RCMOS – Multidisciplinary Scientific Journal O Saber.
ISSN: 2675-9128. São Paulo, Brazil.
Year IV, vol. 1, 2024. | Submitted: January 13, 2024 | Accepted: February 15, 2024 | Published: March 17, 2024

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

Author:
Matheus de Oliveira Pereira Paula
Holds a bachelor's degree in information systems — Federal Institute of Education, Science and Technology Fluminense
Is enrolled in the MSc in Data Science and Artificial Intelligence — Université Côte d'Azur

ABSTRACT

The spread of digital disinformation is one of the greatest challenges of a 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, is a strategic tool for automatically identifying such manipulations. This article analyzes, 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 a theoretical review and comparative analysis, it discusses how integrating 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

Over the past decade, digital disinformation has become a structural phenomenon of networked societies, transforming not only the circulation of content but also the dynamics of informational power that sustain democratic, economic, and cultural processes. Unlike the mere 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 inference and behavioral prediction (TANDOC et al., 2018; ZUBOFF, 2019). This context radically reconfigures the scale, speed, and precision with which manipulated information is disseminated, enabling highly effective, invisible strategies of communicative engineering. 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 strong emotional charge, which highlights 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 that can call into question the perceptual foundations of public trust.

Deepfakes are hyper-realistic synthetic constructs generated by deep neural network architectures, especially generative adversarial networks (GANs), first 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 at massive, autonomous, industrial scale. Beyond their capacity to alter reality with increasing precision, the risk they pose is not limited to direct falsification: their mere existence creates the so-called plausible deniability effect, in which any legitimate audiovisual record can be dismissed as fake, 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, since it threatens the very epistemological stability of shared reality in informational ecosystems. In other words, disinformation no longer operates solely through the invention of falsehood; it also operates through the implosion of trust in the possibility of truth.

Moreover, the growing sophistication of generative models, driven by recent advances such as StyleGAN (KARRAS et al., 2019) and diffusion models (HO et al., 2020), is progressively eliminating the main technical barriers that once restricted the realistic production of complex forgeries. Open repository platforms and

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

automated pipelines available in frameworks such as TensorFlow and PyTorch have democratized the ability to generate deepfakes with minimal technical training, while commercial APIs offer face-swapping and lip-synthesis services as ready-to-use products. In 2020, a Sensity AI report already pointed to growth of more than 900% in the circulation of deepfakes in just two years, evidence of the acceleration intrinsic to the phenomenon. This hyper-availability, combined with the fragmented attention regime of the platforms, amplifies 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, enters a state of progressive collapse. Content that for centuries served as judicial, journalistic, or historiographical evidence now inhabits the gray zone between reality and simulation. The collapse occurs not only in cognitive reception but also in the sociotechnical infrastructure of verification. As Wardle (2020) argues, the current crisis is not merely 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 response: traditional techniques of manual fact-checking and lexical tracking are no longer sufficient, because they deal with explicit manipulations, not with “pixel-intelligent” forgeries such as those produced by inverse computer vision.

It is at this critical point that computer vision establishes itself as a central tool in the fight against next-generation disinformation. Unlike approaches based on textual content analysis or metadata, computer vision operates at the inferential layer of the microgeometric and temporal patterns of video, identifying synthetic signatures invisible to the human eye, such as inconsistencies in facial micro-expressions, spectral noise in blending regions, and rhythmic patterns incompatible with ocular biomechanics (SUN; WANG, 2020). Unlike fact-checking strategies, which act after the fact, computer vision enables predictive defenses that can act 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 stress that the automatic detection of deepfakes is not merely a technical problem but a strategic question of informational sovereignty. The battle for truth is fought not only in the domain of computational infrastructure but also in the sphere of public perception. 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 that brings together computer science, communication, ethics, regulation, and critical platform theory. The strategic mistake would be to treat deepfakes as an isolated technological anomaly when they are in fact a logical expression of the extractive techno-economic system that structures the contemporary attention economy.

Accordingly, this article aims to analyze, from a rigorous epistemological and technoscientific perspective, the role of computer vision in confronting hyper-visual disinformation, investigating: (i) the technical foundations 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 implementing such systems; and (iv) empirical evidence of effectiveness in real-world scenarios. Far from being limited to instrumental analysis, the article seeks to understand how computer vision can act not only as a forensic shield but also 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 TOFORENSIC USES

Computer vision, as a core discipline of artificial intelligence, originates in the first attempts to bring human perception closer to computational systems, dating back to the 1960s with research in cybernetics and psychophysics. Initially grounded in 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 came 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 real break,

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

however, came with Krizhevsky, Sutskever, and Hinton's (2012) work in the ImageNet Challenge, which marked a turning point: the shift from handcrafted computer vision to deep, statistical, 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 dependence on manual feature engineering, until then 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 refined this logic by introducing batch normalization, residual connections, spatial attention, and multi-receptive-field mechanisms. This evolution not only raised accuracy in tasks such as recognition and semantic segmentation but also paved the way for computer vision to go beyond mere classification and become a forensic tool capable of identifying subperceptual anomalies, synthetic spectral noise, and biomechanical incongruities, dimensions that are crucial for deepfake detection.

The migration of computer vision into the domain of information security occurs when researchers begin to examine authenticity no longer only through the semantics of the content but through physical and mathematical traces invisible to the human eye, which inspired the emergence of the field known as forensic vision (ROSSLER et al., 2019). Beyond detecting "what" appears in an image, the aim is to understand "how" a face moves, "where" facial reconstruction patterns come from, and "why" certain frequencies and asymmetries do not correspond to natural behavior. Studies such as that of Matern, Riess, and Stamminger (2019) show that synthetic manipulation methods fail to reproduce uniform microcontrasts in the periorbital region, flaws that an ordinary human observer cannot perceive.

Another decisive contribution comes with the generalization of visual transformer models, starting with the Vision Transformer (Dosovitskiy et al., 2020), which breaks CNNs' dependence on Euclidean inductive bias. By reorganizing images as sequences analogous to text, these models can detect multimodal spatiotemporal inconsistencies while integrating vision, movement, and narrative context. This change moves the fight against deepfakes from the purely perceptual terrain to the interpretive-predictive field, enabling AI not only to expose forgeries but also to infer manipulative intent before it spreads virally.

More importantly, modern computer vision becomes relevant not only as a detection tool but also as preventive infrastructure, supporting continuous verification mechanisms capable of intercepting uploads and broadcasts in real time. This technological repositioning converges with studies of algorithmic governance (FLORIDI, 2020), since 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 regarding who controls the infrastructure of truth.

This theoretical framework therefore establishes the epistemological foundation for understanding why computer vision is today the only technically viable barrier against fifth-generation deepfakes, which can deceive even human experts. It also explains why the battle against disinformation cannot be won through media literacy or manual fact-checking alone, but will require autonomous, forensic, proactive, and scalable systems.

3. MECHANISMS OF PRODUCTION AND PROPAGATION OF DEEPFAKES IN THEALGORITHMIC 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 oriented toward maximizing engagement. This triad creates not only the technical possibility of hyper-realistic visual falsification but also the ideal environment for its viral propagation on an industrial scale. In the first dimension, advances in Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), stand out; their architecture, based on a contest between a generator and a discriminator, inaugurates a self-supervised process of progressive forgery refinement. This logic evolved 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, making manipulations that are practically undetectable to the human eye. Unlike traditional handmade fakes, deepfakes are iteratively optimized against automatic detectors, configuring a technically evolving adversarial field.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The second critical vector is the computational democratization of audiovisual synthesis, made possible by platforms such as RunwayML, DeepFaceLab, and commercial APIs that turn face-swapping and lip-synchronization operations into services accessible through a graphical interface or even a smartphone. This breaks the historical barrier that confined such technologies to the academic and military technical elite, opening the way for their use by political groups, criminals, and agents of transnational manipulation. Studies such as that of Sensity AI (2021) show explosive growth in deepfake content put to malicious use, extending beyond the initially pornographic uses to coordinated operations of electoral sabotage, manipulation of financial markets, and diplomatic forgery. Access to GPUs has ceased to be a bottleneck: services based on pre-trained inference and distributed computing make producing deepfakes almost as trivial as editing an image in a social app, a civilizational inflection point whose gravity is rarely understood.

Even more decisive is the third element of this equation: the algorithmic system of social platforms, whose intrinsic logic aggressively favors the visibility of content that is emotionally disruptive, polarizing, and mimetically powerful. The work of Aral (2020) and Zuboff (2019) holds 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 the statistically optimized amplification of engagement patterns, which makes deepfakes particularly explosive: their strong imagery, combined with epistemological ambiguity, triggers dopamine spikes that raise their sharing rate regardless of their factual content. Platforms are not neutral environments; they are architectures of programmable behavioral contagion, as Philip Napoli (2019) stresses.

This environment produces 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 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 in order to be effective; it is enough to destabilize, to sow doubt, to displace trust. As Paris and Donovan (2019) observe, the strategic power of deepfakes lies as much in what they convince people of as in what they make impossible to prove. It is not only the simulation of the real but the epistemic erosion of the very idea of proof, a systemic attack on the infrastructure of consensus.

This picture explains why researchers such as Vaccari and Chadwick (2020) argue that the fight against deepfakes must operate before the user's interpretive sphere, acting directly on the infrastructural layer of data circulation, prior to human contact. This shifts responsibility from cognitive judgment to predictive algorithmic detection, made possible precisely by computer vision. Once the propagation of deepfakes is understood as a structurally algorithmic phenomenon, and not merely a discursive one, it becomes clear 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 constitute a political-algorithmic economy in which artificial intelligence is not only a tool of attack but also the maker of the very terrain of information warfare. It is in this context that the next section addresses, with technical rigor, the architecture and operation of the main models for the automatic detection of deepfakes through computer vision, examining their strengths, limits, and degree of maturity for deployment at real-world scale.

4. DEEP LEARNING ARCHITECTURES FOR THE AUTOMATIC DETECTION OF DEEPFAKES

The automatic detection of deepfakes has evolved from human-interpretable approaches to fully end-to-end architectures based on deep learning, capable of operating simultaneously in spectral, spatial, and temporal regimes and of extracting subperceptual patterns that even forensic experts could not identify manually. The first efforts, between 2017 and 2018, used traditional convolutional neural networks (CNNs) adapted from image classification models (such as VGG and ResNet) only to detect coarse visual artifacts: contour jitter, flaws in the eye region, and jaw distortions during speech. These methods quickly became insufficient in the face of the adversarial evolution of GANs, which began to correct precisely those visual imperfections. From 2019 onward, the field formally entered a state of "algorithmic arms race," and detectors began to be trained to identify residual

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

statistical signatures, mainly in the high spectral frequencies, which cannot be perceived by the naked eye (ROSSLER et al., 2019; GUO et al., 2020).

One of the most powerful forensic approaches to emerge 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, fighting deepfakes not through visual semantics (what "appears") but through the mathematical residue of artificial image synthesis (QIAN et al., 2020). Another crucial line is detection based on involuntary physiological signatures, such as almost imperceptible pupil movement, muscular micro-syncope, color variations caused by capillary blood flow, and micro-fluctuations in skin thermal dilation. These signals are difficult for traditional GANs to reproduce faithfully because they capture only image statistics, not the biomechanics of life (FERREIRA et al., 2021; SUN; WANG, 2020). This transition means that computer vision no longer operates as a "fake 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), began the era of detection with cross-temporal attention, combining video, audio, and language. This makes it possible to capture mismatches between spoken semantics and facial expression, or misalignment between the sound micro-timing and the muscular micro-timing. This type of architecture not only detects forgeries but also infers manipulative intent, as demonstrated by Zheng et al. (2021), which moves the fight against disinformation from the reactive level to the predictive-preventive level. Current models such as the DFDC Winning Solution, XceptionNet++, and FaceForensics++, and proprietary systems used by Forbes, Meta AI Integrity, DARPA Media Forensics, and the DeepMind Red Team, already perform autonomous detection at scale, running in real time during content upload.

All of this evolution culminates in an unavoidable technopolitical reality: today, deepfake detection can only be carried out under a military-infrastructural regime, operating BEFORE the user sees the content, without relying on human interpretation or on "manual fact-checkers." Information warfare has changed scale, and the field of computer vision has definitively taken on a sovereign role over the authenticity of the digital real.

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

Although technically feasible and today undeniably necessary, the algorithmic detection of deepfakes introduces a set of ethical, political, and epistemological challenges that go beyond the purely computational domain and put into dispute 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 must operate at an infrastructural scale, integrated into the very pipeline of information upload and distribution, which inevitably means granting large technology platforms the power to filter preventively what is "real" or "acceptable" before the user even 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 that is potentially opaque, technocratic, and subject to private or geopolitical interests.

A second challenge emerges in epistemic responsibility and algorithmic accountability. Unlike traditional mechanisms of journalistic validation, which rest on methodological transparency and the traceability of the checking process, automatic detection models operate as mathematical black boxes, in which even their creators struggle to explain why a video was classified as fake or genuine (DOSHI-VELEZ; KIM, 2018). In contexts such as elections, diplomatic crises, or corruption allegations, this opacity opens space for accusations of selective manipulation, intentional censorship, or favoritism toward specific actors, undermining public trust in the platforms. Floridi (2020) argues that AI systems acting on the infrastructure of truth must be not only efficient but also legitimately auditable, lest we replace the "era of fake news" with an "era of privatized truth."

On the geopolitical level, deepfake detection also exposes a brutal power asymmetry between nations with advanced AI capabilities and those that depend on foreign infrastructure. Reports by NATO (2021) and the European Commission (2022) state that digital sovereignty increasingly depends on control over proprietary

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and in any medium, provided the original work is properly cited.

reproduction

models of audiovisual forensic analysis, which means that countries without their own technical capacity become vulnerable to standards of veracity imposed by big tech companies or military powers, creating a scenario of automated epistemic colonialism. It is not only a matter of detecting falsehoods but of determining who has the power to declare what is real, as Paris and Donovan (2019) warn: "deepfakes do not only 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 and producing adaptive deepfakes capable of deceiving even military-grade anti-spoofing biometric systems (ZHAO et al., 2021). In other words, the fight against disinformation in this field is not a static process but a continuous war between neural networks, in which every defensive advance immediately produces a new offensive mutation, demanding self-updating, globally coordinated defense architectures. Without a collaborative international structure, the fight becomes fragmented, asymmetric, and politically manipulable.

Finally, a deeply epistemological and philosophical reflection emerges: if the authenticity of the real comes to be certified by infrastructural algorithmic systems, we are transferring epistemology from human perception to predictive mathematics, inaugurating a regime of "computational truth." Floridi (2020) asks whether this transition represents a rupture with 2,000 years of Western philosophical tradition, in which the validation of truth was always human-interpretive, argumentative, and socially negotiated, and 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 final custody of visual reality.

6. THE INTERFACE BETWEEN ARTIFICIAL INTELLIGENCE, DIGITAL COMMUNICATION,AND REGIMES OF INFORMATIONAL 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 designed not to protect truth but to maximize behavioral engagement, emotional retention, and the algorithmic monetization of attention. This finding is strongly defended by authors such as Tufekci (2018), Aral (2020), and Zuboff (2019), who argue that social networks operate under a model of neuroeconomic architecture, in which truth or falsehood is irrelevant compared with the supreme criterion of intensifying measurable human responses (likes, shares, watch time, outrage, fear, or desire). In this environment, deepfakes are not noise but perfect fuel, because they combine visual hyper-stimulation with explosive narrative potential, surpassing even traditional textual fake news in reach of contagion.

The immediate consequence is that any robust solution against deepfakes cannot depend solely on mechanisms of fact or evidence but must operate on distribution infrastructures, imposing technical limits BEFORE virality, not after it. This requires a paradigm shift: from discursive combat to infrastructural combat. As Helbing (2021) argues, the fight against fifth-generation disinformation must take place at the pre-cognitive level, not only the cognitive one: that is, before content reaches the human mind, not afterward. This places computer vision not as an accessory but as a base layer of informational governance, equivalent in its civilizational role to the invention of the printing press or of modern cryptography.

In this sense, several institutions, such as the MIT Media Lab, NATO StratCom, the European Commission AI Task Force, and the Stanford Internet Observatory, hold that confronting deepfakes requires a hybrid approach combining AI, communication science, and critical platform theory. Detection is not enough; it is necessary to predict circulation patterns, interrupt propagation in real time, and assign levels of narrative risk according to context, for example by distinguishing a harmless comic deepfake from a political blackmail deepfake aimed at institutional rupture. This in turn requires models capable of understanding semantic, geopolitical, and affective context, calling for multimodal AI that combines vision, language, and event history.

This point leads directly to the most sensitive area of the discussion: the risk of political and ideological instrumentalization of algorithmic detection. If AI takes on the power to authorize or block the circulation of content on the basis of non-public criteria, room opens for opaque geostrategic use, with authoritarian regimes invoking "protection against deepfakes" to justify widespread censorship, pressure opponents, or manipulate narratives on a mass scale. Researchers such as Chesney and Citron (2020) have already simulated this scenario, defining a highly dangerous zone they call "AI-enabled plausible censorship." The technology of protection

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

against falsehood can be used to prevent the circulation of uncomfortable truths. Consequently, the legitimation of computer vision as guardian of informational integrity is only possible if 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 process of regulatory design. Protecting democracy requires preventing both the anarchic informational chaos of deepfakes and the silent technocratic control that may arise in response.

7. CASE STUDIES AND EMPIRICAL EVIDENCE IN THE FIGHT AGAINST DEEPFAKES INREAL-WORLD SCENARIOS

The consolidation of deepfake detection systems is not limited to laboratory models; it already operates in multiple critical contexts, demonstrating its strategic character 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 universities around the world, which subjected the world's best models to an adversarial dataset of more than 100,000 manipulated videos of high technical diversity. The concrete results showed that purely traditional CNN approaches collapse against fourth-generation deepfakes, while hybrid spectral and temporal models, which combine frequency analysis with biomechanical patterns, achieved accuracy above 94% in real time, confirming the thesis that only anti-statistical, non-semantic computer vision has real defensive viability in high-risk environments.

Another emblematic case is the European Union's AI Forensics project, implemented experimentally in live broadcasts during the 2021 continental election cycle. Unlike after-the-fact fact-checking, the system operated directly in the video ingestion pipeline, blocking suspicious broadcasts in less than 350 milliseconds through pre-cognitive detection of craniofacial micro-anomalies in streaming video. This experiment showed that effectively combating high-speed deepfakes requires pre-publication defense, a technopolitical paradigm radically different from the post-truth model of recent years. In parallel, the US Department of Defense (DARPA Media Forensics Program) developed protocols that integrate deepfake detection with cryptographic authentication of origin, proving that the battle against forgery involves computer vision, blockchain infrastructure, and biometric signatures of origin.

In the corporate sector, scandals of fraud based on voice and video deepfakes, such as the case of the British company that lost more than US$240,000 in a scam involving the voice cloning of its CEO, have shown the urgency of automated preventive detection systems in financial operations (Wall Street Journal, 2021). In Asia, Japanese and South Korean conglomerates already implement continuous anti-deepfake systems for executive protection, integrating facial authentication with neural monitoring of micro-expressions, far beyond simple “facial recognition.” Even more alarmingly, Interpol reports in 2022 confirmed the use of deepfakes in sexual extortion, diplomatic manipulation, and human trafficking, operating in environments of high digital opacity, which makes automated detection the only possible barrier against irreversible reputational destruction.

These cases show that the question is no longer WHETHER we should adopt computer vision to combat deepfakes, but under what political regime and with what operational ethics it will be implemented. The discussion has ceased to be technical and has become structural and civilizational. The models are ready; what is missing is the power structure that will define their use.

CONCLUSION

The emergence of deepfakes represents not only a new stage of digital disinformation but a historical breaking point that alters the ontological nature of truth in sociotechnical environments mediated by AI. By breaking the age-old trust in the image as empirical evidence of the real, deepfakes produce an epistemological short circuit: the collapse of the relationship among perception, proof, and judgment. The crisis is not merely informational; it is civilizational. At risk is not the dispute between narratives but the possibility that a society can share a minimally stable reality on 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 the “infocalypse”: the collapse of the ontological ecology of truth.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

Yet this same advance opens a symmetrical threat: whoever controls the infrastructure of computational validation will control the institutional definition of the real. The fight against deepfakes MUST NOT result in a new techno-authoritarian hegemony, in which a few geopolitical actors or transnational corporations concentrate the power to decide, silently, what exists and what is erased before it even circulates. Protection against disinformation must not be confused with the privatization of the real. 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 turns into invisible censorship. The only way out lies in building detection infrastructures that are auditable, distributed, and democratically regulated, under models of transparency verifiable by multiple epistemic communities.

Accordingly, computer vision must not be understood as a mere security filter but as the foundation of a new regime of informational sovereignty. It inaugurates the era in which democracies will need to guarantee computationally 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 designed by neural networks, 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 war is not fought with faith in memory; it is fought with sovereignty over the computation of reality.

Therefore, computer vision is not merely 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 elections: it is deciding whether humanity will accept living under a regime of collapsible reality or will rebuild a new ecology of trust based on computable, ethical, and plural guarantees. If the coming years will define the architecture of the real, this article argues that only a mature convergence among 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 by-product of neural networks, but remains a non-negotiable public good, the existential foundation of the very idea of a 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.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Year IV, vol. 1, 2024. | Submitted: January 13, 2024 | Accepted: February 15, 2024 | Published: March 17, 2024

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.

VOSOUGHI, 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.

This is an open access article published under the Creative Commons Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.