Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.7969

Research Paper

From Deepfake Detection Research to Public Policy: A Survey on Generalization Challenges in Synthetic Media for the Brazilian Electoral Context

Thauan de Souza Tavares da Silva [Federal University of Paraná | *tstsilva@inf.ufpr.br*]
Diego Addan Gonçalves [Federal University of Paraná | *diego@inf.ufpr.br*]
David Menotti [Federal University of Paraná | *menotti@inf.ufpr.br*]

Department of Informatics, Universidade Federal do Paraná, Centro Politécnico, Curitiba, PR, 80240-021, Brazil.

Abstract. Deepfake detection has advanced rapidly in recent years, yet its practical effectiveness remains limited when models are exposed to real-world conditions that differ from controlled benchmarks. This gap is particularly critical in high-stakes scenarios such as electoral processes, where synthetic media can influence public perception at scale. This paper investigates the generalization limitations of current deepfake detection approaches, with emphasis on their applicability in the Brazilian context. Rather than focusing solely on algorithmic performance, we frame deepfake detection as a socio-technical problem, where issues of trust, interpretability, and user response are central to system effectiveness. Through a critical analysis of existing datasets, evaluation protocols, and detection strategies, we show that current research paradigms often fail to account for distribution shifts, cultural variability, and adversarial adaptation. By bridging detection research with policy-oriented perspectives, this work contributes to a more actionable understanding of how synthetic media challenges can be addressed in Brazil and similar socio-political contexts. Finally, we explicitly derive implications for public policy in the Brazilian electoral context, connecting observed technical limitations to challenges in regulation, platform governance, and the use of automated systems in institutional decision-making.

Keywords: Public Policy, Digital Governance, Deep-Fake

Edited by: Vagner de Santana | **Received:** 31 March 2026 • **Accepted:** 03 August 2026 • **Published:** 21 August 2026

1 Introduction

In recent decades, Artificial Intelligence (AI) has evolved from a distant concept into an everyday tool. The rise of generative models, such as Generative Adversarial Networks (GANs) proposed by Goodfellow *et al*. [2014] and Diffusion Models introduced by Ho *et al*. [2020], has inaugurated a new era in the creation of synthetic digital media. More recently, systems such as the Sora model [Brooks *et al*., 2024] have achieved unprecedented levels of realism, making it increasingly difficult for humans to distinguish between real and artificially generated content.

This rapid technological progress introduces challenges beyond the computational domain, including privacy, public trust, and democratic integrity. The volume of deepfake videos has also grown rapidly, with evidence suggesting exponential increase over recent years [Tolosana *et al*., 2020; Mirsky and Lee, 2021]. These trends highlight the urgency of the problem, particularly in high-stakes contexts such as elections.

Beyond technological advances, deepfakes pose a direct threat to democratic processes, leveraging the "economy of emotions" [Bakir and McStay, 2018]. Synthetic faces can even appear more trustworthy than real ones [Nightingale and Farid, 2022]. Recent cases show how deepfakes can simulate speeches and amplify disinformation at scale [Mirsky and Lee, 2021]. Despite growing responses, regulatory and technical efforts remain misaligned, creating enforcement gaps on digital platforms. Effective mitigation requires integrating detection with governance and platform-level enforcement.

Given this scenario, this survey seeks to address the following central research question: how has the evolution of generative architectures impacted the generalization capability of deepfake detection methods, and what specific gaps hinder the mitigation of these risks in the Brazilian electoral context?

Beyond the technical challenges of detection accuracy and generalization, this work frames deepfakes as a socio-technical problem involving user perception and interaction. Prior studies indicate that individuals often fail to reliably distinguish synthetic from real content and may not consistently verify suspicious media, particularly in high-volume information environments [Nightingale and Farid, 2022; Vaccari and Chadwick, 2020]. This behavioral dimension directly impacts the effectiveness of detection systems, as their outputs depend on user trust, interpretation, and integration into platform workflows.

In this context, this paper aims to bridge the gap between deepfake detection research and public policy, particularly in the Brazilian electoral scenario. This gap is further reinforced by operational limitations identified in prior studies, which show that even when detection models achieve high performance in controlled settings, their deployment in real-world environments is constrained by scalability, adversarial adaptation, and lack of integration with institutional workflows [Dolhansky *et al*., 2020]. These limitations highlight a broader disconnect between algorithmic advances and their practical adoption in governance and monitoring systems. By analyzing the limitations of current datasets and models, we derive evidence-based insights that can inform policy-making, regulatory frameworks, and the design of socio-technical infrastructures for combating synthetic media misuse.

The main contributions of this work are as follows: (i) a systematic and policy-oriented analysis of deepfake detection methods, with emphasis on generalization limitations; (ii) an examination of the Brazilian context, highlighting challenges related to data availability, institutional capacity, and socio-political risks; (iii) the identification of critical gaps between academic benchmarks and real-world deployment scenarios; and (iv) a set of evidence-based recommendations aimed at bridging the gap between HCI research and public policy in the context of synthetic media; (v) a discussion of the implications of deepfake detection limitations for public policy and electoral regulation. In addition, this work adopts a policy-oriented perspective, aiming to translate technical findings into evidence that can support decision-making in digital governance and electoral regulation. This positioning aligns the paper with ongoing efforts in HCI to bridge the gap between research and public policy. This perspective is consistent with the challenges in Ethics and Responsibility and Human-Data Interaction [Pereira *et al*., 2024].

This work identifies concrete pathways through which the proposed approach can inform regulatory practices, public digital infrastructure design, and evidence-based policymaking processes. In doing so, we position this research not only as a technical contribution, but as an actionable bridge between Human-Computer Interaction research and public policy formulation in Latin America. From a governance perspective, effective responses to synthetic media require alignment between detection systems and institutional processes. However, prior work highlights that content moderation and platform governance are often reactive, fragmented, and not systematically integrated with technical detection pipelines [Gillespie, 2018; Gorwa *et al*., 2020]. This creates operational gaps in detection, attribution, and enforcement, particularly in large-scale platforms where decisions must be made under uncertainty and time constraints.

2 Methodology of the Review

This section presents the methodological procedures adopted to conduct this survey. It details the search strategies used in scientific databases, as well as the criteria for study selection and the overall workflow.

2.1 Search Strategy

The search strategy was designed to map the state of the art in deepfake generation and detection, with emphasis on generalization, in-the-wild data, and electoral implications. In this work, we adopt the term “deepfake” as a specific subset of synthetic media, referring to AI-generated or AI-manipulated audiovisual content produced using deep learning techniques. While the broader term “synthetic media” encompasses a wider range of generative approaches, we use “deepfake” consistently to denote realistic human-centered manipulations. Queries were conducted across the following academic databases and scientific repositories:

- IEEE
- ACM
- Google Scholar
- SOL (Brazilian Computer Society’s OpenLib)
- arXiv

The searches were carried out between January and February 2026. Combinations of keywords in English were used, considering that most relevant literature is published in this language. The main search terms included:

- "deepfake detection";
- "in-the-wild deepfakes";
- "deepfake generalization";
- "deepfake benchmark";
- "Vision Transformer deepfake";
- "deepfake elections";
- "synthetic media political disinformation";
- "deepfake dataset";
- "synthetic media";
- "AI-generated videos".

To ensure reproducibility, we defined the following Boolean query adapted to each database:

("deepfake" OR "synthetic media" OR "AI-generated media") AND (detection OR generation OR generalization OR benchmark OR elections OR disinformation OR dataset)

The query was applied to titles, abstracts, and keywords. A backward snowballing procedure was also conducted using highly cited papers. Additionally, materials from Brazilian electoral authorities and fact-checking organizations (e.g., reports from the Superior Electoral Court, public fact-checking archives from Agência Lupa and Aos Fatos) were obtained through manual search of official websites and public repositories. These materials were used solely for contextualization of the Brazilian scenario (Section 6) and were not treated as scientific evidence, in accordance with the exclusion criteria. No journalistic materials were retrieved from the academic databases (IEEE, ACM, Google Scholar, arXiv), which were reserved exclusively for peer-reviewed scientific sources.

2.2 Inclusion and Exclusion Criteria

We considered studies published primarily from 2020 onwards, a period that marks the consolidation of diffusion-based models and the expansion of Transformer-based approaches in computer vision. Earlier works were included only when deemed essential for understanding the historical development of the field.

Inclusion Criteria

- Studies presenting technical methods for deepfake generation or detection;
- Works introducing or evaluating relevant benchmarks;
- Studies investigating generalization or robustness issues;
- Research discussing social, electoral, or regulatory impacts of deepfakes;
- Publications in recognized journals, conferences, or scientific repositories.

The first four were combined using the OR operator, meaning that a study meeting any one of them was considered eligible. The fifth criterion was combined with the preceding criteria using the AND operator and applied to all candidates.

Exclusion Criteria

- Purely opinion-based articles without empirical grounding;

- Studies lacking a clear methodological description;
- Duplicate publications or preliminary versions superseded by revised works;
- Non-scientific journalistic materials (except when used for contextualization of the Brazilian scenario).

After applying these criteria, the selected works were thematically organized for comparative analysis, as described below.

2.3 Selection Process

The study selection process followed a systematic approach based on the previously defined inclusion and exclusion criteria. Initially, more than 80 studies were identified from searches across IEEE, ACM, Google Scholar, and arXiv. After removing duplicates and applying filters based on relevance, scope, and methodological quality, a final set of 36 sources was obtained, including scientific articles and contextual materials (such as reports and news relevant to the Brazilian scenario).

The filtering process followed three successive stages. First stage (title/abstract screening): works on audio synthesis, text generation, or unrelated tasks were eliminated, removing approximately 40% of candidates. Second stage (full-text reading): we applied three criteria: (i) reproducible experimental protocol; (ii) direct engagement with deepfake generation/detection; (iii) novelty not subsumed by a later version. Studies failing any criterion were excluded. Third stage (thematic contribution): remaining works contributed to at least one of four axes: (i) generative models (GANs, diffusion); (ii) datasets; (iii) detection architectures (CNNs, Vision Transformers, hybrids); (iv) social, political, and psychological impacts. Selected works were organized into these axes to identify trends and generalization gaps.

3 Generative Models and the Evolution of Realism

The evolution of generative models over the past few years has profoundly transformed the way synthetic media is consumed, making it accessible to a large portion of the population. A fundamental milestone occurred with the introduction of Generative Adversarial Networks (GANs) by Goodfellow *et al*. [2014], which established the concept of adversarial learning between a generator and a discriminator. This architecture enabled, for the first time, the creation of facial images with high enough quality to deceive human observers, giving rise to the phenomenon known as *deepfakes*.

In the last decade, these advances were intensified by the transition from simpler statistical models to highly sophisticated artificial intelligence architectures capable of creating visual content with a high degree of realism. More than just a technical leap, synthetic media has come to represent a significant shift in how images are produced, consumed, and verified. In this context, this chapter presents the evolution of synthetically generated images, discusses the datasets used in the field, and analyzes the main detection methods, with a focus on recent developments.

3.1 GANs and the First Generation of Deepfakes

Based on the foundational GAN architecture introduced in Section 3, this architecture established an adversarial learning paradigm in which two neural networks, the Generator (G) and the Discriminator (D), are trained simultaneously in a competitive game: while the generator seeks to approximate the distribution of real data, the discriminator attempts to recognize authentic samples and flag synthetic ones. Subsequent works, such as DCGAN [Radford *et al*., 2016], demonstrated the ability of GANs to learn high-quality hierarchical visual representations.

Based on this paradigm, the first generation of *deepfakes* emerged, largely grounded in techniques such as *face swapping*, autoencoders, and adversarial models [Thies *et al*., 2016; Korshunova *et al*., 2017]. These methods allowed for the mapping of facial features from a source individual to a target, enabling identity manipulation in videos and images in an unprecedented manner.

Despite promising results reported in the literature, most deepfake detection models are evaluated under controlled conditions that do not reflect the variability of real-world data [Arjovsky *et al*., 2017; Salimans *et al*., 2016]. Training and testing datasets often share similar generation pipelines, compression artifacts, and semantic distributions, leading to an overestimation of model robustness [Li *et al*., 2020; Afchar *et al*., 2018]. Recent studies indicate that deepfake detection models exhibit a measurable degradation in generalization when evaluated across datasets. For instance, models trained on FaceForensics++ often experience performance drops exceeding 20 percentage points when tested on more challenging datasets such as Celeb-DF, highlighting the impact of dataset bias and limited representativeness [Rossler *et al*., 2019]. This discrepancy reinforces the need for more diverse and ecologically valid benchmarks.

Detection approaches typically exploit characteristic imperfections that generative models introduce into synthetic content. These imperfections can be grouped into three main categories, each corresponding to a class of artifact that detection methods have learned to identify: **Geometric inconsistencies**: failures in preserving the three-dimensional structure of the face, especially during rotations or pose changes, resulting in visible misalignments (the "mask" effect) [Thies *et al*., 2016]; **Spectral artifacts**: anomalous patterns introduced during the up-sampling process, such as so-called *checkerboard artifacts*, detectable in the frequency domain [Odena *et al*., 2016]; **Physiological inconsistencies**: the inability to reproduce natural biological dynamics, such as realistic blinking patterns and lip-syncing [Li and Lyu, 2019]. Despite these limitations, subsequent advances led to a significant increase in the quality of synthetic images, progressively reducing visible artifacts and making detection based solely on perceptual cues increasingly challenging.

3.2 The StyleGAN Era

Between 2018 and 2021, the StyleGAN architecture, developed by researchers at NVIDIA, represented a significant breakthrough in photorealistic image synthesis [Karras *et al*., 2021b]. This progress was built upon foundations established by previous architectures, such as DCGAN [Radford *et al*.,

2016], which demonstrated the viability of using deep convolutional networks for image generation. StyleGAN introduced a fundamental reformulation of the latent space by projecting the input vector (Z) into an intermediate space (W), allowing for better separation between high-level attributes (such as identity and facial structure) and low-level features (such as texture, lighting, and color). This decomposition enabled unprecedented control over the generation process, facilitating more interpretable and stable semantic manipulations. Subsequent advances with StyleGAN2 [Karras *et al*., 2020] mitigated characteristic artifacts of previous versions, such as "water-droplet" patterns, through improvements in normalization and convolutional architecture. Later, StyleGAN3 [Karras *et al*., 2021a] introduced mechanisms to address aliasing issues, promoting greater spatial consistency and invariance to transformations such as translation and rotation. As a result, models in this family reached a level of realism where synthetic images became virtually indistinguishable from real photographs for casual human observers, significantly raising the stakes for detection methods based solely on visual artifacts.

3.3 Diffusion Models and the New Generation of Realism

Starting in 2020, a new paradigm began to eclipse GANs: Denoising Diffusion Probabilistic Models (DDPM), introduced by Ho *et al*. [2020]. Unlike GANs, which learn data distribution through an adversarial process, diffusion models are trained to reverse a gradual process of stochastic data degradation.

The diffusion mechanism consists of two main stages. In the *forward* process, Gaussian noise is progressively added to an image until it approaches white noise. In the *reverse* process, a neural network is trained to predict and remove this noise in multiple steps, reconstructing the original image from a random distribution [Sohl-Dickstein *et al*., 2015; Ho *et al*., 2020].

This approach offers significant advantages over GANs, including greater training stability and better data distribution coverage, mitigating issues such as *mode collapse* [Dhariwal and Nichol, 2021]. Consequently, diffusion-based models have begun to produce images with higher perceptual fidelity, characterized by:

- Better global lighting consistency;
- More natural textures;
- Reduced presence of high-frequency artifacts;
- Greater fidelity in facial details.

Subsequent advances, such as guided diffusion models [Dhariwal and Nichol, 2021] and latent models [Rombach *et al*., 2022], enabled the efficient generation of high-resolution images with reduced computational costs. In practice, modern systems such as DALL-E, Stable Diffusion, and Midjourney demonstrate an unprecedented visual synthesis capability, making synthetic images increasingly indistinguishable from real content.

This evolution also directly impacts the field of deepfake detection, drastically reducing the effectiveness of methods based exclusively on visual artifacts or spectral inconsistencies.

3.3.1 Recent Advances in Multimodal Diffusion

More recently, diffusion architectures have been extended to multimodal and temporal generation scenarios. Models like Stable Diffusion incorporate efficient textual conditioning mechanisms [Rombach *et al*., 2022], while more recent advances explore the use of transformers to jointly model visual and linguistic representations.

In this context, recent works indicate a convergence between diffusion models and transformer-based architectures, allowing for greater semantic adherence to the prompt and better structural consistency. Additionally, techniques such as terminal noise optimization and refinements in the sampling process have significantly reduced the number of steps required for generation, increasing computational efficiency.

Finally, the evolution toward video generation models, as discussed in Brooks *et al*. [2024], suggests a transition from purely generative models to systems capable of simulating real-world dynamics, further broadening the challenges for the detection and verification of synthetic content.

3.4 Video Generation and Spatio-Temporal Coherence

Recent advances in generative models have increasingly focused on video synthesis, where the challenge extends beyond spatial realism to include temporal consistency and physical dynamics. Unlike static image generation, video models must preserve coherence across consecutive frames, including motion, lighting, and object identity over time.

3.4.1 Sora: Video Models as World Simulators

The introduction of the Sora model by OpenAI in 2024 marked a fundamental shift in how artificial intelligence models real-world dynamics [Brooks *et al*., 2024]. Sora is a diffusion-based model that operates over spatio-temporal representations, treating videos as structured sequences within a compressed latent space.

Its emerging capabilities suggest behavior resembling a "world simulator", including:

- **Object persistence**: elements that leave and re-enter the scene maintain consistent identity and characteristics over time.
- **Plausible physical interactions**: simulation of effects such as marks, deformations, and persistent changes in surfaces.
- **Complex camera motion**: the ability to perform dynamic transitions while preserving the geometric integrity of the scene.

Despite these advances, limitations remain in modeling complex physical phenomena, such as non-linear material interactions (e.g., fluids or fractures), indicating that these systems do not yet constitute fully accurate physical simulators.

3.4.2 Alternative Architectures and Recent Advances

In parallel, alternative approaches have been proposed to explicitly address spatio-temporal coherence. The Lumiere model, developed by Google Research, employs a *space-time U-Net* architecture that generates videos holistically, avoiding

the need for frame-by-frame reconstruction [Bar-Tal *et al*., 2024]. This strategy significantly improves global motion consistency from the early stages of generation.

Additionally, recent commercial models, such as those developed by Runway [Runway Research, 2023], explore transformer and diffusion-based architectures for video generation with refined semantic control, including text-guided editing and motion manipulation.

Overall, this new generation of models demonstrates substantial improvements in:

- Temporal continuity across frames
- Global coherence of lighting and texture
- Lip synchronization and facial expressions

This evolution addresses one of the main weaknesses of early deepfake systems: temporal inconsistencies. As a result, detection approaches based on abrupt frame-to-frame variations are becoming increasingly ineffective, requiring the development of more sophisticated methods capable of capturing subtle inconsistencies at semantic and dynamic levels. A synthesis of this technological evolution and its impact on detection can be seen in Table 1.

Table 1. Evolution of Generative Models and Impact on Detection.

Generation	Architecture	Key Advance	Impact on Detection
1st	GANs	Initial facial realism	Exploitable artifacts
2nd	StyleGAN	High fidelity and control	Reduction of visible artifacts
3rd	Diffusion	Realistic texture and lighting	Challenges spectral analysis
4th	Video	Spatiotemporal coherence	Reduces inconsistencies

4 Evolution of Datasets

The development of effective methods for deepfake detection directly depends on the quality and representativeness of the datasets used for training and evaluation. In this context, the evolution of datasets directly tracks the advances in generative models, reflecting a tendency toward greater representativeness of real-world scenarios. This relationship holds because datasets built around newer generative architectures tend to include manipulations that are closer in style, quality, and compression profile to what actually circulates on social media platforms. However, this correspondence is not automatic: updating the generative model used to produce a dataset does not in itself guarantee that the dataset captures the distribution of content encountered in the wild, as real-world dissemination involves additional factors such as re-encoding, platform compression, and organic capture conditions, as the example discussed in section 4.2 [Rossler *et al*., 2019; Li and Lyu, 2019].

4.1 The First Wave: FaceForensics++ and Celeb-DF

Datasets introduced between 2018 and 2020, such as FaceForensics++ (FF++) [Rossler *et al*., 2019] and Celeb-DF [Li *et al*., 2020], were fundamental to the initial progress of deepfake detection. These datasets were primarily designed to capture manipulations based on classic *face swapping* and reenactment techniques.

FaceForensics++, for instance, provides manipulated videos through multiple generation pipelines, allowing models to learn specific artifact patterns introduced by each method. In contrast, Celeb-DF sought to increase the realism of forgeries by reducing evident visual artifacts and approaching scenarios that are more challenging for detection.

Despite their relevance, these datasets present significant limitations. In general, they lack diversity in terms of ethnicity, lighting conditions, and backgrounds, in addition to using lower resolutions than those found in real-world applications. Consequently, models trained exclusively on these data tend to overestimate their performance in controlled environments but show a significant drop in real-world scenarios [Li *et al*., 2020]. This overestimation is corroborated by systematic reviews indicating that while state-of-the-art algorithms can report accuracy rates exceeding 99% on controlled academic platforms (e.g., Kaggle), they are rarely validated against real-time, dynamic data extracted directly from social media, where the bulk of disinformation actually circulates [Villela *et al*., 2023]

4.2 The Transition to Diffusion and Realistic Scenarios

With the advent of diffusion models, characteristic GAN, artifacts have become less apparent. In response, a new generation of datasets has shifted focus toward more realistic content aligned with the state-of-the-art in generators.

The VideoDi dataset [Battocchio *et al*., 2025], for example, is specifically designed for videos generated by diffusion models and *text-to-video* systems. One of its main contributions is the simulation of the *media laundering* phenomenon, in which content undergoes multiple stages of compression (e.g., H.264), reproducing the dissemination cycle on social media and digital platforms.

In parallel, recent benchmarks such as Deepfake-Eval-2024 [Chandra *et al*., 2026] aim to capture the complexity of the *in-the-wild* environment. Collected from actual data on online platforms, this benchmark includes dozens of hours of multimodal content (video and audio), multiple languages, and a wide variety of generation techniques.

Experimental results in these scenarios reveal a significant performance drop in detectors trained exclusively on traditional academic datasets. This phenomenon, widely discussed in the literature as the generalization problem, highlights the gap between controlled research environments and the actual landscape of digital disinformation [Wang *et al*., 2020; Coccomini *et al*., 2022].

Thus, the evolution of datasets not only follows the progress of generative models but also continuously redefines the challenges faced by detection systems, demanding more robust, generalizable, and adaptive approaches.

5 Detection Architectures

The detection of synthetic media has evolved from approaches based on explicit visual inconsistencies to models capable of capturing global semantic relationships and complex temporal dynamics. This advancement directly reflects the sophistication of generative models, especially with the transition from GANs to diffusion and video generation models.

5.1 Artifact-based Methods and CNNs

Early detection approaches exploited natural limitations of initial generative models, such as physiological inconsisten-

cies and visual artifacts. A classic example is the work of Li and Lyu [2019], which identified anomalous blinking patterns in synthetic videos. Other strategies included frequency analysis, lighting inconsistencies, and failures in facial edges.

Given the limitations of these heuristic approaches, the field evolved toward supervised deep learning methods. Convolutional models, such as MesoNet [Afchar *et al*., 2018] and deeper architectures like XceptionNet [Chollet, 2017], began to automatically learn discriminative patterns between real and synthetic content.

Despite showing high accuracy in controlled benchmarks, subsequent studies demonstrated significant structural limitations, such as dataset overfitting and poor generalization [Wang *et al*., 2020]. Furthermore, these models are highly sensitive to compression and re-encoding, which compromises their performance in real-world scenarios.

5.2 Vision Transformers and Global Modeling

The introduction of Vision Transformers (ViT) by Dosovitskiy *et al*. [2021] marked a significant shift in the detection paradigm. Unlike CNNs, which operate with local receptive fields, Transformers model global relationships through self-attention mechanisms.

In practice, this allows models to detect semantic inconsistencies distributed across the image—such as incoherence between lighting, reflections, and facial geometry—aspects that are frequently ignored by local convolutional filters.

However, ViTs also present challenges, including high computational costs and a dependency on large volumes of data for efficient training.

5.3 Hybrid Architectures and State of the Art

To mitigate the limitations of both CNNs and Transformers in isolation, hybrid approaches have emerged as the state of the art. Models such as GenConViT [Deressa *et al*., 2025] combine convolutions (ConvNeXt) with global attention mechanisms, allowing for the simultaneous capture of local artifacts and long-range structural inconsistencies.

Recent studies also explore Vision Transformer architectures combined with EfficientNet and Swin Transformers, demonstrating greater robustness in *in-the-wild* scenarios and under compression [Coccomini *et al*., 2022].

Results reported in modern benchmarks, such as Deepfake-Eval-2024 [Chandra *et al*., 2026], indicate that hybrid architectures can achieve accuracies exceeding 95% in controlled settings. However, when evaluated on highly diverse real-world data, the same models show a significant performance drop, decreasing from over 95% to values between 53% and 68% in cross-dataset and in-the-wild evaluations [Wang *et al*., 2020].

5.4 Video Detection and Temporal Embeddings

With the advancement of video generation models, detection has come to require explicit modeling of the temporal dimension. Recent approaches utilize embeddings extracted by Vision Transformers across frame sequences, combined with temporal attention mechanisms.

In this context, frameworks such as Video-ViT and architectures based on Q-Former [Battocchio *et al*., 2025] propose the integration of spatial and temporal information through learnable query tokens. This mechanism allows for the condensation of multiple frames into compact representations, capturing dynamic patterns and motion inconsistencies.

These approaches are particularly effective against state-of-the-art models, such as those based on spatiotemporal diffusion, as they analyze not only the static appearance of frames but also their temporal coherence.

Despite these advances, deepfake detection remains an open problem, especially in real-world scenarios where factors such as compression, device diversity, and hybrid manipulations impose additional challenges on current models.

6 Challenges and Gaps in the Brazilian Context

Despite significant advances in deepfake generation and detection, the Brazilian context presents a set of structural challenges that hinder the effective mitigation of synthetic media risks. These challenges emerge from the intersection of technical limitations, data scarcity, institutional constraints, and socio-technical factors that shape how such technologies are produced, disseminated, and interpreted in real-world scenarios.

This scenario aligns with recent research on the sociotechnical dimensions of misinformation, which highlights that high exposure to digital platforms, combined with limited media literacy and fragmented institutional trust, significantly increases societal vulnerability to manipulated content [Puska *et al*., 2024; Shoaib *et al*., 2023; Al-Naeem *et al*., 2025]. In particular, deepfake technologies have been shown to amplify existing weaknesses in information ecosystems, affecting not only individual perception but also collective trust in media and institutions. In this context, the effectiveness of deepfake mitigation strategies depends not only on detection accuracy, but also on trustworthy external sources and the integration of detection systems within broader information verification frameworks [Sagar *et al*., 2026].

6.1 Data Scarcity and Generalization Limitations

One of the most critical limitations in the Brazilian context is the absence of large-scale, representative datasets for deepfake detection. The literature consistently highlights that the lack of high-quality datasets is a major bottleneck for the development of robust detection systems [Li *et al*., 2020; Rossler *et al*., 2019].

This issue is even more pronounced in low-resource languages such as Portuguese, where the availability of annotated data remains limited [Chandra *et al*., 2026]. Existing initiatives in Brazil, such as FakeRecogna [Garcia *et al*., 2022] and datasets derived from fact-checked content shared on social media platforms [Reis *et al*., 2020], represent important steps toward addressing misinformation. However, these efforts are still primarily focused on textual or image-based fake news, rather than multimodal deepfake scenarios.

As a result, datasets specifically designed for deepfake detection in the Brazilian context remain limited in scale,

modality, and diversity. Current datasets often focus on specific domains or contain a restricted number of samples, making them insufficient for training models capable of handling real-world variability.

Moreover, most widely used datasets in deepfake research are developed in controlled environments, with limited demographic diversity and artificial generation pipelines [Rossler *et al*., 2019]. This creates a significant domain mismatch when models are deployed in real-world Brazilian scenarios, characterized by diverse phenotypic traits, varying recording conditions, and strong compression artifacts due to social media platforms.

These data limitations do not affect only the development of new models, they directly undermine the generalization capability of existing ones. Several studies show that models achieving high accuracy in benchmark datasets experience significant performance degradation when evaluated on unseen data [Wang *et al*., 2020; Chandra *et al*., 2026].

This suggests that the generalization gap should not be understood solely as a limitation of model architectures, but rather as a structural issue rooted in data distribution. In practice, detection systems are often optimized for specific datasets rather than for real-world deployment conditions.

Additionally, biases present in existing datasets, including limited demographic diversity, can lead to unequal performance across different population groups, raising concerns about fairness and reliability in sensitive applications [Xu *et al*., 2024]. Recent demographic analyses of deepfake detectors confirm this critical vulnerability, revealing that unbalanced training data causes models to systematically misclassify authentic images of specific marginalized demographics, such as young black individuals, as synthetic fakes [Cunha *et al*., 2024]. In a highly diverse country like Brazil, deploying algorithms that carry such pronounced intersectional biases could inadvertently penalize legitimate users and severely undermine the credibility of automated moderation systems.

6.2 Operational Regulatory Gap and Institutional Response

Beyond technical challenges, the Brazilian context reveals a critical gap between regulatory discussions and practical implementation. Two legislative initiatives are central to this debate. PL 2630/2020, known as the "Fake News Bill," proposes obligations for platforms to moderate content, maintain user traceability, and cooperate with electoral authorities. However, its provisions directly conflict with end-to-end encryption protections on messaging applications, a tension that affects the monitoring of deepfake dissemination on platforms such as Telegram [Benevenuto and Melo, 2024], where much of Brazil's electoral disinformation circulates. PL 2338/2023, Brazil's draft AI Act, establishes a risk-based classification for AI systems and explicitly categorizes electoral manipulation as a high-risk use case. However, it lacks specific operational requirements for deepfake detection, leaving a gap between regulatory intent and the technical infrastructure needed for enforcement. Both bills share the challenge of balancing freedom of expression and innovation with the need to safeguard democratic processes, a trade-off that, in the specific case of deepfakes, results in a framework broad enough to protect fundamental rights, but insufficiently specific to mandate or fund detection capacity at the institutional level.

In the electoral context, the Brazilian Superior Electoral Court (TSE) has introduced measures such as mandatory labeling of AI-generated content. However, the effectiveness of these initiatives is limited by technological constraints and the presence of malicious actors intentionally exploiting generative AI systems.

In practice, there are no standardized operational protocols addressing key questions such as: who is responsible for verifying suspected deepfakes, which tools should be used, how results should be validated, and what actions should follow detection. This institutional vacuum becomes especially critical in light of the performance figures presented in Section 5.3: even state-of-the-art models achieve AUC values as low as 0.53 on real-world benchmarks, meaning that any formal detection protocol relying solely on automated tools would carry an unacceptably high error rate in electoral contexts where misclassification has direct democratic consequences [Gillespie, 2018; Gorwa *et al*., 2020]. This absence of clear procedures creates a disconnect between technological capabilities and their real-world application.

Furthermore, state-of-the-art detection methods, including those based on diffusion models and temporal consistency [Battocchio *et al*., 2025], remain largely confined to academic environments. The lack of robust and accessible tools for institutional actors, such as the judiciary system, prevents timely and effective responses to large-scale malicious uses of generative AI.

6.3 Implications for Electoral Integrity

These structural limitations become particularly critical in the context of elections. The Brazilian information ecosystem, strongly influenced by messaging platforms such as WhatsApp and Telegram, has been shown to facilitate the rapid dissemination of misinformation at scale [Reis *et al*., 2020; Benevenuto and Melo, 2024].

To illustrate these risks, we generated a synthetic image depicting Brazilian presidential candidates in the 2022 election, in an informal social setting (e.g., a barbecue), a scenario that could plausibly be interpreted as real by the general public. This image was created using widely accessible generative AI tools (Gemini), requiring only a few user inputs and no technical expertise. The generated example and its evaluation by a state-of-the-art detection model are shown in Figure 1.

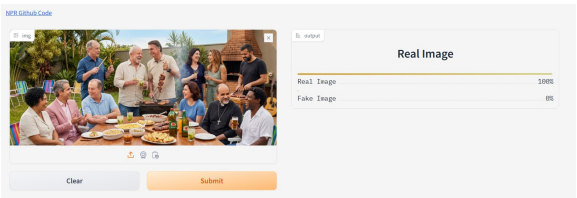

Figure 1. AI-generated image depicting Brazilian presidential candidates in a fabricated informal social scenario (created with Gemini). The image was evaluated using the NPR (Nearest Peak Rethinking) detector proposed by Tan *et al*. [2024]. Despite being entirely synthetic, the image was classified as authentic (100% real), illustrating a generalization failure under out-of-distribution content.

Despite being artificially generated, the image was classified as authentic by an online deepfake detection tool (NPR de-

tector [Tan *et al*., 2024], online demo), highlighting a critical failure in real-world detection systems. Furthermore, while models such as Tan *et al*. [2024] report high performance in controlled settings (AUC of 0.98 and precision of 0.95), their effectiveness drops significantly in more realistic benchmarks such as Deepfake-Eval-2024 [Chandra *et al*., 2026], where performance can decrease to AUC values around 0.53 and precision near 0.69.

This discrepancy reinforces the generalization gap discussed earlier and demonstrates how synthetic content, even when created without malicious intent, can be easily repurposed to produce misleading narratives or fake news. The combination of high platform penetration, low digital literacy, and the absence of institutional detection mechanisms creates a highly vulnerable environment for the spread of such content.

This example illustrates that the barrier to generating potentially misleading synthetic media is extremely low, often requiring only a few clicks in publicly available tools, while the mechanisms to reliably detect such content in real-world scenarios remain insufficient.

6.4 Digital Literacy and Institutional Trust as Socio-Technical Barriers

The Brazilian scenario thus represents a complex socio-technical challenge. The asymmetry between the increasing ease of generating highly realistic synthetic media and the difficulty of detecting such content in real-world, highly compressed environments reinforces deepfakes as a persistent threat.

A critical and widely observed dimension of this challenge is the digital literacy gap in the Brazilian population. Brazil exemplifies a structural paradox observed in several emerging economies: according to the TIC Domicílios 2024 survey, 80% of Brazilian Internet users access social media regularly, with 141 million people connected as of 2024, one of the highest rates in the world [Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), 2024]. Yet, according to the Inaf 2024 report, 29% of Brazilians aged 15 to 64 remain functionally illiterate, meaning they are unable to interpret simple texts or perform basic reasoning tasks despite being able to read and write at a rudimentary level [Ação Educativa and Conhecimento Social, 2025]. Critically, the same report finds that 95% of functionally illiterate individuals have very limited digital skills, creating a population segment that is both highly exposed to social media content and largely unable to critically evaluate it. This challenge is further exacerbated by well-documented human limitations in distinguishing synthetic from real visual content, which persist even among digitally experienced users [Nightingale and Farid, 2022]. As emphasized by the Brazilian HCI research agenda, this mismatch between technology access and functional literacy constitutes one of the central challenges for Human-Data Interaction in Brazil [Coleti *et al*., 2024].

Evidence from the Brazilian news consumption context suggests that content verification habits remain limited among the population. The Reuters Institute Digital News Report 2024 identifies Brazil among the countries with the most significant increases in selective news avoidance, and reports that 24% of users of platforms such as TikTok and X find it difficult to distinguish trustworthy from untrustworthy content [Newman *et al*., 2024]. This difficulty is compounded in the deepfake context, where manipulation operates at the visual and audiovisual level rather than the textual level, a dimension for which Brazilian users have even fewer established verification routines.

This institutional trust gap represents a critical bottleneck for the deployment of detection technologies in real-world scenarios. Recent studies indicate that deepfake proliferation contributes to the erosion of trust across multiple layers, including individual perception, media credibility, and institutional legitimacy [Peng and Zhao, 2025; Shoaib *et al*., 2023]. Even when technically robust detection mechanisms are available, their practical impact remains limited if users do not trust the entities responsible for verification. This reinforces the need for integrated approaches that combine technical robustness with institutional credibility and transparent communication mechanisms.

From a digital literacy perspective, the dimensions most relevant to deepfake resilience are: (i) *technical literacy*, awareness that synthetic media exists and a basic understanding of how it is produced and distributed; (ii) *critical media literacy*, understood as the disposition to question the authenticity of viral content before sharing or acting on it; and (iii) *institutional trust calibration*, the ability to identify and rely on credible verification sources, including public bodies and fact-checkers. Research indicates that AI-synthesized faces are perceived as more trustworthy than real ones [Nightingale and Farid, 2022], suggesting that even users with some technical awareness may remain susceptible to deepfake manipulation if critical media literacy and institutional trust are not jointly developed. Taken together, high social media exposure, persistent functional illiteracy, and limited verification habits place the Brazilian population at particular risk of synthetic media misuse in electoral settings.

Overcoming this challenge requires not only advances in detection architectures, but also the development of local data infrastructures and public policies focused on digital literacy, taking into account the technological and social realities of the country.

While deepfake detection is often approached as a computer vision task, its real-world impact is deeply rooted in Human-Computer Interaction (HCI). Users interacting with digital platforms are required to interpret, trust, and make decisions based on audiovisual content, often without awareness of manipulation risks.

From an HCI perspective, deepfakes introduce challenges related to trust calibration, cognitive bias, and information overload. Even highly accurate detection systems may fail in practice if their outputs are not interpretable or accessible to end-users, policymakers, or institutions. Furthermore, the design of interfaces for content verification, warning systems, and media provenance directly influences how effectively detection technologies can be utilized.

The results suggest that policies relying solely on automated deepfake detection may be inadequate in practice. Instead, regulatory strategies should consider hybrid approaches that combine automated tools with human oversight, as well as requirements for transparency in training data and system

limitations. These considerations are particularly relevant in the Brazilian electoral context, where misclassification can directly affect public trust and democratic processes. This motivates the design of human-centered systems that integrate technical detection with interaction mechanisms aimed at enhancing user understanding and trust calibration.

7 Conclusion

This survey presented a comprehensive analysis of the evolution of synthetic media, from early GAN-based approaches to recent advances in diffusion and video generation models, highlighting how this rapid progress has directly reshaped the challenges of deepfake detection. The findings indicate a clear trend: as generative models become increasingly realistic and semantically consistent, detection methods based on low-level visual artifacts progressively lose effectiveness, requiring more robust approaches capable of capturing global dependencies and spatio-temporal dynamics.

Our analysis reveals that the evolution of generative architectures has significantly widened the generalization gap in deepfake detection. Models that excel in controlled benchmarks consistently fail in real-world scenarios, particularly under compression, re-encoding, and heterogeneous data sources, exposing a structural disconnect between academic evaluation and practical deployment.

This survey reframes this gap as a central challenge, not a marginal limitation. We consolidate evidence that current detection approaches remain fragile under real-world variability, and argue for a paradigm shift in system development and evaluation. In Brazil, reliance on foreign datasets, limited phenotypic diversity, and the dominance of compressed content on messaging platforms intensify domain shift effects. This not only degrades performance but may also introduce demographic biases, undermining trust in critical applications like electoral processes and judicial analysis.

Beyond technical limitations, this work identifies a gap between academic advances and their practical adoption in Brazil. Although recent Transformer-based and hybrid detection methods achieve high performance in benchmarks, their integration into institutional workflows remains limited. In practice, electoral and judicial institutions still rely on manual verification, platform reporting, and fact-checking agencies rather than automated detection pipelines.

This gap is reinforced by the lack of operational protocols for handling synthetic media in high-stakes contexts such as elections. While initiatives exist, the adoption of AI-based forensic tools in real-time decision-making remains limited, creating an asymmetry where large-scale generation contrasts with slow, reactive verification. This reflects a broader misalignment between technological capabilities and institutional readiness, as effective deployment requires not only technical maturity but also governance frameworks and operational integration [Gorwa *et al*., 2020; Veale *et al*., 2018; Brundage *et al*., 2018].

Based on these findings, we argue that addressing deepfake detection requires moving beyond purely technical solutions toward an integrated socio-technical approach. In practical terms, we highlight the following actionable directions: (i) the development of national, in-the-wild datasets that reflect Brazilian demographic and media conditions; (ii) the design of detection models explicitly optimized for robustness and generalization under real-world distortions; (iii) the integration of detection tools into institutional workflows with accessible interfaces; and (iv) the establishment of public policies focused on digital literacy and effective regulation of synthetic media.

Beyond technical advancements, this work underscores the urgent need to translate deepfake detection research into actionable public policies. Addressing the generalization gap requires coordinated efforts between researchers, policymakers, and platform providers. By framing deepfakes as a socio-technical problem, this paper contributes to a broader understanding of how emerging technologies can inform evidence-based policy-making in Brazil and Latin America.

Finally, we emphasize that the current scenario does not represent a stable endpoint, but rather a point of inflection. The rapid evolution of generative models suggests that detection approaches relying exclusively on visual analysis may become insufficient in the near future. Therefore, future research should explore complementary strategies, such as media provenance tracking, authenticity verification mechanisms, and platform-level interventions, aiming to build a more resilient ecosystem against AI-driven misinformation. By explicitly connecting these findings to policy-relevant challenges, this work contributes to ongoing efforts to bridge the science–policy gap in Human–Computer Interaction.

Declarations

Acknowledgements

The authors would like to thank the Department of Informatics at Federal University of Paraná (UFPR) for all their support and the project SEI 23075.023372/2025-01.

Authors' Contributions

TT contributed to the conception of this study and is the main contributor and writer of this manuscript. DAG and DMG has added further information and reviwed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The complete list of reviewed works will be available in the datasets section at: `web.inf.ufpr.br/vri/databases` (last access: 21 August 2026).

Further relevant information

Artificial intelligence tools were used only for spell-checking and template formatting.

References

Ação Educativa and Conhecimento Social (2025). Functional literacy indicator (inaf) 2024: Literacy from the analog to the digital world. Technical report, Ação Educativa and Conhecimento Social. URL: https://fundacaoitau.org.br; Accessed: 11 August 2026.

Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Informa-*

tion Forensics and Security (WIFS), pages 1–7. IEEE. DOI: https://doi.org/10.1109/WIFS.2018.8630761.

Al-Naeem, M., Hafizur Rahman, M. M., Banerjee, A., and Sufian, A. (2025). Bsm-dnd: Bias and sensitivity-aware multilingual deepfake news detection using bloom filters and recurrent feature elimination. *IEEE Access*, 13:163218–163234. DOI: https://doi.org/10.1109/ACCESS.2025.3609831.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 214–223. JMLR.org. DOI: https://dl.acm.org/doi/10.5555/3305381.3305404.

Bakir, V. and McStay, A. (2018). Fake news and the economy of emotions. *Digital Journalism*, 6(2):154–175. DOI: https://doi.org/10.1080/21670811.2017.1345645.

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., Li, Y., Rubinstein, M., Michaeli, T., Wang, O., Sun, D., Dekel, T., and Mosseri, I. (2024). Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3680528.3687614.

Battocchio, J., Dell'Anna, S., Montibeller, A., and Boato, G. (2025). Advance fake video detection via vision transformers. In *Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSEC '25, page 1–11, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3733102.3733129.

Benevenuto, F. and Melo, P. (2024). Misinformation campaigns through whatsapp and telegram in presidential elections in brazil. *Commun. ACM*, 67(8):72–77. DOI: https://doi.org/10.1145/3653325.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Scholz, A., Janner, M., Chu, A., *et al.* (2024). Video generation models as world simulators. Technical report, OpenAI. https://openai.com/research/video-generation-models-as-world-simulators; Accessed: 11 August 2026.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. `https://arxiv.org/abs/1802.07228`. Accessed: 11 August 2026.

Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) (2024). Survey on the use of information and communication technologies in brazilian households: Ict households 2024. Technical report, Cetic.br / NIC.br. URL: https://cetic.br/media/analises; Accessed: 11 August 2026.

Chandra, N. A., Lee, H., Murtfeldt, R., Qiu, L., Karmakar, A., Tanumihardja, E., Farhat, K., Caffee, B., Lee, C., Choi, J., Paik, S., Kim, A., and Etzioni, O. (2026). Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. https://doi.org/10.48550/arXiv.2503.02857.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA. IEEE Computer Society. DOI: https://doi.org/10.1109/CVPR.2017.195.

Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection. In *Image Analysis and Processing – ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*, page 219–229, Berlin, Heidelberg. Springer-Verlag. DOI: https://doi.org/10.1007/978-3-031-06433-3_19.

Coleti, T. A., Divino, S. B. S., Salgado, A. d. L., Zacarias, R. O., Saraiva, J. d. A. G., Gonçalves, D. A., Morandini, M., and Santos, R. P. d. (2024). Grandihc-br 2025-2035 - gc5 - human-data interaction data literacy and usable privacy. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*. DOI: https://doi.org/10.1145/3702038.3702058.

Cunha, Y. M. B. G., Gomes, B. R., Boaro, J. M. C., Moraes, D. d. S., Busson, A. J. G., Duarte, J. C., and Colcher, S. (2024). Learning Self-distilled Features for Facial Deepfake Detection Using Visual Foundation Models: General Results and Demographic Analysis. *Journal on Interactive Systems*, 15(1):682–694. DOI: https://doi.org/10.5753/jis.2024.4120.

Deressa, D. W., Mareen, H., Lambert, P., Atnafu, S., Akhtar, Z., and Van Wallendael, G. (2025). Genconvit: Deepfake video detection using generative convolutional vision transformer. *Applied Sciences*, 15(12). DOI: https://doi.org/10.3390/app15126622.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.. DOI: https://dl.acm.org/doi/10.5555/3540261.3540933.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. https://doi.org/10.48550/arXiv.2006.07397.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. `https://arxiv.org/abs/2010.11929`. Accessed: 11 August 2026.

Garcia, G. L., Afonso, L. C. S., and Papa, J. P. (2022). Fakerecogna: A new brazilian corpus for fake news detection. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 57–67, Berlin, Heidelberg. Springer-Verlag. DOI: https://doi.org/10.1007/978-3-030-98305-5_6.

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions

that shape social media. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, pages 1–288. DOI: https://doi.org/10.12987/9780300235029.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press. DOI: https://dl.acm.org/doi/10.5555/2969033.2969125.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945. DOI: https://doi.org/10.1177/2053951719897945.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.. DOI: https://dl.acm.org/doi/abs/10.5555/3495724.3496298.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021a). Alias-free generative adversarial networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.. DOI: https://dl.acm.org/doi/10.5555/3540261.3540327.

Karras, T., Laine, S., and Aila, T. (2021b). A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228. DOI: https://doi.org/10.1109/TPAMI.2020.2970919.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, Los Alamitos, CA, USA. IEEE Computer Society. DOI: https://doi.org/10.1109/CVPR42600.2020.00813.

Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast face-swap using convolutional neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705. DOI: https://doi.org/10.1109/ICCV.2017.397.

Li, Y. and Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. https://doi.org/10.48550/arXiv.1811.00656.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213. DOI: https://doi.org/10.1109/CVPR42600.2020.00327.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1). DOI: https://doi.org/10.1145/3425780.

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., and Nielsen, R. K. (2024). Reuters institute digital news report 2024: Brazil. Technical report, Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk; Accessed: 11 August 2026.

Nightingale, S. J. and Farid, H. (2022). Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119. DOI: https://doi.org/10.1073/pnas.2120481119.

Odena, A., Dumoulin, V., and Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3. DOI: https://doi.org/10.23915/distill.00003.

Peng, N. and Zhao, T.-F. (2025). Unveiling risk evolution mechanisms driven by deepfake technologies. In *2025 12th International Conference on Machine Intelligence Theory and Applications (MiTA)*, pages 1–8. DOI: https://doi.org/10.1109/MiTA66017.2025.11100255.

Pereira, R., Darin, T., and Silveira, M. S. (2024). Grandihc-br: Grand research challenges in human-computer interaction in brazil for 2025-2035. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*. Association for Computing Machinery. DOI: https://dl.acm.org/doi/10.1145/3702038.3702061.

Puska, A. A., Baroni, L. A., and Pereira, R. (2024). Decoding the sociotechnical dimensions of digital misinformation: A comprehensive literature review. https://doi.org/10.48550/arXiv.2406.11853.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. `https://arxiv.org/abs/1511.06434`. Accessed: 11 August 2026.

Reis, J. C. S., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020). A dataset of fact-checked images shared on WhatsApp during the Brazilian and Indian elections. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 903–908. DOI: https://doi.org/10.1609/icwsm.v14i1.7356.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society. DOI: https://doi.org/10.1109/CVPR52688.2022.01042.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images . In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, Los Alamitos, CA, USA. IEEE Computer Society. DOI: https://doi.org/10.1109/ICCV.2019.00009.

Runway Research (2023). Runway gen-3: Text-to-video generation model. `https://runwayml.com`. Online; accessed 11 August 2026.

Sagar, A. S. M. S., Bennamoun, M., Boussaid, F., Sharif, N., Xu, L., Sahmoud, S., and Kishk, A. (2026). Fact or fake? assessing the role of deepfake detectors in multimodal misinformation detection. *IEEE Access*, 14:86992–87011. DOI: https://doi.org/10.1109/ACCESS.2026.3686437.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Proceedings of*

the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 2234–2242, Red Hook, NY, USA. Curran Associates Inc.. DOI: https://dl.acm.org/doi/abs/10.5555/3157096.3157346.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., and Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In *2023 International Conference on Computer and Applications (ICCA)*, pages 1–7. DOI: https://doi.org/10.1109/ICCA59364.2023.10401723.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 2256–2265. JMLR.org. DOI: https://dl.acm.org/doi/10.5555/3045118.3045358.

Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y. (2024). Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28130–28139. DOI: https://doi.org/10.1109/CVPR52733.2024.02657.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395. DOI: https://doi.org/10.1109/CVPR.2016.262.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148. DOI: https://doi.org/10.1016/j.inffus.2020.06.014.

Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1):2056305120903408. DOI: https://doi.org/10.1177/2056305120903408.

Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA. Association for Computing Machinery. DOI: https://dl.acm.org/doi/10.1145/3173574.3174014.

Villela, H. F., Corrêa, F., Ribeiro, J. S. d. A. N., Rabelo, A., and Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1):47–58. DOI: https://doi.org/10.5753/jis.2023.3020.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701. DOI: https://doi.org/10.1109/CVPR42600.2020.00872.

Xu, Y., Terhörst, P., Pedersen, M., and Raja, K. (2024). Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 5(1):93–106. DOI: https://doi.org/10.1109/TTS.2024.3365421.