

Deepfake and Synthetic Media: An Analysis of the Concept, Technology, and Regulation

Thiago Henrique Pessoa

Fundação Getulio Vargas, São Paulo, Brazil

Email: thiago.pessoa@fgv.edu.br

How to cite this paper: Pessoa, T.H. (2026) Deepfake and Synthetic Media: An Analysis of the Concept, Technology, and Regulation. *Journal of Computer and Communications*, 14, 55-73. <https://doi.org/10.4236/jcc.2026.1410003>

Received: August 25, 2026

Accepted: October 20, 2026

Published: October 23, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The advancement of generative artificial intelligence technologies has expanded the capacity for creating, synthesizing, and manipulating highly realistic audiovisual content. In this context, the term deepfake has come to be used to designate different forms of synthetic content, often associated with fraud, disinformation, and manipulation. This article proposes an analysis of the concept and technology used in the phenomenon, seeking to distinguish synthetic media, as a broader category, from deepfake, as a specific manifestation that can assume different purposes and degrees of manipulation. The analysis starts from the evolution of deep neural networks and examines architectures and models used in the generation and manipulation of content, with emphasis on convolutional neural networks, recurrent networks, Transformers, autoencoders, and Generative Adversarial Networks. Four recurrent modalities of manipulation are also analyzed: face swapping, facial reenactment, generation of talking faces, and editing of facial attributes, in addition to facial synthesis. From the selected literature, it is proposed to understand and identify the phenomenon through three dimensions: identity or reality of reference, intention of use, and granularity of manipulation. Finally, the article examines the Brazilian regulatory response, especially TSE Resolutions No. 23,732/2024 and No. 23,755/2026, arguing that regulation tends to focus on the effects, purpose, and risk of the content, and not exclusively on the technology employed. It is concluded that the distinction between synthetic media and deepfake is relevant to prevent the technology from being confused with the potentially abusive use that is made of it. It is argued that the identity or reality of reference and the granularity of the manipulation constitute the classification criteria, while the intention qualifies the risk associated with the content, without being a necessary condition to characterize it as deepfake.

Keywords

Deepfake, Synthetic Media, Generative Artificial Intelligence, Deep Learning,

1. Introduction

The recent evolution of artificial intelligence has significantly modified the ability of computational systems to recognize, classify, generate, and manipulate information. Although the fundamentals of artificial neural networks date back to previous work, the 2010s marked a qualitative shift in the performance of deep learning systems applied to computer vision. In 2012, the work of Krizhevsky, Sutskever, and Hinton demonstrated the potential of deep convolutional networks in image recognition, contributing to the rapid expansion of Deep Learning [1]. Subsequently, He *et al.* (2015) reported a top-5 error of 4.94% in ImageNet, a result lower than the human performance estimate of 5.1% used by the authors as a reference [2]. In 2016, He *et al.* achieved a 3.57% top-5 error with a residual network in a set of models, consolidating the evolution of deep networks in visual recognition [3].

This technological trajectory is relevant for understanding the emergence of contemporary synthetic media. The same ability to learn visual, sound, and linguistic patterns that allows classifying information can also be used to produce new representations or modify existing content. LeCun, Bengio, and Hinton (2015) highlight that Deep Learning allows the learning of representations at multiple levels of abstraction, while Goodfellow *et al.* (2014) demonstrate, with GANs, the possibility of structuring models aimed at generating new samples [4] [5].

The discussion about deepfake, however, is not exclusively technological. The term itself emerged in 2017, in Reddit communities associated with the production of face-swapping videos, initially in non-consensual pornographic content [6] [7]. The origin of the term contributed to deepfake acquiring a strong negative connotation. This association, however, does not mean that all synthetic media is produced for malicious purposes. Rauchfleisch, Vogler, and de Seta (2025) show that the terms deepfake and synthetic media produce different frameworks in public perception and that synthetic media functions as a broader category [7], encompassing applications that can present risks but also benefits.

At this point, the central problem of this article emerges: should the technology that makes it possible to produce a highly realistic synthetic representation be confused with the purpose for which this representation is produced or used? The question is relevant because the same technical infrastructure can sustain applications of entertainment, education, accessibility, research, and communication, but also fraud, disinformation, impersonation, and political manipulation. Liu, Padmanabhan, and Viswanathan (2026), based on the analysis of 826 academic publications, identify significant heterogeneity in the definitions of deepfake and organize the literature into three dimensions: Identity Source, Intent, and Manipulation Granularity [8].

The objective of this article is to conceptually analyze the term deepfake, its

origin, technological evolution, and relationship with the broader concept of synthetic media, with special attention to the distinction between technology, content, and purpose of use. As specific objectives, it seeks to: 1) differentiate artificial intelligence, Machine Learning, Deep Learning, neural architectures, and generative models; 2) present the main technological forms of synthetic media production; 3) propose a conceptual reading based on content origin, intention, and degree of manipulation; and 4) relate this distinction to the Brazilian regulatory response, especially in the electoral context.

The guiding conceptual hypothesis is that deepfake should not be understood exclusively as a synonym for synthetic media or as the name of a specific technology. Instead, it is proposed to understand synthetic media as a broad category and deepfake as a manifestation that can be identified by the combination of content characteristics, relationship with an identity or reality of reference, intention of use, and degree of manipulation. This proposition is analytical and will be confronted with the literature examined.

To make this analytical path explicit, the article is organized into three movements. The first, developed in Sections 2 to 4, establishes the conceptual vocabulary and technological mechanisms underlying synthetic media, already anticipating why concept, technology, and regulation cannot be treated as a single undifferentiated problem. The second, in Section 5, proposes a broad and empirical definition of deepfake through the analysis of three analytical dimensions—Identity or Reality of Reference, Intent, and Granularity of Manipulation—capable of distinguishing deepfake, specifically, from synthetic media in general. The third, in Section 6, examines the Brazilian regulatory response in light of these dimensions. Sections 7 to 10 then present the methodology, discuss the findings, and present the article's conceptual contribution and final considerations.

2. Theoretical Framework

2.1. The Origin and Semantic Evolution of the Term Deepfake

The term deepfake emerged in 2017, associated with the username of a Reddit community that shared face-swapping videos, especially in pornographic content. De Seta (2021) notes that the term derives from the nickname used by the user and that the first practices were related to replacing the faces of celebrities in pornographic videos. Rauchfleisch, Vogler, and de Seta (2025) also describe the term as a portmanteau of deep learning and fake, consolidated in amateur communities that experimented with face-swapping techniques [7].

The origin of the term is relevant because it helps to explain its semantic load. Deepfake was not born as a neutral denomination of a family of computational techniques; it was born associated with a specific practice of identity manipulation and a socially controversial purpose. With the popularization of the technology, however, the term began to be used for different modalities of audiovisual manipulation. The semantic expansion produced a situation in which an originally specific term began to function, in some contexts, as a synonym for content generated

or altered by AI.

The most recent literature questions this equivalence. Rauchfleisch, Vogler, and de Seta (2025) distinguish deepfake from synthetic media and demonstrate that the choice of term influences the perception of risks and benefits [7]. Liu, Padmanabhan, and Viswanathan (2026) also show that the academic literature presents heterogeneous conceptualizations [8]. Thus, the definition of deepfake cannot be established only by the technology employed.

2.2. Artificial Intelligence, Machine Learning, Deep Learning, and Generative Models

Artificial intelligence is the broadest field in which different approaches to building systems capable of performing tasks associated with perception, reasoning, learning, and decision-making are included. Turing (1950) shifted the discussion to the possibility of machines exhibiting intelligent behavior, while McCulloch and Pitts (1943) provided a pioneering mathematical formulation of artificial neural networks. The field of AI, therefore, is not reduced to statistical learning or neural networks [9] [10].

Machine Learning is an approach to computer learning in which systems extract patterns from data and improve their performance in tasks from experience. Mitchell (1997, p. 2) formalized this idea as experience-based, performance-based, and task-based learning [11]. Among the classic approaches are supervised, unsupervised, and reinforcement learning.

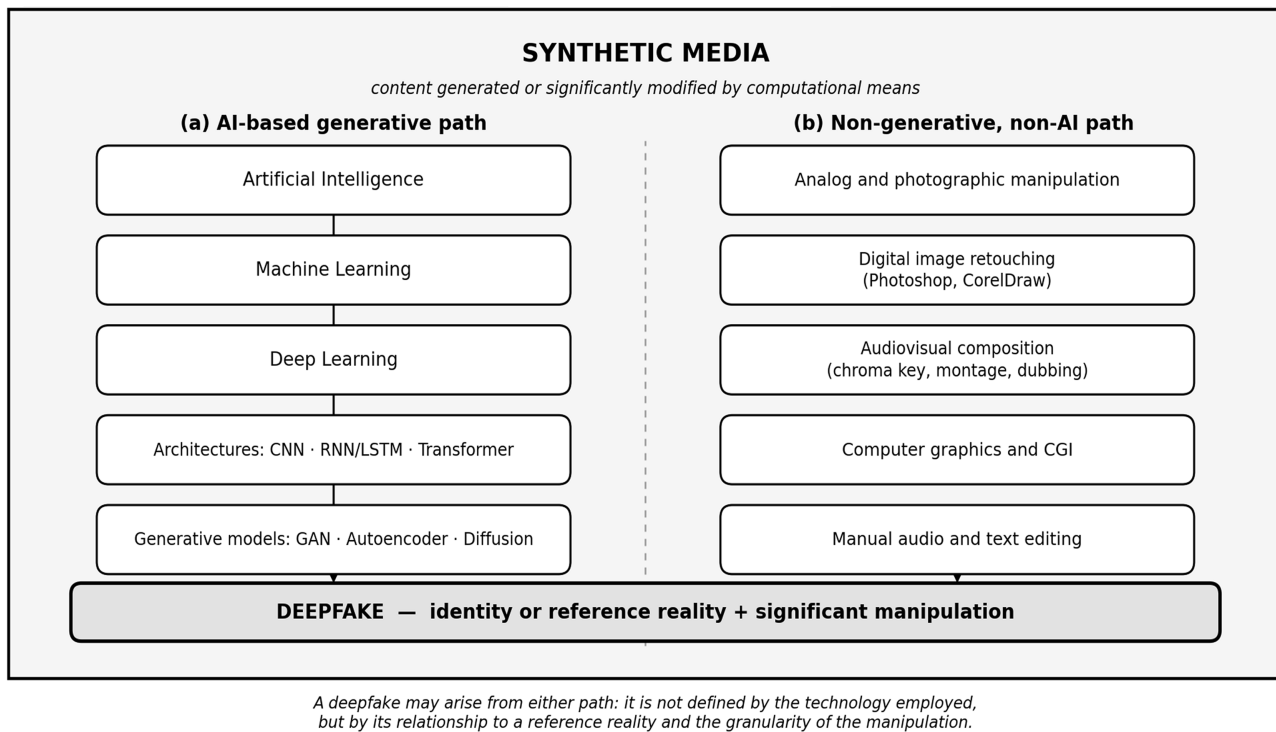
Deep learning is a branch of Machine Learning based on models composed of multiple layers of processing capable of learning representations at different levels of abstraction [4]. This framework has favored advances in visual recognition, audio processing, language, and content generation.

Within Deep Learning, there are different architectures and models. CNNs are architectures especially suited to data with spatial structure, such as images, while recurrent networks have historically been used for sequential data. Transformers constitute an architecture based on attention mechanisms and have become central in language processing and in several multimodal systems. Autoencoders use a coding and decoding structure to learn latent representations and reconstruct data. GANs, in turn, organize two models—generator and discriminator—in an adversarial learning process [5].

This differentiation is fundamental to the argument of this article: CNN, Autoencoder, Transformer, and GAN are not synonyms or successive rungs of a single hierarchy. CNN is an architecture; Autoencoder is a family of representation and reconstruction models; GAN is an adversarial generative framework; and Transformer is an attention-based architecture. All can integrate deep learning systems, but perform different functions.

Figure 1 organizes these relationships. The diagram distinguishes two paths of synthetic media production: 1) the generative path based on artificial intelligence, which starts from the field of AI and goes down to generative models; and 2) the

non-generative and non-AI-based pathway, which encompasses analog and photographic manipulation, digital image retouching, audiovisual composition, and computer graphics, discussed in Section 4. Deepfake crosses both paths, which shows that it is not defined by the technology employed, but by the relationship with a reference reality and the granularity of the manipulation.



Source: prepared by the authors, based on LeCun, Bengio and Hinton (2015), Goodfellow *et al.* (2014), Fanaya (2021), Santaella (2021), [4] [5] [8] [12] [13].

Figure 1. Synthetic media: generative and non-generative ways of production and the position of deepfake.

2.3. Synthetic Media as a Conceptual Category

The concept of synthetic media allows the discussion to shift from the specific technique to content produced or significantly modified. NIST AI 100-4 addresses synthetic content as information, including images, videos, audio, and text, that can be significantly modified or generated by algorithms, including artificial intelligence [14]. NIST emphasizes that the existence of synthetic content does not, in itself, imply harm; the problem involves risks associated with the production, distribution, authenticity, provenance, and use of that content.

From this perspective, synthetic media constitutes a broader category than deepfake. Rauchfleisch, Vogler, and de Seta (2025) show empirically that the choice between the terms deepfake and synthetic media alters public perception of risks and benefits, with synthetic media operating as a euphemism [7]. The distinction is relevant because it allows for separating the technological capacity to generate or alter content from the social purpose attributed to the content.

The proposition of this article is, therefore, that the expression synthetic media

should function as a broad technological and communicational category, while deepfake should be used for situations in which the synthesis or manipulation involves, in a relevant way, the representation of a person, identity, voice, appearance, or manifestation associated with a reference reality, that is, a real and recognizable element—a person, face, voice, object, or event—whose presence confers an appearance of legitimacy and veracity to the other elements and actions created digitally, especially when the manipulation produces the potential for confusion about what actually occurred. The intention to deceive is an important dimension, but it should not be considered in isolation, as there are infinite non-deceptive applications recognized in the literature [8].

2.4. Operational Definitions

Three expressions used throughout this article require prior definition, since the classification rule proposed in Section 5.4 depends on them. The following definitions are operational: they are valid for the purposes of this analysis and are used uniformly in Sections 3 through 10.

Reference reality designates any person, identity, voice, appearance, event, recording, or pre-existing and recognizable manifestation that serves as the basis, total or partial, for the synthetic content produced, and whose presence gives the result a claim to correspond with the world. The expression covers real people, events that actually occurred, and recordings actually made. It also covers fictional representations when they are presented as an authentic record of a fact or manifestation—a hypothesis in which the reality of reference is the simulated event itself. It does not cover entirely fictitious creations, without correspondence with a real person, event, or recording and without pretense of authenticity: in this case, even if the content is generated by artificial intelligence, there is no reference reality to be falsified.

Significantly modified refers to changes that go beyond trivial technical adjustments—compression, color correction, resizing, cropping, noise reduction, stabilization—and that perceptibly modify the identity, speech, expression, action, or context represented to the point of altering an observer's reasonable interpretation of the content. The criterion is the effect of the change on the interpretation, not the ratio of pixels, frames, or samples actually modified.

Potentially confusable with reality indicates that the content is sufficiently plausible for a reasonable observer, without additional information and under the ordinary conditions of circulation of the material, to believe that it records a real person, event, or statement. This is an objective criterion, measured by the properties of the content and the context of circulation, and not by the intention of those who produced it—a distinction resumed in Section 5.4.

3. Synthetic Media Technology

3.1. Convolutional Neural Networks (CNN)

Convolutional neural networks were developed to explore spatial structures pre-

sent in data such as images. LeCun *et al.* (1998) demonstrated the importance of convolutional operations and parameter sharing in the learning of visual features [15]. In the context of synthetic media, CNNs can participate both in the analysis of facial features and in steps of generation, reconstruction, or detection of manipulations.

3.2. Generative Adversarial Networks (GANs)

Goodfellow *et al.* (2014) proposed Generative Adversarial Networks as a framework composed of a generator and a discriminator. The generator seeks to produce samples that approximate the distribution of the training data, while the discriminator seeks to distinguish real samples from those produced. Adversarial training allows the generator to progressively improve the quality of the samples.

It is important, however, to avoid oversimplifying by assuming that every GAN is necessarily composed of two CNNs. The original formulation by Goodfellow *et al.* uses neural networks, and the framework can be implemented with different architectures depending on the application. In imaging applications, convolutional architectures are often incorporated into the components of the GAN, but CNN and GAN remain distinct concepts.

3.3. Recurrent Networks, LSTM, and Transformers

Recurrent neural networks have been developed to work with sequential information, allowing earlier states to influence the processing of later inputs. LSTM architectures have been proposed to address learning difficulties in long-term dependencies. In the context of synthetic media, sequential models can be employed in voice, motion, and video.

Transformers constitute an architecture based on attention mechanisms that became central to language processing and were later expanded to multimodal tasks. Their importance for synthetic media is related to the ability to work with large amounts of data and contextual relationships, contributing to the generation of text, audio, image, and multimodal systems.

3.4. Autoencoders and Latent Representations

Autoencoders organize the processing into an encoding and a decoding step. The encoder transforms the input data into a latent representation, while the decoder seeks to reconstruct the data from this representation. In face-swapping applications, this structure allows one to learn relevant features of identities and reconstruct them in new contexts. The latent representation is therefore an important element in understanding how systems can separate features and later recombine them.

The use of autoencoders also demonstrates why deepfake cannot be defined as synonymous with GAN. Historically, different architectures and methods have been employed for facial generation and manipulation, including autoencoders and adversarial networks. The end result—a synthetic audiovisual medium—does not reveal, by itself, which architecture was employed.

3.5. Deepfake Detection: Accuracy and Technical Limits

Deepfake detection operates, in technical terms, as a binary classification—real or synthetic—applied at the image or pixel level, mostly employing Convolutional Neural Networks and, more recently, attention-based architectures. Models based on Vision Transformers (ViT) have achieved expressive performance, with accuracy close to 94% and a response time of approximately 26.7 milliseconds, including on compressed videos—a predominant condition in social networks [16].

This result, however, hides a relevant structural limitation. CNN-based detectors often incur overfitting: instead of learning to identify the manipulation itself, the model tends to memorize specific characteristics of the training set, such as the video's compression pattern, the lighting scheme, or even the background. As a consequence, the detector's accuracy drops sharply in the face of deepfakes produced by generation methods not represented in the training [16].

This pattern configures an arms race between generation and detection: advances in the mechanisms for identifying synthetic content tend to stimulate the development of generation techniques capable of circumventing them, and vice versa [17]. The practical consequence is that no purely technical detection solution should be considered definitive—which reinforces the relevance of regulatory transparency and labeling mechanisms, discussed in Section 6, as a complement, and not a substitute for automated detection.

4. Main Mechanisms of Creating Synthetic Media

The manipulation of images and visual representations is not a recent phenomenon or exclusive to artificial intelligence. In the Renaissance, painters financed by royalty and the wealthy classes deliberately altered the appearance of their subjects, making them younger, more beautiful, or more heroic than reality would allow. In cinema, techniques such as chroma key—replacing a solid-colored background with another image—and, later, advances in computer graphics significantly expanded the ability to simulate realities that did not exist on screen [12].

In the field of photography, technical manipulation became popular in the 1980s, with the introduction of computer image retouching programs—of which Photoshop is the best-known example. It was this movement that called into question, for the first time in a massive way, the hitherto relatively unshakable belief that photographs could not lie [13]. The difference introduced by contemporary generative artificial intelligence lies in the ability of deep learning models to learn patterns from large databases and produce new representations with a high degree of realism, without relying on manual editing.

4.1. Face Swapping

In face swapping, the facial features of a source person are transferred or reconstructed over a target person. The operation may involve face detection, alignment, feature extraction, generation, and composition. Autoencoders, CNNs, and generative models may participate in different steps. The technical goal is to pro-

duce visual continuity between identity, lighting, texture, pose, and context.

4.2. Face Reenactment

In facial reenactment, movements, expressions, or poses of a source person are used to control the representation of a target person. The technique allows for the reproduction of the head, eyes, mouth, and expressions. Irfan *et al.* (2025) [17] describe reenactment as one of the modalities of deepfake content generation and manipulation.

4.3. Talking Face Generation

Speaking face generation produces video of a person from text, audio, video, or multimodal inputs, seeking to synchronize speech, lip movement, and facial expressions. The technique can combine speech synthesis, motion models, and visual generation, making it possible to animate a photograph or static representation.

4.4. Facial Attribute Editing

The editing of facial attributes modifies specific characteristics without necessarily replacing identity. Attributes such as apparent age, hair, expression, or other visual characteristics can be changed. This modality demonstrates that not all facial manipulation implies a complete change of identity.

4.5. Facial Synthesis

Facial synthesis represents a different situation: instead of necessarily modifying an existing person, the model can produce a face that does not correspond to a specific real person. This ability highlights the need to distinguish synthetic content from deepfake, since an entirely synthetic face can be synthetic media without necessarily constituting a forgery of identity.

4.6. Summary of the Modalities

Table 1 systematizes the five modalities of synthetic media discussed in Sections 4.1 to 4.5, indicating, for each one, its main operation, the identity reference involved, and an example of the resulting content.

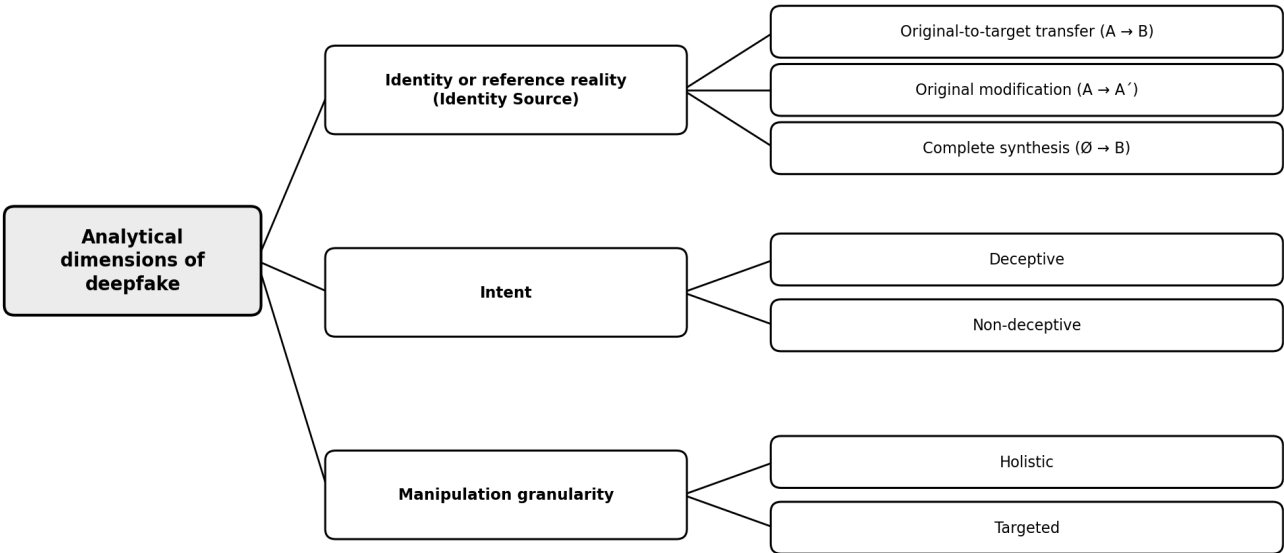
Table 1. Summary of synthetic media modalities.

Modality	Main operation	Identity reference	Result example
Face swapping	Face replacement/reconstruction	Source and target person	A's face in the context of B
Face reenactment	Transfer of movements/expressions	Source and target person	B reproduces A's movements.
Talking face generation	Speech and face synchronization	Target person or image	Photography that speaks
Facial attribute editing	Changing attributes	Existing person	Change of age/hair/expression
Facial synthesis	Generation of a new facial identity	There may not be a specific source person.	Synthetic face

Source: prepared by the author based on Irfan *et al.* (2025) and Pei *et al.* (2026) [17] [18].

5. Dimensions to Differentiate between Synthetic and Deepfake Media

The word deepfake is widely used, but rarely defined systematically. This section adopts as its starting point the survey conducted by researchers at the University of Maryland, who analyzed 826 peer-reviewed publications, published between 2017 and 2025, and identified three recurring dimensions in the conceptualizations of deepfake present in the literature [8]. They are: 1) the identity or reality of reference, that is, the relationship between the preexisting referent and the generated content; 2) the intention or purpose of use; and 3) the granularity of the manipulation. **Figure 2** summarizes the three dimensions and their developments; the following subsections examine each of them.



Classification criteria: identity or reference reality and granularity. Risk criterion: intent.

Source: prepared by the authors, based on Liu, Padmanabhan, and Viswanathan (2026) and ANPD (2026, p. 19).

Figure 2. Analytical dimensions of deepfake definitions.

5.1. Identity or Reference Reality (Identity Source)

The first dimension examines the relationship between the identity or reality of reference and the generated content, that is, the link between factual reality and synthetic creation. Liu, Padmanabhan, and Viswanathan (2026) distinguish three configurations:

1) transfer from original to target ($A \rightarrow B$), in which one identity fully replaces another, as in the complete exchange of faces; 2) modification of the original ($A \rightarrow A'$), in which attributes are altered—apparent age, expression, hair, eye color—preserving the central identity; and 3) complete synthesis ($\emptyset \rightarrow B$), consisting of the creation of an entirely artificial identity, without an individual referent in the real world. This dimension corresponds to what Liu, Padmanabhan, and Viswanathan (2026) call Identity Source, here translated as identity or reference reality to maintain coherence with the operational definitions in Section 2.4.

The more direct the relationship between the content produced and a recognizable identity or event, the greater the relevance of the issue of authenticity and representation. In configuration 3), there is no individual identity of reference to be falsified, which does not exclude the existence of a reality of reference regarding the event, place, or context portrayed.

5.2. Intent—Purpose of Manipulation

The second dimension considers purpose. Content can be produced for entertainment, education, accessibility, research, artistic creation, or communication, without the intention of misleading third parties. It can also be used for fraud, disinformation, impersonation, reputational damage, or other illicit or abusive purposes. Liu, Padmanabhan, and Viswanathan (2026) show that literature is not homogeneous and that there are applications that are not exclusively threatening. It should be noted at the outset that this dimension is not part of the classification rule proposed in Section 5.4: it qualifies the risk associated with the content, and not its characterization as deepfake.

5.3. Manipulation Granularity—Degree of Change

The third dimension examines how much of the content has been altered. Manipulation can be localized, such as a facial attribute change, or comprehensive, involving identity, voice, expression, context, and narrative. This dimension is important because the mere use of AI does not indicate the degree of intervention carried out on the original content.

5.4. Analytic Proposition

The combination of the three dimensions allows for the proposal of a more precise conceptual matrix. Content can be highly synthetic and still not constitute deepfake in the strict sense when there is no real identity or reference event being falsified. Likewise, content that uses simple editing techniques can produce a relevant falsification without employing the most advanced generative models. Therefore, the technology employed should not be the only criterion that guides classification.

The central proposition of this article can be summarized as follows: synthetic media corresponds to the universe of content generated or significantly modified by computational means, including but not limited to artificial intelligence; deepfake constitutes a specific manifestation of this universe. In terms of the classification rule, content should be treated as deepfake when, cumulatively: (i) there is an identifiable identity or reference reality—in configurations $A \rightarrow B$ and $A \rightarrow A'$ and, as for the event or context portrayed, also in $\emptyset \rightarrow B$; and (ii) the granularity of the manipulation is sufficient to significantly modify the perception of this reference reality, making the result potentially confusable with reality, under the terms of the operational definitions in Section 2.4. In the absence of either of the two requirements—because there is no identifiable referent or because the altera-

tion is trivial—the content remains classified as synthetic media in the broad sense.

Intention, the third dimension, is not part of the classification rule: it qualifies the risk associated with the content. The reason is simultaneously practical and epistemic. Classification needs to operate on verifiable properties of the artifact—existence of referent and extent of alteration—while intention is the subjective state of the producer, often inaccessible to the observer, the recipient, and even the judge at the time the content circulates. Conditioning classification on intention would make the category inapplicable precisely in the situations in which it is most necessary: anonymous content, of unidentified authorship, or of contested purpose.

It follows that a transparent and consented manipulation of identity remains, technically, a deepfake. An actor digitally rejuvenated with authorization, a cloned and labeled voice for accessibility purposes, or a declared reenactment in audio-visual production satisfies requirements (i) and (ii): there is an identifiable referent and there is significant manipulation. It is, however, a low-risk deepfake, and not a distinct category. Consent, labeling, and transparency do not mischaracterize the phenomenon; they shift it on the risk scale and alter the appropriate legal response. This is precisely the logic of the regulation examined in Section 6, which does not prohibit synthetic content itself, but imposes duties of identification and prohibits specific uses in certain contexts.

It should be noted that regulatory definitions may adopt a more restricted approach. The ANPD (2026, p. 22), when delimiting the scope of the authority's action, considers deepfakes to be the holistic transfer of identity ($A \rightarrow B$), the manipulation of an individual's attributes ($A \rightarrow A'$), and the facial synthesis of a character ($\emptyset \rightarrow B$) when used with the intention of harming the public [16]. The excerpt is consistent with the purpose of a document oriented toward data protection, which needs to isolate the uses subject to administrative intervention. It does not coincide, however, with the conceptual classification proposed here, in which harmful intent is an indicator of risk and a criterion for regulatory response, and not a constitutive element of the concept. The two readings are compatible as long as the plan in which each operates is kept explicit.

5.5. Model Reach beyond Facial Audio-Visual Media

The modalities described in Section 4 focus on audiovisual facial manipulation, as they are the most documented in the literature and because they are at the historical origin of the term. The three dimensions proposed in this section, however, are not restricted to this domain. Voice cloning without associated facial manipulation, text attributed to a real person—a forged statement or quote—and non-facial images that depict real events, documents, or locations can be evaluated by the same criteria: existence of identity or reference reality, granularity of the alteration, and potential for confusion with reality.

The application only requires identifying on which element the source of iden-

tity in each modality falls. In voice cloning, it operates on the timbre and prosodic patterns of a given speaker; in text, on the authorship attributed to an identifiable sender; in non-facial images, on the event, object, or place portrayed. The classification rule formulated in Section 5.4 remains unchanged in these hypotheses. Section 4 does not, however, intend to offer an exhaustive taxonomy of non-facial modalities, which is indicated as a future research agenda.

6. Legal Basis and Regulatory Response

The regulation of synthetic content faces a structural difficulty: the technology evolves faster than traditional legal categories. For this reason, a strategy based exclusively on the technical description of the algorithm tends to have limitations. The regulatory response may alternatively focus on the purpose, effects, context, and risk of the content.

6.1. Civil Rights Framework for the Internet

In Brazil, Law No. 12,965/2014, known as the Civil Rights Framework for the Internet, establishes principles, guarantees, rights, and duties for the use of the Internet and regulates aspects related to users' rights, applications, and liability. The rule was not specifically designed for deepfakes, but it is part of the legal set that must be considered in the analysis of illicit or harmful content distributed on the Internet [19].

6.2. TSE Resolution No. 23,732/2024

In the electoral sphere, TSE Resolution No. 23,732/2024 amended Resolution No. 23,610/2019 and introduced specific rules for the use of artificial intelligence in electoral propaganda [20]. The TSE now provides, among other measures, for the prohibition of deepfakes and the obligation to identify the use of artificial intelligence in synthetic content used in propaganda. The rule, therefore, does not regulate only one specific technology: it establishes obligations linked to the nature of the content, its electoral use, and the potential to affect the balance of the election [21].

6.3. TSE Resolution No. 23,755/2026

For the 2026 Elections, TSE Resolution No. 23,755/2026 again updated Resolution No. 23,610/2019. The new text maintains the duty to inform, in an explicit, prominent, and accessible manner, that certain synthetic content was manufactured or manipulated and what technology was used. The amendment also established that the publication, republication, and paid promotion of new synthetic content produced or altered by AI that uses the image, voice, or manifestation of a candidate or public figure in the period between 72 hours before and 24 hours after the end of the election, even when labeled and in accordance with the other requirements of the rule [22], are prohibited.

Resolution No. 23,755/2026 also provides for mechanisms related to accounta-

bility and proof, including the possibility of reversing the burden of proof in situations in which the technical proof of manipulation is excessively burdensome, in addition to the possibility of cooperation between electoral courts, universities, entities, and bodies that have professionals trained in digital illicit expertise and artificial intelligence [22].

The normative evolution reinforces the hypothesis of this article: the regulatory focus is not exclusively on the technological architecture that produced the content, but on the characteristics of the content, the identification of its use, the electoral context, the capacity to cause damage, and the obligations of transparency and removal. The norm, therefore, is close to a logic based on risk, purpose, and context.

7. Methodology

This article adopts a qualitative, exploratory, and conceptual approach, combining bibliographic research and documentary analysis. The bibliographic research was oriented toward the identification and analysis of academic works related to the origin and definition of deepfake, synthetic media, Deep Learning, GANs, autoencoders, and audiovisual manipulation techniques. Among the central references are Goodfellow *et al.* (2014), LeCun, Bengio, and Hinton (2015), de Seta (2021), Rauchfleisch, Vogler, and de Seta (2025), Irfan *et al.* (2025), and Liu, Padmanabhan, and Viswanathan (2026).

The bibliographic search was conducted between May and August 2026 in the Google Scholar, Scopus, Web of Science, SciELO, IEEE Xplore, and SSRN databases, complemented by Consensus, a semantic search tool for peer-reviewed literature. The descriptors deepfake, synthetic media, deep learning, generative adversarial network, face swapping, face reenactment, deepfake detection, deepfake definition, and deepfake regulation, in Portuguese and English, were used, separately and in combination.

The following were considered eligible: 1) peer-reviewed articles, chapters, technical reports from public agencies, and normative acts; 2) published between 2014 and 2026, with previous works admitted due to their foundational relevance to the field (McCulloch and Pitts, 1943; Mitchell, 1997; LeCun *et al.*, 1998; Turing, 1950); 3) written in Portuguese or English; and 4) dealing directly with the definition of the phenomenon, generation architectures, manipulation modalities, or regulatory response. Studies focused exclusively on detector performance metrics or the description of datasets, as well as dissemination texts without editorial review, were excluded.

From the recovered set, twenty-two sources were selected, which make up the framework of this article: fifteen peer-reviewed articles and preprints, one reference work, two institutional technical reports—NIST AI 100-4 and ANPD, Radar Tecnológico n. 6—and four Brazilian normative acts. The selection does not intend to exhaust the literature on deepfake, but to support the conceptual construction proposed in Sections 5 and 6.

The documentary analysis focused on normative and technical documents relevant to the understanding of the phenomenon, especially NIST AI 100-4 and TSE Resolutions No. 23,732/2024 and No. 23,755/2026. The objective was not to conduct an exhaustive systematic review of the entire literature on deepfake, but to build a conceptual analysis capable of organizing the relationships between technology, content, intention, manipulation, and regulation.

As an analytical procedure, the materials were organized into four axes: 1) origin and semantics of the term deepfake; 2) technological architecture and generation models; 3) synthetic media modalities and degrees of manipulation; and 4) regulatory response. From these axes, the proposition of the three analytical dimensions—identity or reality of reference, intention, and granularity of manipulation—used in the discussion was constructed.

8. Analysis and Discussion

The analysis shows that the main conceptual problem of the field lies in the overlapping of three elements that should be distinguished: technology, content, and purpose. Technology corresponds to the models and architectures used; content corresponds to the audiovisual, textual, or multimodal object produced or altered; purpose corresponds to the social use of this content. When these elements are treated as equivalent, the expression deepfake comes to simultaneously designate a technique, a product, and a behavior.

The literature analyzed supports the need for this separation. Rauchfleisch, Vogler, and de Seta (2025) show that deepfake and synthetic media are not semantically neutral terms and that affects the perception of risks and benefits. Liu, Padmanabhan, and Viswanathan (2026) demonstrate that academic definitions vary and that the intention of use is only one of the relevant dimensions. NIST (2024) works with synthetic content as a broad category, without assuming that all synthetic content is harmful.

From a technological point of view, the analysis also shows that there is no single deepfake technology. CNNs, autoencoders, GANs, recurrent models, and Transformers can participate in different systems. Therefore, classifying content as deepfake just because it uses a GAN, for example, would be conceptually insufficient. In the same way, content can present identity manipulation without a GAN being employed.

The consequence is that the most appropriate classification must start from the content and its relationship with an identity or reality of reference, combined with the granularity of the manipulation; the purpose of use is then incorporated into the risk assessment. This perspective allows legitimate applications to be included without eliminating the possibility of recognizing malicious uses. It also avoids the inverse problem of considering that all synthetic content is benign just because the technology has legitimate applications.

In the regulatory field, the Brazilian experience offers evidence that the normative response can be structured without depending on the definition of a specific

algorithm. TSE Resolutions No. 23,732/2024 and No. 23,755/2026 use categories such as multimedia synthetic content, AI, manipulation, labeling, prohibition, and electoral context. The move is relevant because it reduces the dependence on a rigid technological classification and allows the standard to keep up with the evolution of generation methods.

The analysis thus allows us to propose a conceptual sequence: AI constitutes the broadest technological field; Machine Learning corresponds to a learning approach; Deep Learning constitutes a strand based on deep neural networks; architectures such as CNN and Transformer structure different forms of processing; models such as autoencoders and GANs can perform synthesis tasks; these technologies result in synthetic content; and certain manifestations of these contents can be classified as deepfakes according to the identity or reality of reference and the granularity of the manipulation, and it is up to the intention to qualify the resulting risk.

This sequence should not be understood as a rigid technological hierarchy, since CNN and GAN belong to distinct conceptual categories. It is an explanatory chain that starts from the general field and arrives at the social and legal phenomenon that constitutes the object of this article.

9. Conceptual Contribution

The main conceptual contribution proposed is the separation between synthetic media and deepfake through three articulated dimensions—identity or reference reality (Identity Source), intent, and manipulation granularity—of which the first and third operate as classification criteria and the second as a risk qualifier. Based on them, it is proposed that the classification of content should not be carried out exclusively by the technology that produced it.

This contribution unfolds on three levels, which should not be confused with the three classificatory dimensions of Section 5. The first contribution is semantic: synthetic media functions as a more comprehensive and technologically neutral term, while deepfake carries a history of association with impersonation and manipulation. The second is technological: deepfake does not correspond to a single architecture, and may result from the combination of different models and techniques. The third is analytical: the intention of use, although relevant to risk assessment, must be analyzed in conjunction with the identity or reality of reference and with the granularity of the manipulation.

The proposal can be represented as a two-axis matrix. In the classification axis, the more direct the relationship with an identity or reference reality and the greater the granularity of the manipulation, the more clearly the content is characterized as deepfake; the absence of an identifiable referent or the triviality of the alteration keeps it in the field of synthetic media in a broad sense. In the risk axis, the intention to mislead third parties, added to the context of circulation and the ability to cause harm, defines the seriousness of the case and the intensity of the appropriate regulatory response. The same content can, therefore, be low- or high-

risk deepfake without changing category.

This contribution is not intended to establish a universal and definitive definition of deepfake. Its goal is to offer a conceptual instrument to reduce ambiguities and allow future research, regulatory analysis, and risk assessments to separately describe technology, content, and purpose.

10. Final Considerations

Before synthesizing the results, it is worth distinguishing two levels in this article, whose overlap may generate ambiguity. The conceptual, technological, and normative analysis announced in the subtitle describes the structure of the investigation and corresponds, respectively, to Sections 2, 3, and 6. The three classificatory dimensions—identity or reality of reference, intention, and granularity of manipulation—constitute the analytical instrument proposed in Section 5 and are used with the same denomination throughout the text. The two sets are not interchangeable.

The development of generative artificial intelligence technologies has transformed the ability to create and modify digital content, especially images, videos, audio, and language. The evolution of deep learning, convolutional networks, autoencoders, GANs, and more recent architectures has expanded the realism and scale of digital synthesis. However, the existence of this technological capacity does not determine, by itself, the purpose or risk of its use.

The article demonstrated that deepfake and synthetic media should not be treated as equivalent expressions. Synthetic media constitutes a broader category of content generated or significantly modified by computational means, while deepfake represents a specific manifestation associated with the representation or manipulation of identities, people, voices, expressions, or reference events—modifications capable of deceiving the senses and the perception of veracity of the common observer. The analysis also demonstrated that deepfake does not correspond to a single technology: different architectures and models can be employed to produce similar results.

Based on the literature analyzed, the article proposes three dimensions for the classification of the phenomenon: identity or reality of reference, intention of use, and granularity of manipulation. This approach makes it possible to overcome an exclusively technological classification and contributes to distinguishing legitimate applications from potentially abusive uses. Among them, identity or reality of reference and granularity operate as classification criteria, while intent qualifies risk, as detailed in Section 5.4.

The analysis of Brazilian regulation reinforced this conclusion. TSE Resolution No. 23,732/2024 introduced specific rules for artificial intelligence and deepfakes in electoral propaganda, while Resolution No. 23,755/2026 expanded and updated these rules for the 2026 elections. Normative evolution demonstrates that the legal response can focus on transparency, context, purpose, effects, and risk, rather than relying on an exhaustive description of the technologies used.

It is therefore concluded that the expression deepfake should be used with greater conceptual precision. Technology is not, in itself, equivalent to harm; the risk emerges from the combination of technical capacity, content produced, identity or reality of reference, purpose, and context of use. This distinction is particularly important for public policies, regulation, governance of artificial intelligence, and preservation of trust in digital environments.

As an agenda for future research, it is recommended to deepen the empirical validation of the proposed matrix, examining real cases of synthetic media and deepfake and comparing its classification among technology, communication, and law specialists. It is also relevant to investigate mechanisms of provenance, marking, detection, and governance capable of following the evolution of generative technologies without relying exclusively on technological categories, which can quickly become obsolete.

Artificial Intelligence Usage Statement

Generative artificial intelligence tools were used to support the screening of the material recovered in the consulted databases, the translation of excerpts, the editorial review of the text, and the formatting according to ABNT standards. They were not used to generate arguments, conclusions, or bibliographic references. All sources cited were directly verified in the original documents, and the author assumes full responsibility for the content of the article.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- [1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**, 1097-1105.
- [2] He, K., Zhang, X., Ren, S. and Sun, J. (2015) Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 *IEEE International Conference on Computer Vision (ICCV)*, Santiago, 7-13 December 2015, 1026-1034. <https://doi.org/10.1109/iccv.2015.123>
- [3] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 27-30 June 2016, 770-778. <https://doi.org/10.1109/cvpr.2016.90>
- [4] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. <https://doi.org/10.1038/nature14539>
- [5] Goodfellow, I., et al. (2014) Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, **27**, 2672-2680.
- [6] de Seta, G. (2021) *Huanlian*, or Changing Faces: Deepfakes on Chinese Digital Media Platforms. *Convergence: The International Journal of Research into New Media Technologies*, **27**, 935-953. <https://doi.org/10.1177/13548565211030185>
- [7] Rauchfleisch, A., Vogler, D. and de Seta, G. (2025) Deepfakes or Synthetic Media? The Effect of Euphemisms for Labeling Technology on Risk and Benefit Perceptions.

- Social Media + Society*, **11**, 1-12. <https://doi.org/10.1177/20563051251350975>
- [8] Liu, Y., Padmanabhan, B. and Viswanathan, S. (2026) What Exactly Is a Deepfake? A Systematic Literature Analysis of Concepts, Risks, and Potentials. 2026 *IEEE Conference on Artificial Intelligence (CAI)*, Granada, 8-10 May 2026, 687-694. <https://doi.org/10.1109/cai68641.2026.11536577>
 - [9] Turing, A.M. (1950) I.—Computing Machinery and Intelligence. *Mind*, **LIX**, 433-460. <https://doi.org/10.1093/mind/lix.236.433>
 - [10] McCulloch, W.S. and Pitts, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, **5**, 115-133. <https://doi.org/10.1007/bf02478259>
 - [11] Mitchell, T.M. (1997) Machine Learning. McGraw-Hill.
 - [12] Fanaya, P.F. (2021) TECCOGS-Revista Digital de Tecnologias Cognitivas. *Digital Journal of Cognitive Technologies*, No. 23, 104-118.
 - [13] Santaella, L. (2021) The Conjoined Sisters Fake News and Post-Truth Expanded in Deepfakes. *Digital Journal of Cognitive Technologies*, No. 23, 15-24.
 - [14] Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P. and Awad, G. (2024) Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4. National Institute of Standards and Technology.
 - [15] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. <https://doi.org/10.1109/5.726791>
 - [16] ANPD (2026) Radar Tecnológico N° 6—Deepfakes. ANPD. https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/rt_deepfakes_anpd.pdf
 - [17] Irfan, M.T., Arora, B., Sandotra, N. and Raza, A.A. (2025) On Machine Learning and Deep Learning Based Deepfake Generation and Detection. *Procedia Computer Science*, **259**, 1927-1936. <https://doi.org/10.1016/j.procs.2025.04.148>
 - [18] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., *et al.* (2026) Deepfake Generation and Detection: A Benchmark and Survey. *ACM Computing Surveys*. <https://arxiv.org/abs/2403.17881>
 - [19] Brazil (2014) Law No. 12,965, of April 23, 2014. Establishes Principles, Guarantees, Rights and Duties for the Use of the Internet in Brazil. Civil Rights Framework for the Internet. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm
 - [20] Superior Electoral Court (TSE) (2019) Resolution No. 23,610, of December 18, 2019, with Subsequent Amendments. Electoral Propaganda. <https://www.tse.jus.br/legislacao/compilada/res/2019/resolucao-no-23-610-de-18-de-dezembro-de-2019>
 - [21] Superior Electoral Court (TSE) (2024) Resolution No. 23,732, of 2024. Amends Resolution No. 23,610/2019 and Regulates, among Other Points, the Use of Artificial Intelligence in Electoral Propaganda. <https://www.tse.jus.br/legislacao/compilada/res/2024/resolucao-no-23-732-de-27-de-fevereiro-de-2024>
 - [22] Superior Electoral Court (TSE) (2026) Resolution No. 23,755, of March 2, 2026. Amends Resolution No. 23,610/2019, which Provides for Electoral Propaganda. <https://www.tse.jus.br/legislacao/compilada/res/2026/resolucao-no-23-755-de-2-de-marco-de-2026>