Int J Semiot Law
https://doi.org/10.1007/s11196-023-09970-5

The Spiral of Digital Falsehood in Deepfakes

Massimo Leone[1]

Accepted: 3 January 2023

Abstract
The article defines the research field of a semiotically oriented philosophy of digital communication. It lays out its methodological perspective, pointing out how the fake has always been at the center of semiotic research. It traces the origin of deepfakes back to the conception of GANs, whose essential semiotic workings it expounds on. It enucleates the specificities of the digital fake, especially in the production of artificial faces. It reviews the deepfake phenomenon, enunciating its most recent statistics, prevalent areas of application, risks, and opportunities. It surveys the most current literature. It concludes by emphasizing the novelty of a situation in which the fake, in human societies and cultures, is produced mostly by machines. It stresses the desirability for a semiotic and interdisciplinary study of these productions.

Keywords Fake · Forgery · Artificial intelligence · Generative adversarial networks · Semiotics

> "When an object is lied about, all objects, not just the one immediately involved, are distorted".
>
> (Picard, Max. 1955. *Der Mensch und das Wort*, Erlenbach-Zürich: E. Rentsch, p. 51; our translation).

1 Introduction

At one point, that was all there was to it. Artificial intelligence was everywhere. It was in everyday conversations, at the bar or in a taxi, and then it immediately became the subject of predictions and clichés, about how it will watch over us, how it will change our lives, the jobs it will eliminate. But it also entered academic discourse, that of the humanities and social sciences, where having read a few popularisation articles made it possible to present oneself as an expert, and to

✉ Massimo Leone
massimo.leone@unito.it

[1] University of Turin, Via S. Anselmo 8, 10125 Turin, Italy

Published online: 19 January 2023

make inferences again about the impact, the dangers, the doubts, with that critical grimace that philosophers now reproduce with ease, or launching into enthusiastic proclamations about the revolutionary breakthrough that the new technology would represent in research. First timidly, then more and more impetuously, an over-excited discourse on artificial intelligence has made its way on a planetary level, merging with the also effervescent discourse on the digital, and intersecting with all the calamities afflicting the world, its concerns, but also its desires, its ambitions.

If, by a play of the imagination, one could picture the moment of invention and diffusion of the first writing techniques, of alphabetical writing for example, perhaps one could grasp a similar fibrillation, although probably slower and with a more jagged and irregular geography: all of a sudden—the humans of the time must have said to themselves—as though out of nowhere came signs that meant something, that were not like the natural signs that humans would already interpret—lightning, clouds, animal footprints, etc.—but signs artfully created by man, which did not mean something to everyone, but only to those who knew how to decipher them, and which were not created by everyone, but only by those who knew the code, and used it to manage memory and communication, i.e., the time and space of information, that which is passed on from the present to the future and that which is transmitted from here to elsewhere.

Writing, therefore, perhaps also generated awe, a sense of mystery and reverential admiration, and an unstoppable proliferation of signs, which has come down to us. It generated, and there are traces of this in the history of writing reflecting on itself, inclusions and exclusions, aversions and desires, and then a new way of managing the construction of memory and culture through language. Ong's studies on orality and writing have taken stock of many of the consequences of this transition [1]; perhaps, we should do the same today with the new writings emerging at the dawn of the contemporary technological horizon.

In fact, it is not a mere question of new genres, as was the case with the invention of the novel, or of new techniques, as was the case with the advent of movable type printing, but of a new kind of writing in the broadest sense of the term, i.e., in the sense of a new textuality. Artificial intelligence disrupts because it does not merely create new meaning, but changes the rules of human meaning, just as in the past, language first did it, as a product of biological evolution, and then writing did it, as its extension in cultural evolution. Language enabled the human species to transform the cognitive articulations of intelligence into shared socio-cultural structures, which were not only internal to organisms but also external to them, expressed in signs that could thus be reproduced and transformed in time and space. Writing, then, allowed a stabilisation of this matrix of intersubjective articulations, constituting the trace of a collective memory no longer entrusted to individual memories and intellects, but still in constant relation with them. Language created shared thought, writing generated collective memory, and artificial intelligence is perhaps giving rise to a partaken elaboration of thought that, just as memory did with writing, is becoming autonomous from bodies. In this imaginary chronology of the history of human signification, language is thought that becomes autonomous from bodies, writing is memory that becomes independent from them, while artificial intelligence allows

the processing of thought and memory to continue not only outside bodies, and not only far away in space and time from them, but also independently of them.

Using a metaphorical image, one could think of the true effect of artificial intelligence as that of a hand that, detaching itself nightly from the body, as in a dream, continues to write independently of the body to which it normally belongs, and of the mind that usually guides it, yet doing so in a way that constantly reminds one of that body, and recalls that mind. This gives rise to the simultaneously fascinating and disturbing aspect of this hand, which is activated far away from us in time and space, perhaps even in a different dimension than that of waking life, in a kind of computational dream—or nightmare?—, but which simultaneously imitates us, resembles us, stutters vagaries reminiscent of the timbre of our voice.

The hype about this new way of making sense is also due to the fact that it expresses itself through variable geographies, and above all through different degrees of visibility. For instance, there are elements of artificial intelligence in the algorithms that determine the results of our Internet searches, with sometimes very disruptive effects on our everyday ways of life and behaviour; the artificial intelligence behind the appearance of faces and profiles on a cascade of Tinder profiles, for example, may have led millions of people to often radical changes in their lives, sometimes culminating in the decision to form a couple or a family with another individual met through these searches. There is much discussion today, sometimes even wearily, about the distortions that these technological panders can introduce, but at the end of the day, one has to admit that perhaps the inclinations and prejudices of a social dating app are not very dissimilar to those of any procurer; digital matchmakers amaze by their quantity and speed, but still remain within the realm of familiar technologies, perhaps because, at the end of the day, the last word always rests with the users, or the consumers, who, albeit influenced and guided by a thousand algorithmic strategies, must ultimately give their libido impulse to the fingertip that decrees the permanence or the elimination of a face and a body in the realm of desire.

There is, however, an artificial intelligence which is far more conspicuous, which fascinates us tremendously, and which at the same time—as we said—sometimes dismays us: in essence, it is that artificial intelligence that either imitates us or overpowers us, or, in the most striking hypothesis, does both. The emergence of language, then of writing, their depositing in not only socio-cultural but also external and intersubjective material structures, have given the human species a feeling of extraordinary elevation and uniqueness; for millennia, we have believed that we are not only a species but that we are special, different and above other species, unique and invested with the power to dominate nature; many religious ideologies since antiquity have corroborated this attitude, which has become more and more acute with modernity, detaching itself from its traditional metaphysical horizon and taking root, instead, in the feeling of a technical supremacy: the world is ours because no other species—or so it seems—is able to transform it as we do, to destroy it as we do.

At a certain point, however—in a course that began with the first vague beginnings of wartime computing—a computer with artificial intelligence managed to beat the world chess champion, and became invincible. Chess is now a domain in

which, incontrovertibly, artificial intelligence overpowers natural intelligence. However, this exploit has astonished and baffled only to a certain extent. Firstly, because chess expresses a limited sphere of human intelligence, which is not necessarily what everyone universally admires and aspires to or bows down to; being a chess champion is a well-regarded merit socially, but one would hardly entrust the government of a country to someone just because he or she moves his or her pieces well on the board, as it is part of common sense that human intelligence expresses and asserts itself in much more complex forms than those required to win a chess game.

Secondly, if artificial intelligence has been recognised as having superiority in this domain, it has done so in the same way as one recognises a horse to run faster than a human being, or a harvester to operate more efficiently than an individual with scythe in hand. In other words, in this as in other fields of application, the superiority of artificial intelligence over the human one has been regarded as essentially quantitative, albeit with extraordinary results, a fact of rapidity and computational capacity rather than a mysterious step, such as to arouse a mixture of deference and dismay.

The landscape began to produce shocks when this same capacity and speed of calculation began to be applied not to the sphere of rational decision-making in a closed system, as chess ultimately is, but to spheres involving different human faculties, in contexts with far more numerous and ambiguous variables. A first level of amazement was determined by the reproduction of recognition, and in particular facial recognition. The fact that a machine is able to recognise one face among a thousand in a crowd moving along the street of a metropolis is certainly surprising, as it enhances a capacity essential to human social life—that of identifying the faces of others—to a dimension that no human being, not even the so-called 'super-recognisers', could tap into.

Ultimately, however, here too there remained plenty of room to reaffirm—ever more nervously, though—the distinctiveness and ultimately also the superiority of the human; on the one hand, facial recognition merely enhanced human memory and vision quantitatively, without significantly improving them qualitatively; on the other hand, even admitting quantitative superiority, the fact remained that facial recognition was very mechanical, in the sense that it would recognize a face but would not know it, and continued to face many difficulties even in the mere recognition function—for instance, humans were still much more effective at recognising faces under difficult contextual conditions, such as poor visibility or movement of the face itself.

Today, technological progress is eroding this margin of safety of human primacy, not only because artificial intelligence recognises faces with a speed and confidence unmatched by humans—in difficult contexts, with a mask on, in poor visibility, moving, etc.—but also and above all because this artificial intelligence recognition is acquiring an ever-increasing human capacity for *cognition*, which passes from face recognition but which also draws a great deal of further information from the face, comparing it with a number of other faces, and arriving at conclusions that, however hypothetical, go far beyond the extent and degree of precision of any human physiognomy. To an exponentially growing extent, indeed, our behaviour in digital networks leaves traces and produces data that, often with our consent, but very

often well beyond it, constructs a digital twin of us that not only has the same face as us, recognisable by a machine, but also lends itself to algorithms that relate our choices, measure them over time, compare them with those of countless other digital twins, and come to conclusions and predictions that sometimes make us smile at their naivety, while sometimes surprise us.

Indeed, on the one hand, the artificial twin that proposes artificial intelligence applied to massive data, especially in large platforms of expression and consumption, resembles the grandmothers of southern Italy, so that if you went to their house for dinner one evening, and they prepared parmigiana for you, and you loved it, and you lavished praise on it, you would then have parmigiana for the next ten years, even though you might have had enough of it, or if that time you had eaten it and liked it was really only because you were so hungry, or if you had praised it just to be nice, etc. Similarly, the algorithm that keeps proposing dog kennels to us even if we do not have a dog, and even if having a dog in the house basically horrifies us, just because we once accidentally stumbled upon a video of puppies, and lingered on it longer than we should, makes us smile a little, and as we know smiling often serves to allay fears about our ontological integrity.

Much more nervous is in fact this grimace of hilarity when we have the impression that "Facebook listens to us", or "Amazon eavesdrops on us", because it seems to make predictions about us and our tastes and desires that would be impossible for a human intelligence, unless it was used, precisely, to eavesdrop on what we secretly confessed to a friend about our most hidden desires. Instead, by cross-referencing digital traces of our online behaviour with massive data through artificial intelligence, my computer now knows me better than a next of kin could, and proposes me gifts that are much more suited to my personality.

There is, however, no domain of artificial intelligence development and application that more fascinates and at the same time disquiets humans than the one where the capacity for calculation combines with the ability of imitation. Until the advent of artificial intelligence, machines were more cognitively powerful than humans, but basically incapable of imitating them in their most distinctive traits; animals such as parrots or monkeys, on the other hand, were able to imitate some traits of human behaviour very well, but with a cognitive capacity that seemed markedly inferior and, in any case, not very versatile. Today, artificial intelligence is beginning to cause concern because it not only surpasses humans in computing power, but also begins to imitate them to perfection, unhinging in ever more spectacular forms that awareness—or perhaps that presumption—of uniqueness and superiority that characterises the species and guides its actions with respect to nature. There are three areas, in particular, where this fusion of cognitive primacy and capacity for imitation is producing results that are both surprising and disturbing.

The first is that of image production. Artificial intelligence recognises images of reality, it interprets them through cross-referencing with massive visual data, but it is also beginning to produce simulacra that are increasingly indistinguishable from the original, first in two-dimensional static images, then in moving images, and now increasingly also through three-dimensional artefacts, and even—in robotics for example—in three-dimensional moving simulacra. We still smile when we come across the deepfake of a famous actor, a simulacrum produced by artificial

intelligence, yet perhaps we would smile less if the same technology were to be used to 'write' duplicates of us that would then roam the digital world independently and undisturbed, like rebellious clones or mischievous twins.

The second domain is that of language; again, it gives us a flicker of superiority to see how a chatbot produces bizarre answers to very normal questions, but perhaps we do not realise how exponentially rapid the improvement of this technology is, and how soon even in this domain we will find it difficult to recognise speech produced by humans from that produced by artificial intelligence, and in an increasing number of discursive contexts. Are we teachers so sure that the papers we receive from our students are solely the product of their human minds? Once, a well-known Italian politician asked me to answer some of her questions on artificial intelligence, and all I did was to turn over to her the answers produced by a chatbot, without the politician in question even noticing the substitution until I revealed it to her, arousing her astonishment and disquiet. In fact, a similar shift is taking place here as between facial recognition and face *cognition*: we are no longer dealing with an artificial intelligence that recognises verbal constructs and translates them into another language—again, with increasingly spectacular results—but rather with an AI that produces its own verbal constructs in interaction with us, often going so far as to make us believe that there is another human and not a machine in front of us.

The third area, which in a way combines and fertilises the first two, is that of creativity; by exploiting massive data on the relationship between images and verbal texts, for instance, today's artificial intelligence produces surprising visual scenarios on the basis of verbal input; the resulting output is still rather stereotypical—some will say—yet creative enough to compete with a contemporary human graphic designer of average inventiveness, perhaps not with Michelangelo, but certainly with a recent graduate in digital graphics, who often has a background and creative disposition that are far more stereotypical than that of a machine. In fact, there would be much to ponder on how this growing creative exuberance of computational power seems today to be matched by a progressive standardisation of human practices of meaning, as if there were a tendency according to which, after having delegated cognition, memory, and elaboration, the next step might consist in delegating to artificial intelligence the tasks that humans perform with greater creativity.

Faced with this scenario, in which computational power and the capacity for imitation give rise to new forms of machinic (pseudo?)creativity, it is not difficult to call upon semiotics, among the human and social sciences, to say something pertinent and specific about the new avenues of meaning opened up by artificial intelligence. Saying something specifically semiotic about AI is indeed perfectly within the bounds of a discipline that, since its foundation, has been concerned with signification, meaning, emulation, simulacra, and even innovation and creativity [2].

Semiotics is the study of signs and symbols and how they are used to communicate meaning. In the context of artificial intelligence (AI), semiotics can be used to analyse the ways in which AI systems communicate and the meanings that are conveyed through their interactions with humans and other systems. One aspect of semiotics that is particularly relevant to AI is the concept of the 'signifier', which refers to the physical manifestation of a sign, such as a word or image. In the case of AI, the signifier could be a computer programme or a machine learning model,

while the signified is the concept or meaning that the signifier represents. Semiotics can also be used to examine the cultural and social contexts in which AI systems are developed and used and how they may be perceived and interpreted by different groups of people. This can be particularly important for understanding the potential biases and limitations of AI systems and how they may be perceived and used by different users. Overall, semiotics can provide a useful framework for understanding the complex interactions between AI systems and humans and how these interactions are formed and shaped by cultural and social norms.

By the way, the above paragraph was created by OpenAI's chatbot ChatGPT in response to the question "What can semiotics say about artificial intelligence?". As it is evident through this witty example, reflecting on the new presence of the face in AI-driven communication is becoming urgent. Semioticians can give their contribution not only in the customary frame that was suggested by the chatbot above, but also through a more philosophical reflection about the role of the fake in human communication, and how artificial intelligence is radically changing its laws of appearance. The article that follows concentrate, in particular, on one of the problematic areas singled out above, those in which the new computational power of AI is leading to new both surprising and disquieting imitations of human behaviors and artifacts. The paragraphs that follow, in particular, will concentrate on the imitation of human face and facial behaviours brought about by AI digital graphics.

2 The field of research

A semiotics-oriented philosophy of digital communication aims to read technologies of meaning into the long history of human meaning systems, to reveal the implicit ideologies that underlie the creation of new devices, processes, and artifacts. Artificial intelligence is no exception, as its development is usually underpinned by specific preconceptions about what intelligence is, how it should work, and what kinds of outcomes it is supposed to produce in the world.

Each culture and each historical epoch are characterized by the semiotic modalities that they adopt in the production of the fake [3]. The human species is endowed with an innate capacity to give rise to representations that intentionally do not correspond to empirical reality. The technologies and languages of the fake, however, change in time and space. With digital technology, with telematic communication, and especially with artificial intelligence and deep learning, the human culture of the fake is crossing a decisive threshold.

In the digital world, human culture enters the domain of the "absolute fake". This is due, firstly, to the material characteristics of this technology: everything that can be the object of digital representations, can also be the object of it without any ontological reference. Any digital image that will be produced of an aged face in a future whose ontology does not yet exist can be reconstructed in the present of the digital simulation. Secondly, the domain of the absolute fake is caused by the power of quantitative accumulation: an image of a rejuvenated face can circulate in social networks in such an intense and viral way that it will end up representing its identity in the web. Thirdly, the domain of the absolute fake is determined by its new

modalities of creation: previously, the stake of the fake was played between forgers and connoisseurs (for example, in the field of art); now, this game is played more and more by algorithms with largely unpredictable results.

Artificial intelligence applied to the creation of the fake has always been exercised towards a particular object, namely, the face, which is the main interface and the most important human device for interpersonal communication (Leone, 2022 Forthcoming).

3 Research methodology

Semiotics is perfectly equipped to carry out a study whose object would be at the crossroads between forgery, face, and digital representation. As for the fake, all the founding fathers of semiotics have addressed the subject [4]: (1) Charles S. Peirce in the American tradition [5]; (2) the main voices of structural semiotics, starting with a special issue of the French journal *Communications* devoted to the concept of "vraisemblable": Tzvetan Todorov, Gérard Genette, Christian Metz, Julia Kristeva, Gérard Genot, Roland Barthes, and others [6]; Baudrillard dealt with the topic [7, 8]); a roundtable on "Post-truth and Democracy" was organized by Jacques Fontanille at the Congress of the French Association of Semiotics in Lyon, June 11–14, 2019 [9]; Umberto Eco wrote extensively on the fake [10], edited a special issue of the semiotic journal *Versus* on "Fakes, Identity, and the Real Thing" ([11]; with essays by Eco, Prieto, Calabrese, and others), and also dealt with the topic in numerous essays and novels (*Foucault's Pendulum*; *The Prague Cemetery*; *Number Zero*); (3) Jury M. Lotman repeatedly addressed the issue of forgery [12, 13]. The ERC research group FACETS has devoted several initiatives to the relation between the fake and the digital face, including a roundtable at the Fondation Maison des Science de l'Homme, in Paris (23–24 June 2021) [14].

4 The genesis of research

Semiotic research on digital representations of the face is increasingly abundant, particularly regarding the representation of the face by artificial intelligence. To develop an analysis of the semiotic ideologies underlying the creation of synthetic faces, however, it is necessary to look at the origin of the algorithms that, in recent years, have revolutionized the practices of meaning in this field. In particular, one must return to their founding text, an article that the young Ian J. Goodfellow published on June 10, 2014—when he was a PhD student at the University of Montreal—with the title "Generative Adversarial Nets" [15].[1] Together with a group friends, all doctoral students in computer science, he proposed a new framework for estimating generative models via an adversarial process, in which two models are

[1] Since then, this researcher has become the world's guru of artificial intelligence and especially of deep learning.

trained simultaneously: a generative model that captures the distribution of data, and a discriminative model that estimates the probability that a sample comes from the training data rather than the generative model. The generative adversarial model has led to revolutionary applications in artificial intelligence and deep learning, including the creation of "artificial faces" [16] and deepfakes.

Semiotics has already been applied to the study of artificial intelligence.[2] It has, however, focused on its results and products, whereas it would be essential to examine, through semiotics, its ideological presuppositions, and the deep structure of its functioning. Goodfellow's productive scheme of artificial intelligence is imagined as an opposition between two instances; the framework of structural semiotics can therefore contribute much to its intelligibility. Semiotics, and especially the structural and generative one championed by Greimas, would see two main actants in the abstract architecture of GANs: the first is a generative actant that examines a data pattern and produces a text that could be derived from it; the second is a discriminative actant that examines the text thus produced and evaluates whether it comes from the data pattern or from the generative actant. From the epistemic point of view, therefore, the generative actant aims at making appear, and thus believe true, what is not, whereas the discriminative actant aims at making appear, and thus believe false, what is not true.[3]

5 The results of research

When one reads through semiotics the founding article of generative adversarial networks (GANs), one is struck above all by two elements: (1) the conception of artificial intelligence that it expresses is based on the idea of antagonism (neither of cooperation, nor of simple competition); (2) the metaphor that best explains the new deep learning architecture is that of the counterfeiter and the connoisseur (especially in the manufacture of money). Both aspects deserve further philosophical and semiotic reflection, because this new architecture of artificial intelligence is now finding applications in many professional and social domains, and in particular in the creation of computer-generated images and videos of static or moving faces, increasingly associated with heads, bodies, and contexts that are also computer-generated, and often expressed by multiple sign systems, such as facial expressions, gestures, movements, fragments of verbal discourse, songs, and dances.

[2] See the seminar "The Semiotics of Artificial Intelligence", organized by Massimo Leone at the University of Shanghai between 21 June and 2 July 2021, whose peer-reviewed proceedings, integrated by other papers submitted in response to a specific call for papers, have been published [17].

[3] Technically speaking, to learn the distribution of the generator pg on the data x, one defines an a priori on the input noise variables pz (z), and then represents a mapping to the data space as G (z; θg), where G is a differentiable function represented by a multilayer perceptron with parameters θg. A second multilayer perceptron D is also defined (x; θd) that produces a single scalar. $D(x)$ represents the probability that x comes from the data rather than from pg. One trains D to maximize the probability of assigning the correct label to both the training examples and the samples in G. One simultaneously trains G to minimize log (1–D (G (z))).

The GAN scheme can be read by the metaphor proposed by the same Goodfellow in 2014: D and G behave like a connoisseur and a counterfeiter, respectively. The counterfeiter examines the currency in circulation and tries to reproduce it; the connoisseur examines the currency produced by the counterfeiter without knowing its origin and tries to figure out whether it is counterfeit or genuine. In doing so, however, the connoisseur gives the counterfeiter information that will be useful in creating counterfeit money that, in turn, will be even more difficult to distinguish from genuine currency. But the connoisseur also learns to discriminate, better and better, between genuine and counterfeit money. The metaphor of the art market can effectively capture the idea of this spiral of generation and discrimination: a forger tries to put fake Modigliani paintings into circulation, while a connoisseur tries to distinguish them from genuine Modigliani artworks; in doing so, however, the latter gives the former information on how to better forge the works; vice versa, the former also learns from the latter how to forge the paintings of the Italian artist.

One must wonder about the nature of the observer actant of this spiral, on how information is distributed throughout it. The products of the generative model are not only sanctioned by the discriminative model but also by a human addressee, who coincides, at least in the first instance, with the addressee of the GANs. The models are programmed by a human addressee, and yet their 'behaviors' are not entirely predictable, notably because of the computational gap between human cognition and artificial intelligence (the 'black box'). The human programmer is, thus, both the sender and the receiver of the products of the interaction between the generative model and the discriminative one. Moreover, beyond this professional observer actant, there is another one which is composed by those who will receive the products of the generative model without having knowledge of their origin. The spiral that has just been described is destined to increase more and more the epistemic uncertainty of this non-professional observing actor.

To put it in simpler terms within the framework of the first metaphor: the competition between counterfeiter and connoisseur puts into circulation money or works of art that are false, but that are increasingly difficult to recognize as such, especially by the observing actant outside the spiral. The massive circulation of a fake that is no longer identifiable as such ends up discrediting authentic works of art, and authentic money. This is perhaps the most important danger of the 'spiral of forgery'. Some researchers have sought to shed a positive light on GANs, suggesting that their internal dialectic should be compared to that between teacher and student. The generative model would, thus, be like a student trying to produce credible representations from a database, while the discriminative model would be like a teacher examining and evaluating those representations. This is partly true, but what makes the difference is that, in the world of GANs, the representations of the generative model begin to flow without reference to the learning context.

That is also the main difference between digital and analogic falsehood. The human species is intrinsically capable of producing, in an intentional way, false representations of reality, representations that, while lacking a (clear) indexical origin, simulate one, especially through the creation of an iconic sense effect. This capacity was probably selected by the biological evolution of the species as adaptive because it allowed it to make a mental experience of potentially dangerous situations without

having to experience them empirically. It also allowed it to protect itself from predators or to trap potential prey. This is a capacity that is not absent in other species, both plants and animals. One of the most remarkable features of minnows, for example, is their ability to imitate sounds, such as those of other birds and of various natural elements, but also the sounds of the human environment, such as the triggering of a camera, a chainsaw, a fire alarm, a hydraulic cylinder, etc. In the human species, however, this ability, expressed in and through language, has given rise to a kind of 'exaptation', consisting in the capacity to attach aesthetic pleasure and value to intentional and false representations, which has triggered an immense production of fictional texts. The digital introduces an essential qualitative and quantitative change in the history of the relation of the human species to the fake.

6 Semiotic properties of the digital fake

In the first place, the digital is endowed with a protean materiality whose semiotic manifestation is entirely programmable, which is never the case in the manifestation of pre-digital texts. This implies that any digital representation having an indexical relation with its object, can be reproduced identically even when this relation is absent; painting can, of course, simulate faces that do not exist, and yet the gap between the ontological face and the painted one will always be evident, which is not the case in the digital. Digital technology absorbs the sense of indexicality that is characteristic of photography and reproduces it in the absence (or rather, remoteness) of indexicality; at the same time, it introduces a total programmability in the construction of the photographic image. Painting can represent non-existent objects but not make one believe in their existence (the reality effect of a trompe-l'oeil is fleeting [18]); analogical photography can make one believe in the existence of the objects it represents but it cannot represent non-existent objects, at least not in an efficient way; digital photography can effectively induce believing in the existence of the non-existent objects it represents.

Secondly, the application of artificial intelligence, and of deep learning by GANs, to the production of the material manifestation of digital representations removes them from human evaluation. Forgery is consubstantial to the human species, but this is the first time in the history of the species that non-human agents are put in a position to produce forgery whose evaluation escapes more and more from humans and is increasingly entrusted, instead, to an evaluation that is in turn conducted by means of artificial intelligence.

Thirdly, the digital fake can be reproduced and circulated with unprecedented ease, and this quantitative aspect also results in a qualitative change: it is as if the authentic art had to defend itself from an infinite number of forgers who work constantly and very quickly in the production of copies.

The digital fake is destined, in the long run, to be indistinguishable from the 'digital real'; in the case of faces, for example, it is only a matter of time before one can no longer know, from the digital photo of a face, whether this photo has been produced from a biological and ontological face or whether it is a synthetic image. Semiotics tends to problematize the logical concept of "truth" as adequacy to reality,

considering, instead, the semiotic conditions that produce a "reality effect". Explaining the rhetoric of a "reality effect" without postulating an ontological reality, however, leads to unavoidable aporias. In an analogous way, one can well problematize the "reality effect" of an analogical photograph, but one must also recognize that the arrival of digital technology, and, in particular, of digital deep learning applied to the creation of images, undermines the possibility of distinguishing between a referential image endowed with a reality effect and a synthetic image producing exactly the same effect.

7 Towards a semiotics of deepfakes

The undetectable nature of digital forgery becomes worrisome as it manifests itself in increasingly complex and socially central texts. From this point of view, the deepfake phenomenon requires urgent reflection. First, because it involves the digital simulation of an object, the face, which is essential to the functioning of human societies; second, because it simulates this object not only in the static image but also in the moving image and, increasingly, in its context and functions, for example through the synthetic representation of lip movements and in combination with the AI simulation of human voices.

In 2019, Deeptrace, an Amsterdam-based cybersecurity company providing deep learning and computer vision technologies for online synthetic media detection and monitoring—since renamed Sensity—published a report titled "Deepfake", claiming that the phenomenon at the time was growing rapidly online, with the number of deepfake videos almost doubling in the seven months of investigation until reaching the figure of 14,678 videos online. A Sensity blog post published by one of its collaborators, Francesco Cavalli, on February 8, 2021, revealed that the number of fake videos online had grown exponentially since 2018, doubling approximately every six months, with 85,047 deepfake videos detected by Sensity in December 2020.

At present, Sensity monitors 516 sources that systematically elaborate deepfakes, resulting in the production, until today, of 118,232 "visual threats", targeting 3,231 public figures. The targets of the deepfakes are mainly in the USA (42%); in the United Kingdom (10.3%); in India (6%); in South Korea (5.7%); and in Japan (5.6%). The most targeted social and professional activities are the entertainment industry (55.9%); fashion (23.9%); politics (4.6%); sports (4.5%); and senior industry executives (3.1%). The 2019 Deeptrace report also identified the prominence of non-consensual deepfake in pornography, which at the time accounted for the 96% of total deepfake videos online. It also found that the top four websites dedicated to deepfake pornography had received more than 134 million views for videos targeting hundreds of female celebrities from around the world. Indeed, the term "deepfake" came into common use after a Reddit user named "Deepfakes" claimed in late 2017 that he/she had developed a deep learning algorithm that allowed him/her to transpose celebrity faces into porn videos. Deepfakes also have a significant impact on the political sphere. In at least two prominent cases from Gabon and Malaysia—which received very little Western media coverage—deepfakes played a central role, specifically in an alleged government cover-up and political smear campaign. The

first case resulted in an attempted military coup, while the second led to a high-profile politician being threatened with imprisonment. In June 2022, amid the Russia-Ukraine war, the mayors of Berlin, Vienna, and Madrid were swindled by a deepfake of Kiev Mayor Klitschko.

The research area traditionally devoted to general media forensics is now directing increasing efforts to detect face manipulation in image and video. Part of this effort builds on previous research on biometric anti-surveillance and modern database-mining deep learning. To standardize the evaluation of detection methods, an automated benchmark for facial manipulation detection is urgent and various options have been proposed, such as those based on DeepFakes, Face2Face, FaceSwap, and NeuralTextures as prominent representatives of facial manipulations at a random level and compression size. Nowadays, it is becoming increasingly straightforward to automatically synthesize non-existent faces or manipulate a person's real face in an image or video, thanks to the public availability of data, and the evolution of deep learning techniques that eliminate many manual editing steps, techniques such as auto-encoders (AE) and, precisely, Generative Adversarial Networks (GAN). As a result, open software, and mobile applications like ZAO and FaceApp now allow anyone to create fake images and videos, with no prior experience in the field required.

Traditional methods of detecting false images in media forensics have generally been based on: (i) "fingerprints" produced internally to the camera, namely the analysis of intrinsic digital fingerprints introduced by cameras, both by devices and by software, such as optical lens, color filter array, interpolation, compression, etc.; (ii) "fingerprints" produced outside the camera, such as the analysis of external fingerprints introduced by an editing software, for instance, the copy and paste operations or the integration of different elements of the image, or the reduction of the frame rate in a video, etc. Most of the features considered by traditional methods of fake CGI detection, however, are highly dependent on the specific training scenario and are, therefore, ineffective against unexpected conditions (see the 2020 paper "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection" [19]).

The social impact of Deepfakes is a recent object of study but one that is attracting the attention of a growing number of researchers. In 2021, Jeffrey T. Hancock and Jeremy N. Bailenson edited a special issue of the journal *Cyberpsychology, Behavior, and Social Networking*, entitled "The Social Impact of Deepfakes" [20]. The state of the art on this issue is still underdeveloped. In 2021, Saifuddin Ahmed, a researcher at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, published an article entitled "Who Inadvertently Shares Deepfakes? Analyzing the Role of Political Interest, Cognitive Ability, and Social Network Size" [21]. Drawing on survey data collected in the United States and Singapore, this study examines the role of political interest, cognitive ability, and social network size in inadvertently sharing deepfakes. The results suggest that users with more narrow political interests and less cognitive ability are more likely to inadvertently share deepfakes. The results also indicate that the relationship between political interest and deepfake sharing is moderated by network size. The likelihood of politically engaged citizens sharing deepfakes, thus, intensifies in larger social networks.

There is still little but growing empirical evidence regarding the psychological and psychosocial effects of deepfakes. Interesting insights, however, can be drawn from the creation of "Doppelgänger" in virtual reality. Watching a simulacrum of oneself in virtual reality causes the encoding of false memories in which participants believe they have performed the actions in which they see themselves represented; other experiments show the influence of these simulacra on brand preference or health behaviors. Already in 2009, Segovia and Bailenson published the article "Virtually True: Children's Acquisition of False Memories in Virtual Reality" in *Media Psychology* [22]; in the same issue, Fox and Bailenson published the article "Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors" [23]; later, in 2014, Ahn and Bailenson published "Self-Endorsed Advertisements: When the Self Persuades the Self" in the *Journal of Marketing Theory and Practice* [24].

The approaches taken to study the social effects of deepfakes are varied. The article "To Believe or not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube", by YoungAh Lee et al. [25] provides a historical overview of 10 current most popular deepfakes on YouTube and analyzes linguistic responses through viewer comments. The article "Popular Discourse Around Deepfakes and the Interdisciplinary Challenge of Fake Video Distribution", by Catherine Francis Brooks [26], mines Reddit in 2018 to gauge the reception of deepfakes and uses this data to suggest possible solutions to unfavorable use cases. "Deepfakes: Awareness, Concerns, and Platform Accountability", by Justin D. Cochran and Stuart A. Napshin [27], surveys students to assess their awareness and concerns about deepfakes, as well as the degree of accountability of platforms in their efforts to regulate this new technology.

Other articles provide some initial insights about the psychological dynamics of deepfakes on self-perception. Wu, Ma, and Zhang examine how young women evaluate their own appearance before and after exposure to a deepfake that blends an image of themselves with that of a celebrity [28]. These experiments demonstrated positive effects on self-perception. Another study [29] investigates how exposure to a reconstructed version of oneself created by an artificial intelligence program influences trust in the AI. Exposure to a talking head with the participant's face reduces affect-based trust in the AI.

8 Beyond the state of the art

Bibliography on deepfakes has considerably grown in 2022, mainly in relation to three factors: (1) the further improvement of technology for the production of deepfakes and their more and more effective integration with platforms, multimedia production, and settings of extended reality, as well as with other techniques of digital forgery, such as those for the recreation of human voices, for instance; (2) the increasing buzz around the metaverse, and the growing tendence of imagining and planning it as a place for the integration of platforms, extended realities, and deepfakes of various kinds; (3) the intensifying worry about the social effects of the massive production of deepfakes, also as a consequence of their popularization through

easily accessible and usable apps. Literature produced in the last months, then, is characterized and articulated in relation to these factors.

First, politics is starting to take deepfakes more and more seriously, also in relation to the ongoing conflict between Russia and Ukraine and the justified fear that deepfakes might be used as instruments of propaganda, fake news, and cyberwarfare. For instance, the article "Deepfakes, Misinformation and Disinformation and Authenticity Infrastructure Responses: Impacts on Frontline Witnessing, Distant Witnessing, and Civic Journalism" [30], by Sam Gregory, deals with the impact that digital forgery of news and deepfakes might have on the "authenticity infrastructure" in which we live, especially in contexts where independent reporting is scarce or wanting. United States as well as other countries are now creating task forces to counter the possible national threats represented by deepfakes (for a survey, see "Cybercrime and Artificial Intelligence. An Overview of the Work of International Organizations on Criminal Justice and the International Applicable Instruments", by Cristos Velasco [31]). The US Deepfake Task Force Act of 2022 explicitly states that

> As the software underpinning these technologies becomes easier to acquire and use, the dissemination of deepfake content across trusted media platforms has the potential to undermine national security and erode public trust in our democracy, among other nefarious impacts. [32: 2]

Consequently, research on deepfakes is also expanding in digital forensics, from both the technical and the legal perspective, to the point that "deepfake forensics" and its counterpart, "deepfakes anti-forensics" have become common expressions in academic articles about the legal and technical implications of digital deep simulation and dissimulation [33]. From the first point of view, several methods have been devised in the attempts to early detect deepfakes; all researchers in this domain, however, agree on describing it as a "cat and mouse race", meaning that new digital methods of detection become fuel for new digital methods of forgery, and vice versa. Contests have been regularly organized to prompt new solutions, but the outcomes of them have been soon integrated into new techniques for producing deepfakes. Currently, most technical, forensic, and legal efforts seem to converge towards the elaboration of standards for the tracing of digital images.

In 2022, several studies have surveyed the technical and forensic state of the art on deepfakes. A major companion has been published, *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, by a multidisciplinary team of specialists, focusing especially on computer vision and pattern recognition, but also on digital forensics [34]. Other more concise and technical surveys have been also published in the form of papers, notably in the Chinese and Middle Eastern academic world. The article "Countering Malicious DeepFakes: Survey, Battleground, and Horizon", by Juefei-Xu et al., for instance, proposes a well-structured taxonomy of current methods for the production and the detection of deepfakes [35] (see also [36–38] for a survey of deepfake "generation, detection, datasets, and opportunities"). 2022 has also been a year of both technical growth in audio deepfakes and in their integration with the visual ones [39] as well as of increased awareness of the importance of audio in the detection of deepfakes [40,

41]); that has spurred, in turn, awareness about the further possible dangers of these technical advances.

The technical literature on deepfakes (together with the grey one) remains predominant [42, 43]; that in the humanities and social sciences, instead, is still waiting for a major synthesis. Interesting studies, though, have been published in 2022; many of them are related to the psychology of deepfakes, both in terms of dangers and opportunities. As regards the first, pornography continues to be the area that most worries researchers as well as policymakers and legislators, especially in relation to the disquieting phenomenon of deepfake in digital pornography involving the representation of children; a pioneer contribution in this field is the article "Sexualization of Children in Deepfakes and Hentai", by Simone Eelma [44], but notable are also those studies that focus on the reaction of viewers to the experience of deepfakes that porn-sexualize the faces of celebrities, for instance the article "Users' Emotional and Behavioral Responses to Deepfake Videos of K-Pop Idols", by Soyoung Wang and Kim Seongcheol, which focuses on the porn deepfakes that target celebrities of the K-Pop world [45]; see also the article "Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging Form of Image-Based Sexual Abuse", also published in 2022 [46], as well as Fido, Rao, and Harper on "Celebrity Status, Sex, and Variation in Psychopathy Predicts Judgements of and Proclivity to Generate and Distribute Deepfake Pornography" [47].

The psychology of deepfakes is certainly the area that is most producing research in the humanities and social sciences about this new avenue of digital creation, focusing on dangers but also on opportunities, for instance that of treating the PTSD of rape victims by having them interact with the deepfakes of their abusers, so as to guide a reconstructive and regenerative approach to the memory of the traumatic past; to this regard, see the article "Initial Development of Perpetrator Confrontation Using Deepfake Technology in Victims with Sexual Violence-Related PTSD and Moral Injury" [48]; see also [49]. Some studies attempt at contextualizing and qualifying the true psychological impact of deepfakes on memory and empathy. The 2022 article "Deepfake False Memories" [50] experimentally introduces the hypothesis that fake images or videos may not impact the recollection of fake news. Worthy of notice are also the contribution [51], on deepfakes and the new ways of "artificial empathy" and, from a more sociological perspective, [52], which constructively proposes the possibility of using deepfakes for building experiments in the social sciences.

From a different perspective, more related to the cultural studies of deepfakes, which is an area that still deserves further investigation, Rob Cover, author of the essay "Deepfake Culture: The Emergence of Audio–Video Deception as an Object of Social Anxiety and Regulation" [53] suggests that general obsession and hype around the authenticity of deepfakes is deflecting the attention of both scholars and policy-makers from more general and substantial issues, like the ways in which societies negotiate and renegotiate the boundaries of trust and common sense also through playing with figments of digital fiction. Along a complementary argumentative line, Craig Hight also proposes a refreshing perspective, showing how "'synthetic media' offer a disruption of the documentary genre, they are also a continuation of long-standing trends within software culture and also clearly augment practices which are deeply embedded within

the documentary genre" [54: 1]. From a less 'integrated' perspective, [55] argues that deepfakes represent a threat to "epistemology of online truth" but qualifies this view through suggesting that "allowing fakes in certain sections of the online environment may facilitate the identification of fakes in others by allowing the distinguishing characteristics of fakes to be studied by internet users" (*ibidem*). An article entitled "Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be Imminent—Yet" [56], by Anna Broinowski, looks for an intermediate view, after surveying both utopian and dystopian discourses about deepfakes.

Semiotics has not been inactive in the field, with several contributions on the phenomenon. A whole issue on the semiotics of deepfakes has been published in the French journal *Interfaces numériques* ["digital interfaces"], devoted to "Images, mensonges et algorithms : La sémiotique au défi du Deep Fake" ["images, lies, and algorithms: semiotics and the challenge of the deepfake"] [57], with articles by Maria Giulia Dondero [58], the author of the present paper [59], Vivien Lloveria [60], Stefania Caliandro [61], and others. Moreover, Marco Viola and Cristina Voto, both researchers in the FACETS ERC research project, are about to publish an article on the cognitive semiotics of deepfake pornography in a forthcoming special issue of *Synthèse* entirely devoted to deepfakes; also, FACETS researcher Remo Gramigna is currently finalizing a monograph on deepfakes, to be published in the series *Facets Advances in Faces Studies* (Routledge). An interesting survey of the semantic fields involved in research publishing on deepfakes in the period 2017–2021 is in [62] which points out "a progressive discursive evolution in which new semantic fields and discursive genres of deepfake emerge, with a tendency towards beneficial uses and not only criminal ones in the sphere of the audiovisual industry, activism, experimental art, commercial uses, medical, advertising, propaganda, and education, among others" (*ibidem*: 1; on advertising and deepfakes, another promising area, see also [63]). The interdisciplinary field of the semiotics of law is becoming key in research on deepfakes, also from the technical point of view of shaping a new legal framework around the issue of protecting the identity of individuals and their faces and bodies in an increasingly digitalized environment. As some authors suggest, then, the thorny task of determining when a violation to the integrity of someone's own image and identity is committed is inseparable from the specific pragmatic context in which the fake is not only created and circulated but somehow also 'performed' (see on this aspect the article "The Identification Game: Deepfakes and the Epistemic Limits of Identity", [64]; the academic journal *Synthèse* has devoted keen attention to the philosophy of deepfakes, from different perspectives (see [65, 66]).

9 Conclusions

For the most part, deepfakes still make people smile, although the hilarity is sometimes related to their disturbing character; but, with a few exceptions, deepfakes still work as trompe-l'oeil: they amuse because one notices their deception. Given the technical conditions of deepfakes production, however, it is only a matter of time before the deception becomes undetectable, even if it means using machines to

reveal the deceptions of other machines, with an algorithmic overkill excluding the human [67]. The face, which several human societies have erected as a bulwark of singularity, will soon be falsifiable at will in all its digital representations [68]. The progress in the production of three-dimensional deepfakes or artificial biological faces connectable to artificial intelligence will make the individuation of fake faces even more complicated [69]. Semiotics, a discipline that, more than any other, has focused on the discourse of the fake, is urgently called upon to reflect on the epistemological drift that a proliferation of 'digital fakes' could entail, particularly regarding the representation of the face as an interface for living together. Semiotics itself will have to renew itself at least in part to take into account the new challenges of digital falsification, in the attempt to grasp in-depth the meaning of the profoundly false.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. This essay results from a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No 819649-FACETS).

Declarations

Conflict of interest The author declares that the submission of the article does not entail any potential conflicts of interest. That the preparation of the article has not involved any human participants and/or animals. That the article is in accordance with the highest EU ethics standards.

References

1. Ong, Walter J. 2002. *Orality and Literacy: The Technologizing of the Word (1982)*. London and New York, NY: Routledge.
2. Leone, Massimo. 2022. Sémiotique et innovation. *Acta Semiotica*, 2(4): 69–77 (26 December 2022); ISSN: 2763-700X. https://doi.org/10.23925/2763-700X.2022n4.2721
3. Leone, Massimo. 2022. Semioethics of the Visual Fake. In *Digital Ethics: The Issue of Images* ["Bild und Recht", 5], ed. Tiziana Andina and Thomas Dreier, 187–205. Baden-Baden: NOMOS.
4. Ousmanova, Almira. 2004. Fake at Stake: Semiotics and the Problem of Authenticity. *Problemos* 66(1): 80–101.
5. Cooke, Elizabeth F. 2014. Peirce and the 'Flood of False Notions.' In *Charles Sanders Peirce in His Own Words: 100 Years of Semiotics, Communication and Cognition ["Semiotics, Communication and Cognition", 14]*, ed. Torkild Thellefsen, Bent Sørensen, and Cornelis De Waal, 325–331. Boston: De Gruyter Mouton.
6. Todorov, Tzvetan, ed. 1968. *Recherches sémiologiques : Le vraisemblable*, special issue of *Communications*, 11.

7. Baudrillard, Jean. 1987. Au-delà du vrai et du faux, ou le malin génie de l'image. *Cahiers internationaux de sociologie*, new series, 82 ("Nouvelles images, nouveau reel"), January-June: 139–145.
8. Baudrillard, Jean. 2000. *The Vital Illusion*. New York: The Wellek Library Lectures.
9. Di Caterino, Angelo. 2020. Fake News : Une mise au point sémiotique. *Actes Sémiotiques*, 123: online; available at https://www.unilim.fr/actes-semiotiques/6445 (last accessed January 1, 2023).
10. Eco, Umberto. 1995. *Faith in Fakes: Travels in Hyperreality (1986)*. London: Minerva.
11. Eco, Umberto. 1987. *Fakes, Identity and the Real Thing, special issue of Versus*, vol. 46. Milan: Bompiani.
12. Andrews, Edna. 2003. *Conversations with Lotman: Cultural Semiotics in Language, Literature, and Cognition [Toronto Studies in Semiotics and Communication]*. Toronto, Buffalo, and London: University of Toronto Press.
13. Makarychev, Andrey S., and Alexandra Yatsyk. 2017. *Lotman's Cultural Semiotics and the Political: Reframing the Boundaries*. London: Rowman & Littlefield International.
14. Thierry, Gianmarco Giuliana, and Massimo Leone, eds. 2023. *Le Futur du visage ["I saggi di Lexia"]*. Rome: Aracne.
15. Goodfellow, Ian J. *et al.* 2014. Generative Adversarial Networks. available at https://arxiv.org/abs/1406.2661 (last accessed January 1, 2023).
16. Leone, Massimo. 2021. Prefazione / Preface. In *Volti artificiali / Artificial Faces, special issue of Lexia: International Journal of Semiotics, 37–8*, ed. Massimo Leone, 9–25. Rome: Aracne.
17. Santangelo, Antonio, and Massimo Leone, eds. 2023. *Semiotica e intelligenza artificiale ["I saggi di Lexia"]*. Rome: Aracne.
18. Leone, Massimo. 2014. Détrompe loeil: Come disfare cose con le immagini. In *Immagini efficaci / Efficacious Images, special issue of Lexia, 17–18*, ed. Massimo Leone, 43–72. Rome: Aracne.
19. Tolosana, Ruben *et al*. 2020. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection, online; available at https://arxiv.org/abs/2001.00179 (last accessed January 1, 2023).
20. Hancock, Jeffrey T. and Jeremy N. Bailenson. 2021. *The Social Impact of Deepfakes*; special issue of *Cyberpsychology, Behavior, and Social Networking*, 149–52; available at https://stanfordvr.com/pubs/2021/the-social-impact-of-deepfakes/ (last accessed January 1, 2022).
21. Ahmed, Saifuddin. 2020. Who Inadvertently Shares Deepfakes? Analyzing the Role of Political Interest, Cognitive Ability, and Social Network Size. *Telematics and Informatics* 57: 1–10. https://doi.org/10.1016/j.tele.2020.101508.
22. Segovia, Kathryn Y., and Jeremy N. Bailenson. 2009. Virtually True: Children's Acquisition of False Memories in Virtual Reality. *Media Psychology* 12(4): 371–393. https://doi.org/10.1080/15213260903287267.
23. Fox, Jesse, and Jeremy N. Bailenson. 2009. Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology* 12(1): 1–25. https://doi.org/10.1080/15213260802669474.
24. Ahn, Sun Joo-Grace., and Jeremy Bailenson. 2014. Self-Endorsed Advertisements: When the Self Persuades the Self. *The Journal of Marketing Theory and Practice* 22(2): 135–136. https://doi.org/10.2753/MTP1069-6679220203.
25. Lee, YoungAh, et al. 2021. To Believe or Not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube. *Cyberpsychology, Behavior, and Social Networking* 24(3): 153–158. https://doi.org/10.1089/cyber.2020.0176.
26. Brooks, Catherine Francis. 2021. Popular Discourse Around Deepfakes and the Interdisciplinary Challenge of Fake Video Distribution. *Cyberpsychology, Behavior, and Social Networking* 24(3): 159–163. https://doi.org/10.1089/cyber.2020.0183.
27. Cochran, Justin D., and Stuart A. Napshin. 2021. Deepfakes: Awareness, Concerns, and Platform Accountability. *Cyberpsychology, Behavior, and Social Networking* 24(3): 164–172. https://doi.org/10.1089/cyber.2020.0100.
28. Fuzhong, Wu., Yueran Ma, and Zheng Zhang. 2021. 'I Found a More Attractive Deepfaked Self': The Self-Enhancement Effect in Deepfake Video Exposure. *Cyberpsychology, Behavior, and Social Networking* 24(3): 173–181. https://doi.org/10.1089/cyber.2020.0173.
29. Weisman, William D., and Jorge F. Peña. 2021. Face the Uncanny: The Effects of Doppelganger Talking Head Avatars on Affect-Based Trust Toward Artificial Intelligence Technology are Mediated by Uncanny Valley Perceptions. *Cyberpsychology, Behavior, and Social Networking* 24(3): 182–187. https://doi.org/10.1089/cyber.2020.0175.

30. Gregory, Sam. 2022. Deepfakes, Misinformation and Disinformation and Authenticity Infrastructure Responses: Impacts on Frontline Witnessing, Distant Witnessing, and Civic Journalism. *Journalism* 23(3): 708–729. https://doi.org/10.1177/14648849211060644.
31. Velasco, Cristos. 2022. Cybercrime and Artificial Intelligence. An Overview of the Work of International Organizations on Criminal Justice and the International Applicable Instrument. *ERA-Forum* 23(1): 109–126. https://doi.org/10.1007/s12027-022-00702-z.
32. Deepfake Task Force Act 2022. *Report of the Committee on Homeland Security and Governmental Affairs, United States Senate, to Accompany S. 2559, to Establish the National Deepfake and Digital Provenance Task Force, and for Other Purposes*. Washington: U.S. Government Publishing Office.
33. Ding, Feng, Guopu Zhu, Yingcan Li, Xinpeng Zhang, Pradeep K. Atrey, and Siwei Lyu. 2022. Anti-Forensics for Face Swapping Videos via Adversarial Training. *IEEE Transactions on Multimedia* 24: 3429–3441. https://doi.org/10.1109/TMM.2021.3098422.
34. Rathgeb, Christian, et al. (eds.). 2022. *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham: Springer International Publishing AG.
35. Juefei-Xu, Felix, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. 2022. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision* 130(7): 1678–1734. https://doi.org/10.1007/s11263-022-01606-8.
36. Rana, Md Shohel, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. 2022. Deepfake Detection: A Systematic Literature Review. *IEEE Access* 10: 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404.
37. Seow, Jia Wen, Mei Kuan Lim, Raphaël C.W.. Phan, and Joseph K. Liu. 2022. A Comprehensive Overview of Deepfake: Generation, Detection, Datasets, and Opportunities. *Neurocomputing* 513: 351–371. https://doi.org/10.1016/j.neucom.2022.09.135.
38. Solaiyappan, Siddharth, and Yuxin Wen. 2022. Machine Learning Based Medical Image Deepfake Detection: A Comparative Study. *Machine Learning with Applications* 8: 100298. https://doi.org/10.1016/j.mlwa.2022.100298.
39. Efanov, Dmitry, Pavel Aleksandrov, and Nikolay Karapetyants. 2022. The BiLSTM-Based Synthesized Speech Recognition. *Procedia Computer Science* 213: 415–421. https://doi.org/10.1016/j.procs.2022.11.086.
40. Almutairi, Zaynab, and Hebah Elgibreen. 2022. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* 15(5): 155. https://doi.org/10.3390/a15050155.
41. Kong, Chenqi, Baoliang Chen, Wenhan Yang, Haoliang Li, Peilin Chen, and Shiqi Wang. 2022. Appearance Matters, So Does Audio: Revealing the Hidden Face via Cross-Modality Transfer. *IEEE Transactions on Circuits and Systems for Video Technology* 32(1): 423–436. https://doi.org/10.1109/TCSVT.2021.3057457.
42. Huang, Yihao, Felix Juefei-Xu, Qing Guo, Yang Liu, and Pu. Geguang. 2022. FakeLocator: Robust Localization of GAN-Based Face Manipulations. *IEEE Transactions on Information Forensics and Security* 17: 2657–2672. https://doi.org/10.1109/TIFS.2022.3141262.
43. Eelmaa, Simone. 2022. Sexualization of Children in Deepfakes and Hentai. *Trames* 26(2): 229–248. https://doi.org/10.3176/tr.2022.2.07.
44. Liang, Buyun, Zhongyuan Wang, Baojin Huang, Qin Zou, Qian Wang, and Jingjing Liang. 2023. Depth Map Guided Triplet Network for Deepfake Face Detection. *Neural Networks* 159: 34–42. https://doi.org/10.1016/j.neunet.2022.11.031.
45. Wang, Soyoung, and Kim Seongcheol. 2022. "Users' Emotional and Behavioral Responses to Deepfake Videos of K-Pop Idols. *Computers in Human Behavior* 134: 107305. https://doi.org/10.1016/j.chb.2022.107305.
46. Flynn, Asher, Anastasia Powell, Adrian J. Scott, and Elena Cama. 2022. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging Form of Image-Based Sexual Abuse. *British Journal of Criminology* 62(6): 1341–1358. https://doi.org/10.1093/bjc/azab111.
47. Fido, Dean, Jaya Rao, and Craig A. Harper. 2022. "Celebrity Status, Sex, and Variation in Psychopathy Predicts Judgements of and Proclivity to Generate and Distribute Deepfake Pornography. *Computers in Human Behavior* 129: 107141. https://doi.org/10.1016/j.chb.2021.107141.
48. van Minnen, Agnes, F Jackie June Ter. Heide, A. de Tilly Koolstra, Sezer Karaoglu Jongh, and Theo Gevers. 2022. Initial Development of Perpetrator Confrontation Using Deepfake Technology in Victims with Sexual Violence-Related PTSD and Moral Injury. *Frontiers in Psychiatry* 13: 1–8. https://doi.org/10.3389/fpsyt.2022.882957.

49. Lucas, Kweilin T. 2022. Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology. *Victims & Offenders* 17(5): 647–659. https://doi.org/10.1080/15564886.2022.2036656.
50. Murphy, Gillian, and Emma Flynn. 2022. Deepfake False Memories. *Memory* 30(4): 480–492. https://doi.org/10.1080/09658211.2021.1919715.
51. Yang, Hsuan-Chia, Annisa Ristya Rahmanti, Chih-Wei Huang, and Yu-Chuan Jack. Li. 2022. How Can Research on Artificial Empathy Be Enhanced by Applying Deepfakes? *Journal of Medical Internet Research* 24(3): e29506–e29506. https://doi.org/10.2196/29506.
52. Eberl, Andreas, Juliane Kühn, and Tobias Wolbring. 2022. Using Deepfakes for Experiments in the Social Sciences: A Pilot Study. *Frontiers in Sociology* 7: 907199–907199. https://doi.org/10.3389/fsoc.2022.907199.
53. Cover, Rob. 2022. Deepfake Culture: The Emergence of Audio-Video Deception as an Object of Social Anxiety and Regulation. *Continuum* 36(4): 609–621. https://doi.org/10.1080/10304312.2022.2084039.
54. Hight, Craig. 2022. Deepfakes and Documentary Practice in an Age of Misinformation. *Continuum* 36(3): 393–410. https://doi.org/10.1080/10304312.2021.2003756.
55. Harris, Keith Raymond. 2022. Real Fakes: The Epistemology of Online Misinformation. *Philosophy & Technology* 35(3): 1–24. https://doi.org/10.1007/s13347-022-00581-9.
56. Broinowski, Anna. 2022. Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be Imminent-Yet. *Journal of Asia-Pacific Pop Culture* 7(1): 109–139. https://doi.org/10.5325/jasiapacipopcult.7.1.0109.
57. Chatenet, Lodovic, ed. 2022. Images, mensonges et algorithms : La sémiotique au défi du Deep Fake, special issue of *Interfaces numériques* https://doi.org/10.25965/interfaces-numeriques.4824.
58. Dondero, Maria Giulia. 2022. Du portrait pictural aux deepfakes : Le visage en tant que totalité. In *Images, mensonges et algorithms : La sémiotique au défi du Deep Fake* ed. Lodovic Chatenet, special issue of *Interfaces numériques* 11(2): online. https://doi.org/10.25965/interfaces-numeriques.4855.
59. Leone, Massimo. 2022. L'idéologie sémiotique des deepfakes. In *Images, mensonges et algorithms : La sémiotique au défi du Deep Fake* ed. Lodovic Chatenet, special issue of *Interfaces numériques* 11(2): online. https://doi.org/10.25965/interfaces-numeriques.4855.
60. Lloveria, Vivien. 2022. Le deepfake et son métadiscours : l'art de montrer que l'on ment. In *Images, mensonges et algorithms : La sémiotique au défi du Deep Fake* ed. Lodovic Chatenet, special issue of *Interfaces numériques* 11(2): online. https://doi.org/10.25965/interfaces-numeriques.4855.
61. Caliandro, Stefania 2022. Fake Art, entre le contrefait et le contrefactuel. In *Images, mensonges et algorithms : La sémiotique au défi du Deep Fake* ed. Lodovic Chatenet, special issue of *Interfaces numériques* 11(2): online. https://doi.org/10.25965/interfaces-numeriques.4855.
62. Bañuelos Capistrán, Jacob Israel. 2022. Evolución del Deepfake: Campos semánticos y géneros discursivos (2017–2021), *La Revista Icono* 14, 20: online. https://doi.org/10.7195/ri14.v20i1.1773.
63. Campbell, Colin, Kirk Plangger, Sean Sands, and Jan Kietzmann. 2022. Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising. *Journal of Advertising* 51(1): 22–38. https://doi.org/10.1080/00913367.2021.1909515.
64. Öhman, Carl. 2022. The Identification Game: Deepfakes and the Epistemic Limits of Identity. *Synthèse* 200: 4. https://doi.org/10.1007/s11229-022-03798-5.
65. Harris, Keith Raimond. 2021. Video on Demand: What Deepfakes Do and How They Harm. *Synthèse* 199: 13373–13391. https://doi.org/10.1007/s11229-021-03379-y.
66. Atencia-Linares, Paloma, and Mark Artiga. 2022. Deepfakes, Shallow Epistemic Graves. *Synthèse* 200: 518. https://doi.org/10.1007/s11229-022-04003-3.
67. Leone, Massimo. 2021. El rostro aumentado: Trayectorias tecnologicas de lo falso. In *Next: Imaginar el Post-Presente: Filosofía, arte y tecnología en la cultura digital*, ed. Humberto Valdivieso and Loreja Rojas Parma, 55–76. Caracas: Universidad Católica Andrés Bello.
68. Leone, Massimo. Forthcoming. *The Semiotics of Artificial Faces* ["FACETS Advances in Face Studies", 1]. London and New York: Routledge.
69. Ding, Feng, Bing Fan, Zhangyi Shen, Yu. Keping, Gautam Srivastava, Kapal Dev, and Shaohua Wan. 2022. Securing Facial Bioinformation by Eliminating Adversarial Perturbations. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/TII.2022.3201572.