RCMOS – Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.
Year II, v.2 2022. | Submission: 12/16/2022 | Accepted: 12/18/2022 | Publication: 12/20/2022

Computer Vision and Algorithmic Ethics: Challenges in Recognition and Detection of Deepfakes

Computer Vision and Algorithmic Ethics: Challenges in the Recognition and Detection of Deepfakes

[1]Matheus de Oliveira Pereira Paula — Bachelor's degree in Information Systems from the Federal Institute of Education, Science and Technology Fluminense. Master's degree in Data Science and Artificial Intelligence from the Université Côte d'Azur.

Summary

The rise of *deepfakes* represents one of the greatest ethical and technological challenges of the 21st century. Based on computer vision and deep learning techniques, these manipulated contents challenge the limits of digital trust and the very notion of truth in contemporary media. This scientific article examines the intersections between the evolution of computer vision and the principles of algorithmic ethics, highlighting how advances in convolutional neural networks and generative adversarial models (GANs) impact the detection and recognition of digital forgeries. The analysis considers not only the technical dimension, but also the social, legal, and moral implications involved in the dissemination of *deepfakes,* pointing to paths for the development of ethical and transparent systems.

Keywords: computer vision; algorithmic ethics; *deepfake;* artificial intelligence; detection.

Abstract

The rise of *deepfakes* represents one of the greatest ethical and technological challenges of the 21st century. Based on computer vision and deep learning techniques, these manipulated contents challenge the limits of digital trust and the very notion of truth in contemporary media. This scientific paper examines the intersections between computer vision evolution and algorithmic ethics, highlighting how advances in convolutional neural networks and generative adversarial models (GANs) affect digital forgery detection and recognition. The analysis covers not only the technical dimension but also the social, legal, and moral implications involved in the dissemination of *deepfakes,* pointing out pathways for developing ethical and transparent systems.

Keywords: computer vision; algorithmic ethics; *deepfakes;* artificial intelligence; detection.

1. Introduction

The advancement of artificial intelligence (AI) technologies, and especially computer vision, has brought with it challenges that transcend the boundaries of engineering and reach ethical, philosophical, and social dimensions. Among these innovations, the phenomenon of *deepfakes stands out,* combining deep neural networks and generative learning techniques to create synthetic videos, audios, and images with an extremely realistic appearance (Goodfellow et al., 2014). The problem intensifies as these technologies, originally conceived to enhance human interfaces and creative applications, have begun to be used in contexts of disinformation, political manipulation, and attacks on individual privacy.

The notion of truth and visual authenticity, which has always been a cornerstone of human communication, now finds itself at a technological crossroads. Computer vision, a branch that seeks to equip machines with the ability to interpret the visual world, has also become an instrument of visual manipulation. According to Cappelletti (2020), the perfect simulation of human faces through generative adversarial networks (GANs) redefines the status of the image and public trust in digital media.

Algorithmic ethics emerges in this context as an essential interdisciplinary field. It seeks to understand and regulate the moral impacts of decisions made by automated systems.
As Floridi and Cowls (2019) state, algorithms are not morally neutral; they carry the values and biases of their designers, reflecting ethical decisions embedded in the code. Therefore, understanding the role of ethics in the detection and use of *deepfakes* is to understand the relationship between power, information, and responsibility in the 21st century.

This research aims to explore the technical foundations of computer vision applied to *deepfake detection,* while also investigating its ethical implications. The approach integrates theoretical and empirical perspectives, articulating specialized literature, case studies, and philosophical reflections. Thus, this article proposes to contribute to the debate on the balance between technological innovation and the safeguarding of fundamental human values.

Finally, the work is structured into seven sections. After this introduction, the evolution of computer vision is discussed; then, the role of neural networks and generative models in the creation of *deepfakes is examined;* next, the ethical challenges of perceptual automation are analyzed; subsequently, AI-based detection methods are explored; and finally, legal and social aspects are discussed, culminating in final considerations on the ethical future of computer vision.

2. The Evolution of Computer Vision and the Foundations of Artificial Perception

Computer vision has established itself as one of the most promising areas of artificial intelligence. Since the first attempts at pattern recognition in the 1960s and 1970s, researchers like Marr (1982) have envisioned the potential of systems capable of replicating human visual processing. The development of edge detection, segmentation, and classification algorithms paved the way for the current era of *deep learning,* which has revolutionized how machines interpret the visual world.

The qualitative leap occurred with the emergence of convolutional neural networks (CNNs), popularized by LeCun et al. (1998) and improved over the following decades. These networks enabled the automation of visual feature learning, making it possible to recognize objects, faces, and movements with accuracy comparable to that of humans. This advance also brought the dilemma of replicating human perception—and, with it, the reproduction of its biases and ethical limitations.

As Russell and Norvig (2010) observe, computer vision transcends mere mathematical coding: it involves the construction of interpretive models of reality. This interpretation, mediated by data, is influenced by the quality, diversity, and representativeness of the training information. Biased data can lead to discriminatory results, distorting automated judgment and, consequently, trust in AI systems.

Furthermore, the evolution of computer vision has expanded into fields such as biometrics, surveillance, medical diagnosis, and public safety. However, the same potential that promotes efficiency and innovation also amplifies ethical risks. According to Suresh and Guttag (2019), the absence of robust ethical guidelines can transform visual algorithms into tools of social control or exclusion when applied without transparency and accountability.

Therefore, understanding the trajectory of computer vision is to understand the genesis of contemporary dilemmas involving the misuse of synthetic images. It is in this context that *deepfakes* emerge —sophisticated products of the fusion between generative learning and visual manipulation —whose ethical and social consequences deserve critical and systematic attention.

3. Generative Neural Networks and the Technical Architecture of Deepfakes

Generative neural networks are emerging as the technological backbone behind *deepfakes,* responsible for synthesizing hyper-realistic images that defy human perception. The fundamental milestone in this advancement occurred with the introduction of *Generative Adversarial Networks.* (GANs), proposed by Goodfellow et al. (2014), which operate through a dual system composed of two models: the generator , responsible for creating fictitious images, and the discriminator , responsible for evaluating the veracity of these images. This competitive struggle allows the system to iteratively improve its ability to simulate reality.
Over the past few years, this technology has evolved significantly, giving rise to variants.

like StyleGAN and ProGAN, capable of manipulating not only faces, but expressions, intonations, and even behavioral aspects with disconcerting precision (Karras et al., 2019). The result is the dissolution of the perceptual boundary between the authentic and the synthetic, inaugurating what many authors call the "era of visual post-truth".

The impact of these generative models goes far beyond facial replacement. Research such as that by Korshunov and Marcel (2018) demonstrates that the sophistication of *deepfakes* allows for the replication of microfacial patterns—subtle muscle movements around the eyes and mouth that, until then, were exclusive to human performance. This ability to reproduce what Paul Ekman (2003) describes as "universal microexpressions" represents not only a technical advance but also a significantly amplified sociopolitical and legal risk. In this sense, it becomes evident that the problem of *deepfakes* is not limited to mere visual falsification, but to the creation of entire narratives capable of shaking legal systems, individual reputations, and democratic processes on a global scale.

Although GANs have historically been the protagonists in the generation of *deepfakes,* recent advances in transformer models *are* taking on a leading role in the field. Architectures such as GPT, Vision Transformers (ViT), and multimodalities like CLIP (Radford et al., 2021) have introduced the ability to integrate vision and language, exponentially expanding the horizons of audiovisual synthesis. This convergence between automated cognitive modalities reveals a new stage of AI, in which forgeries not only mimic appearance but also complex discursive behaviors, linguistic tonalities, and contextual narrative coherence. This means that a *deepfake* ceases to be just a manipulated image and becomes a complete simulacrum of digital identity—endowed with "body," "voice," and "intention."

However, the advancement of these technologies is not without contradictions. There is a fundamental paradox in the very nature of GANs: the more effective they are at creating forgeries, the more necessary it becomes to use the same technology to combat them. This dynamic was described by Nguyen et al. (2019) as the “algorithmic war of trust,” in which defense and attack essentially share the same cognitive raw material: data and statistical optimization. The detection of *deepfakes* therefore depends on understanding the generative mechanisms themselves—

This poses an ethical and strategic dilemma regarding the transparency of the models and access to the databases used in their training.

This duality raises profound questions about power and technological governance. The centralization of generative models in the hands of a few corporations and research laboratories reveals a scenario of global epistemic asymmetry, in which only a minority has the capacity to create and detect forgeries with advanced precision. Researchers such as Crawford (2021) argue that this technological concentration may represent a new power regime—not only economic, but cognitive—capable of defining not only what is true, but what is not.

It is possible to see it as truth. This places *deepfakes* at the epicenter of a dispute that transcends engineering: an ontological war over reality.

4. Algorithmic Ethics and Moral Dilemmas in the Post-Digital Truth Era

Algorithmic ethics emerges as an indispensable discipline in the face of the rise of *deepfakes,* not only as an ex post evaluation tool, but as a structuring principle in the very design of AI systems. Floridi and Cowls (2019) argue that technology has ceased to be neutral and has become morally active, as it acts concretely in the world and therefore demands distributed responsibility among developers, institutions, and society. In this sense, *deepfakes* highlight a critical point: the automation of perceptual manipulation. The ethical harm is no longer restricted to individual privacy or the honor of public figures, but threatens the fundamental civilizational pact based on shared trust in reality. The risk lies not only in the lie, but in the potential for total disbelief: when everything can be false, nothing can be true anymore.

The so-called “ethics of algorithmic ambiguity” (Crawford, 2021) reinforces the need to understand that the effects of *deepfakes* are not limited to factual distortion, but extend to the generation of cognitive landscapes in which systematic doubt becomes a strategic weapon. This concept is close to what Hannah Arendt (1967) described as “the collapse of trust in the public sphere,” where the liar's objective is not to impose a narrative, but to destroy the very possibility of consensus. The contemporary problem, therefore, is not only that of identifying falsifications, but of preserving the civilizational infrastructure of truth.

It is in this context that the debate on algorithmic accountability becomes central. Who should be held responsible for a *deepfake?* The author of the video? The creator of the technology? The platform that disseminates it? The absence of a clear regulatory architecture puts legal systems at a historical lag, as Kuner and Mitsch (2020) point out. The structure of responsibility is fragmented, and regulatory mechanisms tend to react with a delay to events of high social intensity. There is, therefore, a temporal mismatch between technological evolution and institutional resilience—and it is precisely in this interval that the greatest threats emerge.

However, it would be simplistic to treat *deepfakes* as a one-dimensional threat. There are legitimate and highly beneficial uses for these technologies, such as in historical reconstruction, restoration of old films, medical applications, and even in preserving the memory of individuals with degenerative diseases. Vincent et al. (2020) argue that ethics cannot operate through the binary of "permitted or prohibited," but through the structuring of normative environments based on principles such as *transparency, informed consent ,* and *proportionality of risk.* The ethical challenge, therefore, is not to prevent *deepfakes,* but to ensure that they are used within responsible and auditable systems.

Finally, it is worth highlighting that the ethics applied to *deepfakes* cannot be reactive or merely declarative. It must be incorporated into the technical development itself, through...

Frameworks such as “Ethics by Design” (Dignum, 2018) require that moral deliberation be translated into computational decisions from the very beginning of algorithmic engineering. This means that developers must think not only about what the technology can do, but also what it **should do.** to do — shifting the focus of the discussion from "technical ability" to "moral responsibility". Thus, algorithmic ethics becomes less of a belated moral filter and more of a foundational principle for the future of computer vision.

5. Technological Detection and Counter-Attack Models against Deepfakes in Computer Vision

Deepfake detection has become one of the most strategic and challenging fields in contemporary computer vision. Unlike other digital threats, such as malware or phishing, identifying audiovisual forgeries requires systems capable of analyzing subtleties.
physiological, biomechanical, and even behavioral. Pioneering research such as that by Matern et al. (2019) showed that early *deepfakes* exhibited flaws in microvisual patterns, such as irregularities in blinking or imperfect synchronization between speech and lip movement. However, these weaknesses were quickly corrected by the most recent iterations of generative models, initiating a kind of "algorithmic arms race" between creation and detection.
The main difficulty lies in the fact that the opponent evolves at the same speed — or sometimes faster — than the defense systems.

In this context, detection models based on specialized deep neural networks are emerging.
A widely used approach is the use of **autoencoders,** trained to reconstruct real human faces and identify subtle variations when exposed to synthetic images. Other strategies include the analysis of **spectral artifacts** invisible to the human eye, such as imperceptible variations in the frequency histogram or compression patterns resulting from artificial rendering. Marra et al. (2019) highlight that detection based on spectrograms and Fourier transforms has shown promise, especially in videos where vocal manipulation is combined with visual manipulation. This type of approach reveals a new paradigm: combating forgeries not visually, but mathematically—in latent domains invisible to human perception.

However, the major leap in detection strategies occurs with the use of **multimodal models,** capable of simultaneously analyzing audio, video, and even semantic speech patterns. Studies such as those by Mittal et al. (2020) demonstrate that integrating linguistic analysis with visual analysis allows for the detection of inconsistencies between *how* something is said and *how* the face behaves when expressing it. This cross-modality approach yields superior results compared to those that analyze only one dimension of the content. Furthermore, models based on self-supervised learni
Trained on large volumes of unlabeled videos, these tools began to be used to detect "unnatural" behaviors emergently, without relying on previously annotated databases.

However, a structural challenge remains: the need for operational scalability. It is ineffective to require that detection systems be applied only by specialized institutions, as the massive dissemination of *deepfakes* extends to social platforms and decentralized communication environments. Therefore, efforts are underway to incorporate detection algorithms directly into media platforms, operating invisibly and continuously.

Researchers such as Verdoliva (2020) argue that detection should occur *within the distribution infrastructure pipeline,* not after dissemination — a transition that transforms *deepfake* detection from a reactive tool to a preventive tool.

Finally, it is necessary to recognize that detection is only part of the solution. Even when a *deepfake* is correctly identified, reputational, political, or emotional damage may already have been done. This reinforces that technology alone is not capable of solving the problem: it must be combined with information governance mechanisms, media literacy, and regulatory compliance. The fight against *deepfakes* is, therefore, a challenge that is not limited to engineering, but requires a systemic, interdisciplinary, and anticipatory response.

6. Social, Political, and Geopolitical Impacts of Deepfakes on the Infrastructure of Public Trust

Deepfakes represent not only a technological threat, but above all a structural disruption in the social ecosystem of trust. Trust has always been the pillar that sustains human interactions, whether interpersonal, institutional , or media-based. However, when visual authenticity becomes questionable, the entire symbolic architecture of truth—which previously depended on audiovisual evidence as "irrefutable proof"—collapses. Chesney and Citron (2019) call this phenomenon "The Liar's Dividend": it is not only the falsification that threatens public order, but the fact that, faced with widespread doubt, any real individual can claim to be a victim of manipulation—including those guilty. The consequence of this is paradoxical: the more advanced the detection, the greater the room for maneuver for liars.

From a political standpoint, *deepfakes* represent an unprecedented vector of democratic destabilization. In electoral contexts, they can be disseminated en masse seconds before elections, exploiting emotional and cognitive mechanisms more quickly than the institutional capacity to refute them. Studies by Vaccari and Chadwick (2020) demonstrate that fake videos with strong emotional appeal are consumed and shared 70% more intensely than neutral content—even when users suspect their authenticity. This reveals that *deepfakes* operate not only at the level of deception, but also at the level of passion. They affect the field of perception, not argumentation. And this leads governments, courts, and journalistic systems to a time crisis—the truth always becomes reactive and delayed.

In the geopolitical sphere, *deepfake* technology is already being treated as a strategic tool of information warfare. Nation-states, business conglomerates, and even extremist groups have understood its potential for silent destabilization. Synthetic manipulation can be...

Used not to explicitly convince, but to generate widespread doubt—rendering entire societies inoperative in their capacity for discernment. Martínez-Pérez (2021) describes this phenomenon as “high-precision cognitive disorientation,” a modern and more sophisticated form of what is conventionally called *propaganda.* Power ceases to be the imposition of a truth and becomes the destruction of the shared notion of reality.

However, the impacts are not limited to macropolitical spheres. The legal and medical fields also face serious implications. There are already documented records of *deepfakes* being used for extortion, corporate fraud, and the destruction of private reputations. Citron (2020) warns of the exponential growth of *deepnudes* —non-consensual pornographic forgeries— affecting mainly women and adolescents, this is symbolic violence with irreversible psychological, social, and even financial impacts. In terms of mental health and the right to privacy, the threat of *deepfakes* is comparable to psychological weapons of identity destruction.

Ultimately, the greatest threat may not be the false content itself, but the erosion of collective trust. When the entire population begins to distrust everything, states enter a state of civic paralysis and markets collapse due to a lack of public predictability. This informational entropy has the potential to corrode democracies from within—silently. As Harari (2018) states, the 21st century will not be dominated by territorial wars, but by wars of perception. And *deepfakes* are not just another tool in this scenario—they are the perfect weapon.

7. Future Projections, Regulation, and Ethical Guidelines for the Responsible Development of Generative Artificial Intelligence

The future trajectory of computer vision and synthetic image generation technologies points to a scenario of even greater sophistication, in which the distinction between human and artificial may become ontologically irrelevant. The emergence of models such as *foundation* models— Capable of operating in multiple cognitive modalities simultaneously—suggests that future generations of *deepfakes* will not only mimic appearances and behaviors, but will also construct coherent and contextually adaptable narratives in real time. Researchers such as Bommasani et al. (2021) argue that generative AI is advancing from an imitative function to a **strategic creative function,** capable of projecting content that is no longer derived from the past, but anticipatory. In this sense, the ethical struggle ceases to be against the falsification of the past and becomes against the fabrication of the future.

Given this, international regulation emerges as a civilizational imperative. Isolated attempts at national legislation, such as the early European guidelines present in the *AI Act,* while significant, prove insufficient to address a phenomenon that transcends legal boundaries and moves at algorithmic speed. Authors such as Kuner and Mitsch (2020) argue that traditional normative systems, based on ex post punishments, need to be superseded by **predictive governance** frameworks capable of intervening preventively before the damage occurs.

Let's make this concrete. This leads us to the need for regulation based on *distributed accountability,* where platforms, developers, public institutions, and users share explicit and auditable responsibilities.

In parallel, research and industry organizations are beginning to adopt principles such as *AI Ethics by Design* and *Internalizable Computational Transparency* (Dignum, 2018), in which ethics is not a peripheral element, but a programmatic mechanism integrated into the very development of the technology. This conception establishes that generative systems must contain built-in mechanisms for verification, traceability, and intervention—ensuring that no model operates as a “sovereign black box” immune to public scrutiny. Furthermore, there is a growing demand for **irremovable digital watermarks** *(neural watermarking),* which do not prevent creation but ensure authentication of origin and unequivocal accountability for authorship.

Another essential development lies in raising global media literacy, not as mere “critical consumption,” but as **cognitive education for the age of simulation.** This implies forming citizens capable of understanding that human sensory perception is no longer an absolute guarantee of truth, shifting trust from “seeing is believing” to “verifying is trusting.” This is an epistemological shift that redefines not only the use of technology, but the very formation of contemporary consciousness. As Crawford (2021) points out, the struggle of our time is not only about informational security, but about the right to continue being interpretive beings in a hyper-authorial world.

In short, the future of computer vision depends on humanity's ability to govern it before it governs us. Technology itself is not the enemy—the risk lies in its dissociation from the civilizational project. The necessary response is clear: to develop artificial intelligence that not only produces efficiency but also preserves dignity; that expands human capacity without supplanting its autonomy. If deepfakes *represent* the pinnacle of simulation, it is up to ethics to ensure that truth is not extinguished—but rather evolved.

Conclusion

The rise of *deepfakes* constitutes not only a technological phenomenon, but a civilizational watershed that repositions the relationship between truth, perception, and power in the 21st century. Through in-depth analysis of the fundamentals of computer vision, generative architectures, and the resulting ethical and political dilemmas, it becomes evident that we are not only facing a digital threat, but an ontological transformation of the very notion of reality. The ability to synthesize people, narratives, and events with high precision inaugurates the post-authenticity era, in which public trust—a structuring element of democracy, law, and human relations—becomes the scarcest and most vulnerable resource.

This study demonstrated that detecting *deepfakes* is technically possible, but its isolated effectiveness is insufficient to contain the socio-cognitive effects amplified by their dissemination. The solution requires an integrated response ecosystem, in which four pillars converge.

Fundamentals include: **technology,** to identify and prevent algorithmic manipulation; **international regulation,** to ensure distributed accountability and normative standardization; **advanced media literacy,** to train citizens capable of critically interpreting digital reality; and above all, **algorithmic ethics embedded in the very design of AI,** ensuring that innovation is guided by principles before being conditioned by consequences.

It can be concluded, therefore, that *deepfakes* represent not only a technical risk, but an ethical and civilizational test. The central question is not whether we will be able to detect the falsehood of the world, but whether we will be able to preserve the moral integrity of truth as a collective value. Artificial intelligence will continue to evolve—it remains to be seen whether humanity will be able to evolve along with it, not only in computational capacity, but in ethical responsibility. The future of computer vision will depend less on the quantity of data it processes, and more on the quality of the values it preserves.

References

ARENDT, Hannah. *Truth and Politics.* New York: The New Yorker, 1967.

BOMMASANI, Rishi et al. *On the Opportunities and Risks of Foundation Models.* Stanford University, 2021.

CAPPELLETTI, Ivan. *The post-photographic image: simulation and truth in the algorithmic age.* Famecos Magazine, São Paulo, 2020.

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

CITRON, Danielle. *Sexual Privacy.* Yale Law Journal, 2020.

CRAWFORD, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven: Yale University Press, 2021.

DIGNUM, Virginia. *Ethics in Artificial Intelligence: Introducing the Ethics by Design Framework.* Ethics and Information Technology, 2018.

EKMAN, Paul. *Emotions Revealed.* New York: Times Books/Henry Holt, 2003.

FLORIDI, Luciano; COWLS, Josh. *A Unified Framework of Five Principles for AI in Society.* Harvard Data Science Review, 2019.

GOODFELLOW, Ian et al. *Generative Adversarial Nets.* In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montréal, 2014.

HARARI, Yuval Noah. *21 Lessons for the 21st Century.* London: Jonathan Cape, 2018.

KARRAS, Tero et al. *A Style-Based Generator Architecture for Generative Adversarial Networks.* IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

KORSUNHOV, Pavel; MARCEL, Sébastien. *Deepfakes: A New Threat to Face Recognition?* IEEE International Conference on Biometrics Theory, Applications and Systems, 2018.

KUNER, Christopher; MITSCH, Judith. *AI Regulation in a Transnational Context.* European Data Protection Law Review, 2020.

MARR, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: WH Freeman, 1982.

MARTÍNEZ-PÉREZ, M. *Algorithmic Geopolitics and Informational Warfare.* Journal of Strategic Analysis, 2021.

MARREN, Irene et al. *Exposing Deepfake Videos by Detecting Eye Blinking.* IEEE Conference on Computer Vision Workshops, 2019.

MITTAL, S. et al. *Emotional Reasoning in Deepfakes Detection via Multimodal Networks.* AAAI Conference on Artificial Intelligence, 2020.

NGUYEN, T. et al. *Adversarial Machine Learning in the Age of Deepfakes.* Journal of Machine Learning Research, 2019.

RADFORD, Alec et al. *Learning Transferable Visual Models from Natural Language Supervision.* OpenAI, 2021.

RUSSELL, Stuart; NORVIG, Peter. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River: Prentice Hall, 2010.

SURESH, Harini; GUTTAG, John. *A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle.* Conference on Fairness, Accountability, and Transparency, 2019.

VACCARI, Cristian; Chadwick, Andrew. *Deepfakes and Disinformation: An Experimental Test.* New Media & Society, 2020.

VERDOLIVA, Luisa. *Media Forensics and Deepfakes: An Overview.* IEEE Signal Processing Magazine, 2020.

VINCENT, Nicholas et al. *Ethical Applications of Synthetic Media.* ACM Conference on Fairness, Accountability, and Transparency, 2020.