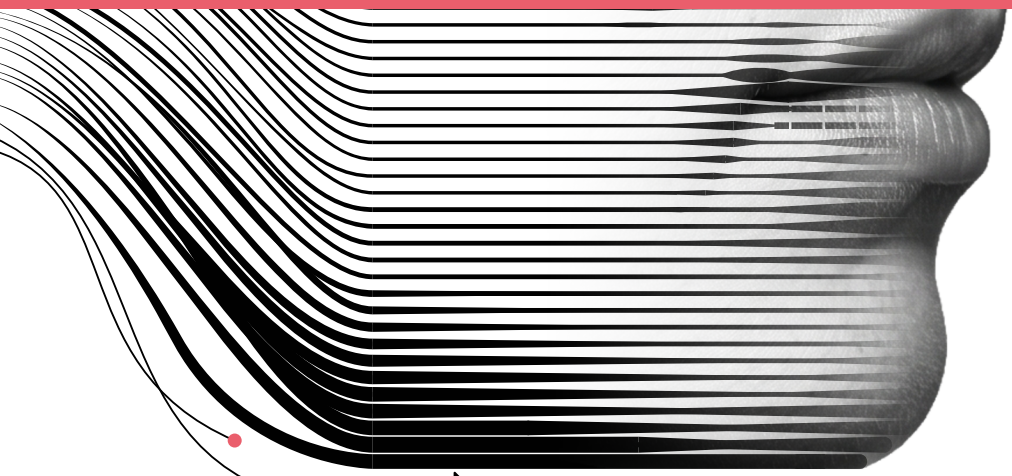


‹ technology radar ›

deepfakes



ANPD

National Data Protection Agency

‹ technology radar ›

no. 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
National Data Protection Agency

Director-President
Waldemar Gonçalves Ortunho Junior

Directors
Iagê Zendron Miola
Lorena Giuberti Coutinho
Miriam Wimmer

Project Coordination and Administration
Superintendency of Technological Innovation (SITEC)
Ticianne de Gois Ribeiro Darin
Angela Halen Claro Franco
Lucas Costa dos Anjos

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

Translators
Ticianne de Gois Ribeiro Darin
Gabriela Campos Alkmin
Ingrid David Alves de Carvalho

Graphic Design / Cover
André Scofano Maia Porto

Desktop Publishing
André Scofano Maia Porto
Bruna Assi Hernandes
Júlia Câmara Cunha

Graphic Design
Ewerton Luiz Costadelle

Cataloging in Publication (CIP)

A2656d

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

158 p.il.: Color. (Technology Radar; n. 6)

ISBN 978-65-82658-03-7

Includes bibliographical references.

1. Deepfakes. 2. Generative artificial intelligence. 3. Personal data protection. I. Agência Nacional de Proteção de Dados - ANPD. II. Title.

CDU 004.62:004.8

Cataloging data prepared by librarian
Angela Halen Claro Franco, CRB1- S023

How to cite this publication:

ANPD. **Deepfakes.** Brasília, DF: ANPD, 2026. (Technology Radar, n. 6). Available at: include link to the document. Accessed on: include date of access to the document.

ANPD
SCN, Block 6, Complex A,
Venâncio Building 3000, Block A, 9th floor
Brasília, DF · Brazil · 70716-900
t. +55 (61) 2025-8101
www.gov.br/anpd

‹ about the series ›

The "Technology Radar" series is a periodic publication by ANPD. It provides concise analyses of emerging technologies that affect, or are expected to affect, data protection in Brazil and abroad.

The series does not aim to cover every aspect of each topic or to establish an institutional position. Instead, it offers information to support the public debate on data protection in Brazil, using accessible, educational texts.

For each topic, the main concepts, potential applications, and prospects are addressed, always with an emphasis on data protection in the Brazilian context. ▪

‹ list of acronyms ›

ANN – Artificial Neural Network
ANPD – National Data Protection Agency
API – Application Programming Interface
AVAMEC – Virtual Learning Environment of the Ministry of Education
BNCC – National Common Curricular Base
C2PA – Coalition for Content Provenance and Authenticity
CDC – Consumer Protection Code
CNN – Convolutional Neural Network
CPF – Individual Taxpayer Registry
CSIM – Cosine Similarity for Facial Embedding
DCGAN – Deep Convolutional Generative Adversarial Network
EBM – Energy-Based Model
EBEM – Brazilian Media Education Strategy
ECA – Statute of the Child and Adolescent
ELBO – Lower Bound of Evidence
ENISA – European Union Agency for Cybersecurity
U.S. – United States of America
FID – Fréchet Inception Distance
GAN – Generative Adversarial Network
GPS – Global Positioning System
GPU – Graphics Processing Unit
AI – Artificial Intelligence
IS – Inception Score
KL – Kullback-Leibler (Kullback-Leibler Divergence)
LFW – Labeled Faces in the Wild
LGPD – Brazilian Data Protection Law
LLM – Large Language Models
LSE-C – Lip Sync Error – Confidence
LSE-D – Lip Sync Error – Distance
MEC – Ministry of Education
MPF – Federal Public Prosecutor's Office
NeRF – Neural Radiance Field
NLP – Natural Language Processing
UN – United Nations
PE – State of Pernambuco

PII – Personally Identifiable Information
PSNR – Peak Signal-to-Noise Ratio
RIPD – Personal Data Protection Impact Report
RNN – Recurrent Neural Network
Senacon – National Consumer Secretariat
SSIM – Structural Similarity Index
STF – Supreme Federal Court
TSE – Superior Electoral Court
TTS – Text-to-Speech
UNICEF – United Nations Children's Fund
VAE – Variational Autoencoder
VAEBM – Energy-Based Variational Autoencoder
VC – Voice Conversion
ViT – Vision Transformers
VPN – Virtual Private Network
XAI – Explainable Artificial Intelligence

‹ summary ›

11
1 › introduction

14
2 › key concepts

44
3 › how deepfakes work

74
4 › deepfakes and personal data protection

93
5 › deepfakes in the brazilian context

117
6 › prospects

140
7 › final considerations

143
references

1 introduction › 11

2 key concepts › 14

2.1 Umbrella concepts: synthetic content and media › 15

2.2 Deepfakes › 17

2.2.1 Definition › 17

2.2.2 Classification by use › 23

2.2.3 Cheapfakes and shallowfakes › 30

2.3 What is not a deepfake › 33

2.4 Non-malicious use of synthetic media › 35

2.5 The role of technical definitions in data protection › 40

3 how deepfakes work › 44

3.1 Foundations of the algorithms that produce deepfakes › 47

3.1.1 Machine learning › 48

3.1.2 Deep learning › 49

3.2 Key technical approaches for deepfake generation › 53

3.2.1 Generative Adversarial Networks (GANs) › 54

3.2.2 Variational Autoencoders (VAEs) › 57

3.2.3 Diffusion models › 60

3.2.4 Discriminative Neural Networks (DNN) and Convolutional Neural Networks (CNN) › 63

3.2.5 Neural Radiance Field (NeRF) › 65

3.3 Deepfakes detection › 66

3.4 Technical limitations, uncertainties, and biases › 68

4 deepfakes and personal data protection › 74

4.1 Identification of personal data processed › 76

4.2 How is personal data collected, processed, and shared? › 81

4.3 Connection to the principles of the LGPD › 83

4.4 Specific risks related to technical operations › 87

5 deepfakes in the brazilian context › 93

5.1 Pornographic deepfakes targeting women and girls › 95

5.2 Deepfakes and racism › 100

5.3 Deepfakes in the electoral context › 103

5.4 Deepfakes as a tool of political gender-based violence › 106

5.5 Deepfakes used in financial scams › 108

5.6 Deepfakes for deceptive advertising › 114

6 prospects › 117

6.1 Legal and regulatory perspective › 118

6.2 Information and media literacy › 132

6.3 Mitigation tools and possible solutions › 136

7 final considerations › 140

references › 143



‹ 1 introduction ›

Have you ever watched a video thinking it was a real situation, only to be told afterwards that the content had actually been generated by artificial intelligence (AI)?[1] Or have you ever received an audio message on WhatsApp with the voice of someone you know, only to find out later that it wasn't actually from them? Or have you ever come across an intimate video repeatedly shared on social media, involving a famous person who claims the content is fake? Or have you ever come across a viral audio clip of a political candidate claiming that something they said was digitally manipulated? Situations like these have become increasingly common with the advent of deepfakes — artificially created or manipulated media that, due to their high degree of authenticity, appear real to the general public.

The creation and editing of fake content, whether analog or digital, is nothing new. The use of photomontages and image manipulation as early as the beginning of the 20th century, especially in political and propaganda contexts (King, 1997), is an example of this. However, in recent years, driven by advances in artificial intelligence and greater access to tools capable of generating synthetic content, these activities have taken on a new scope. Due to the techniques used to create them, as well as their realism, deepfakes are increasingly being used for fraud, scams, identity theft, data breaches, mass dissemination of false information, and production of pornographic content, which accounted for 98% of all deepfakes in circulation in 2023 (Security Hero, 2023).[2] Due to these capabilities and their potential to amplify social engineering attacks, disinformation campaigns, and other forms of digital manipulation, deepfakes have become a major concern for international organizations and global cybersecurity reports, such as the World Economic Forum's Global Cybersecurity Outlook 2025, the 2025 ENISA Threat Landscape, and Europol's 2024 report on the challenges of deepfakes for law enforcement.

1 *Artificial intelligence is the field of human knowledge that studies the development of systems capable of improving their own performance based on the data they receive. These systems generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (ANPD, 2024).*

2 *In 2023, the organization Security Hero analyzed 95,820 deepfake videos, 85 channels dedicated to this type of media on various platforms, and 100 websites linked to its ecosystem. The report indicates that 98% of the deepfakes in circulation were pornographic in nature and that 99% of the victims were women. Security Hero is not an academic or scientific research institution. Even so, researchers and international regulators widely use its report as a reference for mapping the uses of this media. That is why it is cited here.*

Initially limited to academic settings or experimental applications, these resources began to circulate widely on social media, often as manipulated videos or audio clips, featuring public figures. The popularization of the term and the practice on social media originated with an anonymous Reddit user who used the term "deepfake" to describe the use of deep learning techniques in a code he shared to create fake content. This code enabled users to insert images of celebrities into adult videos without their consent. Thereafter, interest among users of the same platform in producing this type of material led to the mass production of fake content.

The first targets of deepfakes included actors (e.g., Emma Watson and Scarlett Johansson), singers (e.g., Katy Perry), and politicians (e.g., U.S. Presidents Barack Obama and Donald Trump), whose faces were superimposed — without their permission — onto those of other people (Kietzmann *et al.*, 2020, p. 136). This phenomenon has, on the one hand, helped raise public awareness of this technology and, on the other, intensified concerns about misinformation, the manipulation of public opinion, and damage to individuals' reputations.

This scenario makes critical reflection and public awareness about deepfakes essential, particularly when it comes to digital security, duty of care, and information integrity. Understanding deepfakes and recognizing the risks associated with them helps develop critical thinking when consuming digital content, contributing to digital literacy, and fostering ethical reflection on the use of technology. This understanding is also urgent for personal data protection. Manipulating faces and voices involves processing sensitive personal data and demands a regulatory approach that safeguards fundamental rights. Thus, disseminating this knowledge helps strike a balance between innovation and responsibility, protecting society against malicious uses — particularly regarding vulnerable groups, such as women and children.

This Technology Radar therefore aims to present and explain the main concepts related to deepfakes, as well as to identify potential risks posed by these tools to privacy, data protection, and the creation of secure digital environments. To map the debates, technologies, and regulatory needs surrounding this topic, we reviewed articles, technical documents, and reports published between 2017 and 2026 — a period spanning from the term's first use to the current stage of consolidation of these technologies. In addition to this introduction, the study presents a technical characterization of deepfakes and related concepts, followed by an analysis of cases and examples from the Brazilian context for reflection, and finally discusses prospects for the application and regulation of deepfakes.

This study has a technical, informative, and prospective purpose. It systematizes concepts, risks, and trends related to deepfakes from the perspective of personal data protection. The document does not constitute regulatory guidance, an administrative decision, or a conclusive statement by the National Data Protection Agency (ANPD) regarding specific cases. The Agency's activities in this area are limited by its legal authority. They focus, in particular, on the processing of personal data under the Brazilian Data Protection Law (LGPD), without prejudice to institutional cooperation with other competent agencies and authorities.

‹ 2 key concepts ›

The manipulation of visual and auditory media is a documented practice throughout the evolution of the media, ranging from analog photomontages in the 19th century to digital techniques for real-time image and voice enhancement, such as social media filters and software that corrects voice intonation (Auto-Tune). The emergence of deepfakes, in turn, marks a tipping point in terms of the scale, accessibility, and realism of creating doctored content. This technology has spread as a result of recent advances in artificial intelligence and machine learning, which have enabled automated procedures to generate high-fidelity images, videos, and audio, making it difficult for human observers to detect forgeries (Kietzmann *et al.*, 2020, p. 135). Examples of these advances include the increasing maturity of generative AI[3] itself and the development of technologies that enable the creation of highly realistic synthetic content[4], such as autoencoders, Generative Adversarial Networks (GANs), and diffusion models, which we will discuss later.

The expansion of access to tools for generating synthetic images and videos has lowered the technical barriers to this process, enabling people without specialized training to create such content quickly (Irfan *et al.*, 2025, p. 1928). The indiscriminate use of this technology, though, can have substantial impacts on society, including risks related to identity fraud, security threats, the widespread dissemination of misleading content, and privacy violations (Pei *et al.*, 2026, p. 273). On the other hand, technologies that enable the creation of deepfakes have the potential to be applied in sectors such as entertainment, film production, and the creation of digital human images (Pei *et al.*, 2026, p. 273).

Given this scenario, it is essential to conceptually structure this ecosystem. This section aims to describe and organize the fundamentals of various types of deepfakes and their

3 *Generative artificial intelligence (AI) is a subfield of deep learning. Generative models describe how a dataset is generated in terms of a probabilistic model, and can create data through sampling (ANPD, 2024, p. 7).*

4 *Synthetic content is the set of information or data artificially generated by computer systems, in direct contrast to real data, which originates from physical reality or direct human interactions (ANPD, 2024).*

ramifications, to reduce theoretical and classificatory ambiguities. To this end, we will begin by addressing the concept of synthetic media, and then move on to defining deepfakes, presenting their classifications in terms of applications and types of use. We will also discuss what does not constitute a deepfake, distinguishing algorithmic alterations from simpler audiovisual manipulations or basic contextual edits. This technical distinction will allow us to clearly map the different forms of highly complex forgery, with special attention to voice cloning.

Defining these structural boundaries will help distinguish the non-malicious use of synthetic media from deceptive and unlawful manipulations. We will then discuss the limits that, when violated, render the use of deepfakes abusive, and the ways in which an understanding of these technical definitions can impact their regulation. We will thus demonstrate that conceptual clarity can contribute to how the law and public institutions should structure frameworks for protection, accountability and governance in cyberspace.

2.1 Umbrella concepts: synthetic content and media

According to the Technology Radar on Generative Artificial Intelligence, we can understand synthetic content as a product of generative AI models (ANPD, 2024, p. 11). The publication explains that, technically, these models do not simply copy the data used to train them. Instead, they identify patterns and statistical characteristics present in real datasets. Based on these patterns, the system uses probabilistic sampling processes to generate content. This content retains properties similar to those of the original data but consists of entirely new examples that did not exist before.

Thus, **"synthetic content"** is a broad technical term that refers to any artificially generated or manipulated content — that is,

content that does not stem directly from a real event or data, but rather from an artificial production process. Such content can be created using different tools and techniques, depending on the purpose. In this document, we focus on synthetic content generated by computational means, as this is the relevant domain for the phenomenon of deepfakes. Specifically in the context of deepfakes, this content is produced using generative AI techniques, regardless of whether it consists of structured data — such as tables and records — or media content — such as images, audio, text, and videos (ANPD, 2024, p. 11). The term **"synthetic media,"** on the other hand, refers to the application of this content in audiovisual and communicative formats, such as images, videos, audio, and other materials intended for communication. Deepfakes are, therefore, a subcategory of synthetic media (CNTI, 2025).

In 2021, the European Parliament conducted a study and, based on it, published a report titled "Tackling deepfakes in European policy," which focuses on deepfake videos, voice cloning, text synthesis, and 3D animations (European Parliament, 2021, p. 5). According to the report, the evolution of media creation and manipulation capabilities resulted from three main developments: the creation of algorithms capable of automatically mapping facial features; the widespread availability of audiovisual data on the internet; and the corresponding increase in digital forensic capabilities in response to the growing need to detect forgeries. This scenario has enabled the emergence of artificial intelligence models that use a continuous learning cycle to create realistic simulations and learn to correct flaws based on techniques developed for forensic detection (European Parliament, 2021, p. 5).

Traditionally, creating highly realistic synthetic media required large amounts of data and significant processing time. However, advances in generative models and discriminative networks — which validate data by distinguishing real records from synthetic content — have substantially expanded the capacity to generate images and videos, making it possible to

produce content that is increasingly convincing and similar to authentic records (Pei *et al.*, 2026, p. 3, 5). Despite this technological progress, expert analyses indicate that, in practice, the most immediate risks often stem from decontextualized content or simpler forms of manipulation, presented in a way that supports misleading narratives (CNTI, 2025). In this context, we outline in the next section the different ways of defining deepfakes, as well as their main types.

2.2 Deepfakes

2.2.1 Definition

Simply put, the term "deepfake" is used to describe audio, photos, and videos that have been modified or artificially created using artificial intelligence (CNTI, 2025; Twomey *et al.*, 2024, p. 1). However, what truly characterizes deepfakes and distinguishes them within the broader category of synthetic media is that the result is a hyper-realistic simulation of human identity and behavior (European Union, 2024; Witness, 2022).

There are different types of deepfakes. Voice cloning, also known as audio deepfakes, uses artificial intelligence to create highly realistic imitations by combining voice tone reproduction with text synthesis to mimic a person's vocabulary and style, thereby increasing the clone's credibility (European Parliament, 2021, p. 6). Text synthesis, on the other hand, is used to create deepfakes by applying Natural Language Processing (NLP) techniques to reproduce a person's communication style based on the analysis of large volumes of text associated with the individual (the target), such as transcripts, in which the systems identify linguistic patterns, expressive nuances, and intentions. Based on this data, artificial intelligence learns to mimic a person's characteristic way of expressing themselves to generate new content (European Parliament, 2021, p. 6).

Deepfakes can arise both from synthetic generation processes via prompts — in which there is no direct correspondence with existing data — and from processes involving **the manipulation of real data** (Meding; Sorge, 2025, p. 5), such as face swapping, altering facial expressions, or modifying contexts. To highlight the extent of these alterations in real-world records, the Witness Organization provided examples of various forms of content manipulation, such as: the addition or removal of objects in videos; the alteration of the setting, such as weather conditions; the simulation of facial or bodily movements; the artificial synchronization of speech with lip movements; the generation and modulation of voices to imitate specific individuals; and the creation of realistic images of nonexistent people or events, including those based on textual descriptions (Witness, 2022).

According to the organization Witness (2022), face swapping in videos is the most common form of deepfakes. Altuncu, Franqueira, and Li (2024, p. 2) complement this view, extending the classification beyond video media to include static images of fake faces and speech forgeries. The authors also highlight hybrid videos, which combine different types of forgery by integrating manipulated images and synthetic speech (audio) into a single file.

Although the term "deepfake" is very popular, the scientific community is still debating the exact boundary between what is and what is not a deepfake (Altuncu; Franqueira; Li, 2024, p. 2; CNTI, 2025). In a survey of 826 published scientific articles, researchers at the University of Maryland identified three dimensions that help define deepfakes in a more comprehensive way, depending on the source of the identity, the purpose, and the extent of the alteration, as illustrated in the diagram in Figure 1 (Liu; Padmanabhan; Viswanathan, 2025, p. 4). According to these authors' perspective, deepfakes can be created in three ways: Original-to-Target Transfer (A → B), in which one identity completely replaces another (e.g., face swapping); Modification of the Original (A → A'), in which

attributes are altered while preserving the core identity (e.g., aging, change in facial expression); and Complete Synthesis (ø → B), which involves the creation of entirely artificial identities with no real-world counterparts (e.g., synthetic actors).

Liu, Padmanabhan, and Viswanathan (2025, p. 5) add two further dimensions to this distinction. The first of these is "intent," which distinguishes the purpose behind the creation of this type of content. Intent can involve deceptive applications designed to mislead people about authenticity (e.g., disinformation, fraud) or non-deceptive uses created for legitimate purposes, with transparency regarding their artificial nature (e.g., entertainment, research, education). The second dimension is the "granularity of manipulation," which characterizes the extent of the alterations made. These can be holistic alterations, that is, involving comprehensive changes that affect the entire identity or scene (e.g., complete replacement of the identity) — or more targeted modifications, involving selective alterations to specific attributes or characteristics (e.g., changing only eye color or facial expression).

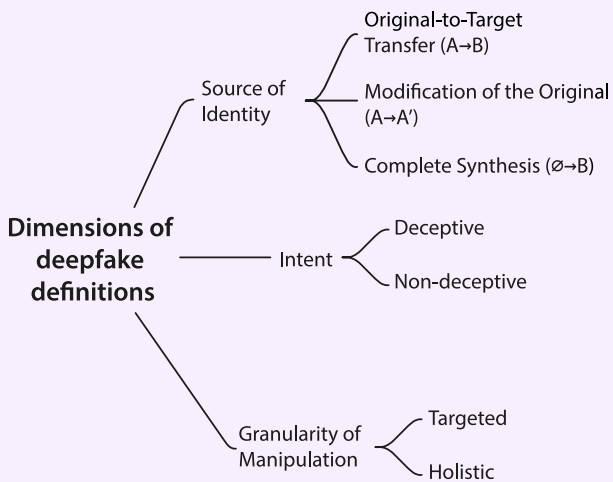


Figure 1 *Three main dimensions of deepfake definitions*

Source: Prepared by the authors, based on the conceptualization by Liu, Padmanabhan, and Viswanathan (2025).

The intersection of these three variables yields different possible definitions in the academic literature. Table 1 summarizes how researchers combine these criteria to classify and characterize deepfakes.

Table 1 *Examples of different definitions of deepfakes by dimension*

Source	Intent	Granularity	Applications	Summary Definition
A → B	Deceptive	Holistic	Disinformation and identity theft (Nguyen *et al.*, 2019)	AI-driven technology designed to deceive, in which entire identities are transferred from an original source to a target using holistic manipulation, with the aim of inducing individuals to believe in their authenticity.
A → B	Deceptive	Targeted	Targeted manipulation attacks (Prezja *et al.*, 2022)	AI-generated content in which specific target attributes from a source are manipulated to create misleading media, focusing on precise alterations of specific characteristic, such as voice, facial tone, facial features, or expressions.
A → B	Non-deceptive	Holistic	Research and entertainment (Eberl *et al.*, 2022)	Technology that transfers complete identities from original sources to targets with non-deceptive intent, used for legitimate purposes, including academic, research, entertainment, and educational applications.
A → A′	Deceptive	Targeted	Attribute-level forgery (Khochare *et al.*, 2021)	Technology that generates synthetic content by modifying specific attributes of the original data, designed to deceive by targeting specific characteristics, such as facial biological features.

Source	Intent	Granularity	Applications	Summary Definition
A → A′	Non-deceptive	Holistic	Research enhancement (Thambawita *et al.*, 2021)	Technology that modifies original data to generate synthetic content without the intent to mislead, focusing on creating realistic alternatives for medical, educational, or research fields.
A → A′	Non-deceptive	Targeted	Content enhancement (Herring *et al.*, 2022)	Technology that creates synthetic content by modifying specific attributes of the original data for legitimate purposes, such as beauty filters and entertainment.
Ø → B	Deceptive	Holistic	Synthetic identity creation (Xie *et al.*, 2024)	Technology that creates entirely synthetic identities without relying on a real-world counterpart, designed to deceive viewers into believing that the fabricated content represents authentic individuals or events.
Ø → B	Non-deceptive	Holistic	Legitimate synthetic content (Kaate *et al.*, 2024a)	Technology that creates entirely new synthetic content with transparent and non-misleading intent, designed for legitimate applications in entertainment, research, or creation.

Source: Prepared by the authors based on Liu, Padmanabhan, and Viswanathan (2025, p. 7–8).

In technical terms, the various definitions of deepfake share fundamental commonalities. The survey conducted by Liu, Padmanabhan, and Viswanathan (2025, p. 39) shows that all the scientific articles analyzed characterize deepfakes as the modification or creation of synthetic content using artificial intelligence. The common objective is to alter media to generate alternative representations of individuals, voices, or events.

There is no consensus, though, among the definitions regarding the origin of the identity, the creator's objectives, and the extent of the modification. This multiplicity of ways of viewing and defining deepfakes poses various regulatory challenges, as we will discuss in **Section 6.**

The definition that best aligns with the perspective adopted in this document is the first one in the table above (Nguyen *et al.* (2019), **which focuses on simulating or replacing the identity of real people** and, potentially, causing widespread deception due to high technical verisimilitude. However, choosing this approach does not imply that altering a personal attribute (A → A'), or even the facial synthesis of a fictional person (Ø → B), is not, technically speaking, a deepfake.

Since the regulatory, legal, and social urgency is concentrated on techniques that reuse, extract, and manipulate biometric data from real individuals for deceptive purposes, this report also adopts, in a complementary manner, the perspectives of Khochare *et al.* (2021), which focus **on the malicious modification of specific attributes**, and of Xie *et al.* (2024), which refers to **facial synthesis**. The former relies on data captured from real individuals and has a high potential to simulate authentic individuals for deceptive purposes; the latter, while not corresponding to any specific person, has a high potential to simulate events and facts as if they were authentic, contributing to the same type of collective deception.

Thus, given the scope of the ANPD's activities, throughout this document we consider that deepfakes involve both the holistic transfer of identity (A → B) and the manipulation of an individual's attributes (A → A'), as well as the facial synthesis of a character (Ø → B), all used with the intent to harm the public. Cases addressing each of these perspectives will be discussed in **Sections 4, 5,** and **6**.

2.2.2 Classification by use

The literature classifies deepfakes into different categories based on the type of alteration or generation that the system performs on digital media (Irfan *et al.*, 2025, p. 1930). This approach organizes these technologies according to the practical objective of the manipulation, that is, according to the visual element the algorithm is designed to alter or generate. Within this framework, deepfake technologies can be grouped according to the behavior of synthetic media into the following operational categories: face swapping, face reenactment, facial attribute manipulation, and facial synthesis.

In the face-swapping category, the system replaces one person's facial features with those of another, while preserving the body and the setting from the original recording (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 4). In face reenactment, the individual is not replaced, but their actions and attitudes are falsified. In the manipulation or editing of facial attributes, the tool edits specific physical characteristics, such as aging the skin, changing gender, or modifying the hair (Irfan *et al.*, 2025, p. 1931; Pei *et al.*, 2026, p. 5). Finally, in facial synthesis, the tool performs a generation process that does not require the use of any base image, enabling it to either fabricate entirely new facial features and identities for fictional human beings, or create photorealistic representations of people who exist in the real world (Irfan *et al.*, 2025, p. 1930).

This section describes the characteristics and functioning of each category, as well as examples of risks associated with applications designed to deceive people in different contexts. Some synthetic media applications serve lawful and socially beneficial purposes, which are addressed in **Section 2.4**, while more detailed examples of malicious uses of these technologies will be discussed in **Sections 4** and **5**. The technological foundations that enable the different types of deepfakes presented below will be explained in detail in **Section 3**.

Face Swapping

Face swapping is the best-known category of deepfake manipulation and the easiest to visualize in practice. The system extracts a person's face — referred to as the source — and transfers it onto the body of another person — referred to as the target — in a photograph or video. In this process, the background, clothing, and body movements of the original subject remain intact; only the facial identity of the person appearing on screen is replaced (Pei *et al.*, 2026, p. 4). Face swapping can be illustrated by the following figure:

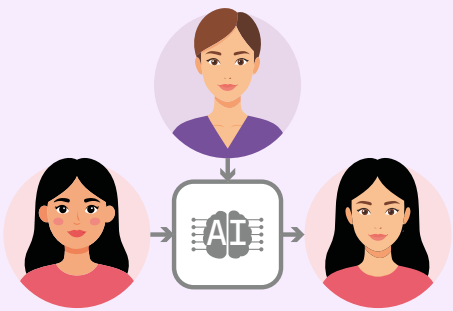


Figure 2 *Schematic illustration of face swapping*

Source: Prepared by the authors.

In this case, the level of realism generated by the algorithm is so high that the new face adapts to the scene's lighting and mimics the original expressions, blinking and smiling in sync with the target video. To illustrate this, Kietzmann *et al.* (2020, p. 137) cite the famous case of the video in which actor and comedian Jim Carrey's face was superimposed onto actress Alison Brie's body during an interview on the television show *Late Night with Seth Meyers*. Anyone watching the modified video might have been under the illusion that it was Jim Carrey giving the interview, as the algorithm mapped and adapted the actor's facial features to the actress's gestures with high precision. Although it has legitimate applications in the audiovisual and entertainment industries, face swapping is a technology widely used for criminal and abusive practices.

In the advertising market, unauthorized manipulation facilitates cyber fraud. The faces and voices of celebrities, journalists, and well-known TV hosts have been cloned by algorithms and inserted into fake ads on social media to sell deceptive products or promote scams (Fukuha; Archegas; Fürst, 2026, p. 187). In these situations, the actor or public figure not only loses control of their digital identity but also faces serious risks of having their credibility and reputation associated with crimes (Fukuha; Archegas; Fürst, 2026, p. 187).

Among the various serious and harmful practices made possible by the use of face-swapping technology, the creation of non-consensual pornography warrants particular attention given the severity of its impact on victims and its rising prevalence. In this type of conduct, which is considered a crime in Brazil[5], perpetrators replicate the faces of famous actresses — or even ordinary women — and superimpose them onto the bodies of adult film actresses (Kietzmann *et al.*, 2020, p. 142; Kirchengast, 2020, p. 2). The result is misleading media that severely violates the victim's image and privacy and may be mistaken for authentic content by the public.

Face reenactment

In face reenactment, put simply, the system begins the process by observing a video of the source person (the one performing the action). It then captures the gestures that the person makes, such as turning their head, blinking, and moving their mouth. This creates a sort of template of the expressions that will be used in the editing process (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). Next, the system analyzes the image of the target person (the one being manipulated) and separates the physical features that identify them from the expressions they display in the original photo. This ensures that the target person's face is not distorted and retains its unique characteristics when the new movements are applied.

5 *Since 2018, Article 218-C of the Brazilian Penal Code has criminalized the acts of "offering, exchanging, making available, transmitting, selling or displaying for sale, distributing, publishing, or disseminating, by any means — including through mass media, computer systems, or telecommunications — photographs, videos, or other audiovisual recordings that contain scenes of rape or the rape of a vulnerable person, or that glorify or encourage such acts; or, without the victim's consent, scenes of sex, nudity, or pornography" (Brazil, 1940).*

Finally, the system applies the gestures copied from the source person directly to the target person's face. Thus, the target person's image begins to reproduce the actions from the other video. To make the result appear realistic, the system reviews and corrects the image several times. It eliminates visual flaws until the movements appear natural and the generated video looks authentic. A diagram illustrating the face reenactment process is shown in the following figure:

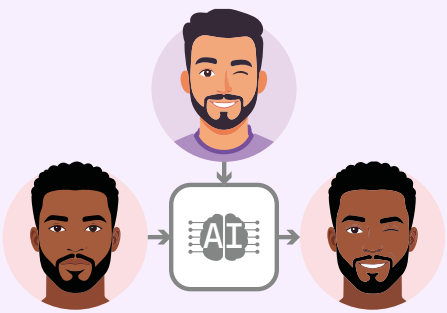


Figure 3 *Schematic illustration of face reenactment*

Source: Prepared by the authors.

Face reenactment has legitimate and very positive applications, especially in the fields of education, healthcare, entertainment, and film industry (Fukuha; Archegas; Fürst, 2026, p. 177), but there are also several risks associated with the use of this technology. For instance, in the healthcare sector, the ability to create hyper-realistic videos is often used to carry out scams and spread medical misinformation. Fraudsters clone the faces and mannerisms of renowned doctors, journalists, or celebrities and feature them in fake advertisements recommending dangerous treatments or selling "miracle" weight-loss products that lack any scientific evidence (Fukuha; Archegas; Fürst, 2026, p. 187).

In the political arena, this technology can be used to spread misinformation, carry out scams, and manipulate public opinion (Fukuha; Archegas; Fürst, 2026, p. 181). Because the system can generate highly realistic false narratives, it can be used to simulate false statements or behaviors attributed to

public officials, political leaders, journalists, or other prominent figures, which can influence electoral processes, damage individuals' reputations, and undermine trust in democratic institutions (Fukuha; Archegas; Fürst, 2026, p. 182; Kietzmann *et al.*, 2020, p. 141).

In a video created in 2018 by American comedian Jordan Peele, the face reenactment technique is used to transfer his own lip movements and voice onto the face of former U.S. President Barack Obama, making it appear as though Obama were severely insulting a political opponent (Irfan *et al.*, 2025, p. 1928; Fukuha; Archegas; Fürst, 2026, p. 182). Although this video was created specifically to alert the public to the risks of fake news, it demonstrated how easy it is to deceive people through synthetic content (Kietzmann *et al.*, 2020, p. 141).

Facial attribute manipulation

In facial attribute manipulation, the system alters only specific details of the image or video — such as age, hair color, or facial expressions — without replacing the person's face (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). To perform this edit, the algorithm maps the face and separates the individual's identity from the specific feature to be modified, ensuring that the rest of the appearance remains intact. After applying the change to that isolated point, the algorithm reviews the material to eliminate minor visual defects and correct continuity errors, ensuring that the new detail follows movements smoothly and that the result appears completely natural and authentic (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 5). Attribute manipulation is illustrated in Figure 4.

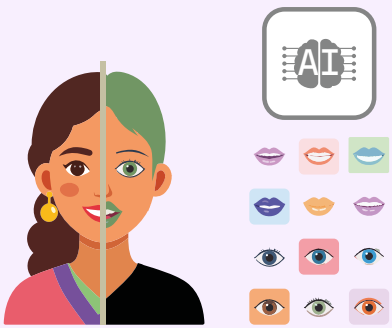


Figure 4 *Illustrative diagram of attribute manipulation*

Source: Prepared by the authors.

Facial synthesis

Facial synthesis is the process of creating hyper-realistic images or videos of human faces that do not actually exist in the physical world. Instead of altering a photograph of a real person, the technology works by constructing a completely new appearance, generating the image of an individual who looks authentic but is entirely fabricated by technology (Irfan *et al.*, 2025, p. 1930; Lazzarini, 2025, p. 12).

To do this, the algorithm begins by analyzing a massive number of photos of real people available on the internet. By being trained on these images, it learns all the characteristics that constitute the human face, such as the distance between the eyes, the proportions of the nose, and the way light reflects off the skin. The algorithm does not create a composite in the sense of "cutting and pasting pieces" from existing photos. Instead, it understands the visual pattern of what a face should look like and produces it without necessarily reproducing any of the images with which it was trained (Lazzarini, 2025, p. 10). Facial synthesis is illustrated schematically in Figure 5.

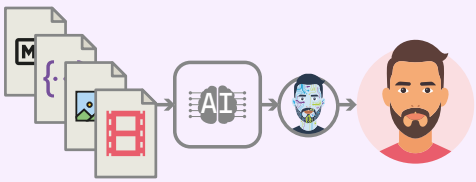


Figure 5 *Schematic illustration of facial synthesis*

Source: Prepared by the authors.

Voice Cloning

Acoustic manipulation, commonly known as voice cloning, encompasses text-to-speech (TTS) and voice conversion (VC) systems, and represents the auditory aspect of synthetic media and deepfakes. Unlike visual formats, which process facial geometries, this technology focuses on isolating and transferring the biometric characteristics of a target individual's speech (Dehghani; Saberi, 2025, p. 24–26).

According to Dehghani and Saberi (2025, p. 24–25), the literature divides these systems into parallel training methods — which require identical utterances recorded by both parties — and non-parallel methods, which use diverse and misaligned audio collections, providing greater practical flexibility for cloning.

The process begins with collecting voice recordings from the person whose voice is to be replicated. These samples serve as a reference for the system to learn the characteristics of the individual's speech. The amount of available audio directly influences the quality of the result, as very short samples may make it difficult to accurately reproduce a voice, while adding larger volumes of recordings yields diminishing returns beyond a certain point (Dehghani; Saberi, 2025, p. 26).

Next, the algorithm analyzes the recordings to identify the elements that make that voice unique. These aspects include characteristics such as speech patterns, speech rhythm, into-

nation, and other distinctive features that help differentiate one person from others. This step is important to ensure that the synthesized voice maintains as much similarity as possible to the original voice (Dehghani; Saberi, 2025, p. 26).

After learning the characteristic patterns of the voice, the system begins to generate new audio clips that reproduce these characteristics. This can occur in two main ways. In the first, the user provides a written text, which is converted into speech using the target person's voice. In the second, the system receives a recording made by another person and alters its vocal characteristics so that the result sounds as if it were spoken by the individual whose voice was learned, while preserving the original message (Dehghani; Saberi, 2025, p. 24).

2.2.3 Cheapfakes and shallowfakes

Deceiving people online does not require only advanced technologies, artificial intelligence, or high-powered computers. Although there is much talk about deepfakes, the most common everyday danger may lie in much simpler forms of fraud (Elliott, 2023). It is in this context that the concepts of cheapfakes and shallowfakes emerge in the literature, both of which are quite different from deepfakes.

This difference begins with the origin of the terms and extends to the technology used in each type of manipulation. The term "deep" refers to the use of deep learning techniques, while "shallow" and "cheap" indicate that the editing was done superficially and at low cost. Thus, while deepfakes are created by artificial intelligence systems capable of generating highly convincing forgeries (Altuncu, Franqueira, and Li, 2024), cheapfakes and shallowfakes are produced by humans using conventional, free editing software to cut, speed up, or alter the context of a video (Brady, 2020, p. 2; Stoll, 2020), without the use of deep learning techniques (Çetin, 2022; Weikmann; Egelhofer; Lecheler, 2025, p. 2; Weikmann; Lecheler, 2022, p. 3).

To better understand these differences, it is helpful to view the ecosystem of digital manipulations as a continuum: at one end, highly complex techniques created by artificial intelligence systems; at the other, simple edits that can be made by any user on a smartphone. As illustrated by Brady (2020), this distinction becomes clear when we analyze three main points: the level of sophistication, the technology used, and the ease of detection (Figure 6).

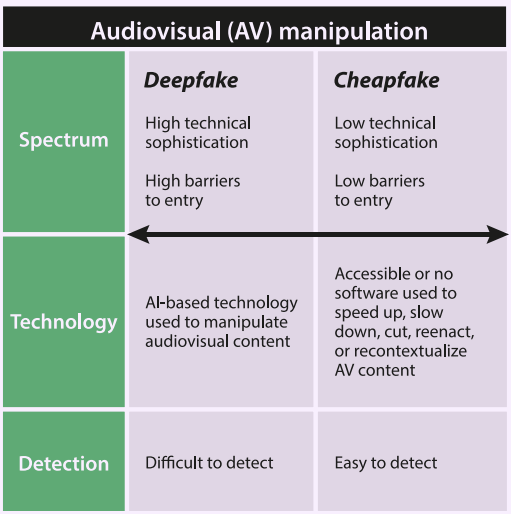


Figure 6 *Audiovisual manipulation analyzed based on level of sophistication, technology used, and ease of detection*

Source: Prepared by the authors based on Brady (2020, p. 3).

Although experts can easily detect such content, it still deceives thousands of people because of its rapid dissemination.

Some studies use the terms "cheapfakes" and "shallowfakes" as synonyms, given the low technical sophistication of both. However, Twomey *et al.* (2024, p. 7) propose clearer distinctions, arguing that the division between deepfakes and shallowfakes is based on the level of technology employed: AI-based tools would belong only to the first group, while media in the second group would derive from reposts in new contexts or from basic structural alterations to the original material, such as cuts and the overlaying of audio tracks.

The same authors propose a gradual classification for cheapfakes, without rigid divisions. The cheapfake-deepfake spectrum organizes the techniques along a continuous line defined by cost, barriers to entry, and sophistication. Deepfakes occupy the top of the scale, with complete digital modifications, while cheapfakes and shallowfakes occupy the bottom, ranging from the recontextualization of authentic files to manual edits, such as rotoscoping techniques for facial replacement and changes in playback speed. Thus, the proposed framework indicates that the **immediate informational danger lies in simplicity and recontextualization**, not merely in algorithmic sophistication.

Weikmann, Egelhofer, and Lecheler (2025, p. 1020) reinforce this interpretation, arguing that the simple visual disinformation of cheapfakes generates a persistent psychological and behavioral impact, often surpassing deepfakes in terms of disinformation effects — evidence that the cognitive risk of such content does not depend on the technical sophistication employed. Brady (2020, p. 3) complements this perspective, arguing that the social impact of these low-cost fabrications is amplified by direct communication channels between users. This is because the dissemination of information from one person to another via private messaging apps obscures the traceability of disinformation. This process can reinforce cognitive biases that validate the manipulated content even before any technical fact-checking takes place, especially in contexts of political polarization.

In this regard, Çetin (2022) emphasizes that everyday tools and social media filters exemplify how accessible these rudimentary manipulations are. Nevertheless, underestimating the operational simplicity of these media poses a critical risk to the cybersecurity ecosystem. From the practical perspective of corporate authentication, Çetin (2022) warns that if an organization's security mechanisms are weak, even these simplified techniques can easily circumvent liveness tests in remote identity verification procedures.

Clearly understanding the terms and the differences that distinguish and connect deepfakes, cheapfakes, and shallowfakes is essential for creating effective defenses against today's technological threats. The threat to informational integrity and digital security does not stem exclusively from sophisticated AI algorithms.

2.3 What is not a deepfake

To precisely define the scope of the debate on synthetic media, it is essential to draw a clear line between what constitutes a deepfake and what are merely common editing tools. First, a deepfake should not be confused with assistive edits, routine edits, or trivial adjustments to image and sound. Everyday interventions such as color correction, audio cuts, background noise removal, or the application of common lighting filters do not fall into this category. The reason for this exclusion is that **these alterations do not change the message's semantics, context, or original meaning.**

At the European level, this distinction is directly established in Article 50 of the AI Act (European Union, 2024). The European regulatory text establishes strict transparency and labeling obligations only for artificial intelligence systems that generate or manipulate content capable of misleading the public. However, the legislation explicitly protects creative freedom and routine digital practices by stipulating that these labeling rules do not apply to AI tools that are used purely for routine editing or that do not alter the actual meaning of the content (European Union, 2024), such as technical corrections or assistive filters.

Second, the boundaries of what counts as a deepfake remain unclear, leading to conceptual ambiguity and terminological instability. Interdisciplinary studies indicate that this lack of definition is associated with the multiplicity of frameworks originating in fields such as computer science, law, and

communication, which contribute to diverse interpretations of the phenomenon (Twomey *et al.*, 2024).

In response to this conceptual limitation, some of the literature proposes distinguishing deepfakes as a subset of the broader category of synthetic media, which encompasses various forms of content generated or manipulated by artificial intelligence, including those not based on real individuals (CNTI, 2025; Twomey *et al.*, 2024; Roe *et al.*, 2025). At the same time, empirical research shows that the use of the term "deepfake" tends to evoke perceptions associated with risk and misinformation, whereas alternative expressions, such as "synthetic media," may foster less negative interpretations that are more aligned with legitimate applications of the technology (Rauchfleisch; Vogler; De Seta, 2025).

In this context, Ding (2020) analyzed the report by the Tencent Research Institute, which proposes the use of more comprehensive and semantically neutral terminology — such as "deep synthesis" — to replace or serve as an alternative to the term "deepfake," which is considered reductionist and potentially stigmatizing. However, this proposal has not become an established standard in international scientific literature, where the terms "deepfake" and "synthetic media" remain predominant, even though their conceptual limitations are recognized.

Ultimately, what distinguishes permissible synthetic media from media created for harmful purposes is the combination of intent and granularity of manipulation, beyond the deep learning technique itself, as discussed in **Section 2.2.** As noted by the CNTI (2025), modern legislation seeks to reconcile non-deceptive, purely artistic, or utilitarian uses with the prohibition of criminal practices.

The focus of effective policies is on isolating and punishing harmful uses intended to deceive the public. This ranges from collective deception due to high verisimilitude in holistic identity transfer to the malicious modification of specific attri-

butes and the facial synthesis of characters for the purposes of fraud or non-consensual pornography. If AI-generated content is transparent and lacks the intent or effect of misleading the public, it constitutes creative or utilitarian expression and falls outside the restrictive and punitive scope applied to deepfakes used for deceptive purposes. These regulatory specifics will be addressed again in **Section 6.1.**

2.4 Non-malicious use of synthetic media

In the context of deepfakes, non-malicious use refers to the creation and application of hyper-realistic synthetic media (whether videos, audio, or images generated by artificial intelligence) without the intent to deceive, defraud, defame, or cause harm to third parties. Thus, this technology has potential for legitimate uses and innovation across various sectors, including artistic interventions, film production, entertainment content creation, and the construction of digital human models (Pei *et al.*, 2026, p. 28), as described below. It is important to note, however, that these uses are not always free from risks or ethical concerns.

Among legitimate uses, a notable example in commerce is the creation of virtual mannequins that customers can use to try on clothing or test cosmetics. In the audiovisual industry, the applications are diverse. This technology enables automatic visual dubbing, adjusting the actor's mouth movements to synchronize with the voice actor's audio in a foreign language. It also lets producers correct mispronounced words without calling the actors back to reshoot, which saves resources and expands the studio's creative work (Kietzmann *et al.*, 2020, p. 141). Face swapping, for example, allows producers to superimpose a lead actor's face onto a stunt double's body during the filming of particularly dangerous action scenes, producing a highly realistic visual result that is imperceptible to viewers (Kietzmann *et al.*, 2020, p. 142). The use of deepfakes in the creative industry can also spark

debate, as even in permitted uses there are ethical concerns regarding authenticity, originality, and authorship (Santos Junior, 2025). Altering a script or copying an artist's style, for example, may conflict with copyright laws (Kietzmann *et al.*, 2020, p. 145). Altering the original performance also raises questions about actors' lack of consent to the use of their image (Kietzmann *et al.*, 2020, p. 142). An example of this is the case involving actress Scarlett Johansson, whose voice characteristics were replicated by an artificial intelligence system developed by OpenAI without her consent (Lazzarini, 2025, p. 7). Because of these legal and artistic conflicts, the cost savings achieved by production companies have drawn criticism.

There is also the delicate and controversial issue of "digital resurrection," which occurs when artificial intelligence is used to recreate the faces and voices of artists who have passed away (Lazzarini, 2025, p. 23). This has already happened in major Hollywood productions, such as the recreation of actor Peter Cushing in the Star Wars franchise (Fukuha; Archegas; Fürst, 2026, p. 175; Kirchengast, 2020, p. 2), and in Brazilian advertising, with the recreation of singer Elis Regina in a television commercial (Lazzarini, 2025, p. 23). The creation of these "posthumous deepfakes" raises discussions about the limits of technology and respect for the memory of deceased individuals. The current legal debate focuses on defining who is entitled to authorize such commercial recreations. For example, there is debate over whether heirs and family members possess this right, or whether legislation should require that the artist themselves give prior and express permission, while still alive, so that a computer could generate new images of them in the future (Lazzarini, 2025, p. 44).

This same ability to bring images from the past back to life, when guided by ethics and transparency, is also beginning to be explored in artistic and cultural initiatives aimed at historical preservation. The artist Guilherme Bretas, for example, used this kind of technology to combine artificial intelligence with archival records, bringing portraits of Black people from

the 19th century to life. By animating faces preserved in museums and projecting these images in public spaces, the artist sought to break the silence and control inherited from the colonial era (Santos Junior, 2025).

In another example, the Museum of Black Memory in AI (@ museudamemorianegraemia) uses artificial intelligence to recreate historical images and produce depictions of figures who were not portrayed in photographs or paintings during their lifetimes. In these projects, the use of algorithms has opened new narrative possibilities and reinterpretations, offering a glimpse into lives and stories that official records erased or omitted, while also serving as an educational tool to prevent the reproduction of racist stereotypes.

Another application related to the preservation of memory is the restoration of visual archives, in which artificial intelligence can reconstruct missing parts of a face, enhance low-quality photos, and even colorize black-and-white photographs in order to preserve old films (Pei *et al.*, 2026, p. 273; Santos Junior, 2025).

In the field of education, this approach of bringing archives and memories to life has concrete pedagogical applications. Face reenactment can promote accessible and inclusive education by supporting the creation of videos tailored to students with dyslexia, hearing impairments, or other learning difficulties through adapted communication styles and precise synchronization of mouth movements with subtitles or sign-language interpretation (Fukuha; Archegas; Fürst, 2026, p. 176–177). Such technology also makes it possible to "bring to life" figures from the past: algorithms can create interactive videos in which historical figures such as scientists, writers, and political leaders tell their own stories and teach concepts to students in a modern and accessible way. A striking example of this application took place at the Illinois Holocaust Museum in the United States, which used artificial intelligence to create holograms of real survivors of the

tragedy, allowing museum visitors to interact and "converse" directly with them (Fukuha; Archegas; Fürst, 2026, p. 176).

These initiatives illustrate how the creation of synthetic media can serve as a technical tool to foster reflection and fill gaps in history, pointing to a path that can be adopted in similar contexts, provided that the use of technology is guided by ethics and transparency in information (Assis Junior; Hessel, 2021, p. 81). The danger, however, lies in the distortion of facts: just as technology can "revive" a historical figure to teach something accurate, it can be used to falsify history, create narratives that never existed, and manipulate the opinions of students and society. For this reason, the literature warns that, for these innovations to be beneficial in healthcare and education, they must be applied with complete transparency, informing the public that artificial intelligence is being used, with the consent of the people depicted and under strict supervision, to ensure that the line between personalized education and malicious manipulation is not crossed (Fukuha; Archegas; Fürst, 2026, p. 177–178).

As a means of promoting accessibility, voice cloning technologies have the potential to restore and create voices for individuals who have lost their ability to speak (Dehghani; Saberi, 2025, p. 2). In clinical practice, these technologies can also be applied to assistive devices for patients with severe conditions such as amyotrophic lateral sclerosis (ALS), or those about to undergo laryngeal surgery. Patients with progressive conditions may clone their voices before losing the ability to speak to preserve their identity.

At the same time, in the film and entertainment industry, applications validated by the authors include the development of digital avatars, the updating of film segments without the need for reshoots, and the realistic dubbing of foreign productions (Dehghani; Saberi, 2025, p. 2) — a market that is expanding into automated translation capable of preserving the actor's original timbre and restoring historical figures. Furthermore, through speech synthesis, the service sector is

pushing these boundaries to include the customization of virtual customer service assistants, satellite navigation (GPS) tools, audiobooks, and home assistants based on corporate or personalized voice identities.

Meanwhile, the advertising and cultural sectors are able to use media generation to compose music and produce advertisements with new audio and visual features. Although not necessarily deceptive, the use of deepfakes in these areas reignites, among professionals, the previously mentioned concerns regarding the authorship, originality, and authenticity of works — especially when copying a style or using the image of public figures without their consent, a practice that may violate copyright laws (Santos Junior, 2025; Kietzmann *et al.*, 2020, p. 142, 145). In addition to artistic and image-related conflicts, there is a danger that the technology will reproduce prejudices and inequalities based on gender, class, and race, as algorithms learn and replicate the biases already present in large databases (Santos Junior, 2025).

Synthetic media may also have applications that benefit the healthcare sector. In public health communication, face reenactment can help break down language barriers. A world-renowned example was the anti-malaria campaign called "Malaria Must Die," in which a face reenactment system altered the lip and facial movements of former soccer player David Beckham so that he appeared to speak fluently in nine different languages, helping to spread a vital medical message and raise awareness worldwide (Fukuha; Archegas; Fürst, 2026, p. 177). Another possible application is protecting patient privacy by altering a person's identity in a medical diagnostic video, so that teams can share such media without revealing the patient's identity (Pei *et al.*, 2026, p. 11). On the other hand, this technique does exclude potential losses in image accuracy, which may render it unsuitable for certain research objectives or applications that require reliable images.

Such technolohy also drives innovations in the book and video game industries. In the production of audiobooks, for example, systems alter the narrator's voice to portray different roles. They generate younger or older voices, male or female, with different accents for each character in the story (Kietzmann *et al.*, 2020, p. 141).

In video games, face-swapping tools allow players to superimpose their own faces onto characters. However, given the widespread popularity of video games among children and adolescents, this hyper-personalization poses significant risks associated with the vulnerability of this audience, who may have greater difficulty distinguishing synthetic content from reality, leaving them more exposed to privacy violations, cyberbullying, and grooming, when offenders make use of fabricated identities and generative tools to their own advantage in order to manipulate children in the virtual environment (Negreiro, 2025, p. 2).

These transformations, using deepfake tools, are carried out with machine learning to automate the process. Thus, the player needs only to provide a few ordinary photos or selfies, and the algorithm uses them to swap faces or perform continuous face morphing (Kietzmann *et al.*, 2020, p. 137). The major difference compared to previous technologies lies in the quality of the alterations and how effortless it is to make them. Previously, creating realistic effects on game characters required expensive equipment and trained professionals. Today, deepfake technologies allow players to do this within seconds, without needing a 3D scanning studio (Kietzmann *et al.*, 2020, p. 137).

2.5 The role of technical definitions in data protection

A shared definition and understanding of technical concepts are essential for guiding regulatory responses that are

compatible with the challenges posed by deepfakes. In the United States, two bills concerning this issue were proposed, but despite their contributions to the conceptual debate surrounding these tools, they were never enacted. The first of these, the Malicious Deep Fake Prohibition Act of 2018, proposed strict rules against digital manipulation. The bill defined a deepfake as an audiovisual recording created or altered in a deceptive manner, with the intent of appearing authentic to an ordinary observer. Such manipulation would aim at imitating exactly an individual's actual speech or manner (United States of America, 2018, p. 3). This bill focused on punishing the malicious use of this technology and prohibited the creation and distribution of synthetic files intended to facilitate crimes or unlawful acts.

This initiative highlights a conceptual challenge in the global regulatory debate, namely the lack of a single, stable technical definition for synthetic media. The rapid and continuous evolution of the generative AI architectures renders static definitions crafted by lawmakers quickly obsolete in the face of new mathematical models for content generation (Fukuha; Fürst; Archegas, 2026, p. 188–189).

The following year, in 2019, the DEEP FAKES Accountability Act introduced new proposals for regulatory strategies. In addition to punishing abuses, the proposal required strict transparency rules to curb misinformation (United States of America, 2019, p. 2). The legislation would mandate the inclusion of digital watermarks and verbal and written warnings in manipulated media that clearly informed the public that the video, audio, or image content had undergone technical alterations.

The text of the bill presented a key technical definition, describing deepfake as any technological representation that was substantially derived from a person's original speech or conduct. This showed that the rule focused on manipulation carried out by advanced computer programs, rather than on physical or vocal human imitation. The offense would occur

when the system made it appear as if the victim had done something that, in fact, they had not done (United States of America, 2019, p. 15).

By comparing the two proposals, we can observe a clear shift in the conceptual framework adopted by lawmakers, which illustrates the volatility of the regulatory debate surrounding synthetic media. The 2018 bill was based on a functional and cognitive definition model, classifying deepfakes essentially by their psychological effect of deceiving a "reasonable observer" regarding the authenticity of the file (United States of America, 2018, p. 2). In contrast, the 2019 bill shifted to a technical-operational definition model, distinguishing algorithmic manipulation from mere physical or verbal human imitation by stipulating that audiovisual forgery must be "substantially dependent on technical means" (United States of America, 2019, p. 15–16).

Right away, this shift in position highlighted a practical challenge: the regulatory focus shifted from the perceptual outcome (the file's ability to deceive) to the computational process of generating synthetic media.

The bills were not passed due to political deadlocks, the complexity of the legislative process, and strong pressure from interest groups (Fukuha; Fürst; Archegas, 2026). In this context, the criteria that define applications as harmless or lawful are intrinsically linked to the right to freedom of information, the creative freedom of individual developers, and the optimization of market processes in technology companies and creative industries. In these operational contexts, cost reduction and the efficiency gains are used as justifications for the adoption and defense of generative AI (Hachkevych, 2025, p. 42) — a different perspective from the one the state must consider when protecting people's rights, especially those of vulnerable groups such as children and adolescents.

The analysis used to distinguish harmless synthetic media from those considered dangerous is typically structured around the conflict of interests among developers, individuals, and the state, rather than based on a purely technical definition. A technical understanding of the mechanics of mass-scale synthetic media generation can directly impact the scope of regulation (Fukuha; Fürst; Archegas, 2026, p. 172–173). This means that government regulation must incorporate precise technical definitions that do not merely prohibit the computational architecture itself but rather focus on the intended use and the harm caused by synthetic media (Fukuha; Fürst; Archegas, 2026; Kirchengast, 2020, p. 222), as will be discussed further in the **Section 6.1.**

In this context, we observe that the way in which lawmakers technically define the technology and conceptualize deepfakes in legislation can shape the law's practical consequences, since static or overly broad concepts can lead to structural discrepancies and the criminalization of beneficial uses (Fukuha; Fürst; Archegas, 2026, p. 188–189; Kirchengast, 2020, p. 415). On the other hand, the pursuit of balanced definitions would enable the State to fulfill its dual mission of ensuring trust in the authenticity of media and safeguarding fundamental rights in the digital environment (Fukuha; Fürst; Archegas, 2026, p. 168). Thus, analyzing the definitions of deepfakes in proposed legislation from other countries can contribute to the regulatory debate in Brazil.

‹ 3 how deepfakes work ›

Now that the conceptual boundaries and distinctions of synthetic media have been established, we will present, in this section, an overview of the technical architecture of deepfakes. The technical study of deepfakes is divided into two processes: one that creates synthetic media and another that attempts to identify them, referred to, respectively, as the generation pipeline and the detection pipeline[6] (Irfan *et al.*, 2025, p. 1928). The first deals with algorithms dedicated to the generation of synthetic media, and the second with models designed for their detection.

The generation pipeline is a conditional transformation operation. This means that, given a target video or image and a control condition (such as another person's facial identity), the network produces new media that preserves the original context but alters the desired attributes. To achieve this goal, generative systems utilize advanced deep learning techniques, such as autoencoders, diffusion models, and Generative Adversarial Networks or GANs (Irfan *et al.*, 2025, p. 1930; Pei *et al.*, 2026, p. 273). The three architectures differ in their mathematical strategy: autoencoders use compression and reconstruction, diffusion models use progressive corruption reversal, and GANs use competition between networks. Each produces distinct results in terms of control, quality, and computational cost.

Although these approaches can be used in a variety of contexts, in the case of deepfakes, the algorithms are extensively trained to extract latent features of the human face and reconstruct them automatically, altering identities, expressions, or movements with high fidelity (Irfan *et al.*, 2025, p. 1930; Kietzmann *et al.*, 2020, p. 144). In this way, the models learn to represent the human face using compact numerical vectors, which capture patterns such as bone structure, skin texture, and muscle dynamics, and which can be modified

6 *A pipeline is a workflow for designing, building, and implementing technological solutions. It can represent the sequential path, consisting of multiple step-by-step stages, that data follows within a system to perform a complete task from start to finish.*

and decoded to generate new faces. These models will be explained below.

On the other hand, **the detection pipeline** acts as a defense and forensic control mechanism and is technically modeled as a mathematical classification operation at the image or pixel level. This means that detection functions as a binary classification: the network decides image by image or pixel by pixel, whether the content is real or synthetic. The granularity (image versus pixel) determines whether the model merely flags manipulation or also locates where it occurs (Pei *et al.*, 2026, p. 4). This macroprocess employs classification neural networks, such as Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), to automatically extract discriminative features and identify artifacts, spatiotemporal inconsistencies, or traces left by synthesis algorithms (Irfan *et al.*, 2025, p. 1932). CNNs analyze spatial patterns in individual images (such as jagged edges and inconsistent textures), while RNNs analyze temporal sequences in videos (such as facial movements that violate natural continuity). Both seek out the traces that synthesis algorithms inevitably leave behind.

Thus, in the field of artificial intelligence engineering, deepfake generators and detectors are locked in a constant competition, in which detection systems evolve to expose the fabrications, while generation methods continually improve to circumvent these security mechanisms and produce synthetic content that is increasingly undetectable (Irfan *et al.*, 2025, p. 1934).

To systematize our understanding of the engineering behind the creation of synthetic media, we can distinguish the underlying computational **infrastructure**, the **operating systems** used, and the **mathematical logic** that governs machine learning. We can present this conceptual framework as follows:

+ **1. Infrastructure:** This refers to the underlying computational architectures and topological structures that provide the foundation for high-volume data processing. Neural networks constitute the logical architecture that processes the data. The physical hardware consists of graphics processing units (GPUs), which are required due to the scale of the calculations involved. Deep learning provides the foundation upon which all models are built: deep neural networks learn hierarchical representations, ranging from the simplest features (edges, colors) to the most complex (facial expressions, identity), by stacking multiple layers of mathematical transformations.

+ **2. Mathematical Models:** These correspond to the functions, statistical equations, distributions, and optimization problems that conduct the internal training of generative models. This is the rigorous mathematical logic that allows algorithms to adjust their parameters to generate realistic synthetic media.

+ **3. Generative Models:** These models represent the specific architectural arrangements, infrastructures, or systems built on top of Artificial Neural Network (ANN) infrastructures to perform the targeted task of content synthesis or manipulation. They define the operational strategy for how the networks interact to create the final file. The main models currently in use are:

  - **Generative Adversarial Networks (GANs):** These train two models simultaneously with opposing objectives: the generator produces synthetic images in an attempt to fool the discriminator; the discriminator classifies images as real or fake, trying not to be fooled. This training process achieves its goal when the generator learns to produce images with the same variety and visual characteristics that are closely resemble those of real faces.
  - **Autoencoders:** They compress an image into a compact numerical vector and then reconstruct it from

that vector. In face-swap deepfakes, two autoencoders share the same encoder but have separate decoders — one for each identity — allowing a person's vector to be decoded with the features of another person.

- **Diffusion Models:** These create realistic images through a gradual "unblurring" process, in which the computer learns to remove, step by step, initial noise resembling a blurred screen until a new, perfectly sharp image emerges (Pei *et al.*, 2026, p. 6).
- **Neural Radiance Fields (NeRFs):** These function as virtual builders that analyze ordinary images to learn the geometric shapes and lighting of a face or scene and use this information to reconstruct highly realistic three-dimensional (3D) figures (Pei *et al.*, 2026, p. 6).

Throughout this section, we will detail the models and frameworks that compose the engineering of artificial intelligence, which enable deepfake tools to function.

3.1 Foundations of the algorithms that produce deepfakes

To understand how deepfake tools work, we need to look at a broader technological landscape: the field of artificial intelligence. Within this field, there is a fundamental subfield called *machine learning*, in which computers move beyond simply following strict programming rules and commands to learning through the analysis of large volumes of data. Within machine learning, we find ANNs and their most advanced form of operation: the deep learning (Fukuha; Fürst; Archegas, 2026, p. 170–171).

This deeper and more sophisticated infrastructure enables generative AI, a branch of technology whose main characteristic is the ability to learn complex mathematical patterns to generate and synthesize previously unseen textual, audio, and visual content (Alves Junior, 2025, p. 10–11).

Each of these concepts — and, in particular, their roles in the production of deepfakes — is detailed below.

3.1.1 Machine learning

A computer program uses a machine learning technique when it learns from experience with specific tasks, and its performance on those tasks improves as it accumulates data (Mitchell, 1997, p. 2). This occurs through repeated cycles of training and testing the AI model, in which the system analyzes large volumes of historical data to identify patterns and adjust its internal parameters, learning the logic of the problem from accumulated experience (ANPD, 2024, p. 7).

Traditional machine learning algorithms do this using **structured**, tabular **data** inputs (e.g., age, number of purchases, duration of use, product category) and employ an attribute matrix to mathematically map the relationship between the data's characteristics and what they represent, thereby learning a function that makes decisions or predictions directly based on these variables. This approach is based on the principle that, by accumulating experience with this historical data, systems can autonomously improve their performance (ANPD, 2024). In this context, human engineers must select which data and details the system should consider when making a decision (AWS, 2023). Therefore, unlike more modern deep learning models, classical models rely on these "handcrafted features" developed by experts to classify information (Dehghani; Saberi, 2026, p. 29).

This is how, for example, email filters that block spam messages or Netflix's personalized content recommendations work. In these cases, the system analyzes structured historical data that users have provided in some way — such as specific words in a text or the genres of movies they've watched — to try to predict their next move (Goodfellow; Bengio; Courville, 2016, p. 3).

Therefore, classical machine learning techniques are valued for their simplicity, efficiency, and interpretability. Since their decisions are based on rules and attributes clearly defined by humans, they perform well on smaller datasets, require fewer computational resources, and are widely used in applications such as financial modeling and fraud detection (Rahnama, 2025).

3.1.2 Deep learning

Machine learning has evolved from a predominantly discriminative approach to the development of sophisticated generative models. The **discriminative approach** focuses on learning decision functions to classify data into known groups or predict probabilities based on attributes that are often selected manually (ANPD, 2024, p. 10–11). As a result, image and video processing using discriminative networks were primarily focused on labeling and identification, such as face detection (Pei *et al.*, 2026, p. 15).

Generative models, on the other hand, have established themselves in tasks such as face swapping, face reenactment, generating speaking faces, and editing facial attributes (Pei *et al.*, 2026, p. 2). This is because, unlike discriminative models — which learn to classify what already exist, generative models learn the probability distribution implicit in the training data, enabling them to synthesize statistically plausible and novel samples (ANPD, 2024, p. 11). By automatically extracting high-level representations directly from this unstructured information, the system becomes capable of synthesizing novel and highly realistic content. And since much of the content in the digital environment consists of unstructured data — such as raw images, audio, and video, this imposes limitations on traditional algorithms.

It is precisely in this regard that deep learning becomes indispensable. While classical machine learning provides the statistical foundation for pattern recognition, deep learning

uses multiple layers to extract knowledge from **raw data** (information that has not yet undergone any type of processing, filtering, or organization), incrementally transforming it to obtain high-level features (ANPD, 2024, p. 7) and generating highly realistic synthetic content.

Thus, deep learning is a subset of machine learning based on **ANNs**, which are mathematical and computational models designed to mimic, in a simplified way, the functioning of the human brain to solve complex problems. According to Haykin (2007, p. 186), the fundamental structure of these networks is the artificial neuron, which acts as a processing unit capable of learning and storing information. The author explains that, in practice, each neuron receives input signals, multiplies this numerical data by "weights" — which indicates the importance of each piece of information — and generates an output response. Before transmitting this result to the next layer, the neuron applies an activation function, which introduces nonlinearity into the system and enables the network to learn complex patterns that go beyond simple proportional relationships between variables.

This approach allows the system, as it builds itself, to also learn complex concepts from simpler representations organized across multiple levels (Goodfellow; Bengio; Courville, 2016, p. 1, 8). The architecture of an ANN is divided into interconnected blocks, in which input signals (stimuli) first reach the "input layer" and are then passed on to the so-called "hidden layers," represented in the image as the first and second hidden layers — locations where most of the system's computations and learning actually take place.

After undergoing all this internal processing, the data reaches the "output layer," which serves to provide the final result or response (Haykin, 2007, p. 186–188). This architecture is designed to function as a data assembly line, in which raw data is transformed, step by step, through multiple computational stages. Each new layer of the network autonomously

extracts and refines higher-level features from the previous layer, with minimal need for human intervention (ANPD, 2024, p. 7; Irfan *et al.*, 2025, p. 1932). Figure 7 below illustrates the structure of an ANN.

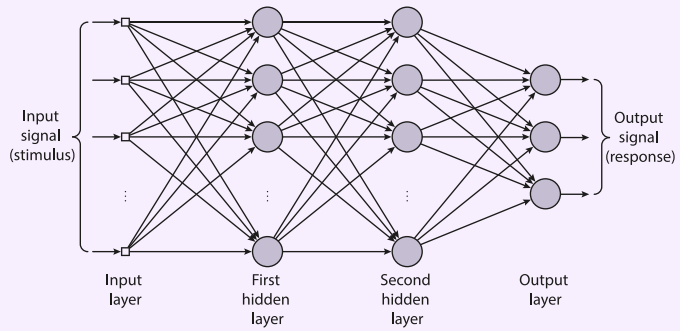


Figure 7 *Example of a Multilayer Perceptron (MLP) Neural Network*

Source: Haykin (2007, p. 186).

Deep learning encompasses several types of networks structured in multiple layers, going beyond a single architecture (Irfan *et al.*, 2025, p. 1932). However, for the generation of synthetic media such as deepfakes, deep learning typically uses a specific neural network architecture known as a Convolutional Neural Network (CNN), which often works in conjunction with other deep structures, such as autoencoders and GANs (Fukuha; Fürst; Archegas, 2026, p. 171). These types of networks are discussed in more detail below.

In neural networks, representations of the processed information are organized in a latent space (or representation space), a mathematical structure that captures the key statistical characteristics of the data (Pei *et al.*, 2026). It is the **latent space** that allows the deep learning algorithm to extract and organize patterns from the data with minimal human intervention (Goodfellow; Bengio; Courville, 2016, p. 4; Tiu, 2020). Thus, it compresses complex and detailed information from unstructured data into a simplified mathematical format, retaining only what is essential for the model (Tiu, 2020). This enables the manipulation and synthesis of new information.

As a simplified example to illustrate this concept, consider a generative deep learning model trained on images of furniture that receives as input an image of a chair (which is raw, unstructured data). When this photo passes through the layers of the neural network, the model discards information that is redundant or irrelevant to its objective — in this case, the background or lighting. However, the system retains essential properties, such as abstract representations of the visual structure contained in the image. In the real world, such representations correspond to visual patterns, such as the ratio of height to width, the density of vertical lines, and the symmetry of shapes. If we were to represent these characteristics on a graph, the vectors for chairs would be clustered by statistical similarity and separated from the vectors for tables (Tiu, 2020).

Thus, this process converts raw data into compact vectors that organize the fundamental patterns of the original dataset, which, in this example, would be pieces of furniture (Goodfellow; Bengio; Courville, 2016; Tiu, 2020). By combining the coordinates of a chair and a table, the system calculates a novel intermediate visual representation (Tiu, 2020). This ability to perform mathematical operations on the structure enables the generation of new images and serves as the technical foundation for high-fidelity media synthesis. In practice, this combination has enabled algorithms that merge or recreate faces with high precision. The creation of deepfakes relies on three main approaches based on deep learning: generative models; neural network architectures; and scene representation methods such as the Neural Radiance Field (NeRF). Each technology has specific operational characteristics in the context of synthetic media creation, as we will see below.

3.2 Key technical approaches for deepfake generation

The algorithmic generation of deepfakes relies on different models, each with its own technical advantages and disadvantages. GANs, for example, offer the advantage of producing images with high visual fidelity through continuous competition between a generator network and a discriminator (EPRS, 2025, p. 1; Pei *et al.*, 2026, p. 273). However, they have notable drawbacks, such as training instability and susceptibility to generating unwanted artifacts, making it difficult to handle severe occlusions and large variations in facial pose (Pei *et al.*, 2026, p. 11–12).

In turn, diffusion models have taken the lead in photorealism due to their superior image quality and resolution, treating synthesis as an iterative process of gradually removing noise to reveal complete images and videos (Pei *et al.*, 2026, p. 273). However, they have disadvantages stemming from their high computational complexity and the extensive processing time required by their various algorithmic stages (Pei *et al.*, 2026, p. 14).

On the other hand, the integration of Large Language Models (LLMs) and multimodal architectures into video processing offers the advantage of combining rigid textual commands with visual generation and acting in a forensic manner, translating the inconsistencies in a fake file directly into analytical explanations in natural language (Qian; Xia, 2025, p. 11–12). Multimodal models are advanced artificial intelligence architectures designed to process, integrate and learn inherent statistical relationships among different types of data modalities simultaneously, encompassing text, images, audio, and video. The use of these models, however, faces drawbacks such as high structural computational costs and the technical difficulty of ensuring refined control over the intensity of continuous emotional expressions on simulated faces in video (Pei *et al.*, 2026, p. 16).

These different approaches are explained in greater detail below.

3.2.1 Generative Adversarial Networks (GANs)

GANs are a revolutionary subfield of deep learning. Introduced by Ian Goodfellow in 2014, they have established themselves as one of the most influential architectures for creating hyper-realistic synthetic media, known as deepfakes (Patel *et al.*, 2023, p. 143301). Unlike reconstruction-based generic approaches, such as traditional autoencoders — which operate through deterministic mapping and compression of individual samples (Dehghani; Saberi, 2025, p. 9; Patel *et al.*, 2023, p. 143301) — GANs generate entirely new data, faithfully mimicking the distribution of the real data on which they were trained (Patel *et al.*, 2023, p. 143297, 143301–143302).

This mathematical refinement and the automation of the process were the major precursors to the massive dissemination of highly realistic synthetic media on the internet, as they enabled users without in-depth technical knowledge to generate ultra-convincing visual and audio fakes on a large scale, using software with simplified interfaces that "hide" the mathematical complexity of these technologies behind intuitive buttons (Dehghani; Saberi, 2025, p. 5; Patel *et al.*, 2023, p. 143297).

The technical operation of GANs is similar to a non-cooperative game of strategic rivalry (Goodfellow *et al.*, 2014, p. 1, 3). This ecosystem consists of two competing ANNs that are trained simultaneously: the Generator (G) and the Discriminator (D) (Dehghani; Saberi, 2025, p. 12; Patel *et al.*, 2023, p. 143297), where:

+ **Generator (G):** Its sole function is to create synthetic content (whether images, videos, or audio) (Patel *et al.*, 2023, p. 143301). It receives a random noise vector as input and attempts to transform it into a plausible visual or audio media sample (Dehghani; Saberi, 2025, p. 12; Patel *et al.*, 2023).

+ **Discriminator (D):** Acts as a binary classifier. It receives, in a mixed manner, both real samples from the original database and synthetic media created by the Generator, estimating the probability that the sample is genuine or fake (Patel *et al.*, 2023, p. 143298, 143301).

Figure 8 below illustrates the technical operation of GANs.

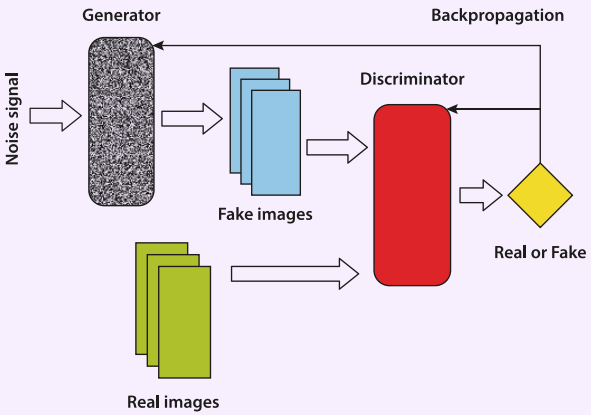


Figure 8 *Technical operation of GANs*

Source: Dehghani and Saberi (2025, p. 12).

The training process is dynamic, involving the continuous adjustment of loss functions based on a "min-max" problem format.[7] The process works as follows: while the Discriminator is trained to maximize its classification accuracy, the Generator is updated to maximize the Discriminator's errors, attempting to deceive it. As the training cycles progress, the Generator learns to map hidden textures, contrasts, and complex statistical patterns. Training reaches the optimal balance when the Generator produces samples so realistic that the Discriminator achieves a classification accuracy of approximately 50% (Patel *et al.*, 2023, p. 143297–143299, 143301–143302).

This basic architecture of GANs has evolved to overcome initial limitations, such as blurry or low-resolution images (Patel *et al.*, 2023, p. 143302). Advanced versions are shaping the current landscape of synthetic media. Among them are DCGAN (Deep Convolutional GAN), StyleGAN, CycleGAN, and StarGAN.

7 *According to Patel et al. (2023), during the training of Generative Adversarial Networks (GANs), the Generator attempts to deceive the Discriminator by creating realistic synthetic data, while the Discriminator attempts to maximize the detection of these forgeries. This "min-max" optimization process promotes continuous refinement, enhancing the model's ability to generate content indistinguishable from the originals.*

With their pivotal role in advancing visual realism, GANs have established a dynamic of co-evolution between deepfake generators and detectors within the digital ecosystem (Dehghani; Saberi, 2025, p. 3–5). As generative models mitigated initial flaws through successive iterations on the data — such as the inconsistencies in blinking patterns described by Patel *et al.* (2023, p. 143308) — traditional forensic systems based on spatial domain analysis became obsolete. This scenario forced a transition to hybrid and multimodal detection architectures based on deep learning (Heidari *et al.*, 2024). Such defenses aim to identify complex anomalies, such as functional dissonances and inconsistencies between audio, video, and lip movements (Patel *et al.*, 2023, p. 143311–143312, 143317).

Detecting these media poses a critical challenge due to the failure of existing models to cope with new generation techniques and the vulnerability of systems to compressed or low-quality content (Heidari *et al.*, 2024; Patel *et al.*, 2023, p. 143314–143315). Given this, the future of combating deepfakes depends on increasingly sophisticated detection algorithms, as well as the development of accessible and user-friendly tools that can be made available to the general public (Dehghani; Saberi, 2025).

Although GANs excel for the high visual fidelity of their samples, the competitive nature of their training often results in mathematical instabilities and problems such as mode collapse (Bond-Taylor *et al.*, 2022). In practical terms, rather than learning to model the full diversity of the original distribution (for example, generating faces of different genders, ethnicities, and expressions), the model enters a vicious cycle in which it ignores most of the modes (clusters) of data and focuses on only one or a few specific formats (Bond-Taylor *et al.*, 2022). This loss of diversity compromises the network's usefulness for realistic synthesis and broad media manipulation, which has motivated the scientific community to seek loss functions based on Wasserstein distance or to migrate

to probabilistic alternatives such as Variational Autoencoders (VAEs) and diffusion models (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

3.2.2 Variational Autoencoders (VAEs)

Among the pioneering generative architectures in deep learning, Variational Autoencoders (VAEs) excel (Dehghani; Saberi, 2025; Bond-Taylor *et al.*, 2022). A VAE is a generative artificial intelligence model designed to compress data and create synthetic versions based on it, widely used in image processing and facial feature editing (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025). Furthermore, this architecture extends to the generation of synthetic audio and music content (Bond-Taylor *et al.*, 2022), as well as to the cloning and synthesis of human speech (Dehghani; Saberi, 2025).

Unlike Energy-Based Models (EBMs), which required slow statistical simulations to generate content, VAEs revolutionized the field by integrating variational inference directly into neural network training. This made it possible to generate new samples quickly and efficiently, in just a single pass through the network (Bond-Taylor *et al.*, 2022).

To understand a VAE, imagine you want to teach a computer to draw human faces. The VAE does this by functioning as a pair of integrated neural networks: the Encoder, which is the inference network, and the Decoder, which is the generative network.

+ **Encoder:** This is the neural network responsible for mapping high-dimensional images into a lower-dimensional latent space, converting visual attributes into continuous probability distributions (usually defined by means and standard deviations) (Bond-Taylor *et al.*, 2022). In the example presented earlier, instead of trying to memorize every pixel in a photo, the Encoder extracts the fundamental features of the face — such as eye color, nose width, or

the presence of a smile — and organizes them statistically within this summarized space. It is worth noting that this direct correspondence between a feature and an isolated dimension of the latent space is a didactic simplification. In a real VAE, these features tend to be entangled across the latent dimensions.

+ **Decoder:** This is the neural network that performs the reverse process. Starting from a sample extracted from this structured latent space, the decoder reconstructs and generates new data samples that share the same mathematical properties as the original distribution (Bond-Taylor *et al.*, 2022). Thus, it uses the list of abstract features generated in the latent space as an "instruction manual" to design a completely new face, ensuring that the result truly resembles a real human face.

Figure 9 illustrates the architecture and operation of VAEs during the training and testing phases.

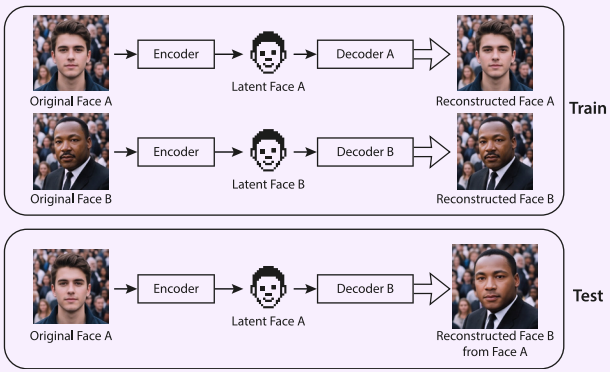


Figure 9 *Architecture and operation of VAEs during the training and testing phases*

Source: Adapted from Dehghani and Saberi (2025, p. 31).

The main conceptual advancement that distinguishes VAEs from traditional autoencoders — whose purpose is limited to the deterministic compression and reconstruction of data — lies in the explicit modeling of the probability distribution of the latent space, enabling the process of sampling and generating new data (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025). In a conventional autoencoder, each input sample is mapped to a fixed, deterministic point in the latent space. This lack of statistical regularization results in a discontinuous and sparse latent space. Consequently, when decoding a random coordinate located in a region not mapped during training, the decoder fails to structure the information, producing only disjointed visual noise.

VAEs overcome this limitation by applying a regularization mechanism that ensures the density, smoothness, and continuity of the latent space (Dehghani; Saberi, 2025). Under this architecture, it becomes possible to navigate through the latent coordinates linearly to perform smooth interpolations — such as the gradual transition between the features of different faces — or to sample any random point to generate novel, visually realistic synthetic data (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

The VAE optimization process is based on balancing a metric known as the Evidence Lower Bound (ELBO) (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025). Because it allows for refined control over its latent space, the VAE has become a tool for precise media editing (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025), serving as one of the models behind face-swapping technologies. In these systems, the encoder isolates features — such as an individual's facial expressions — so that the decoder can project them onto another person's identity. Furthermore, this mathematical flexibility allows the architecture to excel in both voice synthesis and cloning, as well as in the reconstruction and restoration of corrupted images (Bond-Taylor *et al.*, 2022; Dehghani; Saberi, 2025).

Although they offer superior sampling speed and training stability compared to GANs, conventional VAEs face a long-standing limitation: a tendency to generate slightly out-of-focus or blurry images (Bond-Taylor *et al.*, 2022). To mitigate this problem, the field has evolved toward the development of hierarchical VAEs and hybrid architectures that combine the regularization of VAEs with the sharpness of other modeling techniques (Bond-Taylor *et al.*, 2022).

Energy-Based Variational Autoencoders (VAEBMs), for example, are hybrid systems that combine the generation speed of VAEs with the high visual fidelity refinement of EBMs, overcoming the blurriness of traditional approaches (Bond-Taylor *et al.*, 2022). By coupling both structures, the VAE defines the overall structural organization of the data in latent space, while the EBM component acts directly on the data space to correct imperfections and recover high-frequency details (Bond-Taylor *et al.*, 2022).

Although the classic limitations of visual sharpness have driven the development of hierarchical variants and complex hybrid systems such as VAEBMs, the quest for even greater fidelity and more robust generation processes has led the scientific community to explore new avenues (Bond-Taylor *et al.*, 2022). Against this backdrop, the next section explores diffusion models, a class of generative architectures that has redefined the state of the art in media synthesis by approaching data generation from an iterative noise-removal perspective.

3.2.3 Diffusion models

Diffusion models represent the most recent and impactful evolution in the field of generative artificial intelligence, overcoming the historical limitations of their predecessors (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025). While GANs suffer from mode collapse (generating repetitive images lacking variety) and VAEs produce blurry-looking images, these new

architectures have achieved mathematical stability and high visual fidelity (Bond-Taylor *et al.*, 2022; Pei *et al.*, 2026). Instead of attempting to generate complex data all at once, diffusion models break this task down into an iterative process, inspired by principles of statistical physics and thermodynamics (Bond-Taylor *et al.*, 2022).

The mechanism consists of two stages (Bond-Taylor *et al.*, 2022):

+ Forward process: progressively adds Gaussian noise to a real image according to a predefined schedule until it is converted into pure noise (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025).

+ Reverse process: A neural network is trained to estimate the noise introduced at each step and gradually reverse it, reconstructing the image step by step (Bond-Taylor *et al.*, 2022; Liu *et al.*, 2025).

Thus, the generation process starts with a tensor of pure noise, refined step by step until it becomes realistic synthetic data (Bond-Taylor *et al.*, 2022; Pei *et al.*, 2026) — a more stable process that avoids the mode collapse problem of GANs (Liu *et al.*, 2025; Pei *et al.*, 2026).

Unlike VAEs, which compress data into a reduced latent space (resulting in a loss of sharpness), diffusion models preserve the full dimension of the original image, resulting in more realistic textures (Liu *et al.*, 2025) — establishing themselves as the current foundation for hyper-realistic content synthesis and editing (Liu *et al.*, 2025; Pei *et al.*, 2026).

This mechanism functions as a conditional system: the network receives a target facial image and additional information (text, audio, or motion) to generate customized media, applying the four categories of manipulation already introduced (**Section 2.2**). In face swapping, the DiffSwap and DiffFace models reduce artificial cutouts but still fail under

severe occlusion or poor lighting (Pei *et al.*, 2026, p. 10–12); in face reenactment, DiffusionAct allows for greater control over the result, although traditional methods are still more robust for large movements (Pei *et al.*, 2026, p. 12–14); in talking face generation, DAE-Talker and VASA-1 eliminate the need for manual face annotations, at the cost of high computational processing (Pei *et al.*, 2026, p. 15–16); and in attribute editing, DiffusionRig preserves identity and fine details, addressing the challenge of avoiding unwanted changes in other parts of the face (Pei *et al.*, 2026, p. 17–18).

To evaluate these media, the literature employs automated metrics focused on two pillars: visual fidelity and sample diversity (Liu *et al.*, 2025). The main metrics are:

+ **FID:** statistically compares real and synthetic images; the lower the value, the higher the fidelity (Liu *et al.*, 2025; Pei *et al.*, 2026).

+ **IS:** measures sharpness and diversity based on a classifier's confidence; the higher the value, the better (Liu *et al.*, 2025).

+ **SSIM:** evaluates structural similarity (luminance, contrast, structure) on a scale from 0 to 1, which more closely approximates human perception than pixel-by-pixel comparison (Liu *et al.*, 2025; Pei *et al.*, 2026).

+ **PSNR:** compares the synthesized image with the original on a pixel-by-pixel basis; the higher the value, the closer it is to the original (Liu *et al.*, 2025; Pei *et al.*, 2026).

In addition to these global metrics, the literature on facial manipulation uses specific criteria: CSIM, which measures identity preservation via facial recognition, and the LSE-C and LSE-D metrics, which validate lip-sync in speaking faces. According to Pei *et al.* (2026), the convergence of these indicators ensures not only visual fidelity but also temporal consistency and identity integrity.

Despite these advances, the transition of diffusion models from the laboratory to practical, large-scale systems still poses critical challenges regarding robustness (Liu *et al.*, 2025, p. 24–25; Pei *et al.*, 2026, p. 2, 5, 26).

3.2.4 Discriminative Neural Networks (DNN) and Convolutional Neural Networks (CNN)

A CNN is a specialized deep learning architecture designed to process data with a "grid-like" topology, such as the pixels that compose an image.

Technically, it is a feedforward network (Dehghani; Saberi, 2025). This means it is a model in which data flows in a single direction, starting from the input layer and passing through the hidden layers until it reaches the output layer, without the information being returned to previous layers during inference processing. Thus, for a fixed set of input data, the system will produce exactly the same output in every run.

Unlike models such as LLMs, which often operate probabilistically by sampling the next token, CNNs are fundamentally deterministic architectures. Although minor numerical variations may occur in practice due to hardware or software specifics, their deterministic nature makes them a common choice for applications that require high consistency and reproducibility, as is the case with many forensic tasks.

It consists of multiple layers, the main ones being convolutional layers, pooling layers, and fully connected layers. Each of these layers serves to extract and transform features in a hierarchical and incremental manner. Convolutional layers use digital filters (kernels) that slide across the image to detect patterns such as edges and textures; pooling layers perform downsampling reducing the spatial dimension of data to decrease computational load and retain only the most relevant information; and fully connected layers receive the processed

features to perform the final classification or prediction.

The core operation of CNN is based on the mathematical operation of convolution, in which digital filters slide over the input data to extract patterns in a hierarchical manner. Due to weight-sharing mechanisms and local connectivity, CNNs can detect features (such as the shape of an eye) regardless of their position in the image (Dehghani; Saberi, 2025). CNNs are used to automatically identify specific points on the face, such as the corners of the eyes and mouth, enabling the creation of a digital mesh for face reenactment (Pei *et al.*, 2026). Models such as MesoNet and Xception use convolutional architectures to perform binary classification of videos, distinguishing whether a face is authentic or synthetic based on texture anomalies.

Discriminative Neural Networks (DNN), on the other hand, represent a category of models whose goal is to learn the mathematical relationship between the features of input data and the category to which it belongs (ANPD, 2024). They focus on learning a decision function that allows classifying information into known groups or predicting the probability that a data point belongs to a specific class. The model is trained to partition the data space into distinct regions. In classification tasks, it categorizes the data into predefined classes, while in regression tasks, it predicts continuous numerical values (ANPD, 2024).

This means that the system works by mapping the statistical properties of the inputs directly to a label or expected value, acting as a verification mechanism that decides whether a sample meets the technical criteria for a given category based on what was learned during training. Thus, in a GAN, the discriminative component acts as a technical validator that assesses the authenticity of the generated samples, attempting to identify statistical anomalies that distinguish them from real data (Pei *et al.*, 2026). Discriminative models are trained to identify malicious material or physiological

inconsistencies — such as the absence of blinks or lip-syncing errors — that indicate the content has been manipulated (ANPD, 2024).

3.2.5 Neural Radiance Field (NeRF)

Neural Radiance Field (NeRF) is a neural network-based technology for representing three-dimensional scenes that allows for the reconstruction of complex objects and environments from two-dimensional images (Pei *et al.*, 2026). In the context of synthetic media and deepfakes, this approach is used to ensure that the generated content possesses depth and spatial consistency.

NeRF uses neural networks to represent a 3D scene implicitly. Unlike traditional models that use geometric points or meshes, NeRF maps spatial coordinates, and the viewing angle, directly to volume density and color values (Pei *et al.*, 2026). This is achieved through the Implicit Neural Field, which is the data structure used by volumetric rendering — the process by which this data is extracted and transformed into an image.

Unlike traditional formats such as JPG (which stores pixels in a 2D grid), NeRF stores the scene within the weights of a neural network, which serves as a function. Upon receiving a three-dimensional coordinate and a viewing angle, the network instantly calculates the color and density of that specific point. Density refers to opacity. A high density indicates that the point is part of a solid object (such as the skin of a face), while a density of zero indicates empty space. Because it is a continuous function, the system does not suffer from "blurred pixels" when zooming, allowing for high-fidelity reconstruction at extremely high resolution within the modeled space.

Next, through volumetric rendering, complex three-dimensional representations are projected onto a two-dimensional

image, determining the visual value of each pixel on the viewer's screen. To perform this projection, the ray marching method is used, in which "virtual rays" are emitted from the camera's perspective and traverse the volume of the modeled scene. Along the path of each ray, the system performs sampling and querying, selecting multiple spatial points to extract color and volume density information directly from the weights of the neural network that comprises the implicit neural field. The final stage is numerical integration, a mathematical operation that sums up the light contributions from all samples collected along the ray. This calculation considers the light emission from each point and how much of that brightness is blocked by points ahead of the ray, resulting in the final color value of a single pixel.

In deepfakes, this integration enables consistency across multiple angles (multi-view). Since the face's identity is stored volumetrically (in the implicit field), the system can render the face from any new perspective simply by changing the ray paths during the rendering process, ensuring that the lighting and geometry of the synthetic face realistically follow head movements (Dehghani; Saberi, 2025). Several architectures use NeRF for specific tasks in synthetic content generation, for example: AD-NeRF (Audio Driven NeRF), HiDe-NeRF, FENeRF, and FaceDNeRF.

3.3 Deepfakes detection

Unlike manual manipulations, the sophistication of content classified as a deepfake lies in its automated production process (Altuncu; Franqueira; Li, 2024). Deepfake software can accurately map anatomical features and extract the vocabulary and emotional nuances from a target's speech to generate entirely new statements, thereby lowering the technical barriers to this type of manipulation (Witness, 2022). For this very reason, the computational detection of deepfakes remains a challenging task. Most methods based on

deep learning for detecting deepfakes or facial manipulation are unreliable and do not explain the reasoning behind the detection result (Akhtar, 2023). This is mainly because deep learning techniques are, by nature, "black boxes".[8]

Despite the challenge of operating as black boxes — where the internal decision-making process is not visible — current deepfake and facial manipulation detectors provide a label, a confidence percentage, or a falsification probability score (Akhtar, 2023). New technologies based on Transformers demonstrate strong technical capability of automatically identifying deepfakes. For example, the Vision Transformers (ViT) system achieves 94% accuracy with a response time of just 26.7 milliseconds. Furthermore, hybrid models perform well even with compressed videos, which is the typical state of files on social media (Khalid *et al.*, 2025; Charpe; Khokale; Dheeraj, 2025). These technologies handle the most common problems in this context well: loss of file quality, new types of tampering (which the system did not encounter during training), and visual or motion artifacts that occur throughout the video (Saga *et al.*, 2025; Charpe; Khokale; Dheeraj, 2025).

On the other hand, these lab-based studies typically measure the system's processing time in isolation. They do not consider the elapsed time from video submission by the user to the final moderation decision — the actual metric that would trigger the automatic removal of deepfakes in the real world (Charpe; Khokale; Dheeraj, 2025; Raza; Ding, 2022). Thus, applying this technology in real-world contexts for detection and removal remains a critical issue because laboratory models lose accuracy when faced with novel forgery methods or low-quality real-world data, and malicious actors may conceal visual artifacts at key moments in a video to evade detection (Rani *et al.*, 2025; Lu *et al.*, 2025; Wang *et al.*, 2022).

Furthermore, there is no standardized way to assess the explanations provided by detectors regarding how they identified the fraud (Qian *et al.*, 2025). To address this, there

8 *Deep learning techniques are considered to function as a "black box" because their internal decision-making processes are not always easily interpretable. Although it is possible to know which data was used as input and what result was generated, the way in which the model combined thousands or millions of parameters to arrive at that result can be difficult to explain transparently.*

are promising proposals, such as converting the system's raw scores into probability calculations that compare the chances of the file being real or fake (Guo; Li; Tang, 2026). Another approach is the use of hashing, unique identifiers, and processing histories to trace the evidence (Bhondave *et al.*, 2026; Williams *et al.*, 2025). This indicates that the field has advanced further in detecting forgeries than in providing repeatable and legally admissible explanations of how a system reached its conclusion (Kraetzer *et al.*, 2022; Qian *et al.*, 2025).

These detectors only gain trust and legal validity when functioning as support tools for human experts — allowing for verification and contestation — rather than as fully automated systems (Siegel *et al.*, 2024; Sandoval *et al.*, 2024). The scientific literature also points to other solutions currently under development, such as watermarks, blockchain, and origin tracking, but these require more extensive testing, official certification by countries, and adaptation to local laws before they can become established tools (Bhondave *et al.*, 2026; Kraetzer *et al.*, 2022).

3.4 Technical limitations, uncertainties, and biases

Despite recent technological advances, the identification of deepfakes still faces several technical limitations and uncertainties that hinder the accuracy of results. In practice, current analysis tools struggle to process low-resolution media or clips of highly variable lengths, which often leads to inconsistent assessments — such as flagging original content as false positives and videos that have actually been manipulated as false negatives (Moraes; Batista, 2025).

In addition to technical detection flaws, verification systems face a problem related to algorithmic biases. Most systems are trained using highly homogeneous databases, with very limited variety of ethnicities, genders, and ages. This lack of

diversity in the training samples limits the AI's ability to learn facial features and characteristics, leading to significant performance differences when analyzing distinct demographic groups and compromising the reliability of these tools in real-world situations (Prabhu *et al.*, 2026, p. 6).

Since current technological barriers still frequently fail due to variations in resolution and algorithmic biases, false content can bypass verification and spread, causing profound damage that undermines everything from the credibility of media sources to the authenticity of evidence in the justice system (Moraes; Batista, 2025, p. 3; Prabhu *et al.*, 2026, p. 1). Below, we will discuss in greater detail some of the technical limitations and possible algorithmic biases.

Technically, the limitations of detection systems arise from inefficiencies in the mathematical rules that process data within the system, leading to inconsistent assessments in real-world scenarios.

Some deepfake detection techniques fall under the umbrella of traditional forensic methods and biological signal analysis (Kaushal *et al.*, 2025, p. 5; Prabhu *et al.*, 2026, p. 6). Traditional forensic methods are based on signals and manual analysis of file characteristics. Biological signal analysis, by contrast, is a technique in which the system examines natural indicators of life in the video by tracking physiological and behavioral traits, such as blink rate, pulse, and subtle facial expressions. Unlike modern deep learning architectures that automatically discover patterns, these tools operate using statistical classifiers, physical metrics, and mathematically predefined tests. Because they rely on very explicit signals or specific software defects, these models face obsolescence given modern neural networks' ability to regularize and eliminate such imperfections (Kaushal *et al.*, 2025, p. 5; Prabhu *et al.*, 2026, p. 6–7).

In systems based on CNNs, however, the main technical problems are "overfitting" and excessive dependence on

databases. Instead of learning to identify the essence of visual manipulation, the algorithm often memorizes specific defects from its training, such as the video compression level, the lighting pattern, or even the image background. Due to this bias, when the system needs to analyze a deepfake created using an unfamiliar method, its accuracy drops dramatically (Prabhu *et al.*, 2026, p. 8).

Another identification technique is image frequency analysis, which, instead of focusing on the colors, shapes, and details visible to the human eye, analyzes an image as a composition of waves and frequencies. In real life, authentic photographs captured by cameras exhibit a natural, smooth, and continuous flow in these waves (Prabhu *et al.*, 2026, p. 6).

When artificial intelligence attempts to generate a face, it cannot perfectly mimic this harmony. This is because the way the AI system constructs synthetic media leaves mathematical "scars" in the file that are invisible to the human eye. As the system assembles the image in stages and applies artificial filters, it introduces repetitions and abnormal wave peaks that do not exist in reality. Thus, detection programs actively search for these "digital fingerprints" left by the system to determine whether the image is fake (Prabhu *et al.*, 2026, p. 6).

The major vulnerability of this approach is that these digital scars are very fragile and easy to disguise. If the forgery system applies even a minor effect to the image, these scars can disappear. Some ways to do this include blurring the image, applying a blur effect, enhancing the resolution, or posting the video on social media, which compresses the file. Once editing erases these traces, the verification system can no longer detect the deepfake's flaws and gets deceived, rendering the analysis ineffective if used on its own (Prabhu *et al.*, 2026, p. 6–7).

In this scenario, there are more advanced artificial intelligence models that attempt to analyze the overall context of the face

and the image, such as the Transformers architecture, which is a neural network architecture built on a feature called self-attention — or attention mechanism — that teaches the model to analyze all the information at once and focus on the essential details (Prabhu *et al.*, 2026, p. 8).

These Transformer-based verification models divide the image into several small blocks and use the mathematical mechanism called "self-attention" to connect and compare all these parts simultaneously. As a result, the system can evaluate the relationship between distant areas of the photo, verifying, for example, whether the lighting on the face makes sense given the shadow in the background, or whether the geometric proportion between the eyes and the chin is natural. In this way, the system searches for errors in the entire image, focusing on the scene's coherence to detect forgery (Prabhu *et al.*, 2026, p. 8–9).

The vulnerability of this approach lies not in the fragility of digital clues, but rather in its high operational and structural costs. Since the system must cross-reference information from all image blocks simultaneously to understand the overall context, it requires extremely high computational processing power. This technical requirement makes the system very resource-intensive and slow, which makes it impractical for analyzing videos in real time or for using it on everyday devices and cell phones, limiting its use to more powerful computers and specialized laboratories (Prabhu *et al.*, 2026, p. 9).

Other detection approaches also have clear technical vulnerabilities. In addition to machine architecture flaws, commercial verification tools face direct operational limitations tied to the physical characteristics of the file.

Practical tests show that the algorithms lose their analytical capability when processing low-resolution or long-duration videos. In such cases, the systems generate false negatives. This means that the tool flags fake content as real. This error

occurs because the extended duration of the video introduces excessive variability in the analyzed facial data, which compromises the detection of fraud (Moraes; Batista, 2025, p. 3, 6). On the other hand, very short videos do not provide sufficient visual samples for the system to work with, which often leads to uncertain results, conflicts among the tool's own evaluation modules, and false positives in perfectly authentic content (Moraes; Batista, 2025, p. 4–6).

Beyond the technical aspects related to algorithms, the problem of bias is not solely associated with a lack of diversity. Another important issue in the construction of training databases for algorithms is the excessive focus on images of celebrities. When a detection system — trained primarily on the faces of famous people — attempts to analyze videos of ordinary individuals or those with characteristics that deviate from the pattern fed into the machine, it often fails and produces inaccurate results (Kaushal *et al.*, 2025, p. 6).

This highly homogeneous scenario creates what researchers call "spurious correlations," that is, false associations. Since artificial intelligence learns from very limited examples, it may take shortcuts and associate elements that have no direct relation to facial forgery itself (Prabhu *et al.*, 2026, p. 6). For example, the algorithm may link a certain type of lighting or the color of the background in a video from the database to the concept of "fake," coming to regard these irrelevant background details as the true proof that the media has been manipulated (Prabhu *et al.*, 2026, p. 6).

Another form of structural bias occurs when the algorithm becomes "addicted" to the very program that generated the images used in its training (Prabhu *et al.*, 2026, p. 6). Instead of learning to look for the real, fundamental traces of manipulation, the system merely memorizes the specific software's "fingerprint", as if it were memorizing the flaws and visual artifacts unique to a single generator (Prabhu *et al.*, 2026, p. 6).

In this scenario, deepfake detection faces a constant race against the evolution of manipulation techniques (Prabhu *et al.*, 2026, p. 11). Technical limitations and data biases show that there is still no foolproof system. To overcome these barriers, researchers suggest that future systems need to be trained with a wider variety of examples and operate quickly, in real time (Kaushal *et al.*, 2025, p. 6).

It is also necessary to invest in the development of multimodal models. These are more comprehensive tools that analyze a person's audio, image, and movements simultaneously. Another essential advancement is the use of Explainable Artificial Intelligence (XAI). This feature details exactly how the machine reached a given conclusion, bringing greater transparency and trust to the analyzed results (Kaushal *et al.*, 2025, p. 6; Prabhu *et al.*, 2026, p. 11).

Finally, the deepfake technology does not operate in isolation. To ensure a truly secure and efficient analysis, it is crucial to combine the use of these automated tools with critical human evaluation (Moraes; Batista, 2025, p. 6).

‹ 4 deepfakes and personal data protection ›

This section is dedicated to understanding how the emergence and spread of deepfakes pose risks of violating the rules governing the protection of personal data. These risks are linked to three basic purposes for which personal data may be processed, namely: (i) model training, when data is used as input for the algorithm's learning process; (ii) the synthesis or manipulation of visual or audio media, resulting in generated content; and (iii) dissemination in the digital environment. In all these cases, the Brazilian legal system already provides safeguards and enables the identification of data protection risks related to the production and circulation of deepfakes.

Thus, the protection of personal data is at risk during the training of artificial intelligence models for the generation of deepfakes. This occurs, for example, when such processing is conducted without an adequate legal basis, lacking transparency, for secondary purposes, or without proportionate safeguards — especially when involving sensitive data, such as biometric data or data relating to a person's sexual life. Similarly, this category also includes the use of data from children and adolescents without regard for their best interests and without compliance with the requirements applicable to the legal basis used — when the basis is consent, it must be specific, clearly indicated, and provided by at least one parent or legal guardian.

In both cases, the processing of data for training deepfakes entails limitations on data subjects' informational self-determination[9] and restrictions on their privacy, freedom, and the free development of their personality. These rights are protected by constitutional provisions, by Law No. 13,709 of 2018 (Brazilian Data Protection Law — LGPD), by Law No.

9 *Informational self-determination is the principle according to which a person must have the power to control information about themselves and, therefore, their personal data. It ensures that individuals have the autonomy to decide when, how, by whom, and for what purpose their data may be processed. In the Brazilian Data Protection Law, informational self-determination is listed as one of the foundations of personal data protection, as provided for in Article 2, item II (Brazil, 2018)*

15,211 of 2025 (Digital Statute of the Child and Adolescent — Digital ECA), and by other legislation, such as Decrees No. 12,975 and 12,976 of May 20, 2026.

As for the purpose of actually generating synthetic or manipulated media — when the process results in a deepfake — the risks to data protection are amplified. The production of deepfakes using seemingly authentic images or videos of someone without their authorization can pose risks related to the free development of one's personality and the individual's right to their own image. This is because a person's image is linked to characteristics that represent them in social life, such as their honor, reputation, or dignity. A threat to someone's public image, in this sense, can pose risks to them, even in cases where the subject matter of the media is not unlawful in and of itself but may nevertheless harm their public image. This phenomenon becomes particularly concerning when deepfakes are produced for the purpose of spreading false, defamatory, and/or humiliating content in the digital environment, as in cases involving the synthesis and manipulation of non-consensual pornographic media. Data from the organization Security Hero reveal that, in 2023, 98% of deepfakes involved pornography, and in 99% of these cases, the victims were women (Security Hero, 2023). Such cases threaten the victims' dignity and may cause damage to their image, honor, and privacy; they constitute a form of digital violence against women, expressly provided for in Decree No. 12,976/2026,[10] which seeks to ensure the protection of women online.

The widespread circulation of pornographic media, in turn, increases the risk of children and adolescents being exposed to content that is inappropriate, unsuitable, or prohibited under the Digital ECA. As provided for in the 2025 law, suppliers of digital products or services, application providers, websites, and other entities covered by the regulation must adopt appropriate measures to protect children and adolescents and ensure that they access content appropriate for their age group, while respecting their progressive autonomy.[11] In this

10 *The Decree recognizes the seriousness of the risk described above and prohibits internet application providers from generating or modifying intimate content belonging to third parties, requiring that technical and procedural safeguards be implemented to identify and block such requests by providers using artificial intelligence (Brazil, 2026).*

11 *The topic of age verification, which relates to progressive autonomy, was discussed in Technology Radar No. 5, prepared by the ANPD and available at: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar-tecnologico-5-mecanismos-de-afericao-de-idade-em-lingua-inglesa.pdf/@@display-file/file.*

context, the high volume of pornographic deepfake media contributes to violations of regulations designed to protect children and adolescents in the digital environment.

Given this scenario, the following subsections are dedicated to understanding and detailing the risks posed by deepfakes to the protection of personal data. To this end, we analyze the categories of data involved, the methods of collection and processing, and the relationship between these practices and the principles of the Brazilian Data Protection Law (LGPD). Finally, we discuss the social risks arising from the misuse of deepfakes — made possible by their technical functioning — which extend beyond the personal data protection ecosystem and affect other rights.

4.1 Identification of personal data processed

When discussing deepfakes, the first step toward a proper analysis from a data protection perspective is to understand — or, more precisely, to recognize and classify — which personal data are being used throughout this process. This identification is not always immediate or evident. Unlike traditional data processing, deepfakes operate in layers, involving provided, observed, derived, and inferred data.

According to Mussalo, Gain, and Hotti (2018, p. 56–57), these classifications can be outlined as follows: provided data consist of data consciously and actively submitted by individuals, representing the category in which the data subject has the highest level of awareness regarding the collection; observed data are recorded automatically based on the user's actions or environment (such as cameras, clicks, or internet sensors); derived data consists of new elements created mechanically from other data using relatively simple logical, mathematical, or arithmetic operations; and, finally, inferred data consists of the production of profiles, categorizations, or probable trends resulting from complex analytical processes that seek correla-

tions across various databases. The authors emphasize that, as data progresses from "provided" to "inferred," an individual's level of awareness regarding their own data may decrease considerably, and that the data may be significantly transformed.

Therefore, deepfakes involve the processing of multiple categories of personal data. This highlights the central role of data in the functioning of the technology, which can be grouped as shown in Table 2 below:

12 *Brazil's individual taxpayer registration number, similar to the US social security number.*

Table 2 *Categories of personal data*

Data Category	Description	Examples	Regulatory Implications and Risks (LGPD)
Personal Data	Information related to an identified or identifiable natural person	Name, CPF[12], email, ID number, etc.	Processing may only be conducted if there is a legal basis (e.g., consent, compliance with a legal obligation, performance of a contract). In the event of security incidents or improper processing, the controller/processor is required to compensate for financial, moral, individual, or collective damages
Sensitive personal data	Personal data regarding racial or ethnic origin, religious beliefs, political opinions, membership in a labor union or an organization of a religious, philosophical, or political nature, as well as data related to health or sexual life, and genetic or biometric data, whenever linked to a natural person.	Genetic or biometric data linked to a natural person, including facial, vocal, gestural, or behavioral patterns when processed for the purpose of identifying, authenticating, or distinguishing the person.	Data processing requires a legal basis as provided for in Article 11 of the LGPD. The data controller must use only the data strictly necessary for the intended purpose, implement security measures appropriate to the risks, and ensure transparency regarding the processing conducted. Consent, when used, must be specific and clearly indicated for a specific purpose.

Data Category	Description	Examples	Regulatory Implications and Risks (LGPD)
Data Anonymized	Data relating to a data subject who can no longer be identified, considering the use of reasonable technical means available at the time of its processing	Data used in databases for statistical or public health studies managed by research institutions, where the individual's identification has been removed or maintained separately in a controlled environment.	Such data is not considered personal data for the purposes of the LGPD. The main risk is the reversal of anonymization: if anonymization can be reversed through appropriate means or reasonable efforts, the data may once again be treated as personal data, subject to the obligations under the LGPD.
Data on Children and Adolescents	Information relating to children and adolescents, the processing of which must prioritize the best interests of the child, age-appropriate transparency, data minimization, and enhanced safeguards.	Information collected for participation in games, internet applications, and other online activities	The processing of children's data requires specific and prominently displayed consent from at least one parent or legal guardian, when this is the legal basis used. It is prohibited to condition a child's participation in games or applications on the provision of more information than is strictly necessary. Information regarding the processing must be accessible and appropriate for the child's understanding

Source: Prepared by the authors based on the Brazilian Data Protection Law (LGPD).

One aspect that is often overlooked is that deepfakes can **repurpose and recombine** publicly available **content** to create highly realistic simulations. In this regard, given that various jurisdictions have been working to develop regulations governing the protection of personal data, especially sensitive data, how should situations be addressed when the data subject themselves posts a video on social media? When they consciously choose to complete forms, register biometric data, share photos, or express opinions? Even though publicly available content is protected, the central question is whether

the new use created by deepfakes is compatible with the original purpose and the data subject's legitimate expectations.

Recent research shows that training datasets obtained through data scraping[13] often contain personally identifiable information, even after attempts at anonymization. In a forensic study of a popular large-scale dataset, Hong *et al.* (2026) noted a considerable presence of personally identifiable information (PII) despite the platforms' sanitization efforts, corroborating the concern that datasets obtained through web scraping may contain personally identifiable or re-identifiable data, including sensitive data. Applying this logic to the Brazilian context, the public availability of content does not, in and of itself, exempt the processing of personal data contained in that content from the scope of the LGPD; any new processing must comply with the principles of purpose, good faith, public interest, data subjects' rights, and an appropriate legal basis. As Lazzarini (2025, p. 10) adds, the training of these generative artificial intelligence systems depends on the massive ingestion of data collected from the internet, which is incorporated into training models often without adequate transparency or the consent of data subjects.

13 *Data scraping, or web scraping, is the technique of automatically extracting data from websites, transforming unstructured content into organized, tabular information for analysis. Also known as "web data scraping", this process allows a program to collect data quickly and systematically, rather than copying it manually.*

The case of DataComp CommonPool,[14] presented by Hong *et al.* (2026), for example, gained worldwide attention after an academic audit revealed that one of the largest open datasets used to train artificial intelligence models, with approximately 12.8 billion image-text pairs, contained a massive amount of personal data — including sensitive data — collected through web scraping over the course of years. Although the investigation analyzed only a tiny fraction of the dataset (about 0.1%), it found thousands of examples of personal documents such as passports, credit cards, certificates, and driver's licenses, as well as images with clearly identifiable faces.

14 *DataComp CommonPool (https://www.datacomp.ai/dcclip/) is a large dataset of 12.8 billion image-text pairs designed for multimodal AI research.*

The most concerning aspect reported in this article was the identification of documents related to personal and professional life, such as résumés and job applications, including

information on race, disability, address, and other details. Based on this sample, the researchers estimated that the complete dataset may contain hundreds of millions of records containing this type of information.

In the Brazilian context, the debate on deepfakes and the identification of processed personal data is directly linked to the LGPD, especially with regard to the processing of publicly accessible data and compliance with the principles of purpose and transparency, as well as the general duty of good faith that guides the processing of personal data under the LGPD. The recognition that sensitive data — such as race, health, or information that could exacerbate vulnerabilities — can be captured and reused on a large scale highlights the limitations of the provisions governing publicly accessible personal data (under Article 7) and sensitive personal data (under Article 11 of the LGPD), and raises concerns about practices based on indiscriminate data scraping. This is a risk that has already been widely debated in Brazil, with concrete implications for phenomena such as algorithmic discrimination and the misuse of data in fraudulent contexts — a topic directly linked to the central theme of this edition of Technology Radar.

The DataComp CommonPool incident underscores the urgency of more specific guidance from regulators and technology providers themselves, as well as the need to update and revise the regulatory frameworks applicable to AI training datasets. The case demonstrates that this is not a traditional or isolated "data breach," but rather a structural process of collecting, reusing, and redistributing personal data on an industrial scale — a process that is often normalized and occurs without the knowledge or control of the data subjects.

This dynamic thus reveals that the problem lies in the very structural logic of the development of AI and deepfake systems, which normalizes the massive appropriation of personal data under the label of "public data." In this scenario, the identification of personal data is not limited to the original

data but also encompasses the ways in which it is transformed, combined, and re-presented in new digital contexts, requiring a broader understanding of the data lifecycle. This is particularly relevant because the generation of deepfakes relies on models trained on large volumes of real data, capable of extracting biometric and behavioral patterns on a large scale for their continuous refinement.

4.2 How is personal data collected, processed, and shared?

As we have seen so far, the collection of personal data is a fundamental step and occurs through channels that are often combined to feed the systems. These data are indispensable for training artificial intelligence models, such as GANs and diffusion models. In turn, these technologies are responsible for synthetically reproducing faces, voices, and gestures with an impressive level of realism.

Data scraping is one of the most common forms of collection, as it involves the automated extraction of publicly available content from social media, including videos and photos, news articles and public interviews, websites, and stock image databases. Through this process, facial images from different angles, audio recordings of people's voices, expressions, movements, and speech patterns are collected.

Another way to obtain data is through structured databases, such as datasets, as was the case with DataComp CommonPool, which uses pre-organized datasets for training. Other examples of datasets presented by Pei *et al.* (2026) include: **Labeled Faces in the Wild (LFW)** — an image-focused dataset; **VoxCeleb** — a video and audio dataset focused on human speech with over 100,000 samples; and **FaceForensics++** — a dataset for detecting forgeries. These datasets often include personally identifiable or potentially re-identifiable data, including biometric data, which raises

significant controversy regarding the consent of data subjects.

A fourth form of data collection we can consider here is the direct and intentional capture of data through apps and platforms that request selfies, recordings, and/or biometric verification. In this case, there may be explicit consent from the data subject or even misuse for illegitimate purposes. This scenario leads us to consider a fifth possibility for data collection: data obtained from leaked or improperly acquired content, such as through the hacking of user accounts.

In the context of deepfakes, it is important to note that data are not only collected but also extracted, inferred, combined, and reused in new contexts. With regard to processing, personal data undergo technical stages that include storage, data cleaning and organization, labeling to identify specific characteristics, the extraction of biometric attributes, and the iterative training of machine learning models, as we saw in Section 2 of this Technology Radar. Furthermore, processing may involve data augmentation techniques and the transformation of these data into abstract mathematical representations, which makes direct identification difficult but does not eliminate the risk of re-identification.

Once collected and processed, the data and properly trained models can be distributed or shared. The sharing of data and models can occur at different levels of the digital ecosystem, including exchanges among companies, researchers, and technology platforms, and involves both the sharing of datasets and that of pre-trained models, APIs, and AI infrastructure — often on a transnational scale. Such sharing may serve numerous purposes, including academic, commercial, and/or political on the one hand, but also illegitimate or unlawful purposes on the other. In other words, data can be reused in contexts different from those originally intended, which hinders traceability, accountability, and control by data subjects. The final dissemination can be viral and decentralized, originating from social media profiles, subscription-based

apps, and fraudulent websites, for example. There are numerous other risks that will be discussed below, as well as their connection to the principles of the LGPD.

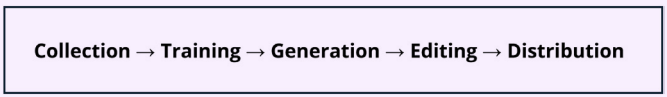


Figure 10 *Data flow in the creation and dissemination of deepfakes.*

Source: Prepared by the authors.

In summary, as illustrated in Figure 10, the flow of data collection, processing, handling, archiving, and sharing proceeds in this manner, with personal data undergoing a continuous and cumulative cycle of processing, in which each stage increases privacy risks, particularly due to the possibility of recombination and unforeseen secondary use.

4.3 Connection to the principles of the LGPD

The previous sections have demonstrated how the risks associated with the deepfake ecosystem extend across the entire personal data processing lifecycle, encompassing the collection, processing, and sharing of personal data for model training and the generation of deepfakes. These risks are directly connected to the foundational principles of the LGPD, which provide important regulatory parameters for assessing the legitimacy and appropriateness of such practices.

A key issue that emerges when examining deepfakes through the lens of the LGPD is that these tools rely heavily on the large-scale collection of personal data, often obtained through data-scraping techniques, as discussed above. This process raises several concerns that may conflict with the principles established by the LGPD.

As previously explained, training AI models on large volumes of data may involve directly identifiable personal data, such as images and voice recordings, as well as sensitive personal data, including information related to race, health, sexual orientation, and biometric data linked to identified individuals. In many cases, these data are collected without the data subjects' knowledge or reasonable expectations, potentially violating the principle of **transparency**, which requires that data subjects be informed in a clear, accurate, and accessible manner about the purposes and methods of processing their personal data.

This context of widespread and automated data collection also frequently involves the reuse of personal data for purposes other than those for which they were originally collected. This is particularly the case with model training based on data originally shared for social, professional, or informational purposes, for example. Such practices may conflict with the **purpose limitation** principle, which requires personal data to be processed for legitimate, specific purposes that have been previously communicated to the data subject. Secondary processing for purposes that are neither clear nor disclosed to the data subject constitutes a violation of this principle.

Furthermore, in automated collection scenarios, such as when data scraping is used for model training, data subjects are often unaware that their data have been collected and, consequently, cannot access information about how they are being used, for how long, or by whom. This may undermine the LGPD's principle of **free access**.

Beyond the issues arising from the further use of personal data obtained through data scraping, the aggregation of large volumes of personal data for model training increases the risk of data breaches, unauthorized access, and improper reuse, raising concerns under the LGPD's **security** principle. In addition, the use of deepfakes to compromise biometric authenti-

cation and age assurance systems, for example, amplifies risks to personal data security across digital ecosystems.

The rapid technological evolution of deepfakes, as discussed throughout this Technology Radar, may also render existing preventive measures insufficient to protect personal data. The absence of effective safeguards against the risks posed by these technologies may facilitate fraud, extortion, digital violence, and other forms of harm that effective preventive measures should seek to anticipate and mitigate, in accordance with the LGPD's principle of **prevention**.

Similarly, the use of deepfakes for improper or unlawful purposes — such as in cases of fraud based on voice and image cloning and multimodal techniques, as well as non-consensual pornography, which will be discussed below — may conflict with the LGPD's principle of **non-discrimination**. Such tools can facilitate abusive practices that disproportionately affect socially vulnerable groups, including older adults, adolescents, and women. This illustrates how malicious use of certain technologies can exacerbate existing inequalities in society.

Finally, the distributed and transnational nature of the deepfake ecosystem makes it difficult to identify those responsible for the improper processing of personal data and the unlawful use of such media. This may hinder the ability to demonstrate compliance and the adoption of effective governance measures associated with the LGPD's **accountability** principle.

It is worth noting that, given the emerging risks associated with practices commonly involved in the creation of deepfakes and their potential to violate LGPD principles, the suitability of consent as a legal basis for the processing of personal data warrants careful consideration. Under the LGPD, consent constitutes "a free, informed, and unambiguous expression by which the data subject agrees to the processing of their personal data for a specific purpose" (Brazil, 2018). Obtaining valid consent may be particularly challenging in

processing operations involving the large-scale and indeterminate collection of personal data available online, especially when such data are subsequently reused for purposes other than those for which they were originally collected.

In this context, reliance on legitimate interest as an alternative legal basis also requires caution, since legitimate interest is not among the legal bases provided for in Article 11 of the LGPD for the processing of sensitive personal data. Thus, when facial images, voice recordings, or behavioral patterns are processed as biometric data linked to an individual, an appropriate legal basis under that article must be identified. Given that facial images and voice recordings are central to the training and generation of deepfakes, their large-scale processing based on legitimate interest raises concerns regarding the proportionality and necessity of such processing and its impact on the fundamental rights of data subjects, as discussed throughout this section.

Finally, interpreting Article 7, paragraph 4 of the LGPD as broadly waiving the requirement for consent whenever personal data — including sensitive personal data — have been "manifestly made public", often because they were shared online by the data subjects themselves, would be inappropriate.[15]

Personal data made available in specific contexts, such as interactions on social media, the submission of documents for recruitment purposes, or the sharing of personal information in digital environments, should not be used for purposes unrelated to the contexts in which the information was originally disclosed. The disclosure of personal data should not be understood as automatically authorizing their use for other purposes, nor should it be considered sufficient to override the safeguards established by the LGPD and other applicable provisions of Brazilian law.

15 *"The consent requirement set forth in the head provision of this article shall not apply to data manifestly made public by the data subject, safeguarding the rights of the data subject and the principles set forth in this Law" (Brazil, 2018).*

It is therefore important to emphasize that the processing of personal data, particularly sensitive personal data, must always comply with the relevant data protection legal requirements. Although deepfakes are relatively recent technological tools, their creation does not take place in a regulatory vacuum. The principles established by the LGPD already provide guidelines applicable throughout the deepfake production process, helping to ensure that the processing of personal data involved does not infringe upon the rights of data subjects.

In higher-risk applications, particularly those involving biometric data, children and adolescents, authentication, profiling, the generation of intimate content, or the large-scale dissemination of synthetic media, enhanced governance documentation, risk assessments, security measures, retention limits, controls over data sharing and access, and mechanisms for enabling the exercise of data subjects' rights are recommended. Where applicable, these measures may also include the preparation of a Personal Data Protection Impact Rapport (RIPD, in Portuguese).

4.4 Specific risks related to technical operations

In addition to the risks associated with the types of data processed to generate deepfakes and the ways in which such data are collected and processed, synthetic or manipulated media also give rise to a range of social risks associated with their use, extending beyond the scope of personal data protection. As described throughout this Technology Radar, the technical processes involved in creating deepfakes often include training artificial intelligence models on large volumes of data, such as images, videos, and voice recordings of real individuals. As a result, the reproduction, combination, and simulation of biometric characteristics are becoming increasingly realistic not only to the human eye but also to security systems that rely on biometric data.

In this context, one harmful use of deepfakes that poses significant social risks is the **bypassing of identity verification mechanisms.** Real-time multimodal deepfakes, that is, those that combine visual and audio elements and simulate movement in real time, can compromise biometric-based authentication mechanisms (Altuncu; Franqueira; Li, 2024, p. 2).

Facial recognition systems and liveness detection systems, for example, seek to verify whether the captured biometric data originates from a natural person who is physically present at the time of authentication, rather than from a static or synthetic reproduction. For this reason, these systems are widely used by financial institutions and banking services to unlock devices and applications, as well as by digital public identity services, e-commerce platforms, and other higher-risk digital services.

Even though they provide high levels of security, these systems may be vulnerable to advanced deepfakes capable of replicating liveness cues in real time (Altuncu; Franqueira; Li, 2024, p. 2; Pei *et al.*, 2026, p. 9). In such cases, deception enabled by deepfakes can cause significant social harm, for example, by allowing unauthorized access to bank accounts or other financial services, facilitating identity theft, or enabling the theft of sensitive personal data, including biometric data.

These same capabilities can also be used to bypass **age assurance mechanisms.** As mentioned earlier, Digital ECA requires the adoption of age assurance mechanisms by providers of technology products or services targeted at or likely to be accessed by children and adolescents. One of the main technical methods used for this purpose is facial biometrics. In this context, deepfakes capable of bypassing identity verification mechanisms may also be used to overcome age restrictions in digital environments, providing children and adolescents with a means of evading safeguards designed to protect them (Colman, 2026; Lucca; Braga, 2026). This may increase their exposure to content that is inappropriate, unsuitable, or prohibited for their age group.

The emerging risks associated with deepfakes go beyond bypassing technical security measures and include their potential use in social engineering schemes. In this context, the growing use of these tools to execute **ultra-realistic financial fraud** is deeply concerning. In such cases, the technical sophistication of deepfakes can be used to simulate video conferences or clone the voices of senior corporate executives, enabling malicious actors to impersonate them and request bank transfers from employees or business partners, for example (Fukuha; Furst; Archegas, 2026, p. 185–186).

Other forms of social engineering involving deepfakes include the **creation of fake advertisements** featuring manipulated and unauthorized videos of Brazilian and international celebrities (Fukuha; Furst; Archegas, 2026, p. 186–187). In these cases, deepfakes are used to simulate the participation of public figures in online advertisements. Seemingly authentic videos of celebrities promoting products or services lend credibility to these advertisements, which typically seek to induce consumers to provide personal data or make fraudulent payments, such as purported shipping costs or fees (Romagna, 2025).

Another growing concern in the context of social engineering is the use of deepfakes to execute **scams targeting older adults by cloning the voices of family members or acquaintances**. In some cases, a voice can be cloned using only a few seconds of audio from the targeted individual, obtained from social media or voice messages, resulting in highly convincing synthetic voices. These scams exploit the realism of cloned voices, lower levels of digital literacy among some older adults, and the strong emotional appeal of the messages, contributing to the growing occurrence of extortion and fraud targeting this group in Brazil (Fukuha; Furst; Archegas, 2026, p. 211; Lazzarini, 2025, p. 13).

Finally, some of the most serious risks associated with deepfakes arise from their use to generate **non-consensual por-**

nography and to facilitate the **sexual exploitation of children** (Kshetri; Voas, 2026, p. 105; Meding; Sorge, 2025, p. 158).

As discussed throughout this Technology Radar, the techniques underlying deepfake generation enable the creation of fake, entirely synthetic images and videos without the participation or knowledge of the person depicted. Models trained on large datasets of images and videos can be used to "insert" a specific person's face into sexually explicit content or to create the appearance of "removing a person's clothing" from an image in which the person was originally clothed, a practice commonly referred to as *deepnude*. The resulting content can be shared with others, disseminated on a large scale, or even used for ***sextortion***, a form of blackmail in which perpetrators threaten to disseminate manipulated intimate content depicting the victim unless a payment is made (Fukuha; Furst; Archegas, 2026, p. 184).

Pornographic content has been closely linked to deepfakes since their emergence and continues to constitute the vast majority of deepfakes currently in circulation. The use of these tools to produce sexual content is facilitated by technical factors, including the increasing accessibility of features for generating pornographic content in deepfake applications. According to 2023 data, one in three deepfake tools allowed users to generate pornographic content (Security Hero, 2023).

The extent to which this technology can be used for discriminatory purposes is documented in specialized literature. A study that collected and analyzed 105 apps available on the Google Play Store in December 2024 found that 89.5% of the deepfake apps examined posed risks to women, meaning that they "normalize objectification, perpetuate gender stereotypes, and/or are potentially exploited to generate sexual content featuring women's bodies" (Cuzcano-Chavez; González-Véliz, 2025, p. 53, our translation).

The study further suggests that the risks to women extend beyond the creation of pornographic deepfakes and involve design choices that may encourage sexist uses of these tools. Examples include features that enable users to simulate kisses, hugs, and intimate scenes involving themselves and third parties, as well as templates explicitly designed to sexualize women. Such templates depict women with exaggerated physical attributes, low-cut necklines, and underwear, while men are shown fully clothed. These findings suggest that the design and features of deepfake applications can facilitate harmful and discriminatory uses of this technology, with significant potential for harm to women. Figure 11 illustrates some of these findings:

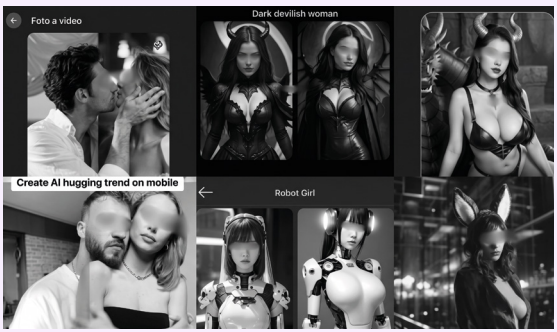


Figure 11 *Screenshots of deepfake applications*

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

Thus, the use of deepfakes to generate non-consensual intimate content constitutes a form of image-based sexual abuse and, more broadly, technology-facilitated gender-based violence (Brigham; Wei; Kohno; Redmiles, 2024, p. 374). Although this form of abuse takes place in digital environments, its consequences extend well beyond them. The creation of non-consensual pornographic deepfakes can have tangible effects on victims' mental health and social lives, including post-traumatic stress, anxiety, depression, social isolation, low self-esteem, trust issues, and a sense of bodily violation (Brigham; Wei; Kohno; Redmiles, 2024, p. 374).

The production of non-consensual sexual deepfakes therefore poses both a quantitative and qualitative challenge to ensure safe digital environments. Against this backdrop, as mentioned in the introduction to this section, Decree No. 12,976 of May 20, 2026, prohibited internet application providers from generating or modifying third parties' intimate content using artificial intelligence.

In addition to sexually explicit content involving adults, there are also cases involving the generation of synthetic content depicting the sexual exploitation or abuse of children (UNICEF, 2026). In such cases, the content may or may not depict real, identifiable victims. However, even when the content is entirely synthetic, its production may rely on the collection and processing of personal data relating to real children and adolescents, such as images, facial features, and other physical characteristics. Once generated, such content can be used to facilitate grooming, extortion, cyberbullying, or other forms of child abuse (Kshetri; Voas, 2026, p. 105), in conflict with the principle of comprehensive protection of children and adolescents established under the Federal Constitution, the Statute of the Child and Adolescent (ECA), and the Digital ECA.

In summary, the uses of deepfakes described in the preceding sections may conflict with the Brazilian regulatory framework and give rise to risks that are directly relevant to personal data protection while also extending into broader areas of social harm. The next section examines how these technologies have been observed in the Brazilian context.

‹ 5 deepfakes in the brazilian context ›

So far, we have discussed what deepfakes are, their core concepts, applications, risks, relationship with the LGPD, and technical limitations. At this point, deepfakes can be understood as one of the most advanced examples of the convergence between technological innovation and the large-scale dissemination of information in today's digital environment.

In Brazil, this technological advancement has already resulted in concrete and measurable cases. The use of synthetic content in digital fraud, disinformation, and identity manipulation has increased, particularly in environments where information circulates rapidly and extensively, such as social media platforms and messaging apps. This scenario demonstrates that deepfakes are no longer a futuristic concern. They have become a current, relevant, and urgent issue with significant economic and social impacts.

Decree No. 12,975/2026, dated May 20, 2026, establishes an accountability, governance, and content moderation framework that indirectly encompasses this type of material in the digital environment. The Decree amends the regulations governing the Brazilian Civil Rights Framework for the Internet (Law No. 12,965/2014) and expands the obligations applicable to internet application providers. It also incorporates the Federal Supreme Court's (STF) revised interpretation of Article 19 of that law, advancing a shift from a predominantly reactive model — such as one based on responding to noncompliance with a court order — to a model centered on the duty of care and risk management.

Decree No. 12,976/2026, in turn, expressly addresses the generation and modification of intimate content through

artificial intelligence or similar technological resources. Under the Decree, which establishes guidelines for the protection of women in digital environments, internet application providers are prohibited from generating such content and are required to implement technical safeguards to prevent requests for its creation.

This regulatory movement is not occurring in isolation. Rather, it is accompanied by initiatives undertaken by various institutions across the country. Notable examples include measures adopted by the Superior Electoral Court (TSE), which has regulated the use of artificial intelligence in election campaigns, as well as initiatives by the ANPD, including guidelines on the processing of personal data in high-risk contexts. Other relevant initiatives will be examined in the following sections and subsections.

Lazzarini (2025, p. 7) points out that the production of artificial images places pressure on traditional civil law frameworks in Brazil. Because these new forms of media are generated through data analysis rather than by capturing actual moments, they require a reassessment of image rights which have historically been grounded primarily in actual photographic and video recordings.

Beyond the conceptual dimension, this tension also manifests itself on a practical level, especially regarding the production of evidence and the verification of the authenticity of digital content. The difficulty of distinguishing real records from synthetic content directly impacts the work of the Judiciary, administrative authorities, and the platforms themselves, requiring the development of technical detection tools as well as legal criteria adapted to this new context.

Brazil's Decree No. 12,975/2026 establishes duties of care, governance, and accountability for digital platforms, classifying certain abusive uses of deepfakes as systemic regulatory risks[16] associated with the operation of service providers, particularly

16 *The concept of systemic risk refers to threats to fundamental rights arising from the architecture or operation of services offered by online application providers, with impacts that extend beyond the individual sphere.*

when related to fraud, illegal advertising, artificial distribution networks, or violations of fundamental rights. This context calls for prevention, detection, and proactive responses from service providers, in accordance with the protection of personal data and personality rights. In practice, abusive uses of deepfakes — such as digital fraud, violence against women, and identity manipulation — are now treated as risk categories that require structural prevention, content moderation, and, where appropriate, the removal of unlawful content.

In this context, we present illustrative cases in this fifth section to support and deepen reflection on the topic. We will discuss cases such as the use of pornographic deepfakes against women and girls; in contexts of racism; in electoral contexts; in political gender-based violence; in financial scams; and in misleading advertising.

5.1 Pornographic deepfakes targeting women and girls

The use of deepfakes to produce non-consensual pornographic content has become one of the most serious and widespread manifestations of the abusive use of artificial intelligence. This phenomenon disproportionately affects women and girls, reproducing structural patterns of gender-based violence in the digital environment (McGlynn; Toparlak, 2025). Recent empirical evidence indicates that the victims identified in national surveys are either exclusively or overwhelmingly female, including children and adolescents (SaferNet Brasil, 2025).

The primary setting for these cases has been the educational environment. A survey conducted by SaferNet Brasil identified, between 2023 and 2025, at least 16 documented cases in educational institutions across 10 states, involving 72 victims and 57 perpetrators (SaferNet Brasil, 2025). These incidents are characterized by the creation — using generative AI tools

— of nude images or sexual content in which real faces — often extracted from social media — are superimposed onto naked bodies, without any consent from the people depicted.

Specific cases illustrate the dynamics of these crimes. In various states, students were investigated for producing and sharing fake nudes of classmates and even female teachers, using publicly available AI applications. At a school in Rio de Janeiro, for example, teenagers created and disseminated fake intimate images of classmates, leading to the opening of an investigation by the Child and Adolescent Protection Unit (Neto, 2023). In another case, in Mato Grosso, four students were held responsible for producing photo montages involving more than 30 victims, including students and a teacher, which were widely circulated in online groups (G1 MT, 2024). These incidents reveal not only the trivialization of the practice but also its role in the dynamics of bullying, public humiliation, and symbolic violence.

In addition to the quantitative aspect, the qualitative aspect of this harm is also noteworthy. Pornographic deepfakes constitute a particularly severe form of violation of human dignity, as they combine elements of sexual exposure, fabricated content, and loss of control over one's own image. Brazilian academic studies indicate that this type of practice should be understood as a contemporary manifestation of gender-based violence, insofar as it instrumentalizes the female body as an object of humiliation and social control (Valente, 2023; Rodrigues, 2025). Furthermore, the synthetic nature of the content does not diminish its effects; on the contrary, it makes it more difficult to contest socially and legally, prolonging the reputational and psychological harm to victims (Rodrigues, 2025).

Another critical aspect is the scale and speed of dissemination of this content. Data from SaferNet Brasil (2025) indicate that links related to the circulation of sexual deepfakes and associated abusive material have been identified since 2023 on social

media, messaging apps, and even dark web forums. These activities often involve organized groups that use automated mechanisms (such as bots) to expand distribution, which contributes to the persistence and replicability of the harm.

Real-Life Case

ANPD Investigation into Pornographic Deepfakes Involving Women and Children/Adolescents

According to Technical Note No. 1/2026/FIS/CGF/ANPD, on January 14, 2026, the ANPD received a Complaint and Petition filed by Federal Deputy Érika Hilton, alleging that the Grok artificial intelligence system violated the LGPD.

On January 20, 2026, the ANPD, the Federal Public Prosecutor's Office (MPF), and the National Consumer Secretariat (Senacon), under the Ministry of Justice and Public Security, issued a Joint Recommendation to the company X Brasil Internet Ltda., which operates the X platform (formerly Twitter). The initiative stems from complaints and evidence regarding the use of the Grok artificial intelligence tool — integrated into the platform — to create potentially illegal synthetic content, particularly sexual deepfakes involving women, children, and adolescents.

In the recommendation, the institutions highlight that Grok has begun to allow the automated editing of third-party images without verifying consent or a legitimate purpose, which enables the production of highly realistic fake content. There are reports of the tool being used to generate sexualized and pornographic images without the authorization of the people depicted — including minors — which have been widely disseminated on the platform. This scenario poses a risk of harm to the victims' dignity, image, privacy, and physical and psychological

integrity. The document is grounded in various national and international regulatory frameworks. Among these, the Federal Constitution, the Statute of the Child and Adolescent (ECA), the Brazilian Data Protection Law (LGPD), the Brazilian Civil Rights Framework for the Internet, and the Consumer Protection Code (CDC) are highlighted. More recent regulations are also cited, such as the Digital ECA (Law No. 15,211/2025), which reinforces the obligation to adopt preventive measures from the very conception of digital services. The text emphasizes that the protection of children and adolescents must adhere to the principle of comprehensive protection and absolute priority, and that the creation of sexualized deepfakes may even constitute crimes under criminal law.

With regard to women, the use of the tool to generate non-consensual sexual content is understood as a potential form of psychological gender-based violence, as defined in the Penal Code, in addition to violating international commitments undertaken by Brazil, such as the Belém do Pará Convention. The document also raises the issue of the platform's liability, arguing that, by offering and structuring the AI tool, the company does not act as a mere intermediary and may be considered a co-author of the generated content.

The institutions emphasize that, even in light of the Brazilian Civil Rights Framework for the Internet, there is a heightened duty of care incumbent upon service providers, especially following recent decisions by the Federal Supreme Court that qualify the immunity provided for in Article 19. Furthermore, they highlight that the company's conduct may violate consumer protection laws by offering a service that is potentially dangerous and inconsistent with its own internal policies, which may even constitute misleading advertising. Given this context, the recommendation established a

series of mandatory measures for the company, including: (i) immediate implementation of mechanisms to prevent the generation of sexualized content without consent, particularly involving minors; (ii) creation of effective procedures for identifying and removing such content already present on the platform; (iii) strict enforcement of the company's internal policies, including the suspension of accounts in violation; (iv) provision of accessible channels for data subjects to file complaints and exercise their rights; and (v) preparation of a detailed Personal Data Protection Impact Rapport (RIPD) regarding Grok's activities.

Finally, on February 11, 2026, the institutions involved in the Joint Recommendation deemed the reported measures insufficient, and the ANPD issued a preventive measure based on Articles 32, III, and 35 of the Enforcement Regulations, in order to prevent the Grok artificial intelligence tool from generating content that depicts children, adolescents, or adults who are identified or identifiable in sexualized or eroticized contexts without authorization. The primary objective of the measure was to prevent violations of fundamental rights and ensure the effective protection of Brazilian users against the risks associated with the use of generative artificial intelligence on the platform.

Cases in Brazil demonstrate that pornographic deepfakes constitute an emerging form of digital violence, characterized by a strong gender dimension, ease of production through technological tools, the vulnerability of adolescents, and the insufficiency of prevention and response mechanisms. The recurrence of such incidents, especially in school settings, suggests that the phenomenon should be treated as a structural trend in the misuse of artificial intelligence, with direct implications for public policy, regulation, and the protection of personal data.

5.2 Deepfakes and racism

It is important to emphasize that the impacts of deepfakes are not limited to the sexual realm, as these technologies have also been used to reproduce and amplify other historical forms of discrimination and violence, including racism. In this context, audiovisual manipulation using artificial intelligence has become part of well-known repertoires of stigmatization, ridicule, and disinformation directed at racialized groups.

Who hasn't received or come across seemingly authentic videos on social media in which Black public figures are portrayed in a caricatured, animalized, or offensive manner? Cases involving manipulated content attributed to former U.S. First Lady Michelle Obama, actress and Duchess of Sussex Meghan Markle, and Brazilian soccer player Vinícius Júnior illustrate how deepfake technology can be used to associate Black individuals with ridiculing behaviors, inappropriate remarks, or humiliating situations that never actually occurred. Although such content is often, framed as humor or entertainment, it reproduces historical patterns of racist representation and reinforces socially harmful stereotypes.

Throughout history, race has been used as a marker of inequality, oppression, and violence, functioning not as a biological or genetic fact, but as an ideological fiction and a sociopolitical construct historically designed to codify human division, create hierarchies, and legitimize exclusion (Mbembe, 2014, p. 71). Among the various manifestations of racism, the deployment of humor as a mechanism for dehumanizing and belittling Black people is particularly salient. While this practice was once prevalent in television shows, comedy acts, and other traditional media productions, it now also manifests through memes, GIFs, short videos, and viral posts on digital platforms. In many cases, Black people and their physical, cultural, or behavioral characteristics are reduced to offensive, pejorative, caricatured, and even hypersexualized representations.

As Dantas (2025, p. 30) notes, this dynamic is not exactly new, but it has been significantly transformed by digital technologies. According to the author, "the difference is that this type of humor is now just a click away, goes viral in minutes, and takes on increasingly sophisticated technological layers" (Dantas, 2025, p. 30, our translation). Deepfakes represent precisely one of these new technological layers, allowing discriminatory content to be produced with greater realism, engagement potential, and a stronger capacity to engage audiences, and greater dissemination potential.

In this sense, deepfakes can be understood as yet another tool in the service of so-called "recreational racism." As Dantas (2025, p. 30) points out, "recreational racism — a term coined by Adilson Moreira to describe the practice of dehumanizing Black people under the pretext of humor — is nothing new. It has merely gained new contexts and tools" (Dantas, 2025, p. 30, our translation). Technology, therefore, reshapes the ways in which prejudice is manifested and offers new means for its reproduction, dissemination, and large scale amplification.

This phenomenon is also linked to an increasingly common practice on digital platforms so-called "rage bait." The term:

> *emerged in informal online discourse (Scott-Railton, 2022) to describe the manipulative practice of eliciting outrage and aggressive responsiveness to increase profile traffic, subscribers, algorithmic ranking and in some cases online revenue (Cover, 2025, p. 3).*

Unlike spontaneous expressions of opinion, rage bait is strategically planned to generate intense emotional reactions and maximize engagement.

One of the most common forms of this strategy involves the publication of racist or hostile content targeting marginalized groups.[17] As Cover (2025, p. 3) notes, rage bait operates by deliberately provoking anger and outrage, a dynamic that

17 *"Marginalized" is a term that has been used in contemporary literature to refer to groups that, although not always numerical minorities, occupy positions of social, political, or economic disadvantage due to historical processes of exclusion and underrepresentation. Unlike the notion of "minority," which may suggest a quantitative criterion, the term "marginalized" highlights the power relations that produce and perpetuate situations of vulnerability and social marginalization.*

can be amplified using artificial intelligence in automated content production. In such cases, discriminatory content is used as an emotional trigger to transform social prejudices into mechanisms for digital engagement, with its circulation amplified by the reactions it provokes.

To further support this observation, a study conducted by Kolluri *et al.* (2025, p. 8) on the dynamics of hate speech dissemination on the social media platform Parler found that content targeting Black people, along with antisemitic content, exhibited the highest levels of hate speech propagation on the platform. The results demonstrated that comments posted on anti-Black content were 227 times more likely to contain racial hate speech than the other interactions analyzed, according to Kolluri *et al.* (2025, p. 8). These data suggest that racist content targeting the Black population is more likely to attract engagement and stimulate new discriminatory expressions, contributing to the rapid amplification of hostility and symbolic violence in digital environments. Thus, the study demonstrates how online racism spreads easily on social media and tends to generate cycles of reinforcement and intensification of hate speech.

Although the academic literature still contains relatively few studies that specifically examine the relationship between deepfakes, rage bait, and racist discourse, the available evidence points to a significant convergence among these phenomena. As Godulla *et al.* (2021, p. 76) emphasize, "the malicious use of deepfakes is strongly linked to automated propaganda and disinformation campaigns as well as the dissemination of misinformation and conspiracy theories through social media". Thus, synthetic technologies, when used maliciously, contribute to amplifying existing discriminatory narratives, lending them an air of authenticity and greater persuasive power.

In addition to the exploitation of racial prejudice, these same strategies can be deployed in political and ideological dis-

putes. AI-manipulated videos can falsely attribute statements to candidates, political leaders, or public figures to amplify polarization, fuel ideological tensions, and exploit ongoing public debates and conflicts.

5.3 Deepfakes in the electoral context

The use of deepfakes represents one of the major contemporary challenges to the integrity of public debate and trust in democratic institutions in the electoral context. The ability to produce fake videos or audio clips involving candidates, officials, political agents, or leaders increases the potential for highly persuasive disinformation campaigns capable of influencing perceptions, reinforcing polarization, and undermining voters' decision-making, especially during periods of intense information flow.

Beyond fabricating statements or events, deepfakes have been used to create entirely synthetic personas that mimic ordinary citizens, influencers, or political commentators. Two examples that have garnered significant attention are the profiles "Dona Maria" and "Sô Zé da Feira" (Figures 12 and 13), characters generated entirely by AI that appear to be real people. Designed to portray people over the age of 60 and given fabricated backstories depicting simple, everyday lives, the characters use colloquial language to comment on political events and share opinions on topics of public interest. Built around easily recognizable elements of everyday Brazilian life (such as the retired woman, the housewife, or the market vendor), these avatars leverage the characteristics of everyday individuals to foster identification, proximity, and trust among the public.



Figure 12 *Character "Dona Maria," created by artificial intelligence*

Source: Reproduced from the internet.

This strategy sheds light on a significant shift in how disinformation can be produced and disseminated. Instead of relying solely on the manipulation of images of public figures, synthetic content now creates false spokespersons for society, giving the impression that certain opinions represent spontaneous expressions of everyday citizens. When these characters do not clearly and transparently disclose that they were generated by artificial intelligence, many people come to believe they are watching authentic testimonials, which increases the persuasive power of the messages and makes it harder to identify the content as artificial. This mechanism directly interferes with the exercise of citizenship, as it can influence the formation of opinions and political decisions based on an artificially constructed perception of what society supposedly thinks or advocates.

Łabuz and Nehring (2024) highlight concerns regarding the use and advancement of generative artificial intelligence in the electoral landscape and the extent to which it can undermine trust in democratic processes. Although these authors' central argument does not focus specifically on deepfakes, it is possible to extend this concern to that context.

This scenario makes combating disinformation more complex. Unlike traditional fake news, deepfakes and other synthetic content exploit audiovisual resources that make such content

difficult to challenge, even when it is later debunked. In a digital environment marked by the speed of sharing and the fragmented consumption of information, manipulated content can reach millions of people before it is identified, corrected, or removed.



Figure 13 *Character "Sô Zé da Feira," created by artificial intelligence*

Source: Reproduced from the internet.

The impact of this strategy is amplified by a lack of transparency and oversight. A survey by the *Observatório IA nas Eleições*[18] found that 61% of political profiles created by artificial intelligence on social media do not inform users that they are AI-generated, despite the rules established by the Electoral Court requiring the identification of this type of content. The study also found that more than half of the posts analyzed contained some form of misinformation, revealing that these synthetic personas conceal their artificial nature and are frequently used to promote misleading narratives.

Given this context, the debate over deepfakes in elections requires a balance between protecting freedom of expression, promoting technological innovation, and, above all, preserving the integrity of the democratic process. The main challenge, in this case, lies in ensuring transparency regarding the use of artificial intelligence in content production, so that citizens can exercise their right to vote in a conscious, critical, and informed way.

In Brazil, this risk has not gone unnoticed by the TSE, which has been taking a proactive stance to address this challenge

18 *The Observatório IA nas Eleições (AI in Elections Observatory) is a go-to resource for researchers and experts on the use of generative AI to manipulate information that could interfere with electoral processes and the integrity of democracy.*

and adapt to the new digital landscape. In recent resolutions on electoral advertising and the fight against misinformation, the Court has begun to explicitly address content generated or manipulated by artificial intelligence, recognizing its impact on the electoral process. Beyond the legal aspect, the TSE has also invested in cooperation with digital platforms and public awareness campaigns. The Court's regulatory actions will be detailed in Section 6.1, which addresses the legal and regulatory approach to deepfakes.

5.4 Deepfakes as a tool of political gender-based violence

The use of deepfakes in political contexts has introduced a new layer of complexity to the already well-known issue of political gender-based violence. While women in positions of public visibility have historically faced attacks on their honor, reputation, and credibility, the introduction of audiovisual manipulation technologies significantly expands the reach, sophistication, and potential for harm of these attacks (Azevedo, 2025, p. 85). The innovation here lies in these tools' ability to produce highly realistic content on a large scale and disseminate it rapidly.

According to Law No. 14,192/2021, political gender-based violence can be defined as any action, conduct, or omission intended to prevent, hinder, or restrict the political rights of women, whether they are candidates or elected officials (Brazil, 2021). Regarding the digital sphere, examples include the dissemination of false photos or videos during an election campaign or term of office, and the defamation of a female candidate by attributing offensive claims to her. Recent studies indicate that deepfakes constitute a direct extension of the dynamics of gender-based violence in the digital environment and are strategically employed to delegitimize women's political participation and silence of dissenting voices. A UN Women report indicates that coordinated attacks involving artificial intelligence are frequently directed at women in

public office, undermining their professional credibility and personal reputation (UN News, 2026). In this context, the manipulation of images and voices serves as both a mechanism for disinformation and an instrument of symbolic coercion and political exclusion (Azevedo, 2025, p. 106).

Literature and empirical data identify three main patterns in the use of deepfakes against women in politics: i) non-consensual sexualization and pornography, involving the creation of fake intimate images to cause public humiliation; ii) the simulation of statements or behaviors, through manipulated audio and video, in order to provoke artificial scandals; and iii) coordinated disinformation campaigns that use synthetic content to influence the electorate's perceptions.

These patterns often combine, reinforcing gender stereotypes, exploiting specific vulnerabilities associated with women's presence in the public sphere, and reinforcing the narrative that women should not hold positions of public power. The UN itself highlights that such practices have concrete effects on democratic backsliding: approximately 41% of women in public office report engaging in self-censorship to avoid further attacks, while a significant proportion reduce their professional or political activities (UN Women, 2026).

In Brazil, although there is still no specific body of case law on deepfakes in the context of political gender-based violence, specific incidents highlight the phenomenon. One notable case occurred during the 2024 municipal election campaign, when Adilma Lacerda, a mayoral candidate in Ipojuca (PE), was the target of a manipulated video in which her voice was artificially altered to simulate insults directed at voters (Souza, 2024). The content, widely circulated on messaging apps, attributed derogatory statements to the candidate, with clear potential to damage her public image and undermine her campaign. The police investigation identified the use of audio manipulation technology, characterizing the incident as the unlawful use of artificial intelligence for electoral purposes.

Cases like this demonstrate that political violence mediated by deepfakes is not limited to the spread of disinformation but also incorporates elements of symbolic violence, sexualization, and public humiliation, directly affecting women's participation in politics. Another relevant example involves the circulation of fake intimate images of women holding public office. In 2024, the mayor of Bauru (SP) was confronted with the circulation of a deepfake image featuring her face superimposed on the body of a naked woman, prompting her to contact law enforcement authorities to report the misuse of her image (Piassi, 2024). In another instance, congresswoman Tabata Amaral was linked to two manipulated images of an adult content creator, prompting the then-congresswoman and candidate for mayor of São Paulo to take legal action against the authors of the posts (Folha de S. Paulo, 2024).

In this context, institutional action, including that of ANPD, is essential to mitigate these practices. Coordination between data protection, electoral regulation, and legal accountability is becoming increasingly necessary given the cross-cutting nature of these violations.

5.5 Deepfakes used in financial scams

In recent years, the use of deepfakes has become part of a more sophisticated form of digital fraud in Brazil, particularly in financial scams. Unlike traditional phishing[19] or social engineering schemes based on text and links, for example, these scams — though rooted in familiar fraud tactics — gain new sophistication through the use of artificial intelligence, particularly in schemes that exploit trust, urgency, and authority by incorporating images and voices. These situations are carefully designed to create a highly convincing appearance of legitimacy. There are recent reports of manipulated videos circulating on social media and messaging apps, in which supposed experts, public figures, politicians, and even celebrities recommend fraudulent investments. In some

19 Phishing is a type of digital scam in which criminals attempt to deceive people into providing sensitive information, such as passwords, bank details, or document numbers, usually by impersonating trusted institutions (such as banks, government agencies, or companies) through emails, messages, or fake websites.

cases, even the visual identity of well-known companies is replicated to boost the scam's credibility.

Below, we present three Brazilian cases of financial scams that have garnered significant attention in the country over the past two years:

The Brazilian case involving the use of deepfakes featuring Brad Pitt's image gained widespread attention following an incident that occurred in Rio Grande do Sul in late 2025, when a woman claimed to be in a virtual relationship with the actor and even waited for him to arrive at a local airport (Henrique, 2026). According to reports, she claimed to have made several video calls with the supposed Brad Pitt and believed that the in-person meeting would lead to a formal relationship. The victim's conviction caught the attention of authorities, especially because she dismissed the possibility of fraud based on the visual experience provided by the calls, which allegedly involved real-time manipulation technologies.

In the ensuing legal proceedings, the case came under analysis by the Civil Police, who sought to determine whether it constituted romantic fraud (a type of fraud involving emotional exploitation to obtain financial gain) or whether the situation involved other criminal offenses, such as a possible false report of a crime. Subsequently, the woman changed her story, claiming it was a "joke," which raised doubts regarding the possible existence of concealed financial losses or the initial veracity of the facts. Considering these inconsistencies, authorities kept the investigation open to examine both the possible involvement of third parties (scammers) and the conduct of the woman herself, in accordance with the Brazilian Penal Code.

The same scam strategy has been used in other countries. In France, for example, a woman lost more than € 830,000 (equivalent to just over R$ 5 million) (Woman..., 2025). As a result, Brad Pitt issued a statement through a spokesperson

expressing regret over the incident and warning fans about online scams that used his image without consent.

This episode is part of a broader context of rapid growth in deepfake fraud in Brazil, especially following the widespread adoption of generative AI tools, as we have seen here. According to the *Identity Fraud Report 2025–2026*, published by the identity verification company Sumsub, Brazil accounted for approximately 39% of the deepfake cases detected in Latin America in 2025, in addition to recording a 126% increase in attacks involving deepfakes and synthetic identities between 2024 and 2025. These findings underscore the quantitative expansion of these practices and their increasing sophistication, marked by the use of emotional narratives and the simulation of "real" interactions, which intensifies legal risks, regulatory challenges, and vulnerabilities in the context of data protection and digital security.

Another case of a deepfake involving a financial scam is the fraudulent video that used the image and voice of Pedro Moreira Salles, co-chair of the Board of Directors of Banco Itaú (Itaú, 2025), which was widely shared on Facebook in 2025. The content attributed false statements to the executive with the aim of convincing users to sign up for a supposed "miracle" investment. To bolster its credibility, the scam directed victims to a link that mimicked the website of a news outlet, creating an appearance of legitimacy. Banco Itaú clarified that the executive had no connection to the publication and that the material exhibited typical signs of fraud, such as a lack of reliable sources and unrealistic promises of financial returns. Figure 14 illustrates the content that was shared at the time to provide a better understanding of the case.

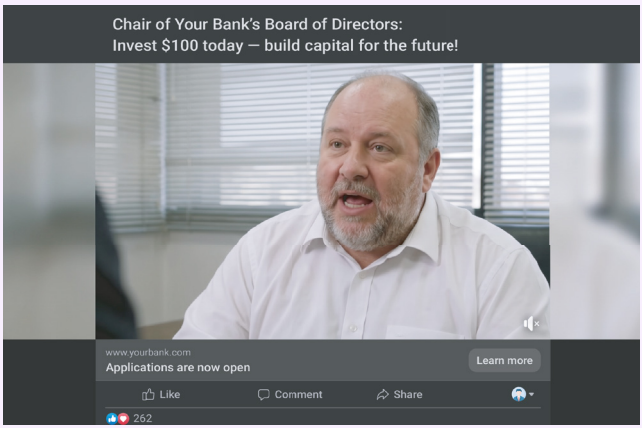


Figure 14 *Image illustrating the financial scam involving a deepfake of a representative of a major bank*

Source: Prepared by the authors with the assistance of the Gemini Flash 3.5 tool.

From a legal standpoint, although this specific case is treated primarily as a preventive warning (without public details of any resulting legal proceedings), it falls under categories already established in Brazilian law, which we will examine in greater detail below. On a broader regulatory level, the incident illustrates a growing problem that has prompted institutional responses and legislative discussions to mandate the identification of synthetic content (such as specific bills on deepfakes), as well as increased pressure on digital platforms to moderate fraudulent advertisements and reduce economic incentives for misinformation.

The third case of financial fraud involves Serasa Experian (Domingos, 2024), a Brazilian company specialized in data analysis for credit decisions. In practice, it operates as a large database that compiles consumers' financial histories based on transactions linked to their CPF, assisting companies in assessing the risk of default before granting loans, financing, or credit.

According to data from May 2026 from Serasa's Map of Delinquency and Debt Negotiation (Mapa..., 2026), Brazil

has approximately 83.5 million people with a negative credit record, which corresponds to roughly half of the country's adult population. In addition, it is estimated that 57 million consumers have debts and are unaware of them because they do not regularly monitor their CPF. In light of this scenario of high levels of indebtedness and the desire to get their financial situation back on track, this group becomes particularly vulnerable to scams that exploit the company's name.

Over the past two years, false content has circulated widely, promising alleged compensation for a data breach, as well as misleading offers to settle debts in exchange for reduced fees or even credit transfers. The severity of the situation was exacerbated using deepfakes, with the creation of fake videos featuring public figures, such as then-Senator Sérgio Moro (Amado; Campos, 2024), who later went public to clarify that the content was false. The materials also reproduced visual elements from television news programs to lend greater credibility to the scam.

The videos claimed that compensation of up to R$ 30,000 would be paid to people allegedly affected by the exposure of their data, such as CPF numbers, full names, addresses, phone numbers, and even salary information, social benefits, and income tax data. At the end, users were encouraged to click on a link to check out a list and see if they were eligible for the benefit. To give the scam the appearance of legitimacy, the scammers used a real 2021 news report about a data breach affecting approximately 223 million people, which at the time was linked to the Serasa Experian database. However, as of the time of this publication, the company has not been ordered to pay any compensation.

In 2026, the aforementioned massive data breach that occurred in 2021 was resurfaced and acquired new legal dimensions. The case became the subject of a class-action lawsuit in the English courts against the Experian group, to which Serasa is affiliated. The lawsuit was filed because the

enterprise's parent company is headquartered in the United Kingdom, which allows for this type of legal action. In Brazil, the issue also remains under judicial review: the Federal Public Prosecutor's Office (MPF) is acting as a co-plaintiff in a public civil action seeking to hold the parties accountable and secure potential compensation for the harm caused. However, to date, there is still no definitive technical evidence regarding the origin of the breach, nor has any court ruling ordered the payment of compensation to consumers.

Another type of scam that has been on the rise involves cloning the voice or image of people who are actually part of the victim's social circle — such as their boss, the CEO of the company where they work, a family member, or someone close to them — to trick them into making payments, bank transfers, or sharing sensitive data. Generally, these requests are accompanied by a sense of urgency, where pressure and immediacy override discernment and caution, leaving the victim with little time to critically assess the request and making them more susceptible to falling for the scam.

In this context, deepfakes act as a catalyst for existing forms of fraud, increasing their level of sophistication and requiring new verification mechanisms — both technical and behavioral— to mitigate risks. In a scenario where it is already difficult to distinguish between what is authentic and what is synthetic — due both to the sophistication of the technologies used to create deepfakes and to the reduced effectiveness of traditional warning signs in situations involving urgency or supposed family emergencies — the risk to individuals and institutions is considerably heightened, especially in environments where quick decisions are valued.

5.6 Deepfakes for deceptive advertising

The use of deepfakes has also expanded rapidly in the field of digital advertising, giving rise to a new form of highly sophisticated deceptive advertising based on the realistic simulation of identities, voices, and reputations. Unlike traditional forms of fraud, content generated by artificial intelligence allows for the creation of visually plausible and emotionally persuasive ads, making it difficult to distinguish between legitimate advertising and fraudulent content.

In Brazil, the spread of these ads has occurred mainly on digital platforms, such as social media and messaging apps, where sponsored content is targeted based on the user's profile. In this context, deepfakes are used to simulate endorsements of products and services by public figures — doctors, journalists, artists — whose credibility serves as a source of trust for consumers.

A pattern of operation can be identified in these cases. Generally, there are three stages: i) creation of synthetic content using the image and voice of public figures; ii) promotion of the content on social media; and iii) redirection of victims to fraudulent environments, such as cloned websites or fake platforms, where they are asked to make payments or provide personal data.

Thus, various digital marketing and psychological manipulation techniques are employed to build trust and influence consumer behavior. Examples include products promising immediate benefits, discounts, free gifts, or miracle cures.

In Brazil, there are numerous reports of deepfakes being used in deceptive advertising involving celebrities and experts. One of the most notable cases involved a criminal organization that used deepfakes of celebrities such as Gisele Bündchen, Angélica, Juliette, and Sabrina Sato to promote nonexistent products. In the case of Gisele Bündchen, for

example, the criminal group sold a "free anti-wrinkle kit" for which shipping fees were paid via PIX[20], but the product was never delivered. The police investigation revealed that the gang had handled more than R$ 20 million, which led to the arrest of its members (Souza, Figueiredo, 2025).

Another recurring tactic is the use of images of healthcare professionals in fraudulent campaigns. Dr. Drauzio Varella, for example, had his image and voice repeatedly manipulated to simulate recommendations for medications and treatments without any scientific basis (Fantástico, 2025). This content, often presented as interviews or news reports, exploits the expert's authority to induce consumers to purchase potentially harmful products, as illustrated in the figure below.

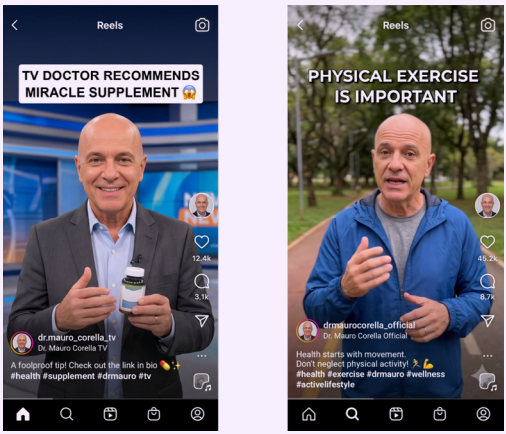


Figure 15 *Fictional depiction of a deepfake used in misleading advertising (left), misusing an image posted by a healthcare professional (right)*

Source: Prepared by the authors with the assistance of the Gemini Flash 3.5 tool.

The image shows two videos, with the one on the left (identified as the @dr.mauro_corella_tv account) illustrating an example of a deepfake. In this fraudulent representation, the healthcare professional's image is used without authorization to promote false claims about a "miracle supplement", a type of content that is dangerous from a health perspective and misleading from an advertising standpoint. The panel on the right, on the other hand, shows the authentic video of the

20 *Brazil's instant payment system.*

same professional (from the official and legitimate account), addressing a legitimate public health topic, which could have been used as the basis for creating the manipulated media.

The main characteristic of these scams is their massive scale, combined with small individual losses. In many cases, each victim loses relatively small amounts, which reduces the likelihood of reporting and investigation, allowing criminals to operate with high aggregate profitability (Romagna, 2025).

In addition to the direct economic impact, these schemes have significant effects in other areas, such as: i) risks to public health, especially when they involve unapproved medical products or unproven therapies; ii) undermining the credibility of professionals and institutions; iii) violation of personality rights, particularly the misuse of images and voices; and iv) exposure of personal data, since victims often provide information on fraudulent *websites*.

Therefore, the use of deepfakes in deceptive advertising represents a convergence of disinformation, economic fraud, and the misuse of personal data. The technological innovation introduced by artificial intelligence significantly increases the effectiveness of these scams by enabling the convincing simulation of figures of authority and trust.

‹ 6 prospects ›

The evolution of technologies for the synthesis and manipulation of visual and audiovisual media, which constitute the deepfake ecosystem, raises contemporary concerns and poses technical, regulatory, and social challenges for the future of the digital environment. To address these challenges, it is essential to examine emerging regulatory approaches to deepfakes in Brazil and other jurisdictions, analyzing the definitions of deepfakes adopted by different legal systems and the scope of restrictions or prohibitions regarding the creation and dissemination of such content in each country.

Furthermore, due to their complexity and increasing sophistication, the governance of these tools requires approaches that go beyond the strictly regulatory sphere and incorporate the promotion of information and media literacy among users, as well as the development of technical risk mitigation mechanisms. These strategies become particularly relevant in light of the challenges discussed throughout this Technology Radar, whether regarding the control of synthetic content that has already been created and is circulating, or regarding the distributed and transnational nature of digital infrastructures.

In this context, this section explores future developments related to deepfakes, organized around three main themes: i) the legal and regulatory treatment of these technologies in Brazil and other countries and jurisdictions; ii) the importance of digital literacy as an element of social resilience in the face of harm associated with the misuse of deepfakes; and iii) the development of tools capable of mitigating the risks related to this ecosystem.

The following analysis does not aim to exhaust these discussions but rather to provide an overview of emerging issues that are likely to shape the regulatory and societal debate in the coming years. Let us now turn to them.

6.1 Legal and regulatory perspective

The previous sections discussed the risks posed by the creation and dissemination of deepfakes to personal data protection and the harm these tools can cause when used to generate non-consensual pornography, to perpetrate fraud and financial scams, to spread misinformation in electoral contexts, and to promote gender-based political violence, for example. For all these reasons, it is clear that the misuse of these technologies threatens a range of fundamental rights, such as the rights to privacy, image, reputation, and freedom, as well as human dignity.

Considering this scenario, authorities around the world have sought to mitigate these and other risks through specific regulation of deepfakes. This regulatory effort, however, faces significant challenges, which are recognized by regulators in different jurisdictions. Some of these challenges are summarized in Figure 16 below and discussed throughout this section.

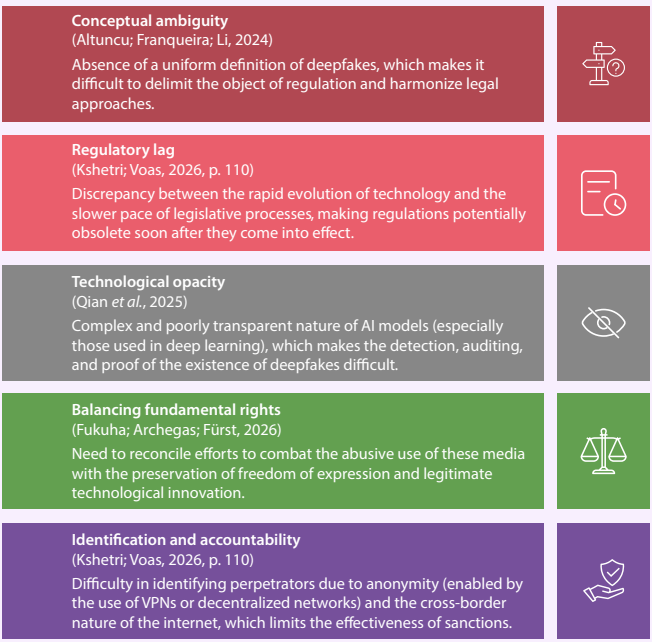


Figure 16 *Regulatory challenges related to deepfakes*

Source: Prepared by the authors.

The **conceptual ambiguity** surrounding deepfakes directly impacts the design of the regulations intended to govern them. Broadly speaking, it is possible to distinguish two ways of conceptualizing deepfakes, corresponding to a different type of regulation.

On the one hand, defining deepfakes as specific technological tools implies establishing **technical** regulations. An example of this is the project *DEEP FAKES Accountability Act*, introduced in 2019 in the United States, which defined a deepfake as any technological representation substantially derived from a person's speech or conduct. The focus of this legislation was on the technological manipulation of human characteristics through computer programs, that is, on the application of advanced technological tools to simulate the features of a specific person, unlike forms of human imitation not mediated by technology, for example.

On the other hand, interpreting deepfakes based on the potential harm they may cause leads to the development of **risk-based** regulations. These approaches place less emphasis on how deepfakes are generated and focus instead on the purposes for which they are produced.

This perspective has become the predominant approach in deepfake laws currently in force around the world, as will be discussed below, because it allows for addressing, at the same time, two important governance challenges linked to this ecosystem. First, there is the challenge of **regulatory lag**, an inherent gap between the time required for the legislative process and the rapid technological advancement of these tools.

Second, there are the barriers imposed by **technological opacity**, which make it difficult to precisely identify the technologies used to generate deepfakes and may prevent the enforceability of laws designed only for certain types of technologies — and not for others. In this context, regulations based on the (malicious) applications of deepfakes shift the regulatory focus from

the technologies themselves to the social risks associated with them, making such regulations more adaptable and effective in mitigating these threats to rights.

Within the risk-based regulatory framework, international laws have followed three main regulatory approaches regarding deepfakes: i) **transparency** requirements, which mandate the labeling of synthetic content; ii) prohibitions related to **elections**, which ban the use of deepfakes in situations considered particularly risky from a political standpoint; and iii) **criminalization** of abuses, which impose penalties for the production or dissemination of deepfakes containing sexual or pornographic content.

An example of the first approach, focused on **transparency**, is found in the European Union's AI Act, which states in Article 50(2) that "Providers of AI systems [...] generating synthetic audio, image, video or text content, shall ensure that the outputs […] are marked in a machine-readable format and detectable as artificially generated or manipulated" (European Union, 2024).

In addition, Article 50(4) further requires that "Deployers of an AI system that generates or manipulates image, audio or video content constituting a deepfake, shall disclose that the content has been artificially generated or manipulated [...]" (European Union, 2024).

In the European regulations, we observe the central importance given to information regarding the manipulation or complete synthesis of artificial media, which addresses concerns related to the potential for deception that deepfakes create in society, due to their high level of realism. According to the regulations previously mentioned, any deepfake content, regardless of its context of use and circulation, must be accompanied by explicit labeling regarding its nature — watermarks, persistent on-screen text, or warnings displayed prior to playback — whether issued by the AI system provider or by the person responsible for creating or disseminating the content.

The second regulatory approach, characterized by restrictions related to **elections**, has been observed in various countries around the world. An example of this is South Korea, which stipulates in Article 82-8 of its *Public Official Election Act* that "no one may produce, edit, distribute, screen, or post deepfake videos for election campaigning purposes from 90 days before the election day to the election day." (South Korea, 2024, our translation).

In Brazil, the Electoral Court has taken a leading role in regulating deepfakes. This is because the dissemination of misleading or false deepfake content poses significant risks to integrity of democratic processes and to the proper conduct of election campaigns in the country. In light of the importance of the issue, since 2024 the Superior Electoral Court (TSE) has been explicitly addressing scenarios involving the restriction and prohibition of the use of these tools in the electoral context.

Thus, Brazil established, in TSE Resolution No. 23,732/2024 — which amends the wording of TSE Resolution No. 23,610/2019 —, the prohibition on the use, whether to benefit or harm candidates, of "synthetic content in audio, video, or a combination of both formats, that has been digitally generated or manipulated, even with authorization, to create, replace, or alter the image or voice of a living, deceased, or fictional person ('deepfake')" (Brazil, 2019, our translation). Notably, the Electoral Court's explicitly uses the term "deepfakes" when referring to tools for the digital synthesis and manipulation of media.

The same Resolution stipulates that "the use, in election advertising, [...] of content fabricated or manipulated to disseminate facts that are notoriously untrue or taken out of context, with the potential to harm the balance of the election or the integrity of the electoral process, is prohibited" (Brazil, 2019, our translation). In this case, national legislation has addressed the risks associated with deepfakes in the spread of disinformation, which become particularly serious

during election periods, as such content may influence voters' decisions.

It is worth noting that these regulatory approaches do not always operate in isolation in practice. Electoral prohibitions such as those outlined above are often combined with transparency requirements, for example. This can be seen in another article of the same Resolution, which requires the disclosure of the fabrication or manipulation of content in campaign advertisements:

> *The use in electoral advertising, in any form, of multimedia synthetic content generated through artificial intelligence or equivalent technology to create, replace, omit, merge, or alter the speed of, or superimpose images or sounds, imposes on the person responsible for the advertisement the duty to explicitly, prominently, and accessibly inform that the content has been fabricated or manipulated and which technology was used (Brazil, 2019, our translation).*

As can be seen from the previous excerpt, Brazilian electoral law permits the use of artificial intelligence to make alterations to multimedia content that fall under the concepts of "cheapfakes" or "shallowfakes" discussed in this Technology Radar. However, even in cases involving media with alterations that are more easily identifiable by the average internet user, explicit disclosure of the content alteration is required to ensure that viewers are aware of the manipulation.

In 2026, the TSE adopted a new Resolution on electoral advertising (Resolution No. 23,755/2026), which introduced even stricter restrictions on content manipulated or generated using artificial intelligence:

> *The publication and republication, even when free of charge, as well as the paid boosting of new synthetic content produced or altered by artificial intelligence or equivalent technologies that uses the image, voice, or statements of a candidate or a*

> *public figure are prohibited, even if such content is labeled and complies with the other requirements set forth in this Article, during the period beginning 72 (seventy-two) hours before and ending 24 (twenty-four) hours after the conclusion of the election (Brazil, 2019, our translation).*

The new Brazilian rule thus recognizes the particular sensitivity of the period immediately before and after the election and prohibits the dissemination of deepfakes during this critical period. Once again, the objective is to prevent highly realistic synthetic or manipulated media from being circulated and thereby influencing voters' decisions through misleading information.

The United States also provides examples of this regulatory approach. Given its decentralized federal structure, the U.S. landscape of deepfake governance and regulation has evolved in a fragmented and heterogeneous manner (Fukuha; Fürst; Archegas, 2026, p. 201–202). In the absence of a unified, comprehensive national law regulating deepfakes, different states have adopted their own regulations to address the misuse of these technologies.

Thus, between 2019 and 2025, 28 of the 50 U.S. states enacted laws regulating the use of synthetic or manipulated media in election campaigns. States such as California and Texas have taken the lead in these efforts, adopting particularly stringent regulatory frameworks (Fukuha; Fürst; Archegas, 2026, p. 202; Kshetri; Voas, 2026). These laws generally required internet service providers and those responsible for election campaigns to use tools to detect and clearly label synthetic content (Kshetri; Voas, 2026), demonstrating, once again, the link between regulatory standards related to transparency and the prohibition of improper electoral uses of deepfakes.

Finally, regarding the third regulatory approach, namely **criminalization**, various countries around the world have

sought to impose criminal penalties for deepfakes in cases of significant social harm, typically related to sexual offenses.

The criminalization approach raises particular concerns in the context of deepfakes. Since they are a tool that can be used for both improper and lawful purposes, their regulation must always **strike a balance between fundamental rights**. In this regard, to preserve freedom of thought and encourage technological innovation, legal systems around the world have avoided criminalizing preparatory acts involving the use of deepfakes, or even the mere intent to commit unlawful acts through such media. This is because the anticipation of punishment increases the risk of state overcriminalization and pose threats to the free exercise of freedom of expression (Kirchengast, 2020, p. 6, 10).

On the other hand, the absence of explicit criminal provisions addressing deepfakes may prevent extremely harmful conduct from being punishable under criminal law. In South Korea, for example, prior to the 2020 and 2024 reforms, the lack of clear laws against these tools led to the acquittal of a man who distributed non-consensual synthetic intimate images (Kshetri; Voas, 2026, p. 6). A similar problem occurred in Mexico, where the courts acquitted a man who created fake sexual images of more than 1,000 women. The perpetrator went unpunished for digital fraud because there was no specific provision to convict him (Kshetri; Voas, 2026, p. 2).

In this context, legal frameworks around the world have been updated to address this form of technology-facilitated gender-based violence in their criminal codes. South Korea is a clear example of this process. According to data from Security Hero, 53% of pornographic deepfakes in circulation in 2023 depicted South Korean victims, and 8 of the 10 people most frequently featured in pornographic deepfakes during that period were South Korean women singers (Security Hero, 2023).

In light of this situation, in 2020 the country expanded its criminal legislation on sexual offenses in digital environments, establishing in Article 14 of the *Act on Special Cases Concerning the Punishment, etc. of Sexual Crimes* a prison sentence for recording videos or photographs of a sexual nature without the consent of the person depicted, as well as for the distribution, sale, dissemination, possession, and viewing of such material. In 2024, Article 14-2 of the same legislation was further amended to include explicit prohibitions on the editing, synthesis, processing, and publication of photos, videos, or audio recordings that include the face, voice, or body of specific individuals in a sexually explicit context (South Korea, 2025). These reforms were therefore designed to directly punish the creation and circulation of pornographic deepfakes.

In the United Kingdom, the criminalization of pornographic deepfakes occurred gradually, through amendments to existing laws. Initially, the *Online Safety Act 2023* amended the Sexual Offenses Act 2003 to criminalize the sharing of non-consensual intimate images, including the artificially synthesized ones (United Kingdom, 2003). In addition, *The Data (Use and Access) Act 2025* criminalized the creation and solicitation of intimate images synthesized without consent (United Kingdom, 2025). Under British law, the prohibitions apply to purported images, which reflects the country's concern with the apparent authenticity of such media, rather than with how they are generated.

Similarly, France has included a provision in its Penal Code aimed at addressing pornographic deepfakes by focusing on their appearance of authenticity. Article 226-8 of the French law prescribes imprisonment for the dissemination of "a montage created using a person's speech or image without their consent, if it is not evident that it is a montage or if this is not expressly indicated" (France, 2026, our translation). The same penalty applies to the dissemination of visual or audio content "generated by algorithmic processing and depicting

the image or speech of a person without their consent, if it is not evident that it is a montage or if this is not expressly indicated" (France, 2026, our translation).

Australia also amended its Criminal Code in 2024 targeting deepfake sexual material. Australian law provides for imprisonment for anyone who uses a communication service to transmit material created by a third party that depicts or appears to depict "(i) the other person engaging in a sexual pose or sexual activity (whether or not in the presence of other persons); or (ii) a sexual organ or the anal region of the other person; or (iii) if the other person is female — the other person's breasts" (Australia, 2024), provided that the material was transmitted without the knowledge or consent of the person depicted.

The amendment further provides that it is irrelevant whether the material is transmitted in its original form or whether it was created or altered using technology. Although it does not define what a deepfake is, a note accompanying this provision clarifies that the material covered includes:

> *images, videos or audio depicting a person that have been edited or entirely created using digital technology (including artificial intelligence), generating a realistic but false depiction of the person. Examples of such material are "deepfakes" (Australia, 2024).*

Following the proliferation of state-level deepfake related laws, the United States also criminalized the intentional distribution of non-consensual intimate visual depictions through its *Take It Down Act* 2025. In the text of the U.S. legislation, which applies nationwide, a deepfake is defined as digital forgery, as stated below:

> *The term "digital forgery" means any intimate visual depiction of an identifiable individual created through the use of software, machine learning, artificial intelligence, or any other computer-generated or technological means, including by*

> *adapting, modifying, manipulating, or altering an authentic visual depiction, that, when viewed as a whole by a reasonable person, is indistinguishable from an authentic visual depiction of the individual. (United States of America, 2025, p. 2–3).*

U.S. law provides for imprisonment for the publication of deepfakes depicting adults without their consent, for the publication of deepfakes depicting minors, and for threatening to disseminate intimate deepfakes with the intent to intimidate, coerce, extort, or cause suffering to the victim depicted in the content. In addition to establishing criminal penalties for practices involving the misuse of these tools, the law also requires online platforms to remove non-consensual content upon notification.

In Brazil, there is still no specific criminal offense addressing the abusive use of deepfakes, nor is there a specific provision criminalizing pornographic deepfakes, such as those discussed above. Such content may nonetheless fall within the scope of Article 218-C of the Penal Code, which addresses the dissemination of scenes of rape of adults and of vulnerable individuals, or scenes of sex, nudity, or pornography without consent. This provision was added to the Code in 2018, and the applicable penalties were increased in 2025 (Brazil, 1940).

Also in 2025, Law No. 15,123 amended Article 147-B of the Penal Code, which addresses psychological violence against women. Under the amended provision, penalties are increased when the violence is committed using artificial intelligence or other technological means capable of altering the victim's image or sound (Brazil, 1940), reflecting growing legislative attention to the risks associated with deepfakes. Furthermore, when sexual media involves children and adolescents, the Digital ECA stipulates, in the civil rather than criminal sphere, that providers of information technology products or services must remove and report content that appears to involve the sexual exploitation or abuse of children and adolescents (Brazil, 2025).

***

The laws described above offer important insights into the legal and regulatory treatment of deepfakes in Brazil and around the world, both today and soon. First, given the conceptual heterogeneity of the deepfake ecosystem, the predominant international regulatory approach has been to restrict or prohibit **harmful uses** of these tools, based on the risks they pose to fundamental rights. Accordingly, most deepfake-related legislation enacted around the world has focused on particularly critical applications of these technologies, especially in electoral contexts and in relation to sexual or pornographic content.

Second, recognition of the technical diversity of deepfakes has also led to regulations focused on the increasing **realism** of these media, which are becoming more convincing and more difficult to distinguish from authentic content. As a result, judicial and administrative authorities in various countries have sought to mitigate the harm arising from the deceptive potential of synthetic or manipulated media, rather than focusing solely on the technologies used to create deepfakes.

Third, the international regulatory efforts have generally prioritized addressing uses of deepfakes that pose **significant risks** to fundamental rights and democratic processes, rather than imposing blanket prohibitions on the technology itself. This can be seen in the absence of general bans on the creation of deepfakes and in the attention devoted to their use in areas such as elections and non-consensual pornography, which have become important subjects of regulatory concern in numerous countries.

Significant regulatory challenges remain to be addressed particularly with regards to the risks associated with deepfakes. One of the main bottlenecks in the governance of these tools is the difficulty of **identifying and holding accountable to** the

various actors involved throughout the deepfake lifecycle, from model training to content dissemination (Kshetri; Voas, 2026, p. 110). In this context, anonymity-enhancing tools and techniques used by creators of synthetic media, such as VPNs and decentralized networks, can hinder the identification of those responsible for unlawful conduct. Furthermore, the inherently cross-border nature of the internet, characterized by the free flow of data and differing regulatory regimes across jurisdictions, limits the effectiveness of national regulations.

In light of this scenario, regulatory approaches capable of addressing the challenges posed by deepfakes are likely to be those based on integrated and multi-layered governance models. On the one hand, this involves limiting criminal penalties to cases involving serious offenses, such as extortion and the misuse of intimate images (Kirchengast, 2020, p. 11). On the other hand, the incorporation of civil and administrative accountability mechanisms is an important way to address less serious, yet still harmful, uses of these tools.

In Brazil, this trend has included, for example, the creation of accountability mechanisms for actors other than the creators of deepfakes, such as internet service providers. In this regard, Decree No. 12,976/2026 prohibits the generation or modification of "a third party's intimate content through the use of artificial intelligence or any other technological resource capable of altering the victim's image or voice", a prohibition applicable to internet application providers (Brazil, 2026b, our translation).

The same Decree, which establishes guidelines for the protection of women in the digital environment, stipulates that "providers of internet applications based on artificial intelligence functionalities or equivalent technological resources must implement technical and procedural safeguards to identify and block requests to generate" such content (Brazil, 2026b, our translation). Under this framework, the synthesis and manipulation of third parties' intimate content are therefore prohibited, requiring application providers capable of

facilitating such practices to technically block these requests. Under the LGPD, these safeguards are particularly relevant when the generation or modification of intimate content involves the processing of the victim's image, voice, biometric data, or other personal attributes, especially in light of the principles of prevention, security, necessity, non-discrimination, and accountability.

In turn, Decree No. 12,975/2026, which regulates the Brazilian Civil Rights Framework for the Internet, provides that the liability of an internet application provider is presumed when unlawful content is disseminated through advertisements or paid promotions. Furthermore, this liability is presumed if the content in question is disseminated through artificial content distribution networks, regardless of prior notification (Brazil, 2026a). As a result, platforms that carry paid deepfake advertisements explicitly intended to promote scams or fraud may be held liable for their dissemination.

In both cases, there is a shift from a regulatory model centered on individual liability, as seen in many of the regulations cited earlier, toward a model that also imposes obligations directly on internet application providers. A similar approach was adopted by the Digital ECA, which prohibits such entities from monetizing or promoting content that sexualizes children and adolescents and requires providers to remove and report content involving apparent sexual exploitation and abuse (Brazil, 2025). Other international laws addressing deepfakes, such as the European *AI Act* and the U.S. *Take It Down Act*, have likewise incorporated mechanisms aimed at holding service providers accountable.

The global movement toward the regulation of deepfakes, briefly discussed in this section, forms part of a broader effort to govern artificial intelligence, which is evident across a growing number of countries and jurisdictions. Although some legal provisions governing AI also apply to deepfakes, the unique characteristics of these highly convincing forms of

synthetic media require specific legal responses to mitigate their risks. Consistent with this understanding, the Report of the Working Group on Deepfakes of the Secretariat for Digital Policies of the Office of the President recently proposed that Bill No. 2338/2023 — which establishes rules for the use of artificial intelligence and is currently under consideration by the National Congress — should include provisions specifically addressing deepfakes (Brazil, 2026c).

The Working Group's proposals include, among other elements: the inclusion or refinement of the concept of "synthetic content" set forth in the legislative proposal; the expansion of the rights of individuals or groups affected by AI systems in cases involving the unauthorized use of images, voices, or other personal attributes; the establishment of transparency and traceability obligations for synthetic content; and the requirement that developers of generative AI systems to conduct specific risk assessments related to deepfakes. With regard to pornographic deepfakes, the Report proposes multiple requirements for the generation of synthetic content depicting a natural person in a state of nudity or engaged in a sexual act, including the identification of the user responsible for generating the content and the express consent of the person depicted (Brazil, 2026c).

The Report's approach is consistent with the proposals described throughout this section. Rather than simply banning deepfakes, it recommends that developers of generative AI systems be subject to obligations related to transparency, traceability, risk assessment and mitigation, in addition to specific safeguards against sexual abuse. Although this is a non-binding recommendation, the document highlights proposals currently under discussion in the legislative process, whose wording and scope should be verified at the time of publication.

Strengthening measures to curb the misuse and highly harmful applications of deepfakes therefore requires a multidimensional legal and regulatory approach. This may include

specific criminalization and prohibitions; mechanisms for civil redress; and administrative oversight to ensure compliance with privacy and data protection laws and guidelines applicable to internet application providers. In this context, it is essential to view the deepfake ecosystem in all its complexity, holding creators accountable while also recognizing the role of companies in promoting safer digital environments.

Furthermore, these measures can also be complemented by actions taken by internet users and by internet application providers themselves. From the users' perspective, it is essential to promote information and media literacy among the population in order to strengthen the public's ability to identify misleading and fraudulent content. From the perspective of internet application providers, it is important that the technological advances that enable deepfakes also be used to detect them, block requests to generate unlawful content, and restrict the circulation of media that pose high risk of harm. The following sections discuss these aspects in greater detail.

6.2 Information and media literacy

Considering the risks and regulatory challenges posed by deepfakes, it is essential to involve users in efforts to mitigate the potential harms associated with the misuse of these tools. This approach is grounded in social resilience and harm prevention and seeks to complement, rather than replace, the regulatory mechanisms for controlling the creation and dissemination of such media, as discussed earlier. In this context, promoting information and media literacy among the general public is an important preventive measure focused on individuals.

Information literacy refers to the development of skills that enable **the critical evaluation of content and information**, including information produced and disseminated through

digital technologies. These skills involve a critical and reflective approach to digital media, which, rather than being assumed to be authentic, are carefully evaluated by considering factors such as the source of the information, signs of manipulation, and the context in which the content is shared and circulated.

Such literacy contributes to mitigating the harm caused by political disinformation, which can be amplified by tools such as deepfakes, and also reduces users' vulnerability to fraud and scams carried out through voice cloning, fake identity verification, or synthetic video calls, among other tactics. Thus, raising awareness about the existence, capabilities, and limitations of deepfakes can reduce the effectiveness of social engineering attacks, which exploit the public's lack of knowledge regarding these tools. Furthermore, such literacy encourages organizations to adopt additional identity verification and authentication procedures, strengthening the security of technical systems that play an increasingly important role in the daily lives of Brazilians.

The benefits of information and media literacy extend across society, but they are particularly important for the **protection of vulnerable groups**. As discussed throughout this Technology Radar, the risks associated with deepfakes are more pronounced for certain populations, such as children and adolescents, older adults, and women. Promoting information and media literacy among these groups may involve, for example, digital literacy initiatives in schools, training programs for caregivers of older adults, community-based digital education programs with modules specifically designed for older adults, and awareness-raising efforts addressing technology-facilitated gender-based violence in a variety of settings.

Similarly, the public needs to understand, even at a basic level, how artificial intelligence works, what these systems can do, what synthetic content is, and what forms of media

manipulation are currently possible. Although this remains a considerable challenge in a country as unequal as Brazil, it is important to recognize that schools, universities, libraries, neighborhood associations, technology and media companies, and other public and private institutions all play complementary and essential roles in expanding digital literacy across the population.

This critical dimension of digital literacy is also connected to a broader understanding of user autonomy in the face of technology. As Darin (2024) observes, digital systems can shape behavior in both positive and negative ways. Consequently, conscious engagement with digital technologies, including an understanding of how applications operate, what interests influence their design, and how they affect everyday decisions, is a prerequisite for preserving individual self-determination in increasingly connected environments. In the context of deepfakes, this perspective reinforces that information literacy is not limited to the occasional verification of suspicious content. Rather, it involves an ongoing habit of critical reflection on the technological mediations through which information is produced, curated, recommended, and consumed.

In this context, one example of a digital literacy initiative is the Brazilian Media Education Strategy (EBEM), which seeks to develop information and media literacy skills by promoting citizens' critical engagement with disinformation and manipulative algorithmic practices. Developed as part of Brazilian Media Education Week, the Strategy aims to foster citizens who are capable of interpreting, producing, and sharing information ethically and responsibly in digital environments. One of EBEM's core pillars is artificial intelligence, with the goal of strengthening people's ability to identify and understand synthetic content to reduce its potential for disinformation and rights violations (Brazil, 2023).

Digital and Media Education is included in Brazil's National Common Curricular Base (BNCC) and seeks to encourage the

effective use of digital technologies in teaching and learning activities while promoting the development of digital literacy skills for their responsible use. In addition to EBEM, several other initiatives seek to strengthen the information and media literacy of different segments of the Brazilian population. Examples include SaferNet Brasil's Digital Citizenship Course, which provides lesson plans and educational materials to help educators implement the BNCC's digital education guidelines (Disciplina..., 2026). The Digital Citizen Program, also developed by SaferNet, promotes dialogue with adolescents and young people on topics such as online safety, combating misinformation, and respectful behavior on social media (Cidadania..., 2026). In addition, the Ministry of Education (MEC) and the Secretariat of Social Communication of the Presidency of the Republic have developed digital education initiatives through the Brazilian Collection of Digital and Media Education and the Virtual Learning Environment of the Ministry of Education (AVAMEC). The debate surrounding information and media literacy is not new; however, it has become increasingly important for the promotion of safer digital environments.

Despite the importance of strengthening information and media literacy, it is important to emphasize that these measures complement, rather than replace, regulatory mechanisms aimed at limiting misuse and preventing abuses involving deepfakes. The effectiveness of digital literacy alone is limited in the face of the increasing realism of synthetic content and the asymmetry of information and resources between AI developers and individual users. For this reason, the promotion of digital literacy should be understood as one component of a broader strategy for governing the deepfake ecosystem, alongside legal, regulatory, and technical risk-mitigation tools, as discussed in the following section.

6.3 Mitigation tools and possible solutions

Advances in generative artificial intelligence have driven the rapid evolution of deepfakes, making them increasingly sophisticated and difficult to distinguish from authentic content. In this context, there is a continuous need to develop mechanisms capable of identifying digital manipulations and providing greater transparency regarding the origin and provenance of content. Mitigating the risks associated with this technology therefore depends both on improving detection tools and on implementing authentication and traceability mechanisms that can strengthen trust in digital environments.

Just like the machine learning and deep learning models used to generate synthetic content, detection tools rely on algorithms trained to identify patterns and inconsistencies that may be imperceptible to the human eye. Among the main indicators analyzed are inconsistencies in lighting, lip synchronization, facial biometric features, metadata, digital artifacts, and other characteristics that may reveal the artificial manipulation of images, audio, and video.

In this context, various solutions are being developed to increase the reliability of verification processes, including:

+ **Automated detection,** based on AI algorithms trained to identify patterns characteristic of manipulated content;

+ **Cryptographic signatures,** which allow the integrity of the content to be verified from its source, ensuring that the file has not been altered since its creation;

+ **Digital watermarking,** which involves embedding visible or invisible markers that identify content produced or modified by artificial intelligence systems, facilitating its subsequent identification and verification;

+ **Traceability standards, such as C2PA (*Coalition for Content Provenance and Authenticity*),** which establish standardized mechanisms for recording information about the origin, authorship, editing history, and production chain of digital content, thereby promoting greater transparency and verifiability.

These mechanisms reflect a broader shift in approaches to mitigating the risks associated with deepfakes. In addition to identifying manipulations after they occur, the objective is increasingly to document the provenance and lifecycle of digital content from its point of origin, thereby facilitating authentication, auditing, and accountability processes.

Over the past year, several online tools designed to detect manipulation in videos and other digital media have become established in the market. Three of them, *DeepWare Scanner, Attestiv*, and *TrueMedia*, were comparatively analyzed from both usability and technical perspectives. The results of this analysis were published in an article examining the challenges and limitations of deepfakes (Moraes; Batista, 2025). At the time, the findings highlighted the need for complementary human investigation, as none of the tools were able to accurately identify all cases of deepfakes or other forms of manipulation. In practice, all three produced uncertain results for both synthetic and authentic content, including lower-resolution materials and content produced in highly controlled environments, such as episodes of the classic television series *Chaves* or clips from broadcasts of *Jornal Nacional*, the country's leading television news program. According to the authors,

> *[...] the tools analyzed have strengths and limitations, and continuous improvements are needed to increase the accuracy and reliability of their results. This variability in performance highlights that, despite technological progress, deepfake detection continues to face significant challenges. Therefore, combining these tools with human evaluation is essential for a more robust and effective verification process (Moraes; Batista, 2025, our translation).*

The findings of this study demonstrate that, although technological advances have significantly improved the ability to identify manipulated content, currently available tools still face important limitations. Deepfake detection continues to involve significant margins of error and varying levels of confidence, particularly when dealing with low-resolution content, archival materials, or recordings produced under specific capture conditions. In this context, verifying authenticity requires an integrated approach that combines technological tools, contextual analysis, and human oversight.

Alongside detection mechanisms, initiatives aimed at promoting transparency and strengthening AI governance are becoming increasingly important. Identifying the origin of content, recording its modification history, and indicating whether a given piece of content was generated or modified using AI systems are measures capable of reducing information asymmetries and strengthening trust in the digital ecosystem. At the same time, traceability and auditing tools help demonstrate compliance with accountability principles and governance requirements established in recent regulatory frameworks.

These initiatives are consistent with a broader agenda of responsible technology design. In this regard, Darin (2024) emphasizes that technological development should not be guided exclusively by metrics of efficiency, engagement, and retention. Rather, developers should consider the ethical implications of their creations and adopt design choices that respect human limitations and promote user well-being. Applied to the deepfake ecosystem, this perspective suggests that technical safeguards, such as labeling synthetic content, blocking requests to generate unlawful material, and implementing traceability mechanisms, should not be viewed merely as regulatory obligations but as expressions of a commitment to informational integrity and user protection by design.

This perspective also supports a more critical legal analysis, such as that developed by Moraes and Batista (2025), which challenges the widespread industry assumption that publicly available data may be freely reused to train AI systems. This understanding is difficult to reconcile with contemporary data protection frameworks, because it fails to adequately account for principles such as purpose limitation, appropriateness, necessity, and transparency. In the context of deepfakes, these principles are particularly relevant, requiring an assessment of whether personal data are being used for legitimate purposes or to facilitate manipulation and deception, as well as whether data processing exceeds what is necessary and whether affected individuals receive adequate information about how their data are used.

Finally, once incorporated into the training of these systems, personal data may continue to pose persistent risks related to retention, re-identification, and propagation across downstream models and systems, making governance more challenging and undermining the effective protection of data subjects' rights. Deepfakes therefore represent not only a technological challenge but also a structural regulatory problem, requiring a reinterpretation of traditional data protection concepts in an environment characterized by the large-scale recombination of information and the potential for significant social harm.

Consequently, mitigating the risks associated with deepfakes requires a multidimensional approach that integrates technological solutions, transparency mechanisms, traceability standards, regulatory frameworks compatible with the pace of innovation, and, above all, qualified human oversight. It is therefore necessary to build a digital environment guided by information authenticity, the protection of fundamental rights, and the accountability of the various actors involved in the production and circulation of synthetic content.

‹ 7 final considerations ›

Deepfakes are among the most emblematic manifestations of the current convergence between artificial intelligence, personal data, and digital communication. Although the manipulation of images, videos, and audio is not a new phenomenon, recent advances in deep learning techniques have considerably expanded the ability to produce highly realistic, widely accessible, and scalable synthetic content. As demonstrated throughout this issue of the Technology Radar series, these technologies have legitimate applications in areas such as entertainment, education, healthcare, historical preservation, and accessibility. However, they also pose significant risks to privacy, personal data protection, digital security, and informational integrity.

The analysis presented throughout this work shows that deepfakes are not merely a technological challenge. Their relevance to the data protection landscape stems, above all, from the fact that these applications rely heavily on the collection, processing, and reproduction of elements capable of identifying individuals or distinguishing them from others, particularly biometric data such as facial images, voices, gestures, and behavioral patterns. In this context, deepfakes challenge traditional notions of authenticity, consent, identity, and accountability, requiring a reassessment of the protective mechanisms applicable to digital environments.

The examples of deepfake misuse examined in this publication demonstrate that the negative impacts of these tools are unevenly distributed across different social groups. Women and girls remain disproportionately affected by non-consensual synthetic pornography and other forms of digital violence. In the political sphere, audiovisual manipulation has been used to misinform voters, damage reputations, and amplify dynamics of gender-based political violence. In the economic sphere, fraudulent advertisements and financial scams based on voice and image cloning exploit relationships

of trust to deceive consumers and users. Deepfakes can also reinforce and reproduce discriminatory and racist practices. These multiple manifestations reveal that the harms associated with deepfakes extend beyond the individual sphere and affect broader social and institutional dimensions.

At the same time, this Technology Radar has highlighted that techniques for generating and manipulating synthetic media are evolving rapidly, often outpacing the ability of traditional detection and response mechanisms to adapt. Automated detection tools, digital watermarks, cryptographic signatures, and content traceability initiatives represent important advances, but they continue to face significant technical and operational limitations when deployed in real-world environments, including social media platforms. In many cases, specialized human analysis remains essential for assessing the authenticity of potentially manipulated content.

From a governance perspective, a variety of regulatory strategies have emerged around the world to address this phenomenon. Some initiatives prioritize transparency requirements and the labeling of synthetic content; others focus on restricting specific uses, particularly in electoral contexts; while others have criminalized abuses involving non-consensual pornography. More recent approaches seek to combine risk-based regulation, technological governance, and accountability mechanisms applicable to the various actors involved. Despite these differences, a growing degree of international convergence can be observed around the need to promote transparency, protect individuals from abusive uses, and strengthen mechanisms for oversight, accountability, and liability.

Finally, a central conclusion of this Technology Radar is that deepfakes should be understood as part of a broader transformation in the ways information is produced, circulated, and validated in contemporary society. Addressing the risks associated with these technologies will require technical

solutions, targeted legal instruments, and, above all, the strengthening of digital literacy, media education, algorithmic transparency, and a culture of responsible Artificial Intelligence use. In a context where visual and audio evidence can no longer, by themselves, serve as guarantees of authenticity, preserving trust in digital environments becomes one of the defining challenges for the protection of rights and the maintenance of informed and democratic societies.

‹ references ›

AKHTAR, Zahid. Deepfakes generation and detection: a short survey. **Journal of Imaging**, v. 9, n. 1, 2023. DOI: 10.3390/jimaging9010018. Available at: https://doi.org/10.3390/jimaging9010018. Accessed on: June 15, 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. 2014. Available at: https://www.metropoles.com/colunas/guilherme-amado/moro-avisa-de-golpe-de-inteligencia-artificial-nao-vendo-credito. Accessed on: July 21, 2026.

ANPD. **Inteligência artificial generativa.** Brasília, DF: ANPD, 2024. (Radar Tecnológico, n. 3). Publicação Digital. Available at: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar_tecnologico_ia_generativa_anpd.pdf. Accessed on: June 19, 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. Available at: https://revistas.pucsp.br/index.php/teccogs/article/view/55980/37928. Accessed on: July 4, 2025.

AUSTRALIA. **Criminal Code Amendment (Deepfake Sexual Material) Act 2024 (No. 78, 2024)**. Canberra: Parliament of Australia, 2024. Available at: https://www.legislation.gov.au/C2024A00078/asmade. Accessed on: Oct. 15, 2026.

AWS. **Qual é a diferença entre machine learning e deep learning?** Seattle: Amazon Web Services, 2023. Available at: https://aws.amazon.com/compare/the-difference-between-machine-learning-and-deep-learning/. Accessed on: June 10, 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. Available at: https://repositorio.unb.br/bitstream/10482/54125/1/IngridBorgesDeAzevedo_DISSERT.pdf. Accessed on: July 6, 2026.

BHONDAVE, Avinash. 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. Available at: https://doi.org/10.1109/TPAMI.2021.3116668. Accessed on: July 6, 2026.

BRADY, Madeline. Deepfakes: a new desinformation threat? Technical report, **Democracy Reporting International**, July 31, 2020. Available at: https://democracyreporting.s3.eu-central-1.amazonaws.com/images/20842020-09-01-DRI-deepfake-publication-no-1.pdf. Accessed on: June 15, 2026.

BRASIL. **Decreto-Lei n. 2.848, de 7 de dezembro de 1940**. Código Penal. Rio de Janeiro: Presidência da República, 1940. Available at: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Accessed on: July 6, 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. Brasília, 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. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed on: July 6, 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. Available at: https://www.tse.jus.br/legislacao/compilada/res/2019/resolucao-no-23-610-de-18-de-dezembro-de-2019. Accessed on: July 6, 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. Available at: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14192.htm. Accessed on: July 6, 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. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Accessed on: July 6, 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. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12975.htm. Accessed on: July 6, 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. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12976.htm. Accessed on: June 11, 2026.

BRASIL. Secretaria de Políticas Digitais. Secretaria de Comunicação Social da Presidência da República. **Relatório Grupo de Trabalho Deepfakes.** Brasília, 2026c. Available at: https://www.gov.br/secom/pt-br/arquivos/2606_gt_relatorio-deepfake-v4-web.pdf/. Accessed on: July 6, 2026.

BRASIL. Secretaria de Comunicação Social. **Estratégia Brasileira de Educação Midiática.** Brasília, 2023. Available at: https://www.gov.br/secom/pt-br/arquivos/2023_secom-spdigi_estrategia-brasileira-de-educacao-midiatica.pdf. Accessed on: July 7, 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. Available at: https://arxiv.org/abs/2406.05520. Accessed on: July 6, 2025.

CNTI. **Synthetic media & deepfakes.** CNTI, 6 out. 2025. Available at: https://cnti.org/issue-primers/synthetic-media-deepfakes/. Accessed on: June 1, 2026.

ÇETIN, Reycan. Differences Between Cheapfake and Deepfake and the Threats They Pose in Identity Fraud. **Techsign**, 2022. Available at: https://www.techsign.com.tr/en-blog-posts/differences-between-cheapfake-and-deepfake-and-the-threats-they-pose-in-identity-fraud. Accessed on: June 15, 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. Available at: https://doi.org/10.22214/ijraset.2025.71021 Accessed on: July 6, 2026.

CIDADANIA Digital para adolescentes e jovens. SaferNet Brasil, cursos online. Salvador, 2026. Available at: https://ead.safernet.org.br/cidadaodigital/. Accessed on: July 6,2026.

COLMAN, Ben. **Real-time deepfakes are rewriting the rules of child safety**. World Economic Forum, Opinion, 27 abr. 2026. Available at: https://www.weforum.org/stories/all/ai-voice-deepfakes-child-safety/. Accessed on: July 6,2026.

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. Available at: https://www.scielo.cl/pdf/veritas/n61/0718-9273-veritas-61-32.pdf. Accessed on: July 6, 2026.

DANTAS, Glenda. Racismo recreativo em tempos de deepfake. *In*: SOUZA, Gustavo; SILVA, Tarcizio (org.). **Enfrentando Deepfakes**. Desvelar: Brasília, 2025. [Livro eletrônico]. Available at: https://www.desvelar.org/enfrentando-deepfakes. Accessed on: July 6, 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**, Jan 7, 2025. https://doi.org/10.48550/arXiv.2501.06227. Available at: https://arxiv.org/abs/2501.06227v1. Accessed on: July 6, 2026.

DING, Jeffrey. CHINAI #104: Tencent 2020 AI white paper. **ChinAI**, July 27, 2020. Available at: https://chinai.substack.com/p/chinai-104-tencent-2020-ai-white. Accessed on: June 12, 2026.

DISCIPLINA de Cidadania Digital. **SaferNet Brasil**, Salvador, 2026. Available at: https://cidadaniadigital.org.br/. Accessed on: July 6, 2026.

DOMINGOS, Roney. É #FAKE vídeo que diz que governo federal condenou Serasa a pagar indenização de R$ 30 mil. **G1**, 30 mar. 2024. Available at: 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. Accessed on: July 21, 2026.

EBERL, Andrea; KÜHN, Juliane; WOLBRING, Tobias. 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. Available at: https://es.wired.com/articulos/cheapfakes-y-por-que-seran-mas-graves-que-deepfakes-en-escena-politica. Accessed on: June 23, 2026.

EUROPEAN PARLIAMENT. European Parliamentary Research Service (EPRS). **Tackling deepfakes in European policy**. Luxembourg: Publications Office of the European Union, 2021. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf. Accessed on: June 1, 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. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Accessed on: July 6, 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. Available at: 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. Accessed on: July 7, 2026.

FRANCE. **Code pénal**. Paris: Parlement, 2026. Available at: 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. Accessed on: July 6, 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. Available at: 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. Accessed on: July 7, 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. Available at: https://revistadedireito.fae.edu/direito/article/view/179. Accessed on: July 6, 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. Available at: https://doi.org/10.5771/2192-4007-2021-1-72. Accessed on: July 6, 2026.

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

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge, MA: MIT Press, 2016. Available at: https://www.deeplearningbook.org/. Accessed on: June 11, 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. **Proceedings [...]**. Montreal: Neural Information Processing Systems Foundation, 2014. p. 2672-2680. Available at: https://papers.nips.cc/paper/5423-generative-adversarial-nets. Accessed on: June 29, 2026.

GUO, Tianli; LI, Jisong; TANG, Yunqi. A score based likelihood ratio framework for deepfake image identification in forensic science. **Scientific Reports**, v. 16, 2026. Available at: https://www.nature.com/articles/s41598-026-42176-w. Accessed on: July 6, 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. Available at: 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. Accessed on: July 7,2026.

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. Available at: https://apir.org.ua/lais/en/article/view/467. Accessed on: July 6, 2026.

HAYKIN, Simão. **Redes Neurais: Princípios e Prática**. 2. ed. Porto Alegre: Bookman, 2007. *E-book*. pág.185. ISBN 9788577800865. Available at: https://integrada.minhabiblioteca.com.br/reader/books/9788577800865/. Accessed on: July 6, 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. Available at: https://www.terra.com.br/diversao/gente/mulher-acreditava-estar-noiva-de-brad-pitt-e-descobre-que-caiu-em-golpe,314ff7b946892c73e5fc85a6dc1d7d52zzodh96o.html. Accessed on: July 21, 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. Available at: https://www.sciencedirect.com/science/article/pii/S1877050925012505 . Accessed on: July 6, 2026.

ITAÚ. Golpe no Facebook usa IA e imagem falsa de Pedro Moreira Salles. **Blog Itaú #ÉFake**, 4 jun. 2025. Available at: https://blog.itau.com.br/efake/e-fake-golpe-no-facebook-usa-ia-e-imagem-falsa-de-pedro-moreira-salles. Accessed on: July 21, 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. Available at: https://doi.org/10.1145/3679318.3685398. Accessed on: July 6, 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. Available at: https://doi.org/10.1109/CVPR.2019.00453. Accessed on: June 29, 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. Available at: https://ieeexplore.ieee.org/document/11135143. Accessed on: July 6, 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. Available at: https://ieeexplore.ieee.org/document/11141063/authors. Accessed on: July 6, 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. Available at: https://link.springer.com/article/10.1007/s13369-021-06297-w. Accessed on: July 6, 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. Available at: https://doi.org/10.1016/j.bushor.2019.11.006. Accessed on: July 6, 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. Available at: https://archive.nytimes.com/www.nytimes.com/books/first/k/king-commissar.html. Accessed on: June 15, 2026.

KIRCHENGAST, Tyrone. Deepfakes and image manipulation: criminalisation and control. **Information & Communications Technology Law**, v. 29, n. 3, p. 308-323, 2020. Available at: https://doi.org/10.1080/13600834.2020.1794615. Accessed on: July 6, 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, Apr./June 2025. Available at: https://doi.org/10.1177/20539517251321752. Accessed on: July 6, 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. Available at: https://www.mdpi.com/1424-8220/22/9/3137. Accessed on: July 6, 2026. DOI: 10.3390/s22093137.

KSHETRI, Nir; VOAS, Jeffrey. The deepfake governance gap. **Computer**, v. 59, n. 3, p. 105-112, Mar. 2026. Available at: https://www.computer.org/csdl/magazine/co/2026/03/11404325/2egupWkE1kA. Accessed on: July 6, 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. Available at: https://link.springer.com/article/10.1057/s41304-024-00482-9. Accessed on: July 6, 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. Available at: https://repositorio.pucsp.br/jspui/handle/handle/45762. Accessed on: July 6, 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. Available at: https://onlinelibrary.wiley.com/doi/10.1155/int/9987535. Accessed on: July 6, 2026.

LIU, Yizhi; PADMANABHAN, Balaji; VISWANATHAN, Siva. What exactly is a deepfake? **arXiv**, arXiv:2510.22128, out. 2025. Available at: https://ui.adsabs.harvard.edu/abs/2025arXiv251022128L. Accessed on: June 15, 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. Available at: https://

link.springer.com/article/10.1007/s00530-024-01655-8. Accessed on: July 6, 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. Available at: https://www1.folha.uol.com.br/cotidiano/2026/04/jovens-compartilham-tutorias-para-burlar-eca-digital-site-porno-ensina-a-fraudar-bloqueio.shtml. Accessed on: July 6, 2026.

MAPA da inadimplência e renegociação de dívidas no Brasil. **Serasa**, 2026. Available at: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/ Accessed on: June 15, 2026.

MBEMBE, Achille. **Crítica da razão negra.** Tradução de Marta Lança. 1. ed. Lisboa: Antígona, 2014. Accessed on: June 30, 2026.

MCGLYNN, 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. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/jols.12527. Accessed on: July 6, 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. Available at: https://doi.org/10.1145/3709025.3712218. Accessed on: July 6,2026.

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. Available at: https://seer.pucgoias.edu.br/index.php/panorama/article/view/15148. Accessed on: July 6, 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. Available at: 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. Accessed on: July 21,2026.

MUSSALO, Petteri; GAIN, Ulla; HOTTI, Virpi. Types of Data Clarify Senses of Data Processing Purpose in Health Care. *In:* MANTAS, John *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. Available at: https://www.cnnbrasil.com.br/nacional/alunos-de-colegio-tradicional-do-rio-usam-ia-para-criar-imagens-intimas-de-meninas-policia-investiga/. Accessed on: July 7, 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. Available at: https://arxiv.org/abs/1909.11573. Accessed on: July 6, 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. Available at: https://brasil.unwomen.org/pt-br/stories/noticia/2026/05/relatorio-ponto-de-virada. Accessed on: July 20, 2026.

ONU NEWS. ONU Mulheres aponta riscos da Inteligência Artificial para igualdade de gênero. **ONU News**, 22 jun. 2026. Available at: https://news.un.org/pt/story/2026/06/1853511. Accessed on: July 7, 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. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10354308. Accessed on: July 6,2026.

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. Available at: 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. Accessed on: July 7, 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. Available at: https://doi.org/10.1145/3801962. Accessed on: July 6,2026.

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. Available at: https://doi.org/10.1016/j.sasc.2026.200476. Accessed on: July 6, 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. Available at: https://www.nature.com/articles/s41598-022-23081-4#citeas. Accessed on: July 6, 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. Available at: https://doi.org/10.3390/cryptography9040061. Accessed on: July 6, 2026.

RAHNAMA, Hossein. AI Glossary/Dictionnary. **MIT Media Lab**, Jan. 8, 2025. Available at: https://www.media.mit.edu/tools/ai-glossary-dictionary/. Accessed on: June 4, 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. Available at: https://doi.org/10.3389/fdata.2025.1521653. Accessed on: July 6, 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. Available at: https://doi.org/10.1177/20563051251350975. Accessed on: June 15, 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. Available at: https://doi.org/10.1007/s41060-021-00302-z. Accessed on: July 6, 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. Available at: https://publicacoes.ibccrim.org.br/index.php/RBCCRIM/article/view/267. Accessed on: July 7, 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. Available at: https://arxiv.org/pdf/2502.18066. Accessed on: June 12, 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. Available at: 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. Accessed on: June 11, 2026.

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

SAFERNET BRASIL. Mapeamento da SaferNet identifica deepfakes sexuais em escolas em 10 dos 27 estados brasileiros. **SaferNet Brasil**, 6 out. 2025. Available at: https://new.safernet.org.br/content/mapeamento-da-safernet-identifica-deepfakes-sexuais-em-escolas-em-10-dos-27-estados. Accessed on: July 6, 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. Available at: https://dx.doi.org/10.14569/IJACSA.2025.0161260. Accessed on: July 6, 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. Available at: https://doi.org/10.1186/s40163-024-00239-1. Accessed on: July 6, 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. Available at: https://www.ram2025.sinteseeventos.com.br/arquivo/downloadpublic?q=eyJwYXJhbXMiOiJ7XCJJRF9BUlFVSVZPXCI6XCIyOTEyXCJ9IiwiaCI6ImM1MmFmZTIwZTU0OWI4NTEyYmI5M2YxOWQ3MTUxNjJmIn0%3D. Accessed on: July 6, 2026.

SECURITY HERO. 2023 State of Deepfakes: realities, Threats, and Impact. **Security Hero**, 2023. Available at: https://www.securityhero.io/state-of-deepfakes/#key-findings. Accessed on: June 11, 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. Available at: https://doi.org/10.3390/jimaging10020046. Accessed on: July 6, 2026.

SOUTH KOREA. **90 Days Until National Assembly Elections:** Election Campaigns Using *Deepfake* Restricted. Seul, Press Releases, June 2, 2024. Available at: https://www.nec.go.kr/site/eng/ex/bbs/View.do?bcIdx=226657&cbIdx=1270. Accessed on: July 6, 2026.

SOUTH KOREA. **Act on Special Cases Concerning the Punishment of Sexual Crimes**. Seul: Assembleia Nacional, 2025. Available at: https://law.go.kr/LSW/lsInfoP.do?lsiSeq=277347&chrClsCd=010203&urlMode=engLsInfoR&viewCls=engLsInfoR#0000. Accessed on: July 6, 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. Available at: 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. Accessed on: July 7, 2026.

STOLL, Ashley. Shallowfakes and their potential for fake news. **Washington Journal of Law, Technology & Arts**, Jan. 13, 2020. Available at: https://wjlta.com/2020/01/13/shallowfakes-and-their-potential-for-fake-news/. Accessed on: June 12, 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. Available at: https://doi.org/10.1038/s41598-021-01295-2. Accessed on: July 6, 2026.

TIU, Ekin. Understanding Latent Space in Machine Learning. **Medium**, Feb. 4, 2020. Available at: https://medium.com/data-science/understanding-latent-space-in-machine-learning-de5a7c687d8d. Accessed on: June 22, 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. Available at: https://www.researchgate.net/publication/385916734. Accessed on: June 15, 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, Apr. 30, 2026. Available at: 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. Accessed on: July 7, 2026.

UNICEF. Artificial Intelligence and Child Sexual Abuse and Exploitation. **Unicef**, Feb. 2026. Available at: https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_2.pdf. Accessed on: June 24, 2026.

UNITED KINGDOM. **Sexual Offences Act 2003**. Londres: British Parliament, 2003. Available at: https://www.legislation.gov.uk/ukpga/2003/42/part/1/crossheading/other-offences. Accessed on: July 6, 2026.

UNITED KINGDOM. **Data (Use and Access) Act 2025**. Londres: British Parliament, 2025. Available at: https://www.legislation.gov.uk/ukpga/2025/18/section/138. Accessed on: July 6, 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: United States Senate, Dec. 21, 2018. Available at: https://www.govinfo.gov/app/details/BILLS-115s3805is. Accessed on: June 10, 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: House of Representatives, June 12, 2019. Available at: https://www.govinfo.gov/app/details/BILLS-116hr3230ih. Accessed on: Accessed on: June 10, 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: United States Congress, 2025. Available at: https://www.congress.gov/119/bills/s146/BILLS-119s146es.pdf. Accessed on: July 6, 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. Available at: https://doi.org/10.1145/3588574. Accessed on: July 6, 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. Available at: https://journals.sagepub.com/doi/10.1177/10776990251357299. Accessed on: 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, Dec. 2022. Available at: https://journals.sagepub.com/doi/10.1177/14614448221141648. Accessed on: June 15, 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, Bulgaria, 2025. p. 1-6. Available at: https://ieeexplore.ieee.org/document/11137140. Accessed on: July 6, 2026.

WITNESS. Backgrounder: deepfakes in 2022. **Witness**, Mar. 2022. Available at: https://lab.witness.org/backgrounder-deepfakes-in-2022/. Accessed on: June 1, 2026.

www.gov.br/anpd





ANPD