Preserving Anonymity: Deep-Fake as an Identity-Protection Device and as a Digital Camouflage

International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique

Digital camouflage: Deepfake as device of identity protection

--Manuscript Draft--

Manuscript Number:	SELA-D-23-00161R1
Full Title:	Digital camouflage: Deepfake as device of identity protection
Article Type:	S.I.: Digital Face
Keywords:	Fakes; Face, Deep-fake, Identity, Anonymity
Corresponding Author:	Remo Gramigna, Ph.D. University of Turin: Universita degli Studi di Torino ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Turin: Universita degli Studi di Torino
Corresponding Author's Secondary Institution:	
First Author:	Remo Gramigna, Ph.D.
First Author Secondary Information:	
Order of Authors:	Remo Gramigna, Ph.D.
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	This paper seeks to shed light on a un unwritten chapter in the history of DeepFake technology, that is, to draw a parallel between DeepFake and its use as protective device to mask the identity of targeted individuals or whistleblowers. Since its inception in 2017, deepfakes are enmeshed in a set of sociotechnical imaginaries. Indeed, DeepFake technology has been generally framed in terms of its potential danger, a threat to people's privacy, and as a weaponized tool in the era of disinformation. This is proven both by the definitions of the term itself, which often stress that deepfakes are "designed" to deceive or generally used maliciously to spread false information, as well as from the history and context of origin of this phenomenon ¾ altered pornographic footage created and spread online with malicious intent. Due to the alleged blurring of the distinction between the evidential and the fictional paradigm, real and virtual, truth and falsehood, these new forms of synthetic media function as a potential source of disinformation and deception, but also as uncharted forms of creativity. The present study is an exploration of DeepFake thought of as digital camouflage, by unpacking the use of DeepFake as protective mask in David France's documentary film: Welcome to Chechnya (2020). I argue that this illustration challenges the mainstream view on DeepFakes by showing an ethical and positive use of this new technology.
Response to Reviewers:	Reviewer 1 did not ask for any amendments. Reviewer 2 asked to take into consideraion 3 critical point, which I have acknowledged and incorporeted into the text. 1) The question of style. I have corrected the style where required, utilizing a proper scientific and academic style of writing. I have included in the footnotes the references that were asked by the reviewer. 2) I have added a section about the legal implications of deepfakes. 3) I have modified the terminology employed in order to accout for the distintion between malicious and benign forms of deepfakes.

Author: DR.Remo Gramigna,
Email: remo.gramigna@unito,it

Affiliation: University of Turin, Department of Philosophy and Educational Sciences

Digital camouflage: Deep-fake as a device of identity-protection

Affiliation: University of Turin, Department of Philosophy and Educational Sciences

Abstract: This paper seeks to shed light on an unwritten chapter in the history of deep-fake technology, that is, to draw a parallel between deep-fake and its use as protective device to mask the identity of targeted individuals or whistleblowers. Since its inception in 2017, deepfakes are enmeshed in a set of sociotechnical imaginaries. Indeed, deep-fake technology has been generally framed in terms of its potential danger, a threat to people's privacy, and as a weaponized tool in the era of disinformation. This is proven both by the definitions of the term itself, which often stress that deepfakes are "designed" to deceive or are generally used maliciously to spread false information, as well as from the history and context of the origin of this phenomenon — i. e. altered pornographic footage created and spread online with malicious intent. Due to the alleged blurring of the distinction between the evidential and the fictional paradigm, real and virtual, truth and falsehood, these new forms of synthetic media function as a potential source of disinformation and deception, but also as uncharted forms of creativity. The present study is an exploration of deep-fake thought of as digital camouflage, by unpacking the use of deep-fake as protective mask in David France's documentary film *Welcome to Chechnya* (2020). I argue that this illustration challenges the mainstream view on deep-fakes by showing an ethical and positive use of this new technology.

Keywords: Fakes; Face, Deep-fake, Identity, Anonymity

1. (Deep)fake it until you make it

Just as there is what has come to be termed 'fake news' there is also deep-fakes and this study is concerned with the uses and misuses of this fairly new technology. This study is an exploration of deep-fake thought of as a form of digital camouflage, by unpacking the uses of deep-fake as a protective disguise in David France's documentary film *Welcome to Chechnya* (France, 2020). Through the discussion of France's documentary, the present study seeks to shed light on a un unwritten chapter in the history and debate around digital face manipulation, that is, to draw a parallel between deep-fake and its application as a protective device to disguise the identity of targeted individuals and whistleblowers and to preserve the anonymity of sources in documentaries

This article is divided into three parts. First, I will provide a literature review on the subject, with a side glance as to how deep-fake is depicted in the "sociotechnical imaginaries" (Taylor 2004; Jasanoff 2016). Such imaginaries are thought of as discourses around the uses of a new technology. In the second part, I will unpack and revise the concept of deep-fake, from the perspective of semiotics. I will discuss some definitions of deep-fakes and, more specifically, I will address the question of whether deception is an essential characteristic of the definition of deep-fake. In the third part, I will present the uses of deep-fakes in the documentary film, *Welcome to Chechnya*, thought of as a corollary and an application of the first, more theoretical part. Based on this case study, I shall lay out the distinction between "malicious" and "benign" forms of deep-fakes as well as the distinction between covert and overt uses of this technology.

There are different varieties of deep-fakes produced for different purposes: technology demonstration deep-fakes, satirical, meme, pornographic and deceptive deep-fakes (Fikse 2018: 30), to mention but a few. Meskys at al. (2019: 4-10) provide a fourfold taxonomy of deep-fakes: revenge porn, deep-fakes used in political campaigns, commercial and creative deep-fakes, and they outlined different legal implications according to each type of deep-fake. It is not uncommon to find news headings hinting that the boundaries between truth and falsity, fact and fiction have been blurred

(Satariano and Mozur 2023). I believe that the blurring of the distinction between authentic and fake videos poses important epistemological questions that semiotics as well as other disciplines should not underestimate.

Whilst image manipulation has a long pedigree and it has been used systematically in warfare, politics and journalism (Brugioni 1999; Chéroux 2003; Farid 2003; Farid 2012; Jaubert 1986; King 1997; Nickell 1994; Smargiassi 2009), today's deep-fakes leverage and capitalize on new ways to use artificial intelligence (AI) to make it look like someone said or did something that they did not, by swapping faces, simulating voices or manipulating scenes.

Today this topic has re-gained momentum. Such matters are at the forefront of discussion arising interest among experts and ordinary people alike. It is in plain sight that terms such as 'truth', 'falsehood', 'credibility', 'verisimilitude', 'reliability', 'certitude' and neighbourhood concepts — key words and terms stemming from the philosophical disciplines — have now entered into everyday life vocabulary, constituting discourses on their own right. Indeed, the problem of truth and lying resurfaces today under many new guises and new terminology in a debate that is as fascinating as it is complex.

"Post-truth", "fake-news", "deepfakes", "digital manipulation", "disinformation" and all the paraphernalia of misrepresentation are phenomena that show the relevance that this subject has reached today. Indeed, a plethora of call for papers, journal articles and book length works devoted to this subject are quickly growing, so much so that, deep-fake is becoming a new and important sub-field of inquiry within the areas of epistemology, journalism, media studies, philosophy, artificial intelligence, computer science and digital forensics, law and semiotics (Cavedon-Taylor 2023; Châtenet 2022; Floridi 2018; Gramigna 2023; Leone 2022; 2023; Santangelo 2022), (and one that deserves attention (Giansiracusa 2012). This also shows the need for an interdisciplinary study of this subject.[1]

There are plenty illustrations of deep-fakes and a set of techniques that define the armamentarium of digital face manipulations (Tolosana et al. 2022: 5-21). Some early examples of deep-fakes went viral and became a much-discussed subject in government, science, academe, and media. For instance, the doctored video created in 2019 by Bill Posters and the *Spectre* project, purporting to shows Kim Kardashian West saying that she loves the process of manipulating people for money, has received thousands of views.[2] Likewise, the deep-fake entitled *You Won't Believe What Obama Says in This Video!*, created in 2018 by Monkeypay Productions and Buzzfeed, featuring Jordan Peele that impersonates the former US president Barack Obama, was reproduced exponentially.[3]

The discussion around this phenomenon, however, has often been characterized with quite alarmistic tones and, generally, is depicted in a quite apocalyptic fashion, portraying a sort of a "doom and gloom" scenario for the days ahead. Whilst deep-fake is a technology that has been growing and fine-tuning at a very fast speed, it still takes quite a lot of technological know-how to create.

Will all this in mind, however, we need to be able to discern the 'wheat from the chaff', as it were. In other words, the uses and misuses of deep-fakes cannot be all put into the same basket since much depends upon how a technology is used, rather than claiming, erroneously, that a technology is good or bad intrinsically in itself. Indeed, deep-fakes seem a case in point.

Given the complexity and the wide range of issues that deep-fakes touch upon, in this study I will be dealing only with one corner of this phenomenon, namely, to discuss deep-fakes as a means for what I term "digital camouflage". To be more accurate, I will discuss the possibility of framing the

[1] See the collection devoted to this subject in the journal *Synthese*, "Designed to deceive? The philosophy of Deepfakes": https://link.springer.com/collections/dhbghgcbha (accessed 11/08/2023).

[2] The video was removed from YouTube after Kim Kardashian fired a complaint: https://www.digitaltrends.com/social-media/kim-kardashian-deepfake-removed-from-youtube/ (accessed 16/10/2023). As it can be gleaned from the Spectre website project (https://billposters.ch/projects/spectre/, accessed 16/10/2023), "Spectre gamifies – and simulates – new forms of computational propaganda including OCEAN (Psychometric) profiling; gamification; algorithmic bias; personalisation; 'dark design'; 'deep fake' technologies; and micro-targeted advertising via a 'dark ad' generator".

[3] The video can be accessed at the Ars Electronica website: https://ars.electronica.art/center/en/obama-deep-fake/ (last accessed 16/10/2023).

phenomenon of deep-fakes from the standpoint of the use of it for anonymity- enhancement and for identity-protection. As we shall see in what follows, deep-fakes can work as a digital mask as well, as a sort of digital disguise that can be used for identity protection as well as an anonymization strategy. In order to support this claim, I will explore the anti-censorship power that can derive from the use of deepfakes in truth documentaries, in particular the documentary film *Welcome to Chechnya* (2020). Thus, the aim and the rationale of the present study is to offer a counter-narrative to the mainstream sociotechnical imaginaries around deep-fakes that depict this technology as essentially deceptive, misleading or utterly bad. It is, however, possible, if not even desirable, to conceive of an alternative way to think about such a technology that, instead, looks not only at the 'dark side' of this technology but also at the positive potential uses of it such as, for instance, the ways in which it can enhance the aspect of identity protection of fragile or targeted people. To discuss this position in more depth I will, then, be presenting the film *Welcome to Chechnya* by David France as a case study. This is a documentary that shows a different use of deep-fake technology which does not necessarily fall in the category of disinformation / deception / non-consensual use of images, thus, providing a fertile ground to conceive of deep-fake from a different and fresh perspective.

2. Legal implications of deep-fakes

Undoubtedly, the widespread use of deep-fake technology poses many new challenges; and the societal, political, ethical and legal implications of deep-fakes should not be underestimated.

Because deep-fake technology can generate hyper-realistic video and audio content that is very difficult to distinguish from the real one, deep-fakes create potential for spreading misinformation and fake news which can be used for malicious purposes. The blurring of the distinction between fact and fiction, illusion and reality, truth and falsehood, pointed out above, is a point on which many have insisted. As philosopher Don Fallis pointed out, "the worry has been raised that, as a result of deepfakes, we are heading toward an 'infocalypse' where we cannot tell what is real from what is not" (Fallis 2020: 623). This view is quite common and acknowledged by many (Leone 2023b). Furthermore, deep-fakes can result in severe privacy and consent violations by purporting people say or do things they actually never said or did, which may potentially damage their reputation and — in case of public figures — their public image and public persona resulting in a permanent damage.[4]

Another area of concern regards the erosion of trust and the shrinking of the epistemic validity of visual media that might derive from the use of deep-fakes on a large scale (Pawelek 2022). The growing use and the access to deep-fake technology on such a scale can, indeed, yield to a reduction of trust in visual media by questioning the authenticity of all videos, regardless of their source, content or motivation. Needless to say, there are also some potential dangers of political manipulation. Deep-fakes can certainly be used as a tool for this by making politicians and public figures appear to say or do things they did not, aiming to influence the outcome of an election or create unrest, havoc, and confusion (Ajder, Patrini, Cavalli, Cullen 2019: 9-10).

Last but not least, there are ethical and legal implications that arise with the use of deep-fake technology, including issues of consent, ownership, liability, and intellectual property rights, and matters related to how actors and individuals should be compensated. As Frederick Mostert and Sheyna Cruz (2023: 161) pointed out, in the last twenty years there has been an evident growth in jurisdiction about what is referred to as "image rights", which are defined as follows:

> legal rights which protect a natural person's interests in controlling their own identity, persona, image and reputation, along with the economic value derived from them. In most cases, image rights serve to protect a person's image against unauthorized use in a commercial context. We use the term 'image' in a broad sense to denote the recognizable attributes of an individual's

[4] See, Viola-Voto (2023) on the issue of the spreading of non-consensual diffusion of intimate images.

> persona which are typically protected by such rights. Usually, such attributes and indicia include an individual's name, nickname, portrait (including photographs), voice and likeness. Collectively these attributes form a person's image which we, the public, perceive (Mostert and Sheyna 2023: 162).

Needless to say, since deep-fakes reproduce exactly what goes under the designation of "image" of a person, there are important legal implications to consider. First of all, there is a crucial issue about consent which, even in the case of creative or artistic deep-fakes, cannot be ignored. Mostert and Cruz (2023: 180) mention the documentary film *Roadrunner: A Film About Anthony Bourdain* (Neville, 2021) in which AI was used to synthetize Anthony Bourdain's voice without the explicit consent and permission of the Bourdain's estate and without actually revealing to the public each time the AI was used to reproduce Bourdain's voice.[5] From a legal perspective, this is considered as "an unauthorized deepfake audio clip of a famous individual's voice" and "a violation of a image rights in most jurisdictions" (Mostert and Cruz 2023: 180) with serious consequences for the victims involved.[6]

Creative or artistic deep-fakes represent, however, only a minimal part as compared to deep-fake pornography, "by far the most prevalent kind of deepfake currently being created and circulated" (Ajder, Patrini, Cavalli, Cullen 2019: 6). The legal implications are plenty.[7] Non-consensual deep-fake pornography videos are thought of as "a form of image-based sexual abuse" (Flynn, Powell, Scott, Cama 2022). Pornographic deep-fakes fall into the category of "noncosensual pornography" which "involves the distribution of sexually graphic images of individuals without their consent" (Citron and Franks 2014: 1). Such images distributed without consent, include both images obtained with consent and without consent (Citron and Franks 2014: 1). The distribution without consent of sexually graphic images obtained with consent is generally referred to as "revenge porn". Whilst "revenge porn" chronologically precedes deep-fake, the two are now intertwined as deep-fake pornographic videos can be used as "revenge porn" by distributing nonconsensual pornographic deep-fakes. While they can operate in tandem, however, deep-fake sex videos and nonconsensual distribution of intimate images are different because deep-fake sex videos not always purport the naked body of the victim (Citron 2019: 1921). Indeed, in the case of face-swapping where the face of celebrities are superimposed on the naked body of people engaging in sexual acts, the deep-fake video does not purport the body of the victim but the body of the pornographic actress or actor. As Citron (2019: 1921) pointed out, "yet even though deep-fake sex videos do not depict featured individuals actual genitals, breast, buttocks, and anuses, they hijack people's sexual and intimate identities".

3. Faking faces. Genesis, history, and meaning of deep-fake technology

Although the term deep-fake has recently entered into everyday language vocabulary, it is worth spelling out its meaning, history, and genealogy from the outset.[8] In a recent study, Loveleen Gaur and colleagues have defined deep-fakes as follows:

> A collection of "deep learning" and "forgery", which employs deep learning algorithms to modify images, acoustic, and video to generate a synthetic/phony media. It is a non-autonomous

[5] "The ethics of a deepfake Anthony Bourdain voice", by Helen Rosner, July 17, 2021, *The New Yorker:* https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice

[6] See, Karen Hao, "Deepfake porn is ruining women's life. Now the law may finally ban it", *MIT Technology Review*, February 12, 2021 (accessed 18/10/2023).

[7] https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (accessed 19/10/2023). According to this Deeptrace report published in 2019, 96% of deep-fake videos circulating online have pornographic content.

[8] The terms "deepfake", "deep fake", or "DeepFake" are generally used interchangeably. Throughout this study, I will emply the term "deep-fake".

process that applies AI algorithms to subject matter, producing doctored images, video, and audio" (Gaur, Mallik, Zaman Jhanjhi 2023: 2).

The term deep-fake results from the collapsing of two words as it is a portmanteau of "deep learning" and "fake". The first head of this duplet refers to deep leaning technology on which the creation of deep-fake relies on for its creation of doctored footage. The second term refers to the counterfeited nature of such videos and their potential to spread false information. As we shall see in what follows, definitions of deep-fakes are not wholly clear as they prove to be somewhat inadequate or misleading; further, they are not exempt from criticisms.

The Oxford English Dictionary provides a more generic definition of deep-fakes as compared to the one outlined above as it stresses the relation between deep-fake and the wider phenomenon of disinformation, to which it interpenetrates. Indeed, deep-fake is thought of as "a video of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information".[9] Deep-fakes are, thus, videos in which human faces have been swapped using Machine Learning and Deep Neural Networks. The output is the creation of very realistic simulations of existing videos. However, Deep-fakes are not limited to videos only as they can mimic recorded audio, sounds and voices, too.

An emblematic case of Deep-fake that has managed to falsify the subject's voice and, hence, the "phonognomic" aspect of one's own identity, is that of the famous Canadian clinical psychologist, Jordan Peterson. On April 2019, a company called *Coding Elit*e, indeed, duplicated Jordan's recorded voice and his manner of speaking with such precision that the artificial intelligence forged-voice was indistinguishable from the authentic one.[10] This 'vocal avatar' sang rap songs by Eminem, perfectly reproducing the intonation, the voice, and the way the words were pronounced by Jordan Peterson, creating very believable fakes. In the same year, another company called *notjordanpeterson.com* created an online artificial intelligence engine that allowed anyone to type in anything and have it reproduced in Jordan Peterson's voice,[11] included swearing, offensive language and non-sense talk.[12]

Other types of deep-fakes simulate the physical appearance and the facial expression of people, thus, purporting and mimicking the physiognomic element of the subject. An emblematic case of deep-fake video is the artificially-created speech delivered by the former US President Barack Obama, already mentioned earlier, that was almost indistinguishable from the real one. In this famous deep-fake, created by the American actor and director Jordan Peel, Obama warns about the dangers of fake news and disinformation.

Yet the illustrations of this technology are now legion and span from pornographic deep-fakes to political satire, from cinematic deep-fakes to applications in various domains such as art, education, psychiatry, to mention but a few. This issue has divided the public opinion in two opposite fronts: the tech–enthusiast, who sees an unprecedented technology that may enhance education, the Arts and the film industry (Kerner and Risse 2021), and the techno–phobic, which worries about the potentially negative outcomes. The phenomenon of deep-fakes, thus, opens up a vast array of issues with ramifications in many fields, from digital forensics to artificial intelligence.

Let us brackets, for the time being, the nitty-gritty of the definitions outlined above and the open issues that may arise from putting this concept under careful scrutiny. I will come back to this point in the reminder of this study. From these aforementioned cursory definitions, however, some benchmarks constituting the foundations of deep-fakes can nonetheless be extracted. I will lay out mainly three: i) the human face as the kernel of deep-fakes; ii) the mechanics behind the production

[9] https://www.oed.com/dictionary/deepfake_n?tl=true (accessed 12/04/2023).
[10] "The deepfake artist must be stopped before we no longer know what's real", by Jordan Peterson, August 23, 2019, *National Post*, https://nationalpost.com/opinion/jordan-peterson-deep-fake (accessed 16/10/2023).
[11] "Make Jordan Peterson say anything you want with this spooky audio generator", by Matt Novak, August 16, 2019, https://gizmodo.com/make-jordan-peterson-say-anything-you-want-with-this-sp-1837306431"
[12] https://www.notjordanpeterson.com/ (accessed 18/10/2023). This website is now disabled.

of deep-fakes and the hidden layers of artificial intelligence; and iii) the concepts of "fake" or "fakery" that is implicit in the definitions of deep-fake.

Firstly, the human face and the outer appearance of the people purported in the video are the kernel (or at least one of the benchmarks) of this new technology. As a matter of fact, an instance of deep-fake intends to create an effect of 'convincingness' (or 'believability') and in order to pass off for something believable, it uses and duplicates the most authentic and unique aspect of the self, that is, one's own face. Much has been written on the importance of the face as the quintessential aspect of identity and this spares me from discussing this issue in greater details (Belting 2013; Gurisatti 2006; Schmitt 2012; Strauss 1969; Tomkins 1997). Interestingly enough, however — and this is perhaps the specific signature that sets deep-fakes aside from older type of media and other forms of visual manipulation — the face not only can be considered the core element of one's identity, but it becomes a simulacrum that can be altered, manipulated, simulated, and duplicated at will, *ad infinitum*. The principle of technical replicability of one's likeness is pivotal in this context. Many scholars have insisted on this point, arguing that this shift, which results in the infinite replicability of images, occurred in the passing from analog to digital photography, but with deep-fake technology this transformation reached its pinnacle and opened up new scenarios. Whilst the technical reproduction of images has been feasible at least since the invention of photography, the emergence of AI and its fast development has given rise to unprecedented forms of digital manipulation and image creation. Undoubtedly, AI has had a huge impact on the creation, processing and circulation of images, as it is shown by *Dall-E*, *Stable Diffusion* and similar inventions.

This is an important aspect that this type of technology has called attention to and has been able to be used as a leverage and to be capitalized on. Indeed, as I will argue next, by following the footprints of Hans Belting, it can be maintained that in the current hyper-digitalized society, there has been a sharp shift from faces to 'cyberfaces'. In this new context, the infinite replicas of one's likeness has perhaps reached its highest point and this has paved the way to unprecedented scenarios where the face, *sensu stricto*, does not necessarily exist any longer as the counterpart of the mask. The present scenario is, indeed, constellated by faceless masks, as it were, in which the nexus between being and appearing is eroded and obliterated, giving rise to 'digital masks' where there is no longer a face beneath the cover or a substance underneath the appearance.[13]

Indeed, in the last chapter of the book *Face and Mask*, the art historian Hans Belting concludes his work by tracing the scenario of facial images in the current technological world. In this epilogue of the face in the internet culture, Belting coined the term "cyberfaces", which he defines as follows:

> Cyberfaces are not faces but rather digital masks with which the production of faces has reached a turning point in the modern media. The total mask is basically no longer a mask, because nothing and no one is there anymore whom it represents or conceals. One could also say: faces can be produced that belong to no one but exist only as images. Cyberfaces no longer represent faces, but only interfaces among an infinite number of potential images. Digital masks no longer need physical bearers and have become disembodied (Belting 2017 [2013]:

It is not difficult to find examples of Belting's 'cyberfaces' in the current digital culture. The website *thispersondoesnotexist.com*, for instance, gained quite a bit of popularity for being able to generate realistic face images that are altogether invented as these are faces that do not exist in real life and do not have a counterpart in the physical world. Likewise, in 2017, The French artist Raphael Fabre managed to fool the French authorities by letting them issue a valid ID card with a synthetic image produced via computer (Conte 2019). More recently, the photographer Boris Eldagsen created an AI-generated image that won the Sony photographic competition. This brings up the age-old and much

[13] For a more historical and genealogical account on how the disjointing between face and mask occurred, see Gurisatti (2012, 15-410).

disputed question of the alleged loss of the referent due to the advent of digital photography or the loss of indexicality in digital media.

The idea in itself is quite simple. There seems to be a gradual shift from a tight connection of signs to their referents to a situation where signs and referents are detached from each other, or where the referent does not seem to exist at all, although the image may give the impression that there is an actual referent. This shift has occurred especially in connection with the passage from analogic to digital texts and has now reached its pinnacle with AI and deep learning techniques. I have decided not to enter into the relevant debate about this issue but direct the reader to the copious literature already existing (Baudrillard 2004; Marra 2006; Gurisatti 2012; 2019).

From what has been said, it can be argued that deep-fakes, whilst capitalizing on mimicking and copying one's likeness in a very realistic fashion, it also nullifies the human face as it was thought of up to now because, in this context, faces become digital masks or "interfaces", infinite replicas and empty signifiers. As Jirsa and Rosenberg (2019: 3) argued, "the current flood of digital images (…) indicates that there is, indeed, a problem of a facial dissolution when the face as a guarantee of a recognizable identity simply disappears". Other scholars have attempted to fathom this phenomenon by coining similar terms such as "artificial faces" (Leone 2021).

I will now turn to the second point to ponder, namely, – how the mechanics of digitally copying one's face actually works. Needless to say, the technical reproduction and the infinite copying of one's likeness calls attention to the methods and the techniques employed for creating altered digital faces. This dovetails with the second aspect that is important to stress at this juncture, namely, the term "deep" that is engrained in the concept of deep-fakes, which has to do with the ways of their production. Indeed, the creation of deep-fakes capitalizes on highly advanced deep learning algorithms, whose mechanisms constitute their unmistakable signature which create at one and the same time a sort of mixture of excitement and incredible verisimilitude or 'uncanniness'. Whilst in fact this technology stretches the limits of imagination creating perfect replicas of the representation of oneself, it also puzzles the viewers since for many, the mechanics behind such creations remain merely as a black box, in face of the claims that deep-fake is a very accessible technology for the vast majority of people.

This aspect requires that I introduce two terms: the concepts of "latent space" and "latent face" to which deep-fakes interlock. "Latent space" may sound like a word with an obscure meaning, but this term is just a buzzword to describe a fairly simple concept. Let us say right away that there is a metaphorical sense and another, more restricted and technical sense of the terms "latent space" and "latency". Latent means "hidden", "occult" and refers to something that cannot be seen. The term latent is often juxtaposed with its opposite in the semantic pair latent/manifest or latent/potential. In semiotics, the latent/manifest nexus is certainly nothing new. In this article, I will not dwell on the genealogy, the semantics and the metaphorical uses of this term, which would require a much lengthier and in-depth discussion. Instead, I would like to recall the technical meaning of the concept of "latent space".

Suppose we have a lot of information at our disposal – and we provide this information to a neural network. By providing a series of information to the neural network, regardless of the quantity, sophistication and quality of the information, we have given an input to the network. From this input, the neural network will extrapolate an output, i.e., a series of transformations of the input previously provided. Suppose, again, that we provide the neural network with a large amount of information regarding a specific object, for example, an entire database of wines in which many variables are present, such as acidity, colour, viscosity, provenance, etc., for a total of twenty variables. Hence, although the information space will have a size of 20, the task of the neural network would be to tell us only whether the quality of the wine is good (1) or bad (0).

The "latent space", then, will be represented by the space of the outputs, namely, all the possible outputs that can be calculated and, in essence, consists of a representation with one variable ('good' or 'bad') in terms of the quality of the wines that were originally described using twenty different variables. The adjective "latent", therefore, should be interpreted to mean that that single variable

('good/bad') hides within itself a sort of summary or skeleton of the previous twenty variables. This is an important point since deep learning makes data compression one of its cornerstones, as compression "allows us to get rid of any extraneous information, and only focus on the most important features".[14]

Latent space is, indeed, based on the so-called manifold hypothesis, that is, on the assumption that the variables we observe are in fact controlled by a much smaller number of variables, which we do not observe directly and are, therefore, "latent" or "hidden". In the case of images or facial images, the process is similar: the variables that can be observed — in this case the pixels of the image, which amount to tens of thousands — are actually controlled by a much smaller number of variables. In the case of images representing faces, for example, if two variables were used, the representation of the latent space is as if they were points on a map and the latent variables longitude and latitude. The closest the facial images are mapped onto the latent space, the bigger their similarity. It is also important to emphasize that thousands of other images are needed in order to reconstruct the latent space of an image, not to mention for images of human faces. This is the process and the technology that lies behind deep-fakes and that allows one to encode features for the formation of "latent faces". Once the "latent face" has been created by encoding all salient features of one particular facial image, hence creating a sort of model of that particular face, then this will be used by the algorithms, in turn, to decode and reconstruct deep-fake faces from that model.

Another point that should be considered is that the generation of deep-fake faces involves a model based on a "generator" and a "discriminator" (fig. 1).

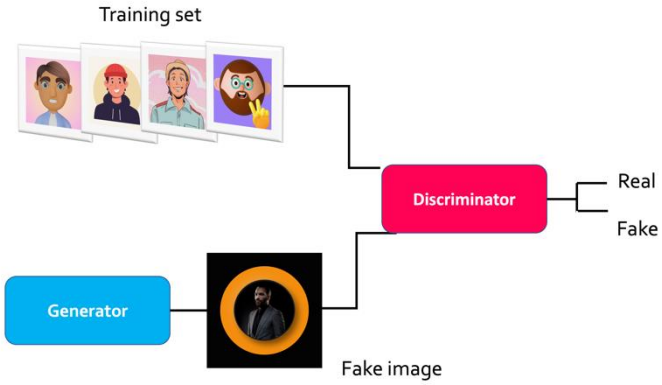

Fig. 1 The generation of deep-fakes images

By using the large data sets (also called "training sets") at its disposal, the "generator" creates a fake image that the "discriminator", in turn, must be able to detect whether its nature is real or fake. The dynamic that goes on between generator and discriminator and their mutual learning is very interesting. It seems, metaphorically speaking, similar to what is going on in any game of acumen, when one party must be able to 'outsmart' the other by leaning from each other moves.

This said, I shall now turn to the third and last point enumerated above, namely, the notion of "fake" that is embedded within deep-fakes.

4. Questioning the definition of deep-fake: Towards an ethical and benign use of deep-fake technology

Undoubtedly, the notion of "fake" is key to the definitions of deep-fakes, as the term itself suggest, although the word "fake", as we shall be seeing in what follows, remains generally vague and ill-defined. The same goes for terminology such as "real", "true", "false" and the like. Moreover, there

[14] https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d (accessed 17/08/2023)

seems to be, as I will detail henceforth, an implicit set of assumptions according to which deep-fakes have a proclivity for deception, that they are designed to deceive and that their degree of deceitfulness is somewhat higher and more refined and sophisticated then, say, other forms of deception. These assumptions are probably due to the milieu in which deep-fakes originated and were created for the very first time – i.e. in the context of nonconsensual pornography and "revenge porn" – and the negative effect that has, regrettably, derived from it.

The history of deep-fakes is, up to a certain extent, the history of a misnomer. Indeed, the genesis of the term deep-fakes goes back to 2017, thus, this technology has been around for quite some time now. The term was coined by a Reddit user, who used this very nickname, "Deep-fake", to post some pornographic footage (both images and videos) onto the forum. In these videos, the faces of celebrities were swapped with faces of the pornographic-actress.[15] The "Deep-fake" Reddit user employed face-swapping software in order to alter the images of the faces depicted in the videos. Subsequently, this user was banned from the platform and Reddit decided to change the policy in use in order to prevent the spreading of this content on the online platform. However, this created a precedent and the altered pornographic footage circulated and was reproduced exponentially for quite some time, creating a lot of heated debate. Afterwards, the term "deep-fake" was used, by extension, to encompasses a host of different content, not only doctored pornographic videos, but as an umbrella-term which provided a 'basket' for all types of deep-fakes.

It is clear, therefore, that from its inception, deep-fakes had a quite unwholesome aura as this technology was used for the first time maliciously by an anonymous user to spread nonconsensual pornographic content. Moreover, this was without the consent of the people depicted and involved in the video, to spread highly sensitive content incurring huge damage for the people involved, with ethical, social, juridical, and financial consequences. Thus "deep-fake", from the spheres of pornography, "revenge-porn", and non-consensual use of images where it originated, has migrated to form a broader spectrum that encompasses a much larger fan of content created for different purposes.

Today, indeed, as has already been pointed out above, deep-fakes have been used in various domains, from the Arts and cinema to museums and education, – and it has been used even in psychotherapy and psychiatry. The stigma and the negative 'aura' surrounding this practice, however, remains, and it is hard to dismantle it, despite the fact that in recent times some commentators argue that deep-fakes can be used for diverse purposes, included some ethical ones.

For this reason, the present study seeks to shed light on a un unwritten chapter in the history of deep-fake technology, that is, to draw a parallel between itself and its uses as a protective device to mask the identity of targeted individuals or 'whistleblowers', thus, showing a potentially positive and ethical usage of this technology. In what follows, I will then proceed with an exploration of deep-fake thought of as digital disguise. In order to support this thesis and to unpack the use of deep-fake as protective mask I will discuss the use of it in David France's documentary film: *Welcome to Chechnya* (2020).

It is my contention that deep-fakes is ill-defined or, at best, too restrictive—as it does not include in its definition a potential ethical and positive use of this technology. Whilst it is obvious that deep-fakes can be weaponized and used perniciously by bad actors, perhaps not enough has been done to show the potential that this technology can have in other contexts, serving more positive ends. The potential harm arising from deep-fakes should derive from its uses and misuses and is not a property of the technology as such (Prusila 2022: 66). Besides the well-documented malicious uses of deep-fakes, the 'officious' or benign use of deep-fakes needs more careful attention.[16]

---

[15] Benjamin Goggin, "From porn to 'Game of Thrones': How deepfakes and realistic-looking fake videos hit it big", Insider, June 23, 2019, available at: https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6?IR=T (last accessed 18/10/2023).

[16] The term "officious" is used by Thomas Aquinas in order to distinguish the malicious from the benign species of lying.

Moreover, the term and its definitions are equivocal or, at best, misleading inasmuch as the term 'fake', embedded in the definition of deep-fake, remains unexplained or is vague To make things clearer, what does the term 'fake' in defining deep-fakes stands for? As pointed out by Umberto Eco, whilst everyone knows what a 'fake' is, "the definitions of such terms as fake, forgery, pseudepigrapha, falsification, facsimile, counterfeiting, spurious, pseudo, apocryphal and others are rather controversial" (Eco 1987: 3).

In this section I argue that the overall definitions of deep-fakes need a reappraisal as they define the term from the point of view of the criterion of the deceptive element engrained in the technology, leaving aside its axiological aspect, that is, the intention — pernicious or benign — behind the use of the technology. In this way, such definitions not only fail to account for the potential ethical and benign uses of deep-fake technology and for deep-fakes that are overt,[17] but also misconceive deception by lumping it together with the medium used for the deception.

Indeed, since its inception in 2017, deep-fakes are enmeshed in a set of sociotechnical imaginaries. Its technology has been generally framed in terms of its potential danger, a threat to people's privacy, and as a weaponized tool in the era of disinformation.

In the social and "sociotechnical imaginaries" (Jasanoff and Kim; Taylor 2004) around this technology, there seems to be a tacit assumption that deep-fake is a threat to humanity in as much as it is easy to deceive with it. The advocates of this position argue that deep-fakes pose a threat to humanity inasmuch as they can be used as a weaponized instrument of disinformation and that they blur the distinction between 'real' and 'fake', fact and fiction. In other words, there seems to be a nexus between deep-fakes and deception as they usually fall into the same category of disinformation.

This is proven both by the definitions of the term itself, which often stress that deep-fakes are "designed" to deceive, as well as from the history and context of origin of the concept — altered pornographic footage created and spread online with malicious intent. Due to the alleged blurring of the distinction between the evidential and the fictional paradigm, 'real' and 'virtual', truth and falsehood, these new forms of synthetic media function as a potential source of disinformation and deception, but also as uncharted forms of creativity.

5. Fakes, 'deep-fakes', and 'shallow-fakes'

Let us pause for a moment and ponder this question: what really sets aside deep-fakes from other kinds of fakes? One recurrent point in the literature on the subject is the distinction between "cheap fakes" and deep-fakes (Paris and Donovan 2019).[18] A famous illustration of "cheap fake" is a video depicting US Speaker of the House Nancy Pelosi seemingly drunk that has received a lot of attention in the media.[19] As Paris and Donovan (2019: 10) showed clearly in a spectrum that charts the differences in terms of technical sophistication between "cheap" and "deep" fakes, the former requires a very small and limited amount of technical sophistication in contrast to the latter, which require very sophisticated techniques and expertise for its creation. In this video depicting Pelosi, the technique used is the speeding-p and slowing-down of the video.

[17] Covert deep-fakes have a potential to deceive; over deep-fakes use labelling and other devices to make sure the audience knows is witnessing to a manipulated video.

[18] Sam Gregory coined the term "ShallowFakes": "By these 'shallowfakes' I mean the tens of thousands of videos circulated with malicious intent worldwide right now—crafted not with sophisticated AI, but often simply relabeled and re-uploaded, claiming an event in one place has just happened in another,". See, "Deepfakes are solvable — but don't forget that "shallowfakes" are already pervasi", by Bobbie Johnson, MIT Technology Review, March 25, 2019: https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/ (accessed 30/10/2023).

[19] See, "Doctored Nancy Pelosi video highlights threat of 'deepfake' tech", May 26, 2019: https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/ (accessed 14/10/2023). See, also, "Distorted videos of Nancy Pelosi spread on Facebook and Twitter, helped by Trump" by Sarah, Mervosh, May 24, 2019, *The New York Times*, https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html

The video was not a deep-fake; nonetheless, it managed very well to fool the audience. This seems to suggest that less sophistication does not necessarily equate to being less convincing or less deceptive. Nancy Pelosi's video is a case in point as it was believed to be genuine, authentic footage. "Cheap fakes", thus, can be as effective as deep-fakes in terms of their persuasiveness —and in terms of the efficacy in deceiving others, since this is a property that is not depending of the type of technology *per se* used for the manipulation and altering on the media content.

There is one more point to consider, however. We have pointed out that deep-fakes rely on a large amount of data, the so-called "data sets" for their creation — which process is very costly. If one considers fakes in the context of Art History, one realizes that one single valuable object, recognized as a piece of art and, therefore, conceived as a *unicum* (Eco 1990)— that is, something that is by definition non-reproducible and unique — is the object of falsification by many skilled forgers, who, eventually will seek to reproduce many fake copies of the artwork. This means that one authentic object generates a host of fakes and copies that seek to mimic and pretend to be like the original.

Today, however, with the emergence of deep-fakes, it seems that the model of faking follows a reverse logic. There is not a one–to–one replica but a model of many–to–one. In fact, the production of deep-fakes, entails the extraction of images of faces through deep learning. This process requires a very large amount of data. In other words, in order to create one deep-fake, the algorithm would need a training-set of thousands and thousands of facial pictures. This model of producing fakes is somewhat a reversal of the model of the past: from a one-to-many to a many-to-one model.

To sum up, one difference between fakes and deep-fakes concerns the logic behind their production. Traditionally, the production of fakes has followed the logic of the one-to-many: one object, an authentic object, considered as *unicum* that has given rise to many copies. It seems that deep-fakes follow a different, — almost reverse —, logic which we may term as many-to-one. In order to produce one deep-fake content or doctored face, it necessitates algorithms and the access to a very large datasets of images.

5. Preserving the anonymity of sources in documentaries through deep-fake technology: the case of *Welcome to Chechnya*

I will now turn to the presentation of the case study.

Welcome to Chechnya (USA, 2020) is a documentary film, released in 2020, directed by David France. France is an investigative journalist and the director of the Oscar-winner film, *How to Survive a Plague* (USA, 2021), as well as the more recent *The Death and Life of Marsha P. Johnson* (USA, 2017).

Welcome to Chechnya offers a rare look at the persecuted LGBTQ+ people in the process of escaping Chechnya, a Republic in southern Russia. Since 2017, the Chechnyan government, led by Ramzan Kadyrov, has been aggressively purging the LGBTQ+ communities in a systematic way. The anti-queer genocide that goes on in this part of the world was left unnoticed until 2018, when *The New Yorker* published a report by Masha Gessen, exposing the work that the activists in Russia were undertaking — in order to rescue homosexual people, to hide them, and possibly sneak them out of the country (Gessen 2018). France became interested in it after reading Gessen's article and asked the LGBTQ+ activists to film the underground pipeline and to document the atrocities.

Welcome to Chechnya revolves around a Russian LGBTQ+ network and their engagement and activism in reaction to the gay-purging in Chechnya. This documentary film, therefore, exposes the persecution of the LGBTQ+ communities in Chechnya and it provides the testimonies of those people who were targeted, oppressed and persecuted by the Chechnyan regime. It offers, thus, a testimony of LGBTQ+ Chechens, who were reporting about the persecution, torture, and execution by the Russian leader Ramzan Kadyrov.

Welcome to Chechnya is remarkable for many reasons. First of all, because it provided an opportunity to the LGBTQ+ people to voice their truth, to tell their stories as victims and their

persecution because of their sexual orientation, and to expose to a wide audience the general purging of LGBTQ+ people in Chechnya. The resonance that this documentary received was huge as it exposed the 'crimes against humanity' committed by the Chechnyan government.

This said, why should one consider this particular documentary as relevant to deep-fake studies? Why did I bring up this specific example? In response, I selected this case study for several reasons. What is extraordinary about this documentary, besides the bravery involved in telling very disturbing and traumatic stories, is the technology used. David France and his team utilized new methods for concealing the people documented — fleeing for their lives. France had to promise them not to reveal their identities, because they knew that, no matter where they landed in the globe, they would still be hunted-down. Thus, the film director had to find a way that would both disguise the identity of the people documented and allow their humanity to be revealed throughout the film.

One of the most important facts is that *Welcome to Chechnya* used deep-fake technology and face-swapping in its footage and was pioneering a new face doubling-technique, called the "digital veil".[20] *Welcome to Chechnya* is, indeed, a pioneering example in the use of deep-fake as it was the first time that this technology was used in a documentary production in order to protect the identity of 'whistleblowers'. The novelty of *Welcome to Chechnya* was, therefore, the use of deep-fake technology so that the 23 subjects filmed were protected, while deciding to tell their stories.[21]

The intention to disguise the people involved in the documentary is clearly stated in the film by a caption that appears at the beginning of the documentary: *For their safety, people fleeing for their lives have been digitally disguised*. This aspect of 'labelling' is important as it characterizes this type of dee-fake as overt. In other words, there is a clear intention that is stated to disclosing the editing and manipulation of the video. The challenge that France was facing should not be underestimated: he wanted to both give the possibility for testimony to the LGBTQ+ Chechens, and, at the same time, he wanted to protect their identities.[22]

If this is not enough to contend with, France needed — for himself and his team, to respect some technical restraints. Indeed, the film director did not intend to use the usual techniques that are generally employed for identity-protection of 'whistleblowers' within the film industry or within journalism, such as blurring, boxes, black bars, dark shadows, etc. In other words, he was quite certain that he did not want to use the techniques generally used to hide one's identity because these prove to be very limited tools.[23]

There are different ways to reach visual anonymity and their effectiveness varies. In documentaries, anonymous sources have generally been reduced to a shadowy, voice-distorted figures or a pixelated blur. The identity of an individual can be hidden in various way: not only the aforementioned blur, but also by a black box, which can be used to disguise by showing the shadow of the person or by distorting the voice of the subject. Likewise, anonymity can be achieved through disguise, travesty, contextual displacement or by wearing a mask.

This said, let us pause for a moment and take a look at this issue from a more abstract level. From what has been said, it can be deduced that all the above-mentioned examples of anonymization, capitalize on a specific function that masks possess. Because masking allows a shift between identity

---

[20]See, "Digital disguise: 'Welcome to Chechnya''s face veil is a game changer in identity protection", by Patricia Thomson, Documentary Magazine, August 26, 2020: https://www.documentary.org/column/digital-disguise-welcome-chechnyas-face-veil-game-changer-identity-protection (accessed 20/10/2023).

[21] https://www.fxguide.com/fxpodcasts/fxpodcast-350-sci-tech-winner-ryan-laney/ (accessed 21/10/2023).

[22]See, "IDA Documentary Screening Series: Welcome to Chechnya | David France, LGBTQ+ Rights, Russia":

https://www.youtube.com/watch?v=ggSw0tqwdPY (accessed 21/10/2013).

[23] This aspect resurfaces in numerous interviews released by France. See the episode 3, in the *Deepfakery* series, "Identity protection with deepfakes: 'Welcome to Chechnya' director David France, https://www.youtube.com/watch?v=2du6dVL3Nuc (accessed 19/10/2023). *Deepfakery* is a series, produced by WITNESS and MIT Open Documentary Lab, of "critical conversations exploring the intersection of satire, art, human rights, disinformation and journalism" (https://lab.witness.org/deepfakery/).

and alterity, masks can serve a variety of purposes, one of which is anonymity. It is well known, that masks, from the point of view of the management of information, have two main functions: a "protective" function, which consists of hiding the identity of the subject either by the substitution of identity (passing off for someone else) or anonymity (being unrecognizable) and an "intrusive" function, which allows one to infiltrate and to gain secret information (Scheibe 1979, p. 67). These two functions of masks are complementary rather than exclusive.

The connection between masking and anonymity is not so obvious because, generally, the elements of the connotation of masking and disguise are dissembling and deceit. This point is so well laid out in a short but well-documented article by Hubert Damisch (1987, p. 776), where he has argued that being incognito can represent a "zero degree" of the mask: "The pleasure of the incognito is not to be mistaken for another, but to go around without being recognized and without being identified with anyone other than a mask, or that mask. While disguise is made to deceive, incognito does not impose any identity substitution, but only wants to nullify it" (Damisch 1987, p. 776).

We can now return to France's documentary. France's challenge was actually more subtle than that, however, as the aim of the film was not just a question of hiding or masking one's identity. Indeed, there is much more to it. France's challenge was to make the subjects unrecognizable — "even by their own mothers", as he has stated in some interviews — thus reaching the highest degree of disguise,[24]and, at one and the same time, he wanted to preserve the original emotional expressions of the subjects. In other words, his intention was to disguise the faces of the people purported in the video, while retaining the emotions of these subjects, thus, preserving the anonymity of the sources while conveying their sense of humanity.

In order to achieve this end, France had a simple but very ingenious idea: first, he got in touch with the Chechens LGBTQ+ people and promised them that their faces would be fully disguised. However, at the time he contacted the persecuted victims, he did not know exactly how he would technically achieve such anonymity. He found a technical solution to this issue only at a later stage of the film-production. Peering up with the engineer Ryan Laney, they started tinkering with AI, DeepFakes, and facial-obscuring technology. Then, he looked for LGBTQ+ volunteers and activists who would be willing to 'trade' their faces so that the face of the volunteers could be swapped with that of the persecuted people. France's intuition ultimately led to the development of a new AI tool: a face-doubling program that uses deep machine-learning to replace the subject's face with an entirely new one that faithfully mirrors every expression of the original face of the subject. They coined the term "digital veil" — in order to refer to this new technique of face-doubling.[25]

However, before arriving at the fully-fledged idea of the "digital veil", France and Laney tried out different options and different ways of camouflaging identities. They considered illustration, animation, rotoscoping as well as various filter-mapping elements, even *SnapChat* technology.[26] France first tried out the "rotomation" technique, that basically covers the subject with cartoon-like effects. Whilst this was an appealing visual effect, it also entailed some major problems. Firstly, it did not really fully disguise people's identity. Moreover, this technique definitely had some illustrious antecedents which should be mentioned. The technique of "rotomation" is reminiscent of the so-called "scramble suit" of which Philip Dick talks about in his *A Scanner Darkly*. In this fictional story, the protagonist wears a "scramble suit" that constantly reflects different physiognomic appearances which render his features a vague, diaphanous, and indefinite blur. A visual rendering of this original idea, can be found in the animated movie, *A Scanner Darkly* (2006), by Richard Linklater.

France and his team also tinkered with the application of facial filters. However, there was an issue with this method of disguise as the face distortion achieved with this type of application produced a strange effect and yielded to a kind of dehumanization effect on the subject masked.[27] As

[24] https://www.youtube.com/watch?v=2du6dVL3Nuc (accessed 11/08/2023).
[25] https://www.fxguide.com/fxpodcasts/fxpodcast-350-sci-tech-winner-ryan-laney/ (accessed 20/10/2023).
[26] https://www.youtube.com/watch?v=2du6dVL3Nuc (accessed 11/08/2023).
[27] https://www.youtube.com/watch?v=2du6dVL3Nuc (accessed 11/08/2023).

France aimed at preserving the human side and empathic elements of the people filmed, this filters option was left out. After having experimented with some of these technical possibilities, France and his colleagues concluded that none of these were a suitable tool for their needs.

In a study entitled, *Trading Faces: Complete AI Face Doubles Avoid the Uncanny Valley*, Welker and colleagues (2020) document the experiments and the process that led to the creation of the "digital veil". The authors tested four different options of face swapping (see, fig. 2):

1) Unmasked face;
2) Full face swap;
3) Partial face swap;
4) Cartoon mask.

In option 1, the subject is unmasked, thus, the footage is original and unaltered and the face of the subject is authentic. In option 2, there is a digital super-imposition of the volunteer's face onto that of the survivor yielding to a full-face swap. In option 3, there is a digital super-imposition of the volunteer's face on that of the survivor, but the actual eyes of the survivor were preserved, thus, creating a "hybrid face" (Welker, France, Henty, Wheatley 2020). In this study, the authors have tested three versions of this technology (options 2, 3, and 4), alongside the original (unaltered) video (option 1), in order to asses which option was the least unsettling to viewers.

The result of the experiment was very surprising. As the authors pointed out, "Surprisingly, the full face swap was rated as the least unsettling (M=33.2, SD=28.22), including significantly less unsettling than the original (unaltered) face (M=41.11, SD=28.99), t(108)=2.14, p =0.034. However, the original face was less unsettling than both the partial face swap (M = 48.72, SD = 26.51; t(108) = -2.71, p = 0.007) and the cartoon mask (M = 49.39, SD = 28.85; t(108) = -2.41, p = 0.017)" (Welker, France, Henty, Wheatley 2020: 6).

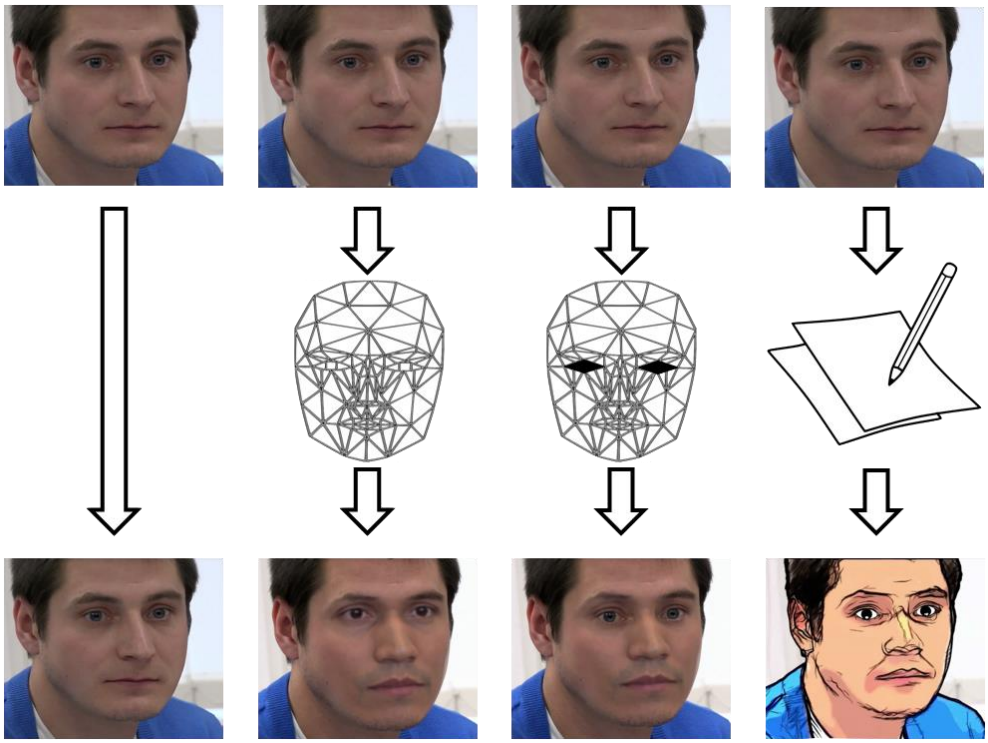

Fig. 2 Face swap options:
original face, full face swap, partial face swap, cartoon mask (Welker, France, Henty, Wheatley 2020: 3)

The documentary featured 23 Chechens, who needed to disguise their identity. Thus, France needed 23 volunteers that would provide their own faces Which he found in the LGBTQ+ activist communities in the United States and Russia. Once he had found the 23 volunteers, he rented a studio in Brooklyn and started a very long data-capture session. The casting of the face-double lasted for around six days. The main task consisted in capturing images of all these 23 volunteers from different angles, filmed against a blue screen. Indeed, for this task, France's team created a grid with nine Lumix GH-5 cameras filming the same subject from different angles — in order to catch all possible facial expressions of the subject. In the end, the casting sessions captured the range of emotions that they would need in the film.

The "digital veil", then, renders the subject fully disguised and unrecognizable, "while preserving dynamic facial expression" (Welker, France, Henty, Wheatley 2020: 1). But the film director wanted the viewers to know who was wearing the digital veil and who was not. The Russian activists who volunteered they faces generally did not. "It's an instant visual communicator that that person is in danger," Laney explained. There are some visual markers that inform the viewers watching the movie, that the people have been disguised. First of all, as pointed out earlier, there is a disclaimer text at the beginning of the film that clearly states this intention. Secondly, there is a sort of a 'halo' or an 'oval blur' (Fig. 3) around the subject's face as a visual-marker that his or her identity has been hidden. Below is a direct quote taken from an interview David France released to Witness: "The last adjustment that we made to the face double usage was to make transparent to the people who are watching the film to know who is wearing one of these face doubles and who isn't. So, we actually add it a king of 'halo' around individuals in the film so you could see that. When you are introduced to them, you know they are hiding and we would want to do that because we wanted to be honest and forthright with the people who are watching the film, but it also underscored the dangers in a way that they were living and made us recognize in scene after scene that they were taking a risk in telling their stories. So, I wanted to be faithful to that as well".[28]

Fig. 3 Application of an oval blur to the face of the survivor

6. Conclusive remarks

[28] https://www.youtube.com/watch?v=2du6dVL3Nuc (accessed 14/08/2023).

The emerge of deep-fake and AI-based image manipulation techniques poses pivotal challenges that should not be underestimated. As laid out in this study the ramifications of this phenomenon are plenty, from privacy and consent over matters of the use of personal images to artistic, creativ and ethical uses of this technology. In this study, I have provided some reflections in order to reappraise the mainstream narrative about deep-fake, as sedimented in the socio-technical imaginaries about this fairly new phenomenon, and to argue that positive implications of deep-fake technology needs consideration as much as the negative and nefarious uses.
In a nutshell, the social imagery and discourses around AI and deep-fake need deconstruction and critique and the definitions of dee-fake needs to pay attention not only to the potential to deceive intrinsic to deep-fakes, but also to its benign and ethical applications Indeed, as the case study discussed in this paper clearly shows, deep-fake technology can be used for preserving the anonymity of sources in documentaries.

Welcome to Chechnya is a case in point. The "digital veil" pioneered by Ryan Laney and successfully used in Frances's documentary clearly shows that enhancing the identity-protection of vulnerable subjects through a 'digital camouflage' generated via deep-fake technology not only is possible, but also effective and desirable in specific contexts such as the one purported in the documentary film. *Welcome to Chechnya*, thus, envisages the use of deep-fakes as protective masks and digital disguise, — thus, showing a potential positive use of synthetic faces that challenges the mainstream narrative. Such benign intention behind the uses of deep-fake in France's documentary testifies that even deception can have a positive end.

Welcome to Chechnya is an example of digital jujutsu, as it were, because it uses digital disguise as a means of the 'weak' to outmanouver the powerful and the strong via anonymity and invisibility that still holds a sense of humanity (avoiding the "uncanny valley" effect). Ultimately, we should ask whether those 'lies that tell the truth', as it were, as in the case of *Welcome to Chechnya*, represent new and uncharted forms of the 'semiotic warfare', that show that what matters is not so much the technology in and of itself but how we use it. This does not mean that deep-fakes cannot be used in order to lie, to deceive or to spread false information; of course, it can, as can any other media, but we should assess each case separately. *Welcome to Chechnya* paved the way to a novel and original understanding of deep-fakes that may lead to new applications in the area of documentary production as well as to a reconsideration of this technology in the days ahead.

Acknowledgements

The author is grateful to the two anonymous referees who were tremendously helpful to improve this paper.

Bibliography

Ajder, Henry; Patrini, Giorgio; Cavalli, Francesco; Cullen, Laurence. 2019. *The State of Deepfakes: Landscape, Threats, and Impact.* Deeptrace.

Baudrillard, Jean. 2004. *Le Pacte de lucidité ou l'intelligence du Mal*. Paris: Galilée.

Belting, Hans. 2013. *Faces: Eine Geschichte des Gesichts*. München: Verlag. (Eng. Trans. *Face and mask. A double history*. Princeton, New Jersey: Princeton University Press.

Brugioni, Dino. 1999. *Photo fakery. The history and techniques of photographic deception and manipulation.* Dulles, Virginia: Brassey's.

Châtenet, Ludovic. (ed.) 2022. Images, mensonges et algorithms. La sémiotique au défi du Deep Fake. Special issue. *Interfaces Numériques* 11(2).

Chéroux, Clément. 2003. *Fautographie: Petite histoire de l'erreur photographaphique*. Crisnée, Belgium: Yellow Now.

Citron, Danielle Keats; Franks, Mary Anne. 2014. Criminalizing revenge porn. *Wake Forest Law Review*, 49, 2014, p. 345+, U of Maryland Legal Studies Research Paper No. 2014-1, Available at SSRN: https://ssrn.com/abstract=2368946

Conte. Pietro. 2019. Mockumentality. From hyperfaces to deepfakes. *World Literature Studies* 4 (11): 11-25.

Craig, Hight. 2022. Deepfakes and documentary practice in an age of misinformation. *Continuum* 36 (3): 393-410, DOI: 10.1080/10304312.2021.2003756

Damisch, Hubert. 1982. La maschera. In: *Enciclopedia Einaudi* Vol. 7: 776-794.

Eco, Umberto. 1987. Fakes and forgeries. *Versus. Rivista di semiotica* 46: 3-29.

Eco, Umberto. 1990. *I limiti dell'interpretazione.* Milan: Bompiani.

Floridi, Luciano. 2018. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology* 31: 317–321.

Gessen, Masha. 2018. A damning new report on L.G.B.T. persecution in Chechnya. *The New Yorker*, December 21, 2018.

Giansiracusa, Noah. 2021. *How algorithms create and prevent fake news: Exploring the impacts of social media, deepfakes, GPT-3, and more*. Acton, MAL Apress.

Gramigna, Remo. 2023. Some remarks on fakes and deepfakes. In A. Santangelo and M. Leone (eds), *Semiotica e intelligenza artificiale*, Roma: Aracne, 45-64.

Gurisatti, Giovanni. 2006. Dizionario fisiognomico. Il volto, le forme, l'espressione. Macereta: Quodlibet.

Gurisatti, Giovanni. 2012. *Scacco alla realtà. Estetica e dialettica della derealizzazione mediatica*. Quodlibet Studio, Macerata.

Gurisatti, Giovanni. 2019. Il delitto perfetto: Strategie dell'immagine tra analogico e digitale. *Scenari: Quadrimestrale di approfondimento culturale* 10 (1): 330–352.

Fallis, Don. 2020. The epistemic threat of deepfakes. *Philosophy and Technology*. DOI: 10.1007/s13347-020-00419-2

Farid, Hany. 2003. A picture tells a thousand lies. *New Scientist* 179 (2411): 38-41.

Farid, Hany. 2012. *Digital image forensics.* MIT Press.

Fikse, Tormod Dag. 2018. *Imagining Deceptive Deepfakes. An ethnographic exploration of fake videos.* Unpublished MA Diss. University of Oslo.

Flynn, Asher; Powell, Anastasia; Scott, Adrian J.; Cama, Elena. 2022. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse, *The British Journal of Criminology* (62): 6: 1341–1358. https://doi.org/10.1093/bjc/azab111

Jaubert, Alain. 1986. *Le commissariat aux archives: Les photos qui falsificient l'historie*. Paris: Editions Bernard Barrault.

Jasanoff, Sheila and Sang-Hyun, Kim (eds.). 2016. *Dreamscape of Modernity*. *Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.

Jirsa, Tomáš and Rebecca Rosenberg. 2029. (Inter)faces, or how to think faces in the era of cyberfaces. *World Literature Studies* 4 (11): 2-10.

Kerner, Catherine and Mathias Risse. 2021. Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics* 8 (10): 81–108.

King, David. 1997. *The commissar vanishes: The falsification of photographs and art in Stalin's Russia*. New York: Henry Holt.

Leone, Massimo. 2022. L'idéologie sémiotique des deepfakes. *Interfaces numériques 11(2). Special issue. Images, mensonges et algorithms: La sémiotique au défi du DeepFake*: 1-16. https://doi.org/10.25965/interfaces-numeriques.4847

Leone, Massimo. 2021. Volti artificiali /*Artificial faces.* Special issue of *Lexia* 37-38: 1-645.

Leone, Massimo. 2023a. The spiral of digital falsehood in deepfakes. International Journal for the Semiotics of Law 36: 385-405. https://doi.org/10.1007/s11196-023-09970-5

Leone, Massimo. 2023b. I compiti principiali di una semiotica dell'Intelligenza Artificiale. In Santangelo A., Leone M. (eds.), *Semiotica e Intelligenza Artificiale*, I saggi di Lexia, 48, 29-44.

Loveleen, Gaur. (ed.) 2023. *DeepFakes. Creation, detection, and impact*. Boca Raton: Tylor and Francis.

Marra, Claudio. 2006. *L'immagine infedele. La falsa rivoluzione della fotografia digitale*. Mondadori: Milano.

Meskys, Edvinas; Kalpokiene, Julija; Jurcys, Paulius; Liaudanskas, Aidas. 2019. Regulating Deep Fakes: Legal and Ethical Considerations. *Journal of Intellectual Property Law & Practice* 15(1): 24-31.

Mostert, Frederick and Sheyna Cruz. 2023. Perspectives on international image rights over the past twenty years. In Bosher, Hayleigh; Rosati, Eleonora (eds.), *Developments and Directions in Intellectual Property Law.* Oxford: Oxford University Press: 161-92

Nickell, Joe. 1994. *Camera Clues: A handbook of photographic investigation*. Lexington: The University Press of Kentucky.

Paris, Britt and Joan Donovan 2019. *Deepfakes and Cheap Fakes. The manipulation of audio and video evidence*. Report. New York: Data & Society.

Pawelek, Maria. 2022. Deepfakes and (democracy) theory. How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *DISO* 1, 19 https://doi.org/10.1007/s44206-022-00010-6

Poulsen, Søren V. 2021. Face-off – a semiotic technology study of software for making deepfakes. *Sign Systems Studies* 49 (3-4): 489-508.

Prusila, Amanda C. 2022. *Truth, Lies, and 'Deepfakes': The Epistemology of Photographic Depictions*. MA Diss. Carleton University.

Santangelo, Antonio. 2022. Il volto del futuro nell'era dei deep fake. In M. Leone (ed) Il metavolto. FACETS digital press, Torino: 19-41.

Satariano, Adam and Paul Mozur. 2023. The people onscreen are fake. The disinformation is real. *The New York Times*, Feb. 7, 2023, https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Scheibe Karl E. 1979.*Mirrors, Masks, Lies and Secrets. The Limits of Human Predictability*. New York: Praeger

Schmitt, Jean-Claude. 2012. For a history of the face: Physiognomy, pathognomy, theory of expression. *Zeitschrift für Kunst- und Kulturwissenschaften* 40: 7-20 Special issue. *EN-FACE. Seven Essays on the Human Faces.*

Smargiassi, Michele. 2009. *Un'autentica bugia. La fotografia, il vero, il falso*. Roma: Contrasto.

Taylor, Charles. 2004. *Modern social imaginaries*. Durham & London: Duke UP.

Tolosana, Ruben; Vera-Rodriguez, Ruben; Fierrez, Julian; Morales, Aythami; Ortga-Garcia Javier. 2022. An introduction to digital face manipulation. In:Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez and Christoph Busch (eds.), *Handbook of Digital Face Manipulation and Detection. From Deepfakes to Morphing Attacks*, Springer 3-27

Tomkins, Silvan. 1995. The Phantasy Behind the Face. In *Exploring Affect: The Selected Writings of Silvan S. Tomkins*, ed. by Virginia E. Demos, 263–278. Cambridge: Cambridge University Press.

Viola, Marco and Cristina, Voto. 2023. Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese* 201 (1): 1-20.

Welker Christopher L., France David, Henty Alice, Wheatley Thalia. 2020. Trading faces: Complete AI face doubles avoid the uncanny valley. DOI: https://psyarxiv.com/pykjr/