Philosophy & Technology (2021) 34:1311–1332
https://doi.org/10.1007/s13347-021-00459-2

RESEARCH ARTICLE

The Distinct Wrong of Deepfakes

Adrienne de Ruiter[1]

Received: 14 September 2020 / Accepted: 21 May 2021 / Published online: 10 June 2021

Abstract

Deepfake technology presents significant ethical challenges. The ability to produce realistic looking and sounding video or audio files of people doing or saying things they did not do or say brings with it unprecedented opportunities for deception. The literature that addresses the ethical implications of deepfakes raises concerns about their potential use for blackmail, intimidation, and sabotage, ideological influencing, and incitement to violence as well as broader implications for trust and accountability. While this literature importantly identifies and signals the potentially far-reaching consequences, less attention is paid to the moral dimensions of deepfake technology and deepfakes themselves. This article will help fill this gap by analysing whether deepfake technology and deepfakes are intrinsically morally wrong, and if so, why. The main argument is that deepfake technology and deepfakes are morally suspect, but not inherently morally wrong. Three factors are central to determining whether a deepfake is morally problematic: (i) whether the deepfaked person(s) would object to the way in which they are represented; (ii) whether the deepfake deceives viewers; and (iii) the intent with which the deepfake was created. The most distinctive aspect that renders deepfakes morally wrong is when they use digital data representing the image and/or voice of persons to portray them in ways in which they would be unwilling to be portrayed. Since our image and voice are closely linked to our identity, protection against the manipulation of hyper-realistic digital representations of our image and voice should be considered a fundamental moral right in the age of deepfakes.

Keywords Deepfakes · Digital ethics · Deception · Misrepresentation · Deepfake porn · Social identity

✉ Adrienne de Ruiter
A.deRuiter@uvh.nl

[1] University of Humanistic Studies, Utrecht, The Netherlands

1 Introduction

Rapid developments in information technology and artificial intelligence (AI) present considerable challenges. Governments, institutions, enterprises, and civilians have become increasingly dependent on digital information systems for sensitive infrastructure, including virtual networks involved in national defence, financial transactions, repositories of personal data, and the healthcare system, thereby heightening the vulnerability to cyber threats (Farwell & Rohonzinski, 2011; Nye, 2017; Valeriano & Maness, 2015; Weimann, 2015). Further issues arise in liberal democracies as a consequence of technological advancements in the field of political communication where the effective functioning of democracy is undermined by the dissemination of fake news (Figueira & Oliveira, 2017; MacKenzie & Bhatt, 2020; Zannettou et al., 2019), micro-targeting (Wilson, 2017; Zuiderveen Borgesius et al., 2018), and cyber subversion and information warfare (Paterson & Hanley, 2020; Polyakova & Boyer, 2018). These developments harm key pillars of the democratic system through the spreading of disinformation, the polarisation of social and political divides, the erosion of trust in electoral results, and the disintegration of a shared public sphere where people with differing views can engage in constructive debate. While the implications of these developments are considerable, a recent breakthrough in information technology and AI is predicted to have an even greater impact on politics and society: the rise of deepfakes.

Deepfake technology refers to machine learning techniques that can be used to produce realistic looking and sounding video or audio files of individuals doing or saying things they did not necessarily do or say. While earlier deepfakes contained tell-tale signs of inauthenticity, such as unnatural eye blinking patterns or inconsistent head poses, the ongoing interchange between experts developing and detecting deepfakes and feedback loops incorporated in the deepfake production process have led to the creation of ever more sophisticated material (Agarwal et al., 2019; Li et al., 2018; Yang et al., 2019). Deepfake footage therefore becomes increasingly difficult to detect.

Over the past years, several articles have appeared that touch on the ethical implications of this technology (Caporusso, 2021; Citron & Chesney, 2019; Chesney & Citron, 2019; Diakopoulos & Johnson, 2020; Fletcher, 2018; Franks & Waldman, 2019; Meskys et al., 2020; Öhman, 2020; Silbey and Hartzog, 2019; Spivak, 2019; Westerlund, 2019). Concerns have been raised about the potential use of deepfakes for blackmail, intimidation, and sabotage (Chesney & Citron, 2019; Citron & Chesney, 2019), ideological manipulation (Fletcher, 2018: 467), and incitement to violence (Chesney & Citron, 2019; Citron & Chesney, 2019). Broader implications in terms of trust, accountability, and plausible deniability have also been discussed (Citron & Chesney, 2019; Fallis, 2020; Fletcher, 2018; Maras & Alexandrou, 2019; Rini, 2019; Vaccari & Chadwick, 2020). While this literature performs an important task in identifying and signalling the potentially detrimental consequences of deepfake technology and assessing various technological, judicial, and administrative strategies for its regulation, notably less attention has been paid to the moral dimensions of deepfake technology and deepfakes themselves. This article will help fill this gap by analysing whether deepfake technology and deepfakes are intrinsically morally wrong, and if so, why.

It seems uncontroversial to claim that deepfakes seem wrong to many people. There is something uncanny about footage that realistically portrays people as saying or doing things they have not necessarily said or done. The term itself furthermore evokes a sense of thorough deception. Nonetheless, it is not directly evident that this feeling of uneasiness also signals a deeper moral wrongness. After all, deepfake technology could potentially be used for good. Companies are working on beneficial applications, such as software that can artificially regenerate the voice of people who are unable to speak due to illnesses, such as ALS (Citron & Chesney, 2019: 1771; Meskys et al., 2020: 28). If deepfake technology and deepfakes can be used for a range of purposes, some of which are neutral, benign, or even morally valuable, our gut reaction of abhorrence may not adequately reflect their moral character. This article therefore seeks to develop a critical vocabulary to describe and analyse the feeling of wrongness commonly linked to deepfakes by considering whether deepfake technology and the products it creates violate any fundamental moral norms.

The central argument that will be developed is that deepfake technology and deepfakes should be seen as morally suspect, but not intrinsically morally wrong. Although beneficial uses of deepfake technology are conceivable, techniques aimed at producing deceptive material are directed towards an ethically questionable objective. Deepfake technology is most relevant in situations where persons are unable or unwilling to appear in footage as saying or doing certain things. While the application of deepfake technology can be beneficial when people are unable to represent themselves in video or audio material as acting or speaking in ways in which they would like to act or speak, it is harmful when it is used to represent persons in ways in which they do not wish to be portrayed as acting or speaking. Since the production of deepfakes requires little in terms of photographic or auditory input, it is relatively easy for third parties to create deepfake footage of people against their will. This technology thus lends itself for harmful applications that use persons as mere means for the ends of third parties, which violates the fundamental moral principle to respect the will of people in treatments that concern them.

Deepfakes, it will be argued, are nonetheless not necessarily morally wrong. The moral evaluation of specific deepfakes depends on whether the represented person(s) would object to the way in which they are represented, on whether the deepfake deceives viewers, and on the intent with which the deepfake was created. The key factor that renders deepfakes morally wrong is the use of digital data representing the image and/or voice of persons to portray them in ways in which they would be unwilling to be portrayed. Deepfakes that feature individuals who would object to thus being deepfaked violate the fundamental right that people have to influence the way in which they are realistically portrayed through the digital representation of their image and voice as central markers of their identity. Deepfakes severely disrupt the interpersonal processes through which a person's identity is socially constituted by presenting hyper-realistic portrayals of individuals over which they were unable to exert any influence. Since our face and voice are closely tied to our social and personal identity, protection against the manipulation of digital representations of our face and voice should be considered a fundamental moral right in the age of deepfakes.

The first part of the article explains what deepfakes are, how they are produced, and what accounts for their singular ability to realistically portray people as doing and saying things they did not necessarily do or say. The second part presents key examples of actual and potential uses of deepfake technology to offer an empirical backdrop for the normative assessment. The third part considers the ethics of deepfake technology and argues that while this technology is morally suspect, it is not intrinsically morally wrong. The fourth part considers when deepfakes are morally wrong. The article ends with final reflections that identify avenues for further research.

2 What are Deepfakes?

Deepfake technology is a relatively recent phenomenon. The wider academic community learned about deepfakes in 2016 when Justus Thies and his colleagues presented their research on real-time face capture and re-enactment at the Conference on Computer Vision and Pattern Recognition (Thies et al., 2016). The technology developed by Thies et al. allows for person A (called the “source”) to control the facial expressions of person B (called the “target”) in a recording in the sense that realistic looking video images are produced of person B “mimicking” the expressions actually made by person A. Outside academic circles, the technology became popular on Reddit in late 2017 after a user under the name “deepfake” posted various videos in which the faces of famous actresses had been inserted into scenes taken from pornographic material (Cole, 2018a; Harris, 2019). A few months later an app was shared on Reddit allowing users to create their own videos using “face swapping” (Chawla, 2019). These two techniques for creating deepfake videos differ in that the first draws from facial re-enactment (through which the visual representation of the target’s face is made to follow the expressions of the source), while the second uses a face swap technique (through which the characteristics of the target’s face are imposed on that of another person in an already existing video) (Diakopoulos & Johnson, 2020: 3). The key similarity between these techniques, classifying both as deepfake technology, is their use of machine learning techniques to produce realistic looking and sounding material that do not correspond to occurrences that have actually taken place as they are presented in the video or audio footage.

While early deepfake materials showed clear signs of manipulation, they soon became more and more convincing due to technological advancements (Agarwal et al., 2019). The creation process draws from two techniques: neural networks and generative adversarial networks (GANs). In the case of face swap videos, the first step in the production process consists in feeding a large amount of visual material to a computer, which allows it to learn about the specific characteristics of the target’s face. This process mimics the way in which learning takes place in the human brain, namely through the creation of neural networks, which allow for knowledge to be activated and reproduced more quickly and effectively when a person has been exposed more frequently to examples or experiences of something. In a similar way, computers learn to offer more accurate representations of the target's face after having "seen" more footage of it.

The second step in the deepfake creation process focuses on evaluating how convincing the produced material is. At this second stage, the created material is evaluated by what is called the "discriminator"—i.e. a network that assesses whether the material produced by the "generator" network is genuine or not. The discriminator bases this assessment on a comparison of the material with real footage it has available of the face of the target and general knowledge it possesses of what a human face looks like, how it moves, etc. The generator and discriminator networks thus work against each other since the former tries to create convincing fake content, while the latter aims to spot any differences with genuine footage, authorising it to reject the material created by the generator. Through this in-build process of automated self-criticism, the software learns to become ever better at creating realistic looking outputs.

The use of GANs thus allows for the continued improvement of the produced footage. When scholars find ways of recognising deepfake content and publish these results, deepfake developers can feed this information back to the discriminator, allowing it to reject any material that displays the feature that exposes it as a deepfake (such as, for example, the unnatural eye blinking patterns or inconsistent head poses mentioned earlier). This two-tier production process allows networks to learn from flaws in earlier material, making new deepfakes ever more difficult to detect (Fletcher, 2018, 459).

The more sophisticated production process and results acquired using machine learning techniques of "deep learning" distinguish deepfakes from other manipulated data and audio files, sometimes called "cheap fakes" (Paris & Donovan, 2019) or "shallow fakes" (Johnson, 2019).[1] Paris and Donovan (2019) argue that the impact of materials tampered with cheap and easily available tools should not be underestimated, while Johnson (2019) points to the effects that even the simple relabelling and re-uploading of videos on social media can have. Citing human rights campaigner Sam Gregory, Johnson (2019) reports, for example, how the reuse and relabelling of a video of a person burned alive incited violence in Ivory Coast, South Sudan, Kenya, and Burma. The related concepts of cheap fakes and shallow fakes draw attention to the fact that deception and misrepresentation do not require the sophisticated means provided by machine learning. Nonetheless, deepfake technology opens up unprecedented possibilities by allowing for the creation of ever more realistic footage of events that did not necessarily occur.

[1] The distinction between deepfakes and cheap fakes may give the mistaken impression that deepfakes are expensive to make. This is no longer the case as the dissemination of deepfake applications on the internet and through mobile phone apps has made the technology freely available to users.

3 Uses of Deepfake Technology

The potential uses for deepfake material are vast and varied. A well-known example of a deepfake video shared widely on the internet shows Barack Obama calling Donald Trump a "total and complete dipshit" (BuzzFeed/YouTube, 2018). After the first minutes of the video, in which Obama appears to make a number of out-of-character statements, the video goes to a split-screen shot, revealing that the claims are not made by the former US president, but by actor, comedian, and director Jordan Peele. As Peele talks, "Obama" continues to lip-sync the words spoken by Peele. The video warns about the risks of deepfakes, emphasising the need to be careful to trust internet sources.

While the literature on the ethical implications of deepfake technology tends to highlight its potentially far-reaching consequences in politics, the most common current use of deepfake technology is for pornographic purposes. A research by cybersecurity company Deeptrace, published in October 2019, found that 96% of deepfake videos online were of a pornographic nature (Ajder et al., 2019). The majority of videos use face swap techniques to produce fake graphic content of celebrities (Ajder et al., 2019: 2). Provided sufficient images of a person are available, face swaps can be used to create deepfake material of virtually anyone. People have used the deepfake applications shared on Reddit to create fake pornographic footage not only of famous people (Cole, 2018a), but also of persons they knew personally, like friends or classmates (Cole, 2018b).

The use of deepfake technology for pornographic purposes clearly demonstrates the harm that this technology can do. It is important to note, however, that the technology has been used for beneficial purposes as well. Thanks to deepfake techniques, LGBT individuals who suffered from persecution in Russia were able to testify of their experiences in the documentary "Welcome to Chechnya" in an unrecognisable artificially induced guise (RD, 2020). Deepfake technology has been used to let David Beckham "speak" nine languages in a video to promote the *Malaria Must Die Initiative* (Meskys et al., 2020: 28; Westerlund, 2019: 41). The Dalí museum in Florida has an artificially induced Salvador Dalí greet visitors for its *Dalí lives* exhibition (Lee, 2019). The use of deepfake technology allows for a surreal experience of "interaction" with the deceased artist, ending with Dalí asking whether the audience would like to take a selfie with him. Another interesting illustration of the potential of deepfake technology is the audio of John F. Kennedy delivering the speech that he was to give in Dallas on 22 November 1963. The audio, produced by the advanced speech synthesis company CereProc and presented in March 2018, synthetically reproduced the speech Kennedy would have delivered had he not been shot (Floridi, 2018: 319–320).[2]

[2] Besides creating video and audio material featuring the images and voices of real people, deepfake technology can also be used to create fake photographic material or video or sound files of non-existing people, animals, and objects. For example, the anonymising of testimonies for the "Welcome to Chechnya" documentary uses deepfake techniques to take away people's identifying features and portray them with the faces of non-existing persons. This points to a distinction between what may be called "real person deepfakes" and "avatar deepfakes". While real person deepfakes attribute digitally produced forms of speech and behaviour to real individuals, avatar deepfakes attribute actual speech and behaviour of real persons to digitally produced avatars. Avatar deepfakes thus signal the use of deepfake technol-

It is also important to consider implementations of deepfake technology that companies are currently working on or that are predicted for the near future. Beneficial applications include research on software that can artificially regenerate the voice of people who are unable to speak due to illnesses, such as ALS (Citron & Chesney, 2019: 1771; Meskys et al., 2020: 28). Similar techniques could also potentially be used to help people with mourning processes by allowing them to engage in virtual conversations with loved ones that died (Meskys et al., 2020: 28; for an early discussion, see: Stokes, 2012: 370). Deepfake technology may cut the costs of film productions that require younger versions of actors for flashbacks and open up possibilities for films to feature actors who are deceased (Citron & Chesney, 2019: 1770). As the example of the Dalí museum shows, deepfake technology can furthermore be used to increase the appeal of museums, or history lessons, with historical figures being brought artificially to live in videos for educational purposes (Citron & Chesney, 2019: 1769).

While research into beneficial uses of deepfake technology is generally proudly announced by the involved companies, the development of potentially damaging deepfakes is likely to be surrounded with secrecy. It is a safe bet that governments around the world are working on their own deepfake implementations, as may be non-state actors engaged in cyberterrorism. To anticipate the potentially destabilising effects of deepfake technology, various scholars have sought to predict the plausible and possible uses of deepfakes in the near future. Chesney and Citron (2019: 147) discuss several bleak scenarios of how deepfake technology may be used to incite armed conflict, such as through “a video depicting the Israeli prime minister in private conversation with a colleague, seemingly revealing a plan to carry out a series of political assassinations in Tehran”, “an audio clip of Iranian officials planning a covert operation to kill Sunni leaders in a particular province of Iraq”, or “a video showing an American general in Afghanistan burning a Koran”. It should be noted that deepfakes may not only have a profound impact on international affairs because they can put strain or diplomatic relations or incite armed conflict, but also because they can produce a general distrust in the reliability of footage, which may heighten the reluctance of the international community to engage in humanitarian intervention on the basis of video or audio material that allegedly proves the perpetration of genocide or other mass human rights violations of which the authenticity may be hard to assess.

Deepfakes could be used to wreak havoc not only in international relations, but also in domestic politics. Diakopoulos and Johnson (2020) draw attention to various ways in which the use of deepfake technology could undermine electoral processes, such as through fabricated videos showing a candidate making sexist or racist comments, audio recordings suggesting a candidate had prior knowledge about questions asked in a public debate, or footage showing a reputed public official misdirecting

Footnote 2 (continued)

ogy to create virtual proxies, rather than digital reproductions of the image and voice of actual people. While avatar deepfakes present ethical concerns of their own, the focus in this article will be restricted to an analysis of deepfake technology and deepfakes that make use of the image and voice of real persons.

voters about the voting procedure. Fletcher (2018: 464) notes that deepfake technology can also be used in more subtle ways to manipulate electoral results, such as through "crowd-turfing"—a technique used in the fields of campaigning and marketing to manufacture the appearance (or absence) of popular support. Fletcher (2018: 467) explains that it may soon be possible to "tweak (in real time) a live feed of facial expressions of those filmed listening to a candidate speaking, retouching facial expressions slightly so that background listeners seem to leer, frown, or look bored during the speech. Such alterations bypass viewers' conscious receptions, working instead on the unconscious cues we use to foster, solidify, or undermine an overall gut impression."

Concerns have furthermore been raised about the potential to use deepfake technology for crime and to destabilise financial markets. Westerlund highlights that advanced AI technologies have already been used to create fake audios of CEOs asking for urgent cash transfers (2019: 42). With the rapid development of deepfake technology, it will soon be possible to use "real-time digital impersonation" for criminal ends (Westerlund, 2019: 43). Digital impersonation could be used not only to demand immediate bank transfers, but also to announce bankruptcy or to portray executives as committing fraud (Westerlund, 2019: 43).

While this overview of actual and potential uses of deepfake technology is not exhaustive, it provides a sense of the ethical issues it poses. The fact that this technology is used to create deepfake porn videos without the target's consent is of immediate concern. Further issues arise from the predicted use of deepfake technology in the near future to undermine trust in the democratic process and institutions, heighten social and political tensions, commit crime, destabilise financial markets, subvert diplomatic relations, and incite violence. Although deepfake technology can also be used for good, the potentially far-researching implications of deepfakes suggest that it is important to look more closely at the ethics of this technology and the artefacts it produces.

4 The Ethics of Deepfake Technology

The available scholarship demonstrates a marked concern for the ethical implications of deepfake technology and its (potential) uses, as illustrated by a rising number of publications in law, information technology, communication studies, and political science (Citron & Chesney, 2019; Diakopoulos & Johnson, 2020; Fallis, 2020; Fletcher, 2018; Franks & Waldman, 2019; Maras & Alexandrou, 2019; Meskys et al., 2020; Rini, 2019; Silbey & Hartzog, 2019; Spivak, 2019; Vaccari & Chadwick, 2020; Westerlund, 2019). While these works importantly identify and signal the potentially detrimental consequences, notably less attention has been paid to the moral dimensions of deepfake technology and deepfakes themselves. To remedy this lack, I will consider here if there is something intrinsically morally problematic about deepfake technology and deepfakes.

Let us consider first if deepfake technology is morally wrong in and of itself. A technology may be considered morally wrong if it violates moral norms (following deontology), has significant adverse consequences that outweigh its positive effects (following consequentialism), weakens virtue and/or promotes vice (following virtue ethics), or undermines interpersonal relations and fundamental social values such as trust and mutual respect (following care ethics). On first sight, deepfake technology seems to qualify as a morally problematic technology. The term itself indicates that it is closely related to deception. A technology that deceives by producing footage that is not genuine appears to be morally dubious. After all, deception violates norms of truthfulness, inspires false beliefs, constitutes a vice, and is considered harmful for social relations and trust.

From the four approaches to ethics mentioned above, deontology is most directly relevant to analysing whether deepfake technology is inherently morally wrong. Consequentialism considers outcomes decisive in determining moral rightness and wrongness, virtue ethics focuses on the ways in which the object under study shapes the character of persons, and care ethics is concerned about its effects on interpersonal relations. By considering the violation of moral norms, deontology contemplates the moral nature of acts, artefacts, or technologies most directly. I will therefore start out from a deontological approach to analyse whether deepfake technology is intrinsically morally wrong, drawing from Kant's categorical imperative, which holds that we should never treat people merely as a means, but always as ends in themselves (1959 [1785]). In the next section, I will complement this analysis with insights from Merleau-Ponty's phenomenological approach to highlight the distinctive wrong of non-consensual deepfakes.

Following deontology, deepfake technology appears to be intrinsically morally problematic because it is directed towards deception and deception violates norms of truthfulness and risks undermining people's autonomy and ability to pursue actions in line with their own will. On closer reflection, deepfake technology turns out not necessarily to lead to deception, however. According to dictionary definition, deception means "the act of causing someone to accept as true or valid what is false or invalid" (Merriam-Webster, 2021). Drawing on the work of Ryle (1949: 130), Mahon (2016) notes that "deceive" is "an achievement or success verb" which entails that "[a]n act of deceiving is not an act of deceiving unless a particular result is achieved." For an act to amount to deception, someone thus needs to accept as true or valid what is presented by the deceiver and what is in fact false or invalid. This means that for deepfake technology to be morally problematic on the grounds that it is deceptive, deepfakes need to convince (some of) the viewer(s) of the true nature of the fake footage. While deepfakes definitely hold the potential to persuade viewers that what they see is real, this effect is not always achieved.

For example, if a husband creates a deepfake of his wife in which she features in a famous late night show (Charleer, 2018), the result will not be deceptive in the sense that she will believe it a genuine representation of affairs. Similarly, when deepfake technology is used to reconstruct the voice of people who are unable to speak due to an incurable illness, their close relations will not be deceived that the artificial sound is the "real" voice of their loved one, even if the effect may be uncanny. This means that while deepfake technology holds a strong potential to deceive, it is not inherently deceptive, as the realisation of a deceptive effect depends on the conditions under which the technology is applied. Particularly when deepfakes are presented with an explicit warning that their content is deepfaked, deception can be avoided.[3]

Such a warning would also offer a partial answer to concerns that deepfake technology is morally problematic because it undermines trust in the reliability of recordings. This risk is discussed by scholars who highlight that the harm of deepfakes should not be seen to lie solely in their ability to portray people as saying or doing things they did not necessarily say or do, but also in casting doubt on the reliability of visual and auditory sources as evidence of the actual occurrence of events (Citron & Chesney, 2019; Fallis, 2020; Fletcher, 2018; Maras & Alexandrou, 2019; Rini, 2019; Vaccari & Chadwick, 2020). This issue is brought up by Citron and Chesney (2019: 1785) who discuss the "liar's dividend" that follows from the wide circulation of deepfakes and makes it easier for liars to deny that they have done or said things when recordings show up. In an extensive analysis of the issue, Rini (2019) argues that the proliferation of deepfakes can undermine our testimonial practices by ruling out recordings as a persuasive source to hold people to account for giving false statements. She contends that, as a consequence of deepfake technology, people may feel at liberty to say whatever suits their purposes, as the regulatory role of recordings is depleted by the possibility to plausibly deny that one has said or done whatever a video or audio recording reports one as saying or doing. Fallis (2020) further points out that the epistemic value of videos as a form of evidence is undermined by deepfakes, as the authenticity of recordings can no longer be taken for granted.

While the effects of deepfake technology on trust in the reliability of recordings, accountability, and plausible deniability are likely to be grave, these effects could potentially be mitigated if deepfake creators would explicitly announce the deepfake nature of their material and/or by including watermarks in genuine footage to make it easier to detect deepfakes (Chesney & Citron, 2019: 152). These steps may go some way to addressing the issue, but their potential should not be overstated. Realistically speaking, it will not be possible to provide all footage produced with authentic recording devices with a seal of authenticity. Furthermore, it should be noted that the explicit acknowledgement of the deepfake nature of material could contribute to the weakening of a sense of trust and accountability in presenting the idea that footage can be digitally manipulated in a way that is not recognisable unless the creator decides to explicitly declare that the material is deepfake. In line with this reasoning, experts have noted that mere awareness of the existence of deepfakes could lower people's trust in the reliability of video and audio materials (Hao, 2019).

While this discussion gives some reason to pause, it does not entail that we should conclude that deepfake technology is intrinsically morally wrong. The fact that a technology can have negative secondary effects is not in itself sufficient reason

[3] A warning may not even always be necessary. In some cases, the content of deepfakes is clearly not deceptive. For example, the video for the *Malaria Must Die* campaign in which David Beckham "speaks" Spanish, Kinyarwanda, Arabic, French, Hindi, Mandarin, Kiswahili, and Yoruba does not convincingly present the football player as speaking all these languages, as it does not use his own voice (Malaria Must Die, 2019). Instead, the clip is easily recognisable as a deepfake through the use of other people's voices with Beckham lip-syncing their words.

to consider the technology morally wrong in and of itself. For example, online banking technology exposes people to theft through hacking or other virtual security breaches but this does not make the technology behind online banking itself morally dubious. By analogy, it would not be right to claim that deepfake technology is inherently morally wrong because it has as an undesirable secondary effect that it contributes to the lowering of trust in the reliability of video and audio materials.

To return to the point about deception, an adjusted version of the argument could maintain that deepfake technology is morally problematic not because it automatically leads to deception, but because it holds a strong potential to deceive. Deepfake technology is prone to deceive because its mechanism is distinctive for using a machine learning technique that makes the forgery especially difficult to detect. It should be noted, however, that potential to do harm is not sufficient to dismiss a technology as intrinsically morally wrong either.[4] One can do great harm with a chainsaw, for example, or a curling iron for that matter, but this does not make the technology behind chainsaws or curling irons inherently morally wrong.

It is important to realise that relevant moral differences exist between deepfakes, chainsaws, and curling irons. Chainsaws and curling irons can be used for bad purposes, even if their main purposes are morally innocent (depending, that is, on how one feels about cutting trees and curling hairs). With deepfake technology, it seems rather to be the other way around; this technology can be used for benign purposes, even if its main end is to produce footage that is not genuine. While the products of deepfake technology may not necessarily deceive, this technology is aimed towards the creation of fakeries and misrepresentations. This in itself may be morally problematic. From a Platonic perspective at least, deepfake technology pulls us to the darkest corner of the cave.

Something similar is allegedly at stake, however, in technology that enables the creation of films, videogames, and other fictional but real-looking footage. Yet, we do not usually morally condemn such technologies on the grounds of misrepresenting reality and presenting falsehoods. There should thus be something peculiar about deepfake technology to warrant a harsher normative assessment. The most relevant distinction here seems to lie in the way in which deepfake technology can be used to manipulate digital representations of the image and voice of real persons, without their consent, to create footage that portrays them as saying or doing things they did not necessarily say or do and that they would not want to be portrayed as saying or doing. Film technology, at least in its classic sense, records acts and speech engaged in by actors. The falsehood there is produced not by what is recorded, but by the story that is played out by the actors. In the case of video games, avatars are created that do not correspond (directly) to persons in real life. Video games are therefore more similar to cartoons or books where a fictional narrative is drawn up. The ground for considering deepfakes intrinsically morally problematic therefore cannot simply lie in the creation of believable fictions. It rather seems to concern the way in which deepfakes use digital data that represent the image and voice of real persons, rather than actors or avatars, to create these fictions.

[4] I am grateful to an anonymous reviewer for indicating this point.

The use of manipulated digital representations of people's image and voice does not necessarily have to be morally wrong either, though. What seems to matter most for assessing the moral quality of deepfake technology is the reason as to why it is used to create footage of people saying or doing things they did not say or do. Deepfake technology becomes particularly relevant in two situations: when people are unable to appear in footage as saying or doing certain things or when they are unwilling to do so. While the application of deepfake technology can be beneficial when people are unable to represent themselves in video or audio material as acting or speaking in ways in which they would like to act or speak (e.g. because they are unable to speak or to go to certain places or enter into particular situations), it is harmful when it is used to represent persons in ways in which they do not wish to be portrayed (e.g. in humiliating or intimate scenes or as engaging in speech or action they do not support). Since the production of deepfakes requires little in terms of photographic or auditory input, it is relatively easy for third parties to create footage of people against their will. It is therefore a technology that lends itself particularly well for applications that use persons as mere means.

Deepfake technology is morally suspect, then, because it is susceptible to be used for actions that violate fundamental moral norms. Particularly, deepfake technology is prone to violate the categorical imperative, which holds that we should treat human beings never merely as means, but always as ends in themselves (Kant, 1959 [1785]: 429). This signifies that we need to respect that human beings have a will of their own and that we cannot treat them as mere instruments to pursue our own goals or satisfy our desires. In the case of non-consensual deepfake pornography, for example, this maxim is clearly violated.

At the same time, deepfake technology cannot be dismissed as intrinsically morally wrong. This technology can be used for good. It can be used to empower and reinforce people's autonomy, as in the case of artificial voice regeneration. It allows persons to see digital representations of themselves in settings that may be unattainable in real life or to virtually explore actions they might not want to engage in for real, but are curious about as a fictional possibility. The products of this technology can be entertaining and humorous without harming the person(s) that are featured in the material. Deepfake technology thus serves as a tool that can be used in a variety of ways. It is therefore important to consider in more detail when its use turns morally problematic and what, if anything, is distinctive of the moral wrong that characterises such morally wrongful deepfakes.

5 The Distinct Wrong of Deepfakes

So far, I have argued that deepfake technology is not morally wrong in and of itself since it can be used for neutral, benign, and even morally valuable purposes. It follows then that deepfakes, as the product of this technology, are not intrinsically morally wrong either. Their moral assessment depends on various factors. The discussion above suggests that deepfakes can be morally acceptable if they are made with the deepfaked person's consent (or at least without their known, foreseeable, or probable disapproval) and with explicit announcement of the deepfake nature of the material to prevent any unintended deceptive effects of footage that may be taken to be genuine by (some) viewers. A third criterion that is important to add for a deepfake to be morally admissible is that it should be made without malicious intent.[5] Intent needs to be included, even if it cannot be worked out in full detail within the limits of this article, because a deepfake can be morally problematic even if the deepfaked person gave or would give consent and the deepfake nature of the material is explicitly announced.[6]

From these three factors, the aspect of consent, in particular, requires further unpacking since it raises fundamental questions about the distinctiveness of deepfakes. Why is it problematic if digital data representing people's image or voice is used in a way to which they would not consent? Is the moral wrong involved here similar to that of other practices in which people are misrepresented against their will, such as defamation, libel, or slander? Or is there something special about the wrong involved in non-consensual deepfakes? To address these issues, it will be helpful to draw from insights developed in the academic literature. Two ideas are particularly relevant here: the consideration that deepfakes may harm people's reputation (Franks & Waldman, 2019: 893; Diakopoulos & Johnson, 2020: 10) and the view that deepfakes involve a form of digital impersonation or persona plagiarism (Citron & Chesney, 2019; Diakopoulos & Johnson, 2020: 10–12).

Franks and Waldman present "reputational injury" as one of the central wrongs of non-consensual deepfake porn videos (2019: 893) in noting how "[l]ike other forms of nonconsensual pornography, digitally manipulated pornography turns individuals into objects of sexual entertainment against their will, causing intense distress, humiliation, and reputational injury." These authors insightfully point to the intimate link between people's image and voice, on the one hand, and their social identity, on the other, by considering the impact that the creation and dissemination of non-consensual deepfake pornography may have on victims through reputational injury.[7]

A similar point is expressed by Diakopoulos and Johnson in their discussion of reputational harm (2020: 10). These authors note that deepfake videos that

[5] The inclusion of intent as a criterion for assessing the moral permissibility of deepfakes poses a potential issue for the moral evaluation of deepfakes. Philosophical discussions on intention point to challenges in knowing and assessing other people's intentions (see, e.g. Anscombe, 2000 [1957]; Bratman, 1987). This issue may well be aggravated in digital contexts where identities are often ambiguous and the intentions of the persons who create digital footage and of those who promulgate it may not align (Phillips & Wilner, 2017). While this indicates a need to consider the role of intent in greater detail, a full discussion of this issue requires more space than is available in the current article. Future research could make a valuable contribution addressing this theme.

[6] Consider, for example, a scenario where the campaign team of a politician of an extremist party secretly orders a deepfake programmer to create an audio file where the politician makes inflammatory statements about a minority with the intent of raising ethnic tensions in the hopes that this will bring the politician more votes in upcoming elections. The material would be produced with the politician's consent and it would be made clearly recognisable as a deepfake so that the politician can easily deny accountability for the claims made. The resulting deepfake would nonetheless be morally problematic, even if it was made with the target's consent and the deepfake nature of the produced material was unambiguously acknowledged.

[7] I will return to the points of distress and humiliation in a moment.

undermine a person's reputation can be framed as a form of defamation (Diakopoulos & Johnson, 2020: 10). The idea that deepfakes can do reputational injury is important because it highlights the effects on people's social identity. At the same time, it is important to realise that the problem does not concern a person's reputation alone. A non-consensual deepfake porn video would wrong the person portrayed even if nobody other than the target were to see it.

Diakopoulos and Johnson identify "misattribution" as a potential further harm of deepfakes in noting how "deepfakes seem to violate ownership rights of the subjects. That is, independent of the reputational harm that might be done to an individual using their image or likeness in a deepfake, there seems to be an additional harm in using and distributing the individual's image or likeness without permission" (2020: 10–11). The specific harm that characterises this misattribution (and distinguishes it from other forms of misattribution such as libel or slander) concerns the use of data representing the image and/or voice of the individuals featured in the deepfake to depict them in a realistic way in digital footage. To make sense of this distinct harm, Diakopoulos and Johnson introduce the notion of "persona plagiarism" which they define as "an inversion of plagiarism focused on the source rather than the content of a message" (2020: 11). The idea here is that whereas the original author is not credited in plagiarism, in deepfakes credits are falsely attributed to the subject portrayed in the deepfake rather than to its creator (Diakopoulos & Johnson, 2020: 11).

Diakopoulos and Johnson's concept of "persona plagiarism" resonates with Citron and Chesney's notion of "digital impersonation" (2019: 1758). Citron and Chesney observe that deepfake technology allows for increasingly convincing digitised impersonations (2019: 1758). This perspective adds to Diakopoulos and Johnson's view a focus on the digital nature of the persona plagiarism involved in deepfakes. Deepfakes, after all, do not allow for persona plagiarism or impersonation in the flesh, but only in the digital realm. While the impact of digital impersonation should not be underestimated, given that our social persona has to a large extent also become a digital persona in the digitalised world most of us live in, the fact that the kind of persona plagiarism enabled by deepfakes is digital is an important fact that should not be overlooked in our normative analysis.

The fact that deepfakes are situated in the digital realm entails that it is not the actual body or voice of people that is at stake, but the data that represent their image and voice in the digital world. Photographic, video, and audio materials are used as input, but what is created through deepfake technology is something different. It is new data that virtually represents a person's image or voice. What needs to be protected is therefore not our image and voice as such, but these digital representations, which function as identity markers for our digital persona. The harm that needs to be understood to develop a persuasive ethics of deepfakes therefore concerns what may be called "*digital* persona plagiarism".[8]

[8] This digital aspect also signals the key moral difference between non-consensual deepfake porn and (other forms of) revenge porn. While lack of consent plays an important role in both cases, the action to which the lack of consent applies is different. In cases of revenge porn, consent is not given for the dissemination (and sometimes the recording) of footage (Citron & Franks, 2014). In non-consensual deepfake porn, consent is not given for the insertion of digital data that represent one's face in pornographic material. While the potential effects of intense distress, humiliation, and reputational injury are likely to

What Diakopoulos and Johnson's notion of "persona plagiarism" grasps more effectively than Citron and Chesney's concept of "digital impersonation", however, is the idea that deepfake creators do not necessarily try to "be" the person who is deepfaked. While this is the case when deepfake technology is used for "real-time digital impersonation" for criminal ends (Westerlund, 2019: 43) or real-time face capture and re-enactment (Thies et al., 2016), this description does not fit well with non-consensual deepfake porn. The point there rather seems to be to exercise control over the other as "other". Non-consensual deepfake pornography involves an exercise of virtual coercion, which allows the creator to decide how a person appears in a video or sound recording by attributing to them speech and action in which the creator (or the person who requested the deepfake) would like to see or hear the deepfaked person engage.

Based on this discussion, we can conclude that what is most distinctive about deepfakes in a normatively relevant sense is the fact that they use digital data that represent a person's image or voice to attribute particular actions and speech to them. This attribution is morally problematic when the featured person would object to the way in which he or she is represented.[9] Deepfakes thus highlight the importance of a right to digital self-representation, which becomes increasingly important as technologies to digitally plagiarise and impersonate persons become ever more sophisticated.

The basis for this right lies in a Kantian understanding of the importance of respecting people as ends in themselves (1959 [1785]: 429). The reason as to why people are not to be treated as mere ends is that human beings are capable of reflecting on their existence, form ideas on what they believe is the right thing for them to do, and have a certain freedom to decide how to direct their actions. When we use other human beings as mere means for the realisation of our own objectives and desires, we thus undermine their intrinsic value as persons who can pursue their own aims.

While the Kantian perspective elucidates the intrinsic moral wrong of deepfakes that fail to respect people's will regarding the way(s) in which they are represented, it is not well-placed for clarifying the more distinct nature of the wrong that is

Footnote 8 (continued)

be the same, as noted by Franks and Waldman (2019: 893), the precise wrong that is done to victims is different.

[9] The centrality of consent in this account of the moral admissibility of deepfakes raises complex questions about deepfakes that feature dead persons. In line with regulations regarding the use of images of dead celebrities in advertising and marketing (Petty & D'Rozario, 2009), the next of kin of the deceased should be asked for consent for the use of their image and/or voice for the creation of deepfakes. The closeness of the heirs to the deceased person place them in the best position to assess whether the project is in line with their former wishes and to protect their reputation, even if disagreements between relatives do occasionally occur, particularly when large financial gains are involved (Petty & D'Rozario, 2009). In the case of Dalí, the use of his image for the *Dalí lives* exhibition was authorised by the Dalí Foundation in Spain, since the artist had no surviving relatives and left his estate to Spain (Lee, 2019). The situation is different when people have explicitly declared that they do not wish for their image to be used after death. The late actor Robin Williams, for example, included a clause in his will that prohibits the use of his voice and image for commercial purposes until 25 years after his death (Panyatham, 2020).

involved in non-consensual deepfakes. This is the case because Kantianism draws from a psychological account of personal identity where the self is identified with the mind. A person is identified by his or her memories, beliefs, desires, and—for Kant, above all—will. This psychological understanding of identity and selfhood does not offer the conceptual means to explain how the digital misrepresentation of people's bodily image and/or voice can constitute a severe moral violation, other than that it fails to respect their will and undermines their autonomy.

The Kantian perspective therefore should be supplemented with insights from the embodied account of identity. According to the phenomenologist Maurice Merleau-Ponty, people do not "have" a body, but "are" a body. The body offers the perspective through which people experience the world and everything and everyone in it (Merleau-Ponty, 1962: 90–97). It is therefore not possible to posit the existence of a self that is independent from this embodied experience. A view that regards personhood as distinct from bodily experience thus misses important aspects of our sense of self.

The relevance of the embodied sense of personhood is emphasised in literature on the ethics of facial graft transplantations, which notes the centrality of the face for human identity (Bound Alberti, 2017; Caplan, 2004; Carosella & Pradeu, 2006; Edgars, 2009; Robertson, 2004; Swindell, 2007). In its report on the admissibility of facial graft transplantations, the Royal College of Surgeons of England (2003, as cited in Swindell, 2007: 451) notes, for example, how "[t]he face is central to our understanding of our own identity. Faces help us understand who we are and where we come from." In his contribution to the debate on the ethics of such transplants, Robertson (2004, as cited in Swindell, 2007: 450) similarly claims that "[f]aces are the external manifestations of our persons […] and help form the image that others have of us. Indeed, our face often provides the image that we have of ourselves." Barker, a surgeon at the University of Louisville, even goes so far as to claim that "[t]he human face is you" (as cited in Swindell, 2007: 451).

These comments suggest that the face is not only important for our personal sense of self but also for our social identity, as others relate to our faces as external manifestations of our person. Considering the body more broadly, Edgars (2009: 127) explains how this works by noting how "the body is not simply mine, for it is an object that will be interpreted, evaluated and given meaning by others. To a significant degree that meaning may be beyond my control. The body is the medium through which others perceive me, and it is the medium of my communication with them." In this communication, the face is of particular importance as it is the part of our body where most of our expressivity derives from. Given the importance of expressivity for our social sense of self, the voice should also be considered a key element of our bodily presence that marks our identity in allowing us to express ourselves and be recognisable to others.

This centrality of the body and voice in the social construction of identity helps us see what is at stake in misrepresentations that use footage of our bodily image and voice to present us in digital material in realistic ways in which we do not wish to present ourselves to others. Non-consensual deepfakes wrong the persons they portray because they manipulate the process through which people's identity is socially constituted by using digital representations of their face and/or voice, as central markers of their identity, to present them in ways that disregard their will and go against their sense of self.

The shock, concern, and dismay at discovering that digital representations of one's face have been manipulated using this technology are reflected in the accounts of women who have been targeted in non-consensual deepfake porn. Kate, whose image was used to create a deepfake sex video, emphasises how the realistic representation of her face in moving images rendered the produced material incredibly convincing: "These things are so horribly believable, and you desperately want to say, 'That's not me!'" (Cook, 2019). She explains: "When it's Photoshop, it's a static picture and can be very obvious that it's not real. But when it's your own face reacting and moving, there's this panic that you have no control over how people use your image" (Cook, 2019). Amy, another target of deepfake porn, similarly observes how "people might actually believe that was me" (Cook, 2019).

The immediate response of wanting to exclaim "That's not me!" emphasises the way in which one's face is intimately tied with one's sense of self. It is difficult to disassociate oneself from footage where digital representations of one's face are inserted in a realistic looking way because the face stands for one's sense of self to an extent that other body parts do not. A similar degree of recognisability and identification could not be realised if any other part of the body would have been used. Tina, another woman interviewed by Cook, explains her reaction to seeing the video for which her image was used, highlighting the sense of dismay that seeing her face in this setting caused: "I was definitely shocked and disturbed. It felt really weird and gross to see my face where it shouldn't be" (Cook, 2019). Maya, another target, similarly notices how "[b]eing violated in such an intimate way is really a weird feeling."

These women testify of the shocking experience of being confronted with graphic material that contains representations of their own face, reacting and moving in ways that make the recording incredibly believable. While disregard for their will is an important aspect of this experience, the wrongness of this type of non-consensual deepfakes cannot be fully grasped without considering the way in which people's identity is tied up with their face and voice. If Barker and other commentators on the ethics of facial grasp transplantation are right in suggesting that your "face is you" (as cited in Swindell, 2007: 451), it follows that it is particularly hard to disassociate oneself from images where a digital representation of "you" is doing things with which you do not identify. Clearly, the manipulated digital data of deepfake footage is not your face, nor is it "you", but the testimonies of victims signal a sense of shock and dismay that we can only begin to understand through an embodied account of selfhood, which is able to make sense of the way digital representations of our voice and face are connected to our identity.

When we combine these insights into the central role that people's face and voice play in the constitution of our social and personal identity with the knowledge that deepfakes can digitally reconstruct our face and voice to present us in footage in a believable way as doing and saying things we did not actually do or say and we would not wish to be presented as doing or saying (or at least not in footage shared with others), the distinctive harm involved in this violation of people's will through deepfakes becomes visible. What is violated in non-consensual deepfakes is people's right to digital self-representation. This right evidently does not entail that people are entitled to fully determine the way in which they are represented by others in the digital realm. It establishes, however, that others may not manipulate digital data that represent people's image and voice, as markers of the self, in hyper-realistic footage that presents them in ways to which they would object.[10]

6 Concluding Remarks

This article has presented an account of the ethics of deepfake technology and deepfakes. The main argument that has been elaborated is that deepfake technology and deepfakes are morally suspect, but not inherently morally wrong. Three factors are central to determining whether a deepfake is morally problematic: (i) whether the deepfaked person(s) would object to the way in which they are represented; (ii) whether the deepfake deceives viewers; and (iii) the intent with which the deepfake was created. The most distinctive aspect that renders deepfakes morally wrong is the use of digital data corresponding to the image and voice of persons to portray them in ways in which they would not want to be portrayed. Since our image and voice are closely tied to our social and personal identity, protection against the manipulation of hyper-realistic digital representations of our image and voice should be considered a fundamental moral right in the age of deepfakes.

While this article has set out important lines for understanding the kind of harm that is involved in non-consensual deepfakes, further work remains to be done. Future research could unpack in more detail the scope and nature of a right to digital self-representation. Another issue that requires further consideration involves the ethics of producing deepfake technology and deepfakes, ordering the creation of deepfakes, disseminating deepfake technology and deepfakes, allowing for the distribution of deepfake technology and deepfakes on online fora or other venues, and watching or listening to deepfakes. A valuable contribution could furthermore be made by a more detailed account of personhood that brings together the psychological, embodied, and digital elements discussed in this work. The present article has established, however, that deepfake technology and deepfakes are not intrinsically morally wrong, even if they are morally suspect, and has made a case for recognising a right to digital self-representation, which protects people against the manipulation of digital data that realistically represent their image and voice as central markers of the self.

[10] Within the limits of this article, only the contours of this moral right to digital self-representation could be sketched. Future research could set out the nature and scope of this right in greater detail and link it to scholarship on the right to one's self-image and publicity (see, e.g. Haemmerli, 1999; McKenna, 2005; Mohebbi, 2016; Post & Rothman, 2020). The related question regarding the legal rights that should be granted to victims of non-consensual deepfakes has been discussed by scholars in law who point out that the current existing legal frameworks are inadequate to address the predicament of persons whose image or voice is deepfaked (see, e.g. Citron & Chesney, 2019: 1788–1803; Harris, 2019).

Acknowledgements I would like to thank Philip Brey for his helpful comments and support in my exploration of the issue of deepfake technology and its ethical dimensions. I also benefited greatly from the insightful feedback I received from the reviewers. The depth of their engagement with earlier drafts of this paper was truly exemplary and their suggestions and ideas helped me rethink and develop my arguments. I am furthermore grateful for the support provided by the University of Humanistic Studies with publishing this article open access.

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest The author declares no competing interests.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. & Li, H. (2019). Protecting world leaders against deep fakes. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 38–45.

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace.

Anscombe, G. E. M. (2000). [1957]. *Intention*. Harvard University Press.

Bratman, M. (1987). *Intentions, plans, and practical reason*. Harvard University Press.

Bound Alberti, F. (2017). From Face/Off to the face race: The case of Isabelle Dinoire and the future of the face transplant. *Medical Humanities, 43*, 148–154.

BuzzFeed/YouTube. (2018, April 17). *You won't believe what Obama says in this video*! Retrieved June 9, 2021, from https://www.youtube.com/watch?v=cQ54GDm1eL0.

Caplan, A. (2004). Facing ourselves. *American Journal of Bioethics, 4*(3), 18–19.

Caporusso, N. (2021). Deepfakes for the good: A beneficial application of contentious artificial intelligence technology. In T. Ahram (Ed.), *Advances in artificial intelligence, software and systems engineering. Proceedings of the AFHE 2020 virtual conferences on software and systems engineering, and artificial intelligence and social computing, July 16–20, 2020* (pp. 235–241). Springer.

Carosella, E. D., & Pradeu, T. (2006). Transplantation and identity: A dangerous split? *Lancet, 368*(9531), 183–184.

Chawla, R. (2019). Deepfakes: How a pervert shook the world. *International Journal of Advance Research and Development, 4*(6), 4–8.

Charleer, S. (2018, February 2). Family fun with deepfakes. Or how I got my wife onto the Tonight Show. *Towards Data Science*. Retrieved June 6, 2021, from https://towardsdatascience.com/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show-a4454775c011.

Chesney, R. & Citron, D. K. (2019). Deepfakes and the new information war. *Foreign Affairs*, January/February, 147–155.

Citron, D. K., & Franks, M. A. (2014). Criminalizing revenge porn. *Wake Forest Law Review, 49*, 345–391.

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*, 1753–1819.

Cole, S. (2018a, January 24). We are truly fucked: Everyone is making AI-generated fake porn now. *Motherboard*. Retrieved June 8, 2021, from https://motherboard.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

Cole, S. (2018b, January 26). People are using AI to create fake porn of their friends and classmates. *Motherboard*. Retrieved June 8, 2021, from https://motherboard.vice.com/en_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes.

Cook, J. (2019, June 23). Here's what it's like to see yourself in a deepfake porn video. Huffington Post. Retrieved June 9, 2021, from https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0faee4b0a3941861fced.

Diakopoulos, N. & Johnson, D. (2020). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 1–27.

Edgar, A. (2009). The challenge of transplants to an intersubjective established sense of personal identity. *Health Care Analysis, 17*, 123–133.

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, 6 August.

Farwell, J. P., & Rohonzinski, R. (2011). Stuxnet and the future of cyber war. *Survival, 53*(1), 23–40.

Figueira, A., & Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. *Procedia Computer Science, 121*, 817–825.

Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal, 70*, 455–471.

Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy & Technology, 31*, 317–321.

Franks, A., & Waldman, A. E. (2019). Sex, lies, and videotapes: Deep fakes and free speech illusions. *Maryland Law Review, 78*(4), 892–898.

Haemmerli, A. (1999). Whose who? The case for a Kantian right of publicity. *Duke Law Journal, 49*(2), 383–492.

Hao, K. (2019, October 10). The biggest threat of deepfakes isn't the deepfakes themselves. *MIT Technology Review*. Retrieved June 9, 2021, from https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/.

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review, 17*, 99–127.

Johnson, B. (2019, March 25). Deepfakes are solvable - but don't forget that "shallowfakes" are already pervasive. *MIT Technology Review*. Retrieved June 9, 2021, from https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/.

Kant, I. (1959). [1785]. *Foundations of the metaphysics of morals* (translated by Lewis Beck). Library of Liberal Arts.

Lee, D. (2019, May 10). Deepfake Salvador Dalí takes selfies with museum visitors. *The Verge*. Retrieved June 9, 2021, from https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.

Li, Y., Chang, M.-C. & Lyu, S. (2018). In ictu oculi: Exposing AI created fake videos by detecting eye blinking. *IEEE workshop on information forensics and security*.

MacKenzie, A., & Bhatt, I. (2020). Lies, bullshit and fake news: Some epistemological concerns. *Postdigital Science and Education, 2*, 9–13.

Mahon, J. E. (2016). The definition of lying and deception. *Stanford encyclopedia of philosophy*. Retrieved June 9, 2021, from https://plato.stanford.edu/archives/win2016/entries/lying-definition/.

Malaria Must Die. (2019, April 9). *David Beckham speaks nine languages to launch Malaria Must Die Voice Petition*. You Tube. Retrieved June 9, 2021, from https://www.youtube.com/watch?v=QiiSAvKJIHo.

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof, 23*(3), 255–262.

McKenna, M. P. (2005). The right of publicity and autonomous self-definition. *University of Pittsburgh Law Review, 67*, 225–294.

Merleau-Ponty, M. (1962). *Phenomenology of perception* (translated by C. Smith). Routledge.

Merriam-Webster. (2021). Deception. In *Merriam-Webster.com dictionary*. Retrieved June 9, 2021, from https://www.merriam-webster.com/dictionary/deception.

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2020). Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice, 15*(1), 24–31.

Mohebbi, S. (2016). The right to one's self-image. In E. Balsom & H. Peleg (Eds.), *Documentary across disciplines* (pp. 280–293). Haus der Kulturen der Welt and MIT Press.

Nye, J. S., Jr. (2017). Deterrence and dissuasion in cyberspace. *International Security, 41*(3), 44–71.

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of deepfake pornography. *Ethics and Information Technology, 22*, 133–140.

Panyatham, P. (2020, March 10). Deepfake technology in the entertainment industry: Potential limitations and protections. *Arts, Management & Technology Laboratory*. Retrieved June 9, 2021, from https://amt-lab.org/blog/2020/3/deepfake-technology-in-the-entertainment-industry-potential-limitations-and-protections.

Paterson, T., & Hanley, L. (2020). Political warfare in the digital age: Cyber subversion, information operations, and "deep fakes." *Australian Journal of International Affairs, 74*(4), 439–454.

Paris, B. & Donovan, J. (2019, September 18). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *Data & Society*. Retrieved June 6, 2021, from https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf.

Petty, R. D., & D'Rozario, D. (2009). The use of dead celebrities in advertising and marketing: Balancing interests in the right of publicity. *Journal of Advertising, 38*(4), 37–49.

Phillips, W., & Wilner, R. M. (2017). *The ambivalent internet: Mischief, oddity, and antagonism online*. Polity.

Polyakova, A. & Boyer, S. (2018). The future of political warfare: Russia, the West, and the coming age of global digital competition. *Brookings Institute*, March, 1–18.

Post, R. C. & Rothman, J. E. (2020). The first amendment and the right(s) of publicity. *Yale Law Review*, *130*(1).

RD. (2020). 'Welcome to Chechnya' uses deepfake technology to protect its subjects. *The Economist*, July 9.

Rini, R. (2019). Deepfakes and the epistemic backstop. Working paper, available at PhilArchive.

Robertson, J. (2004). Face transplants: Enriching the debate. *American Journal of Bioethics, 4*(3), 32–33.

Royal College of Surgeons of England. (2003). *Facial transplantation working party report*. Royal College of Surgeons of England.

Ryle, G. (1949). *The concept of mind*. Hutchinson.

Silbey, J.M., & Hartzog, W. (2019). The upside of deep fakes. *Maryland Law Review, 78*(4), 960–966.

Spivak, R. (2019). 'Deepfakes': The newest way to commit one of the oldest crimes. *Georgetown Law Technology Review, 3*(2), 339–400.

Stokes, P. (2012). Ghosts in the machine: Do the dead live on in Facebook? *Philosophy & Technology, 25*, 363–379.

Swindell, J. S. (2007). Facial allograft transplantation, personal identity and subjectivity. *Journal of Medical Ethics, 33*(8), 449–453.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C. & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RBG videos. *Proceeding of 2016 IEEE conference on computer vision and pattern recognition (CVPR)*.

Vaccari, C., & Chadwick, A. (2020). *Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news* (pp. 1–13). Social Media + Society.

Valeriano, B., & Maness, R. C. (2015). *Cyber war versus cyber realities. Cyber conflict in the international system*. Oxford University Press.

Weimann, G. (2015). *Terrorism in cyberspace: The next generation*. Columbia University Press.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52.

Wilson, D. G. (2017). The ethics of automated behavioral microtargeting. *AI Matters, 3*(3), 56–64.

Yang, X., Li, Y. & Liu, S. (2019). Exposing deepfakes using inconsistent head poses. *IEEE international conference on acoustics, speech, and signal processing*.

Zannettou, S., Sirivianos, M., Blackburn, J. & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, *1*(3).

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B. & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, *14*(1), 82–96.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.