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

Year IV, v.2 2024. | Submission: 12/16/2024 | Accepted: 12/18/2024 | Publication: 12/20/2024

Ethical Training of Neural Networks: Datasets, Bias, and Privacy in Detection of Deepfakes

Ethical Training of Neural Networks: Datasets, Bias, and Privacy in Deepfake Detection

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

Automated deepfake detection, based on the massive use of deep neural networks, has become a civilizational imperative for protecting informational integrity on digital platforms. However, the most sensitive point of this architecture lies not in the algorithmic sophistication of the model, but in the ethical formation of the datasets used in its training—which are frequently marked by structural biases of race, gender, and geopolitics, as well as by privacy violations and non-consensual use of human biometric images. This article critically analyzes, with an epistemological and technopolitical approach, the fundamental ethical dilemmas associated with the creation and curation of datasets to train deepfake detectors.

Based on contemporary studies, this paper explores how errors in dataset composition can reproduce historical forms of algorithmic oppression, amplify social inequalities, and even compromise the democratic legitimacy of technologies used to combat disinformation. It concludes by proposing ethical and technical guidelines for the development of responsible, auditable training pipelines that are compatible with principles of informational sovereignty and algorithmic justice.

Keywords: deepfakes; algorithmic ethics; biases; digital privacy; informational sovereignty.

ABSTRACT

The automated detection of deepfakes through large-scale neural networks has become a civilizational necessity for protecting informational integrity in the digital environment. However, the most critical point in this architecture does not lie in the model itself, but in the ethical formation of the datasets used during training — which are frequently affected by structural biases related to race, gender, and geopolitics, as well as by potential violations of privacy and non-consensual biometric data extraction. This article develops an epistemologically rigorous analysis of the ethical dilemmas embedded in dataset construction for deepfake detection models, examining how flawed curation can reinforce algorithmic oppression, reproduce historical inequalities, and ultimately endanger democratic legitimacy. It concludes by proposing an ethical framework and engineering guidelines for responsible and auditable dataset governance, anchored in informational sovereignty and computational justice.

Keywords: deepfakes; algorithmic ethics; dataset bias; digital privacy; informational sovereignty.

1. INTRODUCTION: THE ETHICAL PARADOX IN THE ALGORITHIC WAR AGAINST DISINFORMATION

Automated deepfake detection, powered by deep neural networks and large-scale computer vision pipelines, has globally established itself as one of the main defense infrastructures against fifth-generation information manipulation. However—and this is the most neglected point in the public debate —the war against deepfakes is not a war between truth and lies, but between adversarial artificial intelligences, both capable of operating at speeds and granularities superior to human cognition. Therefore, the true ethical core of the problem lies not only in identifying false content, but in questioning the power bases and moral criteria that structure the training of the models responsible for declaring reality. This means that a poorly trained deepfake detector—with biased, racially asymmetrical datasets, obtained without legitimate consent or restricted to hyper-specific demographic segments —can produce a new architecture of algorithmic oppression, masked under the rhetoric of social protection. Detection, therefore, is not neutral: it is a political act of the first degree.

The formation of datasets used to train deepfake detectors is now recognized as one of the most critical and dangerous points in the entire information security pipeline. This is because the models do not learn "what is true," but only "what is statistically frequent" within the data provided to them. In other words, an AI does not detect falsehood—it detects the difference in relation to the learned pattern. If this pattern was formed mainly with white European male faces, as occurred with several initial datasets such as FaceForensics++, the model becomes significantly less effective at detecting manipulations in female, Black, Indigenous faces, or those belonging to populations outside the Anglo-Saxon geopolitical axis. This problem, documented by Buolamwini and Gebru (2018) in the field of facial recognition, when applied to deepfake detection, can create a scenario in which certain ethnicities become more vulnerable to manipulation and others more rigidly policed—amplifying historical inequalities within an algorithmic environment.

Beyond demographic bias, there is the problem of privacy and the non-consensual appropriation of human images, often collected silently from social networks and platforms.

from leaked videos or databases, without any authorization from the individuals involved. Detecting deepfakes inherently requires the massive ingestion of real facial data for training—which poses an explosive ethical dilemma: to defend people against visual manipulation, are we violating their very autonomy over their biometric identity?

Authors such as Floridi (2021) warn that the boundary of human dignity is being silently renegotiated in the field of computer forensics: the defense of truth cannot become a justification for total biometric surveillance. Without clear structures of consent, data governance, and sovereignty over facial representations, the very "protective" AI can function as an extractive machine of human identities.

The problem worsens as large platforms and governments begin to claim a monopoly on automated forensic infrastructure, centralizing in a few entities the power to define—in real time—what is legitimately or falsely allowed to circulate in the digital public sphere. Reports from NATO (2021) and the European Commission (2022) already acknowledge that the informational sovereignty of the 21st century depends directly on who controls the deepfake detection infrastructure. But here lies the danger: whoever controls the datasets controls the automated awareness of truth. A model trained with biased data cannot...

It only errs technically—it errs politically. It can reinforce existing power structures, block dissenting narratives, and selectively authorize the circulation of that which serves specific geostrategic interests. Detection technology, if poorly governed, can threaten more than it protects.

Given this scenario, it is clear that the discussion about deepfakes cannot be reduced to a technical or operational dispute. The problem is fundamentally epistemological and ethical, as it involves the transfer of ontological authority over reality—previously mediated by human perception, journalism, science, or the legal process—to automated classification systems trained on databases invisible to society. Datasets thus become invisible mirrors of power, encapsulating not only statistics but also implicit ideological views of what is "normal," what is "acceptable," and what is "real." Without rigorous governance, the algorithm that detects deepfakes can reinforce precisely the colonial, patriarchal, and centralizing logic that contemporary society should overcome. Ethical training, therefore, is not optional—it is a prerequisite for any detection technology to be morally valid.

If algorithmic deepfake detection claims to work in the name of truth, it must be subject to the dual principle of justice and public auditability. This inevitably implies the requirement that datasets be documented, verifiable, representative, geopolitically distributed, and compiled with the informed consent of the people involved. What is at stake is not only technical efficiency, but civilizational legitimacy: a model that

It detects deepfakes with 99% accuracy, but operates with racial or geopolitical biases, or extracts data without consent; it is not technically advanced—it is morally flawed. Society has the right to demand accountability, cross-cutting governance, and protection against structural abuses. in the name of any promise of digital security.

Thus, this first point establishes the central thesis of this article: the fight against deepfakes is only legitimate if it is simultaneously technically effective, democratically auditable, and morally acceptable, which requires that the discussion begin not with the architecture of the model, but with the total ethical integrity of the dataset's formation. The ultimate question, therefore, is not only "how do we detect deepfakes?", but "in whose name, under what criteria, and based on what human images do we train the intelligence that will decide what is real?" Detection, when poorly oriented, can become what it intends to combat: a mechanism for distorting reality, legitimized by superior algorithmic authority. Technology can only be a solution if it is, first and foremost, morally just from its origin—and that origin is the dataset.

2. Ethical and Structural Foundations of Datasets in Deepfake Detection

The construction of datasets to train deepfake detectors involves decisions that are not only technical but profoundly civilizational, as they determine in advance what artificial intelligence will understand as normal, anomalous, legitimate, dangerous, or irrelevant—even before being used in any real-world environment. This means that each data curation is an act of power, selecting which human bodies will become statistically visible and which will be made invisible; which facial expressions will be considered neutral behaviors and which will be treated as noise; which cultures will have their physiology represented and which will be treated as exceptions. This logic is denounced by Crawford (2021), who states that contemporary datasets are “ideological infrastructures disguised as facts.” In the case of deepfakes, the gravity multiplies, as the detectors not only describe reality—they will judge it with forensic authority, defining which visual evidence can or cannot be trusted in judicial, electoral, or geopolitical contexts. Contrary to popular belief, a dataset is not a mirror of reality—it is an epistemological construct that shapes how reality itself will be recognized by AI.

Historically, a large portion of the datasets used for training visual models have been formed from collections centralized in the Euro-North American axis, generating a statistically hypertrophied representation of white male faces, lit under Western photographic standards, with neutral expressions that conform to specific cultural patterns. Datasets such as LFW, CelebA, MegaFace, and even FaceForensics++ have been criticized by experts for repeating the Eurocentric homogeneity of the mainstream internet, almost entirely excluding African, Indigenous, South Asian phenotypes, mixed-race facial features, or non-hegemonic physiological morphological variations. The result is devastating: models

Systems trained on such foundations detect deepfakes with high accuracy only in dominant identities, while leaving dangerous loopholes for manipulation targeting non-European individuals—exposing historically at-risk populations to new types of silent digital abuse. It is no exaggeration to say that, in the algorithmic war for truth, those absent from the dataset are absent from protection.

The problem is not merely quantitative—it's ontological. Datasets define what is considered a "legitimate human face," what is computationally recognizable as a "vital sign," what is treated as a "valid microexpression," and what is discarded as "irrelevant noise." This normative framework directly affects the ability of forensic models to detect authentic physiological signatures (such as dermal pulsation, pupillary dilation, or involuntary microcontractions) when these manifest with physiological patterns different from the Caucasian parameters that saturate Western datasets. This generates a perverse paradox: AI can classify a real face from an underrepresented culture as "suspicious"—while simultaneously considering authentic a deepfake that correctly simulates only the dominant biometric pattern. The dataset, therefore, not only statistically biases AI—it can induce a false sense of scientific certainty, technically sophisticated but morally distorted from its origin.

Another inescapable ethical dilemma lies in the origin of the facial data used for training. The overwhelming majority of faces used in global datasets were collected without explicit, informed, and legally sound consent from the individuals involved—often extracted from public social networks, audiovisual libraries, streaming platforms, and even criminal leaks.
The logic of "what's online is available" has become perversely normalized by the AI industry, resulting in the creation of datasets that, while technically powerful, incorporate a structural violation of digital autonomy, using human faces as disposable raw material. In extreme cases, as denounced by researchers at the AI Now Institute (2021), the faces of deceased individuals were used to train facial spoofing detectors without any consultation with families, demonstrating that the ethical limits of biometric extraction have been surpassed with industrial ease. Deepfake detection cannot be "protective" if it stems from the invasive collection of another's identity.

The problem becomes even more complex when placed within the geopolitics of data. China has created official datasets with over a billion faces to train advanced biometric detection and classification models—almost all of which were obtained domestically under authoritarian legislation.
The United States, in turn, operates under the opposite model: highly fragmented, private, and non-standardized datasets, with a high risk of corporate monopolization of the infrastructure of truth. Europe attempts to impose regulatory ethics, but lacks the computational power and population scale of its competitors. The result is a critical triad: whoever controls the dataset controls the AI; whoever controls the AI controls the detection; whoever controls the detection controls digital reality. In other words, the dataset is ground zero of techno-civilizational power. Discussing it is discussing informational sovereignty—and not just technical tools.

Furthermore, the absence of governance over datasets allows private corporations and authoritarian regimes to use deepfake detection as a geopolitical weapon, not as an instrument of public protection. It is already technically possible for a government to selectively classify true content as false to destroy reputations—or for private platforms to silence political narratives under the pretext of "algorithmic security." When the dataset is hidden from society, democracy loses power over the definition of "valid visual truth." Herein lies the ultimate danger: a world in which the perception of reality is mediated by models that cannot be audited, challenged, or replaced. The ethical battle, therefore, begins not in the source code, but in the radical curation of the human raw material used to manufacture it.

Therefore, training datasets for deepfake detectors represent the area of greatest risk and greatest power in the contemporary information security ecosystem. Controlling them requires applied ethics, critical information science, philosophy of technology, and a robust international legal framework. This article argues that there is no legitimate technology for protection against deepfakes that is not accompanied by explicit ethical governance of dataset collection, curation, and auditing. Without this, any model—however technically impressive—will be a simulacrum of protection built upon silent violation and unconfessed structural bias.

3. Algorithmic Bias and Sociotechnical Disparity in Deepfake Detection

The algorithmic bias present in deepfake detectors is not an isolated side effect, but a direct consequence of how datasets are structured—and, above all, of the historical logic that defines which bodies are considered normative and which are treated as statistical exceptions. This means that AI does not fail "by accident" in certain groups; it fails because it has been trained to reproduce a pre-existing structure of selective visibility of the world. Studies conducted by Buolamwini and Gebru (2018), initially in facial recognition systems, demonstrated that market-leading models achieved 99% accuracy for white men, but less than 70% for black women, evidencing an algorithmic hierarchy of human importance. This asymmetry has also been observed in deepfake detection models, especially in the first public datasets, in which Caucasian faces received more than 70% of the representation—and therefore, were the most protected. The technology, therefore, does not protect equally—it protects first those who are statistically dominant.

This phenomenon creates a paradox of structural injustice: populations that have historically been more exposed to visual and representational violence—women, Black people, Indigenous peoples, politically oppressed groups—are precisely the most vulnerable in the context of deepfakes, whether due to greater exposure to the attack or due to late or ineffective detection of manipulations carried out against them. There are cases documented by the European Digital Media Observatory (2021) in which racialized women suffered deepfake pornographic attacks.

But corporate detectors classified the content as "natural" simply because there were no similar examples in the training dataset. This means that the real risk is not just technical—it's moral and structural. A technology sold as "protective" may be invisibly operating as a mechanism for the continuation of historical algorithmic violence.

Asymmetry also manifests itself in the geopolitical dimension. The vast majority of datasets used to train modern detectors are Westernized and monolingual, with very low representation of contexts from the Global South, such as Africa, Latin America, or Southeast Asia. This generates the effect described by Milan and Treré (2019) as data colonialism, in which "computational truth is defined in hegemonic hemispheres and imported as a global civilizational standard." Consequently, the same models perform highly in detecting deepfakes involving American public figures, but fail miserably with African or Brazilian indigenous politicians. As a result, information security becomes profoundly asymmetrical between civilizations: AI detects more and better what is of interest to the powers with central infrastructures. This proves that datasets are not just sets of images—they are strategic projections of global power.

Another critical phenomenon is the so-called AI-driven confirmation bias. Detection models may be inclined to detect manipulations associated with previously stigmatized ideologies, ethnicities, or political stances with greater accuracy—amplifying belief systems already present in society. In other words, AI is not neutral—it can function as an enlarged mirror of invisible cultural biases, disguised as mathematical objectivity. In extreme scenarios, this could allow the use of detectors as tools for the selective suppression of inconvenient narratives, under the pretext of "forensic risk," "anomaly," or "suspicious content." "Automatic truth" ceases to be merely technology—it becomes an infrastructure of interpretive power.

Combating bias, therefore, cannot be done solely through belated statistical adjustments—such as “mathematical balancing” or “compensatory weights” in the models. The root of the problem demands radical ethical reform at the origin of the dataset, guaranteeing multiracial representativeness, real physiological diversity, gender balance, and the deliberate inclusion of underrepresented social compositions. The ethics of training, therefore, is not an optional department—it is a pillar of the cognitive sovereignty of democracy. If automatic detection is to be the filter of reality, then reality needs to be fully represented from the model's formation.

4. Privacy, Consent, and Biometric Expropriation in AI Training

If there is one point where deepfake detection reveals its most serious civilizational contradiction, it lies at the border between informational protection and massive violation of human privacy. To train neural networks capable of recognizing hyper-realistic visual manipulations, it is necessary to feed them colossal amounts of real facial data, recording expressions, microvascular patterns, pupillary reflexes, bone geometry, and thousands of other physiological variations that...

These elements constitute bodily individuality. The ethical problem arises from the fact that the vast majority of this data is collected without any explicit and informed consent, being silently extracted from social networks, public audiovisual databases, criminal leaks, and even journalistic platforms—often under the tacit logic that "if it's on the internet, it's technically available." In other words, models trained to "defend identities" typically begin by violating identities.

This phenomenon has been described by experts as silent biometric expropriation, a new form of digital colonialism in which human faces are transformed into computational raw material without consultation, compensation, or governance. Unlike generic images, faces carry not only a person's identity but also their existential possibility in the public space—their capacity to be recognized or confused, celebrated or destroyed. Stealing a face is not copying a photo—it is hijacking the right to digital existence itself. And when the same dataset is used to train surveillance and deepfake detection systems, the boundary between protection and control dangerously dissolves. What prevents a "protective" model from becoming, in the same second, a mechanism for mass tracking or prior censorship?

The ethical dimension is aggravated because we are not talking about privacy in the trivial sense (personal taste, preferences, location), but about bioprivacy—that is, the legitimate ownership of bodily signals that cannot be replaced. Passwords can be changed; faces cannot.

Therefore, any system that captures facial features, even under the pretext of security, assumes an irreversible structural power of domination if it is not governed by a legal and ethical protocol of sovereignty. An environment where companies or governments can extract, manipulate, and train models with facial data without authorization is an environment where the human body ceases to be a person and becomes a computational resource.

In this context, the absence of active, specific, and auditable consent ceases to be a legal flaw—it becomes an ontological violation. Following Luciano Floridi's thesis (2020), the digitized body becomes an informational metaphenomenon of human dignity—and any non-consensual extraction of its representation constitutes metacognitive aggression, even if allegedly "to protect it." To assert that "detection technology is necessary to protect democracy" is only true if this protection is not built upon the silent and irreversible violation of the citizen's own rights.

Therefore, this article firmly argues that no deepfake detection model can be ethically validated if its base dataset violates fundamental rights to identity and privacy.

Informational self-determination. The war against the manipulation of reality cannot be won at the cost of destroying the individual's sovereignty over their own digital body. The ethics of consent and governance of biometric data is not a technical detail—it is a definitive civilizational frontier.

5. GUIDELINES FOR ETHICAL TRAINING AND RESPONSIBLE GOVERNANCE OF DATASETS

Given the risk that deepfake detectors could simultaneously become weapons of protection and oppression, the crucial question becomes: how to design and train neural networks based on verifiable, transparent, and auditable ethical principles on a global scale? The answer cannot lie in generic “technical adjustments,” but in governance protocols radically guided by informational justice and biometric sovereignty.
This implies that dataset curation is not a byproduct of engineering—it is a first-order political responsibility, comparable to organizing elections, authenticating documents, and the constitutional protection of fundamental rights. Therefore, it must follow non-negotiable principles.

The first of these principles is that of mandatory demographic and phenotypic representativeness. A dataset that excessively concentrates on European, male, adult faces, normalized by Western aesthetics—is doomed to technical and political injustice from its inception. Therefore, deepfake detection models need to be trained on geographically distributed databases, with broad inclusion of Afro-diasporic, Indigenous, Asian, Arab, Latino, and diverse gender, age, and identity expression. This is not about "decorative diversity," but about ensuring that AI recognizes humans in all their forms, and not just the dominant stereotype. Algorithmic protection will only be democratic if the human experience is fully represented in its training source.

The second principle is that of explicit, granular, and revocable biometric consent, supported by auditable mechanisms for traceability of origin. This means that datasets must be built with the conscious participation of individuals, with a documented right to later withdrawal, and with usage models that must be legibly explainable to any citizen, not just AI specialists. The silent extraction of faces from public social networks is ethically invalid and civilizationally dangerous, even if technically "efficient." A system that promises to protect the user cannot begin by irreversibly violating their autonomy.

The third principle is the mandatory adoption of a documentation and auditability protocol—along the lines of *Datasheets for Datasets* (Gebru et al., 2018)—in which each dataset must explicitly declare its origin, demographic composition, sources, purposes, limitations, and potential risks. This model prevents datasets from becoming geopolitical black boxes, invisible to society, protected by corporate clauses. Documentation should not be an “optional standard”—it should be a legal and international requirement, like a mandatory informational passport for any system that interferes with the public perception of reality.

The fourth principle is multidisciplinary participation before coding. The definition of what constitutes "bias," "acceptable," "fraud," and "normal" cannot be defined exclusively by engineers or corporate investors. It is necessary for philosophers, human rights experts, jurists, communication specialists, and representatives of vulnerable groups to be formally involved in the dataset definition phase, and not just in the final reports. Deepfake detectors are not neutral tools—they are infrastructures of perceptual power. Therefore, they must be...

designed using processes that represent the true plurality of society.

The fifth principle is that of global ethical interoperability—so that detection protocols do not become geopolitical weapons disguised as democratic protection. This requires international standardization, with verifiable rules that prevent the unilateral use of technology by authoritarian governments or monopolistic corporations, as well as the execution of independent audits at a civilizational level. Digital truth cannot be the property of single players.

These principles, taken together, establish the foundation for what this article argues is the structural pillar for ethically legitimate Artificial Intelligence in the fight against disinformation: the radical governance of datasets as an absolute condition for informational justice.

6. Auditing, transparency, and algorithmic sovereignty as non-negotiable axes.

The creation of automated deepfake detectors represents a turning point in the history of digital information governance. While the battle for truth previously took place in the interpretive sphere —between journalists, institutions, testimonies, and public consensus—it now migrates to an infrastructural and algorithmic layer, where decisions are made in milliseconds, often before any human being has access to the content. This shift is so radical that it transforms automated detection not only into a security tool but into a technology of epistemic sovereignty, capable of determining what "comes into existence" in the public sphere and what is intercepted before circulation. Therefore, the fundamental question is not only "does AI detect correctly?", but rather "who controls the AI that detects—and under what visible and verifiable rules?". Without mandatory auditing, the fight against deepfakes could become the largest experiment in invisible algorithmic censorship in history.

Auditability needs to be understood as a democratic right and a civilizational principle, not as an optional resource for corporate honesty. This implies making mandatory the creation of independent and international mechanisms for the continuous review of the performance, bias, and potential abuses of deepfake detectors—with participation not only from a technical perspective, but also from philosophical, legal, journalistic, and representative bodies of groups vulnerable to digital oppression. Algorithms that operate as "guardians of truth" cannot be judged solely by their mathematical accuracy, but by their ethical compliance with fundamental human rights, including freedom of expression, informational integrity, privacy, and the right to public rebuttal.

There is no legitimate transparency without full documentation of datasets, model decision rules, and fallback policies (that is, what happens when the AI is in doubt or detects risk). In other words, it is as important to know why a video was blocked as it is to know why another was approved. The danger lies precisely in the asymmetry: a model can be deliberately trained to "fail for some, protect others." This is why researchers like Brundage and Whittlestone (2020) advocate for the creation of algorithmic accountability systems with inviolable logs—auditable by civil bodies—

Similar to how aviation black boxes work. Without this, any detection architecture can be programmed to appear fair—while silently perpetuating injustices.

The issue becomes more complex when the geopolitical dimension of the problem is recognized. Whoever controls automatic detection controls the most powerful weapon of the information age: the power to authorize the public existence of a truth. If only the US, China, or private conglomerates in Silicon Valley control the world's visual validation pipelines, then countries in the Global South will be reduced to passive consumers of imposed computational epistemologies, unable to challenge false positives, false negatives, or selective censorship. This creates the risk of algorithmic colonialism, where a real video produced in Brazil can be automatically blocked by a model trained in the US that does not recognize Brazilian expressive patterns as authentic. Without algorithmic sovereignty, there is neither national sovereignty nor informational sovereignty.

Thus, this article defends the principle of distributed algorithmic sovereignty as a strategic imperative. This means that detection models should be governed by public and multilateral consortia, with explicit audit rights for universities, the free press, and human rights organizations. Detectors should not operate as "proprietary oracles," but as auditable infrastructures of public trust, with institutionalized challenge mechanisms, accessible logs under controlled legal criteria, and clear protocols for correction and review. AI cannot have the final say on reality—it must be subject to human challenge guaranteed by law.

More than that: the real defense of society lies not only in AI that detects deepfakes—but in society's right to verify the AI that detects deepfakes. In a context where even protection can become a weapon, true power lies in radical transparency, distributed control, and the elimination of opaque privilege over the definition of reality. No democracy is safe if visual truth is outsourced to inaccessible and unquestionable machines.

Therefore, auditing, transparency, and sovereignty are not optional complements to deepfake detection —they are the minimum conditions for any such system to even be morally permissible to exist.

Without auditable governance, all defense technology becomes a potential risk of infrastructural tyranny. This is the tipping point: the war for truth will not be won with stronger algorithms—but with fairer algorithms, radically auditable and politically subordinated to the principle of human dignity.

7. SOCIAL, POLITICAL IMPLICATIONS AND POSSIBLE FUTURES FOR AN ETHICAL DETECTION OF DEEPFAKES

The ethical confrontation of algorithmic disinformation requires understanding that the deepfake crisis is merely the most visible manifestation of a deeper problem: the progressive replacement of human mediation by automated decision-making in truth validation structures.

Automatic detectors are not neutral — they are political systems disguised as technology — and the

The way they are designed, trained, and audited will determine whether we live in an informational democracy or in a regime of epistemic technocracy. History shows that each new communication infrastructure reconfigures power. If the 20th century was marked by the struggle for control of the media, the 21st century will be defined by the struggle for control of the intelligence that decides what is legitimate information. In this scenario, the ethics of datasets and neural networks is not an appendage of governance —it is the very arena where the fate of human cognitive freedom is being decided.

Contemporary society lives with the paradox of depending on technologies it does not understand, but on which it relies to maintain its own security. This dependence creates the risk of what Floridi (2020) calls "algorithmic heteronomy"—the irreversible delegation of moral judgment to automated systems. If deepfake detection is treated as a technical monopoly, it will produce not only new information asymmetries but also new ontological inequalities, in which different human groups will have different degrees of access to public trust. It is the silent return of a digital epistemic caste, formed by those who hold the verification infrastructure. The ethical defense against deepfakes, therefore, cannot be limited to the realm of AI—it must reconfigure the very relationship between technology, truth, and democracy.

In this context, the need arises to incorporate algorithmic education as a public policy. Citizens need to understand what a neural network is, how a dataset is formed, and what the limits are of what an AI can decide. This implies a new type of literacy, not just digital, but epistemic, so that individuals can consciously participate in the decisions that shape their own informational reality. Without this foundation, algorithmic power will remain concentrated in the hands of a few, and the discourse of "protection against deepfakes" will only serve to consolidate new forms of control. Technical transparency must be accompanied by citizen participation—or public auditing will be a fiction.

Furthermore, the ethical governance of deepfake detection requires the establishment of global regulatory and certification institutions, similar to the WHO or the UN, but focused on informational integrity. Such bodies would need to define universal standards for ethical data collection, model auditing protocols, sanctions for misuse, and mechanisms for transnational cooperation in sharing forensic knowledge. Only in this way would it be possible to prevent the privatization of truth and avoid informational defense from turning into an algorithmic cold war.

This governance infrastructure must be accompanied by legislation that recognizes computational authenticity as an emerging human right, ensuring that every person has the right to exist digitally without being falsified, manipulated, or extracted without consent.

The ethical detection of deepfakes, therefore, needs to be understood as an interdisciplinary and transnational field that articulates engineering, communication, law, philosophy, and applied ethics. It is necessary to recognize that each dataset is also a narrative about humanity—and that training an AI is, in a way, writing a new ontology of humanity. This recognition implies historical responsibility: whoever trains a neural network is not merely teaching a machine to...

Seeing—it is teaching the world to recognize itself through it. The future of truth, trust, and even the notion of collective identity will depend on society's ability to establish technical and moral principles equivalent in importance to the civil rights of modernity.

Finally, ethics in neural network training is not an idealistic option, but the only cognitive survival strategy for the informational civilization. The world of deepfakes shows that technology is capable of dissolving the boundaries between false and true faster than our institutional capacity to react. The only possible defense is anticipatory: creating AI infrastructures that incorporate morality into their very architecture—what Floridi calls "ethics by design." Without this, humanity will live eternally under a regime of suspicion and perceptual chaos. The ethics of datasets is, therefore, the new social contract of the digital age: whoever controls the formation of AI controls the form of reality.

CONCLUSION

The deepfake crisis exposes, with unprecedented violence, the collapse of traditional boundaries between truth, image, and power. What was once merely a discursive dispute—waged between journalism, politics, and science—has now become an infrastructural war, where reality itself is mediated by algorithms trained on data that reflect, reproduce, and amplify the world's inequalities. The article demonstrated that the core of this dispute is not technological, but ethical: it is not about inventing smarter machines, but about teaching machines what it means to act with epistemological justice. The ethical training of neural networks, therefore, is not an academic luxury, but a matter of civilizational survival. Protection against deepfakes will only be legitimate when truth ceases to be a technical monopoly and returns to being a shared public good, sustained by transparency, auditability, and consent.

The reflections developed here demonstrate that the notion of ethics in AI needs to be shifted from the realm of intentions to the realm of infrastructure. Injustice doesn't only arise from human decisions about technology—it arises in the code, in the dataset, in what is invisible to the average citizen. Therefore, a deepfake detection system will only be morally valid if it is technically accurate, socially just, and politically auditable. This requires a new social contract between science, the state, and civil society, in which access to datasets, algorithms, and decision criteria is treated as a democratic right, not as an industrial secret. AI that detects falsehoods must also be capable of revealing the truth about itself.

In the geopolitical horizon, this study reinforces that control over datasets is the new axis of national sovereignty. Just as in the past powers controlled territory and energy, in the 21st century power is defined by the possession and ethics of data. A civilization that accepts biased, opaque, and extractive datasets as the basis of its truth infrastructure accepts, in practice, living under epistemological hegemony. Algorithmic sovereignty is, therefore, the new frontier of informational independence: without it, entire countries will be merely consumers of foreign models that silently decide what is authentic and what does not exist. The ethics of

In this sense, training is also a matter of state policy and national defense. However, the real battle is internal to contemporary culture itself: that of reconnecting technology and moral responsibility. The advancement of AI reveals that humanity will only be able to live with intelligent machines if it learns to reintroduce human values at the heart of computing. Dataset ethics is a form of civilizational re-education—the effort to remember that data is not neutral, that images are not innocent, and that automated decisions are not above morality. Ethical training is the antidote to algorithmic indifference. It transforms code into commitment, the model into an institution, and technique into an extension of human dignity.

It can be concluded, therefore, that the ethical detection of deepfakes should be understood as a new chapter of the Enlightenment—a computational Enlightenment—in which reason is reprogrammed with morality. No technical advance will have value if it is not accompanied by ethical reflection and democratic regulation. The future of truth depends on a simple, yet revolutionary gesture: teaching machines to respect the human who created them. Only in this way will it be possible to transform artificial intelligence into an ally of truth, and not an accomplice to forgetting.

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.

BRUNDAGE, Miles; WHITTLESTONE, Jess. *Governance of Artificial Intelligence: Ethics, Law and Policy.* Oxford Internet Institute, 2020.

BUOLAMWINI, Joy; GEBRU, Timnit. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* Proceedings of Machine Learning Research, 2018.

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

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

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

GEBRU, Timnit et al. *Datasheets for Datasets.* Proceedings of the Conference on Fairness, Accountability, and Transparency, 2018.

MILAN, Stefania; TRERÉ, Emiliano. *The Rise of Data Colonialism: Reclaiming Digital Sovereignty.* Social Media + Society, 2019.

NATO STRATCOM. *Deepfake Detection and Information Integrity Report.* Brussels, 2021.

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

ZUBOFF, Shoshana. *The Age of Surveillance Capitalism.* New York: PublicAffairs, 2019.