RCMOS – Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year IV, v.1 2024. | Submission: 03/13/2024 | Accepted: 03/15/2024 | Publication: 03/17/2024

The Role of Computer Vision in Combating Digital Disinformation: Detection AUTOMATIC DEEPFAKES ON SOCIAL PLATFORMS

THE ROLE OF COMPUTER VISION IN COMBATING DIGITAL DISINFORMATION: AUTOMATIC DETECTION OF DEEPFAKES ON SOCIAL PLATFORMS

Author:

Matheus de Oliveira Pereira Paula

Bachelor's Degree in Information Systems — Federal Institute of Education, Science and Technology Fluminense

Master's degree: MSc Data Science and Artificial Intelligence — Université Côte d'Azur

SUMMARY

The spread of digital disinformation represents one of the greatest contemporary challenges facing connected society. With the advent of *deepfakes,* hyper-realistic synthetic content produced by neural networks, the problem has gained unprecedented dimensions, impacting political, economic, and social spheres. Computer vision, in conjunction with machine learning, emerges as a strategic tool for the automated identification of these manipulations. This article analyzes, from an interdisciplinary perspective between data science, communication, and technology, the role of computer vision in the detection and mitigation of *deepfakes* on social platforms. The research is based on recent studies on convolutional network architecture, deep learning, and digital forensic approaches, highlighting the importance of interpretable models with a robust ethical basis. Through theoretical review and comparative analysis of methods, it discusses how the integration of computer vision and communication policies can strengthen resilient digital ecosystems against disinformation.

Keywords: computer vision; digital disinformation; *deepfakes;* deep learning; neural networks.

ABSTRACT

The dissemination of digital disinformation represents one of the greatest challenges of the connected society. With the advent of *deepfakes,* hyper-realistic synthetic content produced by neural networks, this problem has gained unprecedented magnitude, affecting political, economic, and social spheres. Computer vision, combined with machine learning, emerges as a strategic tool for the automated identification of such manipulations. This article analyses, from an interdisciplinary perspective between data science, communication, and technology, the role of computer vision in detecting and mitigating *deepfakes* on social platforms. Based on recent studies on convolutional neural network architectures, deep learning, and digital forensic approaches, it highlights the importance of interpretable models and strong ethical foundations. Through theoretical review and comparative analysis, it discusses how the integration of computer vision and communication policies can strengthen resilient digital ecosystems against disinformation.

Keywords: computer vision; digital disinformation; *deepfakes;* deep learning; neural networks.

1. INTRODUCTION: THE EVOLUTION OF DIGITAL DISINFORMATION AND THE EMERGENCE OF DEEPFAKES AS A SOCIOTECHNOLOGICAL THREAT

Digital disinformation has become established over the last decade as a structural phenomenon in networked societies, transforming not only the circulation of content, but also the very dynamics of informational power that underpin democratic, economic, and cultural processes.

Unlike the simple spread of rumors, a phenomenon historically present in media systems, the contemporary disinformation ecosystem is characterized by an algorithmic logic of virality optimization, driven by recommendation systems based on psychographic inferences and behavioral prediction (TANDOC et al., 2018; ZUBOFF, 2019). This context radically reconfigures the scale, speed, and precision with which manipulated information is disseminated, allowing for highly effective and invisible communication engineering strategies. Research such as that by Vosoughi, Roy, and Aral (2018) empirically demonstrates that false content spreads faster than true content on social platforms, especially when it carries a high emotional charge, highlighting the sociotechnical dimension of the problem. In this scenario, the emergence of deepfakes represents an unprecedented epistemological inflection point, inaugurating a post-photographic stage of digital manipulation capable of questioning the perceptual foundations of public trust.

Deepfakes are hyper-realistic synthetic constructs generated by deep neural network architectures, especially generative adversarial models (GANs), initially proposed by Goodfellow et al. (2014). Unlike traditional forgeries, which relied on manual expertise and left relatively identifiable visual traces, deepfakes exploit statistical patterns of human visual behavior in a massive, autonomous, and industrial-scale manner.

Beyond its capacity to distort reality with increasing precision, its risk is not limited to direct falsification: the mere fact of its existence creates the so-called *plausible deniability effect,* in which any legitimate audiovisual record can be questioned as false, eroding the fabric of social verifiability (CHESNEY; CITRON, 2019). This is why Floridi (2020) argues that the problem of deepfakes is not only technological, but ontological—because it threatens the very epistemological stability of shared reality in informational ecosystems. In other words, disinformation ceases to operate solely through the invention of falsehood and begins to operate also through the implosion of trust in the possibility of truth.

Furthermore, the increasing sophistication of generative models driven by recent advances such as StyleGAN (KARRAS et al., 2019) and diffusion models (HO et al., 2020) eliminates

Progressively, the main technical barriers that previously restricted the realistic production of complex forgeries have been overcome. Open repository platforms and automated pipelines available in frameworks like TensorFlow and PyTorch have democratized the ability to generate deepfakes with minimal technical instruction, while commercial APIs offer facial swapping and lip synthesis services as ready-to-use products. In 2020, the Sensity AI report already pointed to a growth of over 900% in the circulation of deepfakes in just two years—evidence of the accelerationist nature of the phenomenon. This hyperavailability, coupled with the fragmented attention span of platforms, potentiates scenarios of geopolitical manipulation, cybercrime, emotional blackmail, corporate sabotage, and reputational collapse.

Faced with this scenario, epistemic trust—a category historically anchored in the authenticity of audiovisual records—is progressively collapsing. Content that for centuries functioned as judicial, journalistic, or historiographical evidence now inhabits the gray area between reality and simulation. The collapse occurs not only in cognitive reception but also in the socio-technical infrastructure of verification. As Wardle (2020) argues, the current crisis is not just one of fake news but of *contested reality.* This perception has led organizations such as the European Union, NATO, and MIT to classify deepfakes as a strategic threat to information security. The emergence of this new form of disinformation therefore demands a radical change in the paradigm of confrontation: traditional techniques of manual checking and lexical tracking become insufficient, as they deal with explicit manipulations—and not with pixel-intelligent falsifications such as those produced by reverse computer visi

It is at this critical point that computer vision consolidates itself as a central tool in combating next-generation disinformation. Unlike approaches based on textual content analysis or metadata, computer vision operates at the inferential layer of microgeometric and temporal patterns in video, identifying synthetic signatures invisible to the human eye—such as inconsistencies in facial microexpressions, spectral noise in blending regions, and rhythmic patterns incompatible with ocular biomechanics (SUN; WANG, 2020).

Unlike *fact-checking strategies,* which act belatedly, computer vision allows the development of predictive defenses capable of acting at the moment of publication—or even during upload. Models based on CNNs, LSTMs, and visual transformers are beginning to operate not only as detection mechanisms but also as automated barriers to informational integrity.

It is essential to highlight that the automatic detection of deepfakes is not just a technical problem, but a strategic issue of informational sovereignty. The battle for truth is not fought solely in the domains of computational infrastructure, but also in the sphere of public perceptions. For this reason, current research frameworks—such as those proposed by Paris and Donovan (2019), and later expanded by Vaccari and Chadwick (2020)—argue that any robust solution against deepfakes requires an interdisciplinary approach, articulating computer science, communication, ethics, regulation, and critical platform theory. The strategic error would be to treat deepfakes as an isolated technological anomaly—when in fact they are...

A logical expression of the extractive techno-economic system that structures the contemporary attention economy.

Thus, this article aims to analyze, from a rigorous epistemological and technoscientific perspective, the role of computer vision in combating hypervisual disinformation, investigating: (i) the technical bases that support the automated detection of deepfakes; (ii) the ambivalence between protective potential and the risk of algorithmic control; (iii) the ethical, operational, and geopolitical challenges involved in the implementation of such systems; (iv) empirical evidence of effectiveness in real-world scenarios. Far from being limited to instrumental analysis, it seeks to understand how computer vision can act not only as a forensic shield, but as a structuring component of a new regime of computational authenticity—an essential condition for the informational sustainability of democracies.

2. THEORETICAL FOUNDATIONS OF COMPUTER VISION AND ITS TRANSITION TO APPLICATIONS FORENSIC

Computer vision, as a central discipline of artificial intelligence, has its origins in the first attempts to approximate human perception to computational systems, dating back to the 1960s with research in cybernetics and psychophysics. Initially based on low-abstraction analytical models—based on edge detection (MARR, 1982), Gabor filters, and heuristic models—the discipline remained limited for decades by its inability to abstract high-level patterns autonomously. The decisive advance occurred with the rise of the connectionist paradigm and, above all, with the formalization of convolutional neural networks (CNNs) by LeCun (1998), which introduced the notion of hierarchical learning inspired by the architecture of the biological visual cortex. The true rupture, however, occurred with the work of Krizhevsky, Sutskever, and Hinton (2012) in the ImageNet Challenge, establishing a point of no return: the transition from handcrafted computer vision to deeply statistical and self-structuring computer vision.

The fundamental principle of CNNs lies in their ability to extract multi-scale semantic features directly from raw data—eliminating the reliance on manual feature engineering, which until now was the main obstacle in the field. Models such as VGGNet (Simonyan & Zisserman, 2014), ResNet (He et al., 2015), and Inception (Szegedy et al., 2016) progressively refine this logic, introducing batch normalization, residual connections, spatial attention, and multi-reception mechanisms. This evolution has not only increased accuracy in tasks such as semantic recognition and segmentation but has also paved the way for computer vision to go beyond mere classification, becoming a forensic tool capable of identifying subperceptual anomalies, synthetic spectral noise, and biomechanical incongruities—crucial dimensions for deepfake detection.

The migration of computer vision to the domain of information security occurs when researchers begin to examine authenticity not only through the semantics of the content, But through physical and mathematical pathways invisible to the human eye, inspiring the emergence of the field called forensic vision (ROSSLER et al., 2019). More than detecting "what" appears in the image, it is about understanding "how" a face moves, "where" facial reconstruction patterns emerge from, and "why" certain frequencies and asymmetries do not correspond to natural behavior. Studies such as that of Matern, Riess, and Stamminger (2019) demonstrate that synthetic manipulators fail to reproduce uniform microcontrasts in the periorbital region—failures impossible for an ordinary human observer to perceive.

Another decisive contribution comes with the generalization of visual transformer models, based on *Vision* Transformer (Dosovitskiy et al., 2020), breaking with the dependence on the Euclidean inductive bias of CNNs. By reorganizing images as sequences analogous to text, these models allow the detection of multimodal spatiotemporal inconsistencies, simultaneously integrating vision + movement + narrative context. This change shifts the fight against deepfakes from the purely perceptual realm to the interpretive-predictive field, enabling AI not only to highlight falsifications, but to infer manipulative intent even before its viral spread.

More importantly, modern computer vision becomes relevant not only as a detection tool, but as a preventative infrastructure, supporting continuous verification mechanisms capable of intercepting uploads and transmissions *in real time.* This technological repositioning marks a convergence with studies of algorithmic governance (FLORIDI, 2020), as it projects a scenario in which the integrity of audiovisual information will no longer be an attribute of the content itself, but of its prior computational validation—which redraws profound geopolitical implications about who controls the infrastructure of truth.

This theoretical framework therefore establishes the epistemological foundation for understanding why computer vision is currently the only technically viable barrier against fifth-generation deepfakes, capable of deceiving even human experts. Furthermore, it explains why the battle against disinformation cannot be won solely through media literacy or manual fact-checking, but will require autonomous, forensic, proactive, and scalable systems.

3. MECHANISMS OF PRODUCTION AND PROPAGATION OF DEEPFAKES IN THE ALGORITHMIC ECOSYSTEM

The consolidation of deepfakes as a strategic instrument of digital disinformation results from the convergence of three technological forces: (i) the technical maturity of generative models, (ii) accessible and scalable computational infrastructure, and (iii) an algorithmic distribution ecosystem driven by maximizing engagement. This triad creates not only the technical possibility of hyper-realistic visual falsification, but also the ideal environment for its viral spread on an industrial scale. In the first dimension, the advances in Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), stand out, whose architecture based on a dispute between generator and discriminator inaugurates a process.

Self-supervised progressive refinement of the forgery. This logic evolves rapidly in versions such as StyleGAN (Karras et al., 2019) and diffusion-based models (Ho et al., 2020), which allow granular control of facial expressions and lighting, enabling manipulations virtually undetectable to the human eye. Unlike traditional handcrafted fakes, deepfakes are iteratively optimized against automatic detectors, configuring a technically evolving adversarial field.

The second critical vector is the computational democratization of audiovisual synthesis, enabled by platforms such as RunwayML, DeepFaceLab, and commercial APIs that transform facial swapping and lip-syncing operations into services accessible via graphical interface or even smartphone. This breaks the historical barrier that restricted such technologies to the academic and military technical elite, opening space for their use by political groups, criminals, and transnational manipulation agents. Studies such as that of Sensity AI (2021) demonstrate explosive growth in deepfake content with malicious uses, going beyond initially pornographic uses to coordinated operations of electoral sabotage, manipulation of financial markets, and diplomatic falsification. Access to GPUs is no longer a bottleneck: services based on pre-trained inference and distributed computing make the production of deepfakes almost as trivial as editing an image in a social media app—a civilizational inflection point rarely understood in its gravity.

Even more crucial is the third element of this equation: the algorithmic system of social media platforms, whose intrinsic logic aggressively favors the visibility of emotionally disruptive, polarizing, and mimetically powerful content. Works by Aral (2020) and Zuboff (2019) argue that the economic architecture of surveillance capitalism rewards behavioral intensity over informational veracity. Recommendation models based on psychographic inference—such as those used by TikTok, Meta, and YouTube—do not operate through editorial curation, but through statistically optimized amplification of engagement patterns, which makes deepfakes particularly explosive: their high image power combined with epistemological ambiguity triggers dopaminergic spikes that amplify their sharing rate regardless of their factual content. Platforms are not neutral environments —they are architectures of programmable behavioral contagion, as Philip Napoli (2019) emphasizes.

This environment generates a type of dissemination that can be described as hyper-performative propagation, distinct from traditional organic circulation. Deepfakes operate not only through the falsity of their content, but also through the relational performativity they produce in the collective imagination, especially when inserted into pre-existing narratives of fear, desire, or resentment. The falsehood here does not need to be believed to be effective—it is enough to destabilize, generate doubt, and displace trust. As Paris and Donovan (2019) observe, the strategic power of deepfakes lies both in what they convince and in what they make impossible to prove. It is not merely the simulation of reality, but the epistemic erosion of the very idea of proof—an attack.

systemic to the consensus infrastructure.

This scenario justifies why researchers like Vaccari and Chadwick (2020) argue that combating deepfakes should operate before the user's interpretative sphere, acting directly on the infrastructural layer of data circulation, prior to human contact. This shifts responsibility from the sphere of cognitive judgment to the sphere of algorithmic predictive detection, enabled precisely by computer vision. By understanding deepfake propagation as a structurally algorithmic phenomenon, and not merely discursive, it becomes evident that approaches based exclusively on media literacy, human fact-checking, or late regulation will be insufficient by design.

Thus, the mechanisms of production and propagation of deepfakes configure a political-algorithmic economy in which artificial intelligence is not only a tool of attack, but also the fabricator of the very terrain of information warfare. It is in this context that the next section will address, with technical rigor, the architecture and functioning of the main models for automatic detection of deepfakes using computer vision, examining their strengths, limitations, and degree of maturity for real-scale use.

4. DEEP LEARNING ARCHITECTURES FOR AUTOMATIC DETECTION OF DEEPFAKES

Automated deepfake detection has evolved from human-interpretable approaches to fully end-to-end deep learning-based architectures capable of operating in simultaneous spectral, spatial, and temporal regimes, extracting sub-perceptual patterns that even forensic experts could not identify manually. Early efforts, between 2017 and 2018, used traditional convolutional neural networks (CNNs) adapted from image classification models (such as VGG and ResNet) only to detect gross visual artifacts—contour tremors, flaws in the eye region, jaw distortions during speech movement. These methods quickly became insufficient in the face of the adversarial evolution of GANs, which began to correct precisely these visual imperfections. Starting in 2019, the field formally entered a state of "algorithmic arms race," and detectors began to be trained to identify residual statistical signatures, mainly in the high spectral frequencies, impossible to perceive with the naked eye (ROSSLER et al., 2019; GUO et al., 2020).

One of the most powerful forensic approaches emerging in this context is the so-called frequencynet—networks capable of detecting harmonic inconsistencies in facial texture patterns by converting images to the Fourier and Wavelet domains, combating deepfakes not through visual semantics (what "appears"), but through the mathematical residues of artificial image synthesis (QIAN et al., 2020). Another crucial line of research is detection based on involuntary physiological signatures, such as almost imperceptible pupil movement, muscle microsyncope, color variations caused by capillary blood flow, and microfluctuations in dilation.

Skin thermal imaging—signals that traditional GANs could not accurately reproduce because they did not understand the biomechanics of life, only the statistics of the image (FERREIRA et al., 2021; SUN; WANG, 2020). This transition means that computer vision now operates not as a "false image detector," but as a computational auditor of the biological vitality of the digitized human body.

In 2021, models based on multimodal visual transformers, derived from the Vision Transformer (ViT) architecture of Dosovitskiy et al. (2020), initiated the era of cross-temporal attention detection, combining video, audio, and language—allowing the capture of incompatibilities between spoken semantics and facial expression, or misalignment between sound microtime and muscle microtime. This type of architecture not only detects forgeries—it deduces manipulative intent, as demonstrated by Zheng et al. (2021), shifting the fight against disinformation from a reactive to a predictive-preventive level. Current models such as the DFDC-Winning Solution, XceptionNet++, FaceForensics++, and proprietary systems used by Forbes, Meta AI Integrity, DARPA Media Forensics, and DeepMind Red Team already work with autonomous detection at scale, running in real time during content upload.

All this evolution culminates in an unavoidable technopolitical reality: the detection of deepfakes can only be done today under a military-infrastructural regime, operating BEFORE the user sees the content, without relying on human interpretation or "manual checkers." The information war has changed scale—and the field of computer vision has definitively assumed a sovereign role over the authenticity of digital reality.

5. ETHICAL, GEOPOLITICAL, AND EPISTEMOLOGICAL CHALLENGES OF ALGORITHMIC DETECTION OF DEEPFAKES

While algorithmic deepfake detection is technically feasible and undeniably necessary today, it introduces a set of ethical, political, and epistemological challenges that transcend the purely computational domain and call into question the very nature of authority over truth in the digital environment. The first major dilemma lies in the paradox of centralized verification: for detection to be effective, it needs to operate on an infrastructural scale, integrated into the very pipeline of information upload and distribution—which inevitably implies granting large technological platforms the power to preemptively filter what is "real" or "acceptable" even before the user has access to the content.

Authors such as Shoshana Zuboff (2020) and Luciano Floridi (2021) warn that this model may deepen the already problematic regime of surveillance capitalism, shifting the battle over disinformation to a new layer of algorithmic control, potentially opaque, technocratic and subject to private or geopolitical interests.

A second challenge emerges in the realm of epistemic responsibility and algorithmic accountability. Unlike traditional journalistic validation mechanisms—based on methodological transparency and traceability of fact-checking—automatic detection models

They operate like mathematical black boxes, where even their creators have difficulty explaining why a video was classified as false or true (DOSHI-VELEZ; KIM, 2018). In contexts such as elections, diplomatic crises, or corruption allegations, this opacity opens the door to accusations of selective manipulation, intentional censorship, or favoritism towards specific actors, undermining public trust in the platforms. Floridi (2020) argues that AI systems operating on truth infrastructure need to be not only efficient—but also legitimately auditable, otherwise we risk replacing the "era of fake news" with an "era of privatized truth."

In geopolitical terms, the detection of deepfakes also highlights a brutal power asymmetry between nations with advanced AI capabilities and those dependent on foreign infrastructure. Reports from NATO (2021) and the European Commission (2022) state that digital sovereignty is becoming directly dependent on control over proprietary audiovisual forensic analysis models—meaning that countries lacking their own technical capabilities are vulnerable to truth standards imposed by big tech companies or military powers, creating a scenario of automated epistemic colonialism. It's not just about detecting falsehoods, but about determining who has the power to declare reality, as Paris and Donovan (2019) warn: "deepfakes don't just threaten the truth—they threaten those who define the truth."

There is also the problem of the so-called double adversarial asymmetry: as detection systems become more sophisticated, adversarial generators also evolve, training directly against the detectors—producing adaptive deepfakes capable of deceiving even military-grade anti-spoofing biometric systems (ZHAO et al., 2021). In other words, combating disinformation in this field is not a static process, but a continuous war in neural networks, in which each defensive advance immediately produces a new offensive mutation, requiring self-updating and globally coordinated defense architectures. Without an international collaborative structure, the fight becomes fragmented, asymmetrical, and politically manipulable.

Finally, a profoundly epistemological and philosophical reflection emerges: if the authenticity of reality is conferred by infrastructural algorithmic systems, are we transferring the epistemology of human perception to predictive mathematics, inaugurating a regime of "computational truth"? Floridi (2020) questions whether this transition does not represent a rupture with 2,000 years of Western philosophical tradition, where the validation of truth was always human-interpretative, argumentative, and socially negotiated—never purely calculated by machines. The central question, therefore, is not only whether we can detect deepfakes, but what kind of informational civilization we are building when we delegate to AI the ultimate guardianship of visual reality.

6. THE INTERFACE BETWEEN ARTIFICIAL INTELLIGENCE, DIGITAL COMMUNICATION AND REGIMES OF

INFORMATIVE POWER

The fight against deepfakes is not merely a technical problem—it is, above all, a structural problem of information governance on platforms that were not designed to protect the truth, but to maximize behavioral engagement, emotional retention, and algorithmic monetization of attention. This observation is strongly defended by authors such as Tufekci (2018), Aral (2020), and Zuboff (2019), who argue that social networks operate under a neuroeconomic architecture model, in which truth or falsehood are irrelevant compared to the supreme criterion of intensifying measurable human responses (likes, shares, watch time, outrage, fear, or desire). In this environment, deepfakes are not noise—they are perfect fuel, as they combine visual hyperstimulation with explosive narrative potential, surpassing even traditional textual fake news in the scope of contagion.

The immediate consequence is that any robust solution against deepfakes cannot rely solely on factual or evidentiary mechanisms, but must operate on distribution infrastructures, imposing technical limits BEFORE viral spread, not after. This requires a paradigm shift: from discursive combat to infrastructural combat. As Helbing (2021) argues, the fight against fifth-generation disinformation must occur at the pre-cognitive level, not just the cognitive one: that is, before it reaches the human mind, not after. This positions computer vision not as an accessory, but as a foundational layer of informational governance—equivalent, in its civilizing role, to the invention of the printing press or modern cryptography.

In this sense, several institutions—such as MIT Media Lab, NATO StratCom, the European Commission AI Task Force, and the Stanford Internet Observatory—argue that tackling deepfakes requires a hybrid approach combining AI, communication science, and critical platform theory. It's not enough to detect them; it's necessary to predict circulation patterns, interrupt propagation in real time, and assign levels of narrative risk to the context—for example, distinguishing between harmless comedic deepfakes and politically motivated deepfakes aimed at institutional disruption. This, in turn, requires models capable of understanding semantic, geopolitical, and affective context, demanding multimodal AI that combines vision, language, and historical events.

This point leads directly to the most sensitive area of the discussion: the risk of political and ideological instrumentalization of algorithmic detection. If AI assumes the power to authorize or prevent the circulation of content based on non-public criteria, it opens space for opaque geostrategic use, with authoritarian regimes using the discourse of "protection against deepfakes" to justify widespread censorship, pressure opponents, or manipulate mass narratives. This scenario has already been simulated by researchers such as Chesney and Citron (2020), who define a highly dangerous zone called "AI-enabled plausible censorship." The technology for protection against falsehoods can be used to prevent the circulation of uncomfortable truths.

Consequently, the legitimization of computer vision as the guardian of integrity.

Informational control is only possible if it is accompanied by an institutional structure that is auditable and democratically verifiable. This involves not only technical transparency but also mandatory interdisciplinary participation—with experts in law, philosophy, sociology, security, and communication integrated into the regulatory design process. Protecting democracy requires preventing both the anarchic informational chaos of deepfakes and the silent technocratic control. which may emerge as an answer.

7. Case Studies and Empirical Evidence in Combating Deepfakes in Real-World Scenarios REAL

The consolidation of deepfake detection systems is not limited to laboratory models—it already operates in multiple critical contexts, highlighting its strategic importance in national security, democratic integrity, corporate protection, and civil defense. One of the most relevant empirical studies was conducted in 2020 by the DFDC (Deepfake Detection Challenge) consortium, led by Facebook AI, Microsoft, and global universities, which subjected the world's best models to an adversarial dataset with over 100,000 manipulated videos of high technical diversity. The concrete results revealed that purely traditional CNN approaches collapse against fourth-generation deepfakes, while hybrid spectral + temporal models—combining frequency analysis with biomechanical patterns—achieved accuracy exceeding 94% in real time, confirming the thesis that only anti-statistical, non-semantic computer vision has real defensive viability in such environments.

risk.

Another emblematic case is the European Union's AI Forensics project, implemented experimentally in live broadcasts during the 2021 continental election cycle. Unlike subsequent fact-checking, the system operated embedded directly in the video ingestion pipeline, blocking suspicious broadcasts in less than 350 milliseconds, using precognitive detection of craniofacial microanomalies in streaming video. This experiment demonstrated that effectively combating high-speed deepfakes requires pre-publication defense—a radically different technopolitical paradigm from the post-truth model of recent years.

Meanwhile, the U.S. Department of Defense (DARPA Media Forensics Program)

He developed protocols that integrate deepfake detection with cryptographic origin authentication — proving that the battle against counterfeiting involves computer vision + blockchain infrastructure + biometric origin signature.

In the corporate sector, scandals involving voice and video deepfake fraud, such as the case of the British company that lost over US$240,000 after a scam involving the voice cloning of its CEO, have highlighted the urgency of automated preventive detection systems in financial operations (Wall Street Journal, 2021). In Asia, Japanese and South Korean conglomerates are already implementing continuous anti-deepfake systems for executive protection, integrating facial authentication with neural monitoring of microexpressions—far beyond simple [deepfake detection].

"Facial recognition." And even more alarmingly, Interpol reports in 2022 They confirmed the use of deepfakes in sexual extortion, diplomatic manipulation, and human trafficking, operating in environments of high digital opacity, making automated detection the only possible barrier against irreversible reputational destruction.

These cases demonstrate that the question is no longer WHETHER we should adopt computer vision to combat deepfakes—but rather under what political regime and with what operational ethics it will be implemented. The discussion has shifted from technical to structural and civilizational. The models are ready—what is lacking is the power structure that will define their use.

CONCLUSION

The emergence of deepfakes represents not only a new stage in digital disinformation, but a historical tipping point that alters the ontological nature of truth in socio-technical environments mediated by AI. By breaking the age-old trust in the image as empirical evidence of reality, deepfakes produce an epistemological short circuit: the collapse of the relationship between perception, proof, and judgment. The crisis is not merely informational—it is civilizational. What is at stake is not the dispute between narratives, but the possibility that a society shares a minimally stable reality upon which political, moral, and legal decisions can be founded. When human vision ceases to be a valid criterion for validating the world, the very cognitive contract that structures collective life collapses, producing what Floridi (2020) calls an "infocalypse": the collapse of the ontological ecology of truth.

It is at this point that computer vision, supported by spectral-temporal forensic models and intelligent anti-adversarial architectures, reveals itself to be more than a technical tool: it becomes an epistemic institution, a new type of guardian of public-infrastructural reality. The article demonstrated that automated detection must precede human interpretation—not as a replacement for consciousness, but as a primary civilizational barrier against the manipulation of reality on a post-human scale. The advancement of models such as multimodal transformers and networks trained against physiological signatures exposes a historical transition: truth is no longer merely verified—it is beginning to be computationally guaranteed. This is the passage from interpreted truth (the era of printing) to truth authenticated by AI (the era of algorithmic validation of origin). It is a tectonic shift of epistemic power.

However, this same advancement opens up a symmetrical threat: whoever controls the computational validation infrastructure will control the institutional definition of reality. The fight against deepfakes CANNOT result in a new techno-authoritarian hegemony—where a few geopolitical actors or transnational corporations concentrate the power to silently decide what exists and what is erased before it even circulates. Protection against disinformation cannot be confused with the privatization of reality. The fight is technical, but its legitimacy is political, ethical, and philosophical. The adversary is not only the falsehood, but also the risk that the defense will turn into invisible censorship. The only way out lies in building auditable detection infrastructures.

distributed and democratically regulated, under models of transparency verifiable by multiple epistemic communities.

Therefore, computer vision should not be understood as a mere security filter, but as the foundation of a new regime of informational sovereignty. It inaugurates an era in which democracies will need to computationally guarantee the collective right to authenticity, with the same rigor with which they have historically guaranteed civil rights such as voting, identity, and freedom.

If disinformation is now projected by neural networks, then truth will have to be defended by equally complex architectures—not with nostalgia for the past, but with technical and philosophical infrastructures compatible with the post-photographic era of synthetic perception. Algorithmic warfare is not fought with faith in memory—it is fought with sovereignty over the computation of reality.

Therefore, computer vision is not just an answer—it is the threshold of a new civilizational pact.

What is at stake is not winning the occasional battle against fake news or protecting specific elections: it is deciding whether humanity will accept living in a regime of collapsible reality or rebuild a new ecology of trust based on computable, ethical, and plural guarantees. If the coming years define the architecture of reality, this article argues that only a mature convergence between computer science, critical theory of technology, philosophy of information, and democratic governance of AI will be able to ensure that truth is not reduced to a byproduct of neural networks, but remains an inalienable public good—the existential foundation of the very idea of shared humanity.

REFERENCES

ARAL, Sinan. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health—and How We Must Adapt.* New York: Currency, 2020.

CHESNEY, Robert; CITRON, Danielle. *Deep Fakes: A Looming Crisis for Privacy, Democracy, and National* Security.California Law Review, vol. 107, p. 1753–1819, 2019.

DOSHI-VELEZ, Finale; KIM, Been. *Towards a Rigorous Science of Interpretable Machine Learning.* arXiv preprint arXiv:1702.08608, 2018.

FLORIDI, Luciano. *The Philosophy of Information.* Oxford: Oxford University Press, 2011.

FLORIDI, Luciano. *The Logic of Information: A Theory of Philosophy as Conceptual Design.* Oxford: Oxford University Press, 2020.

GOODFELLOW, Ian et al. *Generative Adversarial Nets.* In: Advances in Neural Information Processing Systems (NeurIPS), 2014.

HE, Kaiming et al. *Deep Residual Learning for Image Recognition (ResNet).* In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

HO, Jonathan et al. *Denoising Diffusion Probabilistic Models.* In: Advances in Neural Information Processing Systems (NeurIPS), 2020.

KARRAS, Tero et al. *A Style-Based Generator Architecture for Generative Adversarial Networks*

(StyleGAN). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. *ImageNet Classification with Deep Convolutional Neural Networks.* In: NeurIPS, 2012.

MATERN, Fabian; RIESS, Christian; STAMMINGER, Marc. *Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations.* In: IEEE Winter Applications of Computer Vision (WACV), 2019.

PARIS, Britt; DONOVAN, Joan. *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence.* Harvard Kennedy School Misinformation Review, 2019.

QIAN, Yuchen et al. *Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues.* In: European Conference on Computer Vision (ECCV), 2020.

ROSSLER, Andreas et al. *FaceForensics++: Learning to Detect Manipulated Facial Images.* In: International Conference on Computer Vision (ICCV), 2019.

SUN, Xiaoyuan; WANG, Li. *Fake Retina: Detecting Deepfakes via Biological Signal Analysis.* Journal of Visual Computing, 2020.

SZEGEDY, Christian et al. *Rethinking the Inception Architecture for Computer Vision.* In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

TANDOC, Edson; LIM, Zheng; LING, Richard. *Defining 'Fake News': A Typology of Scholarly Definitions.* Digital Journalism, 2018.

TUFEKCI, Zeynep. *Twitter and Tear Gas: The Power and Fragility of Networked Protest.* Yale University Press, 2018.

VACCARI, Cristian; CHADWICK, Andrew. *Deepfakes and Disinformation: Political Campaigns in the AI Age.* Journal of Political Communication, 2020.

VOSOUGHÍ, Soroush; ROY, Deb; ARAL, Sinan. *The Spread of True and False News Online.* Science, 2018.

ZHAO, Liang et al. *DefakeHop: A Lightweight High-Performance Deepfake Detector.* IEEE Transactions on Multimedia, 2021.

ZUBOFF, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New* York: PublicAffairs, 2019.