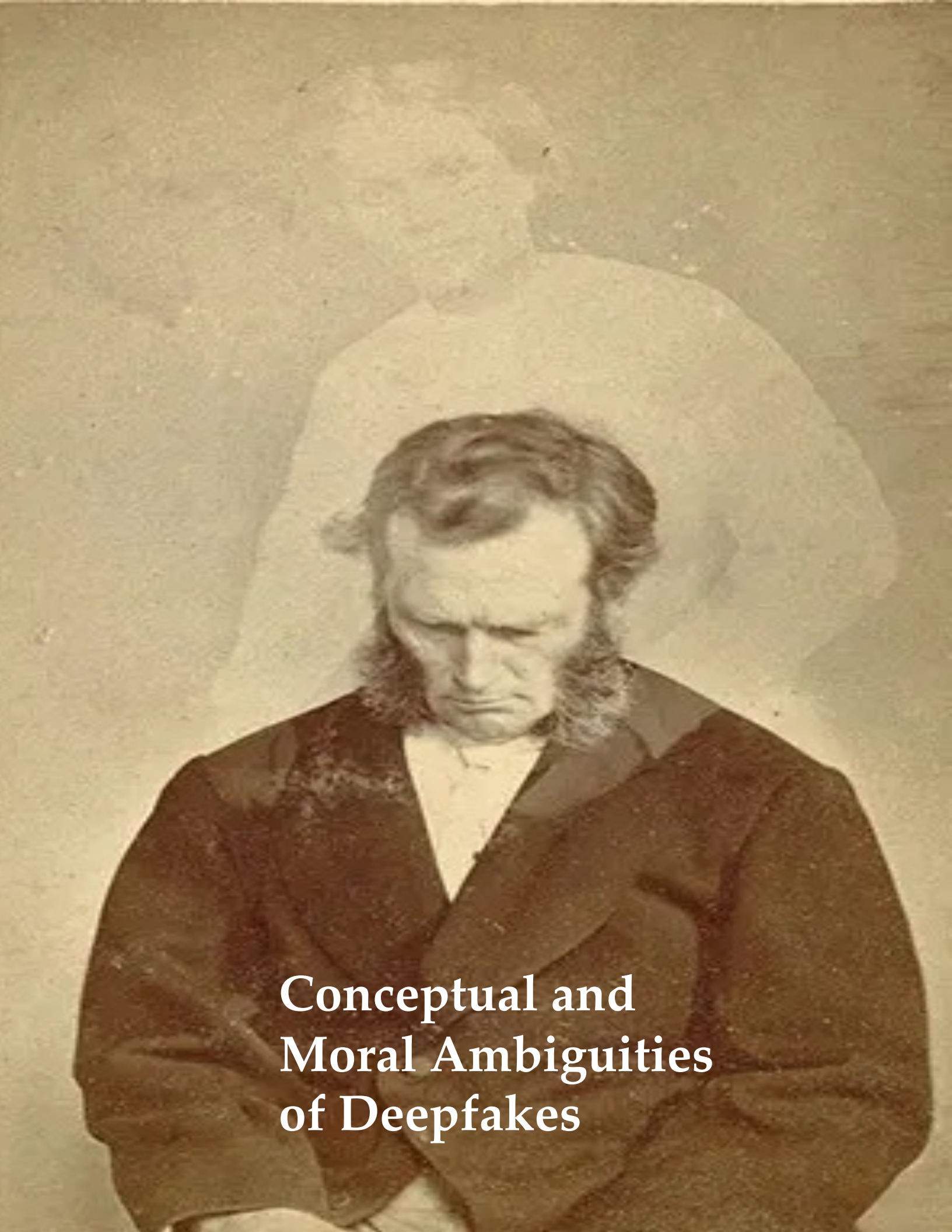

Conceptual and Moral Ambiguities of Deepfakes

Conceptual and Moral Ambiguities of Deepfakes

Abstract

Everyday (mis)uses of deepfakes define prevailing conceptualizations of what they are and the moral stakes in their deployment. But one complication in understanding deepfakes is that they are not photographic yet nonetheless manipulate lens-based recordings with the intent of mimicking photographs. The harmfulness of deepfakes, moreover, significantly depends on their potential to be mistaken for photographs *and* on the belief that photographs capture actual events, a tenet known as the transparency thesis, which scholars have somewhat ironically attacked by citing digital imaging techniques as counterexamples. Combining these positions, this paper sets out two core points: that conceptions about the nature of photography introduce imperatives about its uses; and that (2) popular cultural understandings of photography imply normative ideas that infuse our encounters with deepfakes. Within this, I further raise the question of what moral ground deepfakes occupy that allows them to have such a potentially devastating effect. I show that answering this question involves reinstating the notion that photographs are popularly conceived of as transparent. The rejoinder to this argument, however, is that to take the sting out of deepfakes we must, once again, become skeptical of the veracity of all photoreal images.

Keywords: deepfakes, ethics of technology, philosophy of technology, morality and culture, photography and video, transparency

Introduction

This article explores deepfakes—video manipulations that appear authentic—which have become ubiquitous in recent years. My point of departure considers how everyday uses and understandings of technology suggest certain moral imperatives in this domain. However, there are complicating elements, for the fact that deepfakes make use of lens-based recordings with the intention of mimicking them renders their status ambiguous. Exploring this quandary, my position will be that: (1) conceptions about the nature of photographs—about what kinds of media they are—introduce imperatives about their uses; and (2) popular cultural understandings of photographs imply normative ideas that impact how we encounter deepfakes and other forms of digital manipulation. This is further supplemented with a corollary: as deepfakes increasingly become experienced as decisively other than photographic, *some* of their potential for harm should dissipate, altering moral dimensions of their employment.

Accordingly, my discussion pushes off from a classic notion of photographs, which sees them as supplying a window onto past realities, or documenting things that exist or once existed (Santayana, c. 1907; Bazin, 1951; Sontag 1973; Cavell 1979; Walton, 1984). Known as the

"transparency thesis," the idea originates as much in popular culture as it does in academic circles, despite ample pushback from scholars (e.g., Jarvie, 1987; Mitchell, 1992; Carroll, 1996; Ritchin, 2008; Gaut 2010). My position grants that trickery has long been used to create the photographic appearance of things that did not happen. But it also shows that people have often questioned the legitimacy of calling such material "photographic" to the point of raising moral objections. Even the increasing prevalence of computer-generated images (CGI) in the 1980s and 1990s did not markedly change the scene because it was costly, heavily reliant on expertise and usually intended as explicit artifice, as in science fiction movies. In this regard, deepfakes—as well as other techniques such as photoshopping—instigate a tectonic shift because accessible platforms now allow amateurs to cheaply generate convincing manipulations.[1] Today, as Chesney and Citron (2019, p. 1753) state, it is easy "to create audio and video of real people saying and doing things they never said or did." Although digital technologies have often been used to attack the transparency thesis,[2] what is described here, perhaps rather strangely, vindicates it. After all, one reason informing the problematic nature of deepfakes is this: they use photographic samples to mimic photographically captured realities; and because people understand photographs to be transparent, the impression that the digital counterfeits portray things that actually happened is the consequence.

A broader lesson is that media do not abandon historical legacies when leaving old technologies behind. Again, deepfakes mold lens-based information into photoreal appearances to image events that never occurred. To increase the impact, some digital manipulations go so far as to exhibit facsimiles of lens flare, blurring and overexposure (Crippen, 2016; Strutt, 2019). With occasional stylistic exceptions, these effects were once regarded as undesirable byproducts of the

[1] See Vaccari and Chadwick (2020).
[2] E.g., see Mitchell (1992) and Ritchin (2008).

lens, and the purpose of imitating them is to enhance spectators' sense of witnessing events that really occurred before the camera.[3] Historical legacies therefore affect how artifacts are used, what they mean to us and indeed what they are for us. From this it follows that comprehending photography and deepfakes entails more than examining their physical nature. It necessitates looking at their role, that is, their meaning within human cultures.

A key consideration, then, is that we gain a sense of what human-made artifacts are by looking at their common uses, whereas everyday culture is less helpful in ascertaining the nature of phenomena like quarks and gluons, which were initially defined by specialists. Yet, everyday meanings do not entirely settle things, for although cultural understanding has arguably clarified the meaning of photography, recent technological developments outstrip established connotations. Until meaning catches up, and owing to the presupposition of transparency inflecting encounters with them, manipulated images have an unusual power to do harm. But if in the future—and things appear to be moving in this direction—people do not customarily assume the veracity of what photograph-like images show, then there is little reason to assume the post-truth nightmare many predict (e.g., Chesney and Citron 2019; Fallis, 2020; Rini 2020; Schick 2020; Kerner and Risse, 2021). Instead, circumstances should revert to what they were before the advent of photography, although with some key differences. This means, for reasons to be discussed, that we will likely retain license to regard some images as truthfully documenting the world, while simultaneously recognizing others as ethically problematic infringements on privacy, even if they may become less harmful over time.

[3] See Mullarkey (2009).

Photographic Arts and Meaning

Though not always explicit, Santayana's (c. 1907) and Cavell's (1979) writings suggest that understanding what the photographic arts are involves asking what "photography" means. In this section, I take a cue from these philosophers to introduce an investigative method that explicates why—historically speaking—the transparency thesis has been accepted in popular culture. This is done to set the stage for exploring how widely held ideas about photographic transparency lead to conceptual and moral perplexities in the digital context of deepfakes.

A first step involves articulating the transparency thesis that photographs show us things that exist or once existed. The position captures a popular perspective to the extent, for instance, that most would prefer a selfie over a painting with a celebrity, believing that the former, unlike the latter, allows one to *really* see the event after the fact. Classic scholars (Bazin 1951; Kracauer, 1960; Sontag 1973; Walton 1984) offer interknitted reasons for grasping photographs this way. One is the camera's mirroring ability. This capacity—though insufficient for transparency, as we shall see—is rooted in the causal relationship of photographs to their content, which is somewhat comparable the shadow cast by an object, the fossilized imprint of an animal, or the chiral in a reflective surface. Also central to the transparency thesis is the premise that photographs are viewpoints produced by automated mechanical processes, whereas paintings are interpretations filtered through an artists' mind and body. But this overview falls short precisely when it focuses too exclusively on the physical nature of photography, foregoing the question of what "photography" means as a culturally significant practice.

Grasping the cultural conceptual territory of photography involves recognizing it as an artifact, which signals further considerations. On the one hand, the physical mechanisms by which artifacts are produced connect to the *means* they have been put to, hence their meanings. An

example is that photochemical emulsion (film) was historically used to make films, giving the artform much of its character, so that the means inherently connect to the ends produced and to what we understand film to be (or mean). On the other hand, the material nature of an artifact rarely specifies one kind of use: a polished wooden pole may be deployed as a cane or a weapon, just as things other than movies use light-sensitive emulsions (and films can now be shot without them). This helps unpick an argument offered by Gaut (2010). He discusses using a plaster cast to duplicate Trajan's Column. As in photography, the replication occurs through an automatic mechanical process, yet few would conclude they are seeing the original object when perceiving the cast. Although Gaut thinks this repudiates the transparency thesis (which presupposes that photographs are automatic traces of the world), he misses a key point: not all mechanical traces hold the same meaning within human cultures. Indeed, when not philosophizing, Gaut surely talks as if he sees loved ones in photographs and uses them to show others his family members.

The discussion reveals that capturing the meaning of the photographic arts (understood to include digital, photochemical, still, and moving instantiations) entails an examination of what they have historically meant within specific cultures and languages. Cavell conducts this task in earnest, although some of his harshest critics (e.g., Jarvie, 1987; Carroll, 1996) appear oblivious to this. To begin with, Cavell often follows Wittgenstein (1953) who was also investigating the framework of meaning within human communities. Along these lines, Cavell's (1979) *The World Viewed* begins with a claim inspired by Tolstoy that ontological queries about art are questions on its importance, significance, value, or meaning within human exchanges. And, yet, some challengers of the transparency thesis (e.g., Mitchell, 1992; Alcaraz, 2015) attack it on cultural-historical grounds, claiming that photography was invented at a time when many were questioning or seeking to restore our access to objective reality. Consequently, these challengers assert that

many jumped to the uncritical conclusion that photography offers a mind-independent perspective, hence an ingress to objective reality.

Now, scholars are right to assert that culture, as well as what might be termed "folk beliefs," can generate erroneous views about how things operate. However, it is not hard to grasp the basic mechanisms of photography, whether photochemical or digital, so any ubiquitous confusion about it is not from misunderstanding its physical standing. An added point is that investigations of human artifacts differ from examinations of physical nature *per se*. A fuller comprehension of the latter emerges when the inquiry focuses on physical aspects, with the account aided little by popular culture. By contrast, a physical analysis of artifacts devoid of a cultural-historical framework will explicate little about what they are (i.e., their meanings).

To summarize, knowing what an artifact is simultaneously implies having a sense of its role and significance to/in the culture that produced it. This holds for photographs. So those who defend or attack the transparency thesis premised upon the physicality of photography—thereby neglecting its cultural dimension—adopt mistaken perspectives. The same goes for those challenging the transparency thesis on the grounds that it arises from erroneous beliefs propagated in culture. This is something like noting that a pointy tool that a past culture intended as a writing instrument would have been more effective as a weapon, and then concluding that it is therefore not in fact a writing instrument, but instead a weapon (citation suppressed).

Transparency and Some Ethical Implications

Because the transparency thesis has dominated popular understandings of photography, it inflects our encounters with deepfakes, which is why I have been exploring the parameters of the thesis. A more specific claim relating to this is that many find deepfakes morally and epistemically threatening as a consequence of transferring intuitions about transparent aspects of photography

to digital counterfeits of it. In this section, my aim is to elaborate the transparency thesis and its moral implications in connection with how digital incursions change how we think about images.

I begin by dismissing a standard repudiation to the transparency thesis: that photographers engage in distorting "interpretations" by selecting different film stocks, focal lengths, and lighting, or that there is no retinal disparity or motion parallax in still photography (e.g., Mitchell, 1992; Carroll, 1996). Though true, these objections neglect obvious rebuttals. First, transparency proponents hold that photographers have historically (i.e., mostly) taken imprints of things in front of the camera, avoiding *post hoc* manipulation (interpretation). The above objections do not violate this. Second, there are ways to achieve the aforesaid "interpretations" while nonetheless seeing a given scene: peering through darkened pitted glass may be analogous to grainy monochrome film stock; a telescope has the same effect as a long focal length; going from an artificially to a naturally illuminated setting can indicate a shift in lighting; staring at a static scene through one eye while not moving knocks out retinal disparity and motion parallax. Yet, almost nobody—including critics of transparency—concludes that because these shifts occur, we are not actually observing things in our field of view.

Now, refuting these objections does not directly qualify the legitimacy of the transparency thesis, but a simple empirical experiment—which also works as a thought-experiment—reinforces why many have found the position compelling. The experiment consisted of showing participants two different paintings of Jesus, followed by two cinematic stills of different actors portraying Jesus.[4] When asked who was depicted in the paintings, despite the two paintings portraying men who look different, respondents unhesitatingly identified the individuals as Jesus without noting that the artists may of render the images using human models. For the two cinematic stills, by

[4] See Crippen (2016).

contrast, the participants primarily tried to identify the two actors—the models—while also recognizing that they were portraying Jesus.

An explanation for the disjuncture of the responses may lie in the way paintings and photographs have historically been put to different uses. Compared to painting, which often idealizes portraits and other content, Santayana (c. 1907) claims that photography aims at unadulterated repetition, overstating the case, of course. Cavell (1979) advances a similar argument, observing that it makes little sense to ask what lies beyond the frame of a painting, such as what exists behind or next to the representation of a building. Viewers take for granted that the building is a product of the artist's hand, only having knowledge of its factual existence through external information, as when recognizing it as one visited before. This same question of what lies beyond the frame, however, makes more sense when looking at a photograph of a building—or at least, it did in the era in which Cavell was writing. Thus, when viewing a photograph as opposed to a realist painting, people have historically taken for granted that the subject in the photo exists or at least once existed. This is why moral objections were raised when a photograph of the Giza pyramids underwent *post hoc* editing to better fit a 1982 *National Geographic* cover: people thought that the adjustments violated the nature of photography (citation suppressed; Goldberg, 2016). And today, if we see an image only to be told that a background has been digitally introduced, that the faces of the people have been swapped and certain objects removed, we may wonder whether it can even still be called a photograph.

Returning to the experiment described above, this helps explain why, in the case of viewing a painting, the participants automatically recognize Jesus rather than the model portraying him: they factor in the that the composition is an outcome of imagination. By contrast, in the cinematic stills, the actors playing Jesus are bound up with the very medium used to the extent that classifying

something as a "photograph" means asserting the existence of a subject it shows, which obviously could not actually be Jesus himself.

Although it might seem that the onset of digital photography and its new possibilities of manipulation profoundly changed things, the initial shift was not as radical as some might think because: (1) aggressive *post hoc* alteration is almost as old as photography itself, but has tended to disqualify images from being designated as "photographs"; and (2) digital photography is still understood to document events, selfies being an example. Related to the first point, Atencia-Linares (2012, p. 22) is not quite right to claim that passing a "light pencil" over film emulsions is "a *photo*graphic process." It is rather a photochemical process. Most viewers would likely not deem the resulting product a "photograph," and the same holds for other examples offered by Atencia-Linares, such as Wanda Wulz's "*Cat + I*," a synthesis of her face with a cat's.

All this serves to undermine another line of attack on the transparency thesis, which claims that "although … analog and digital images seem to be very similar or even the same, when perceiving a digital image we can never be sure that it is true" (Alcaraz 2015, n.p.). In fact, the Wulz illustration demonstrates how what is asserted of digital images also holds for photochemical variants almost as far back as the dawn of photography in the 1800s. The point, then, is not that photograph-like images cannot be counterfeited. It is instead that when an image has been judged to be a photograph, this determination has—historically speaking—implied that it has not undergone *post hoc* editing and that it shows what exists or once existed.

Digital Incursions and Changing Ethics

Despite the above discussion, however, there is no serious doubt that deepfakes, photoshopping, and other recent digital advances have changed the status of information gathered through the use

of cameras, such that the meaning of lens-based images take on new, unprecedented roles. My argument will be that if common conceptions about the nature of lens-based images have changed, then this would also bring about an altered normative landscape. This section expounds these issues to support the claim that deepfakes are somewhat like paintings. Only with photographic transparency still lurking in the background, deepfakes are "Janus-faced," which generates a number of moral complexities in need of exploration.

The last section made the case that the emergence of digital photography initially did not vanquish transparency, but technologies have continued to evolve and in recent times trickery has become commonplace. On the one hand, calling an image a "photograph" still insinuates the truth of what it shows such that an earnest mood accompanies it. For instance, after digitally faking an image of a successful ascent, an Indian couple were banned from Mount Everest and lost their jobs as police officers (Safi 2016; Agence France-Presse 2017). Another example is a Reuters editor who was fired for publishing digitally altered images submitted by freelance photographer Adnan Hajj (Cooper 2007). In yet another example, Fox News presented digitally amalgamated photographs, framing the 2020 Seattle protests as more ominous than they were, with negative public sentiment compelling an apology (Brunner 2020). Part of the reason these occurrences were taken so seriously was that they violated the ethical expectation that what we call "photographs" should be used to show things that actually happened.

Moreover, and conforming with the view that photography is still largely grasped as transparent, most people continue to accept digital recordings of a theft as evidence, while a painting would not suffice. And, yet, if used in a court of law—and especially now with the advent of deepfakes—videos would need to be verified. But, if there were also doubts about the authenticity of photochemical images, they would have been subject to similar scrutiny. Even if

deepfakes progress to such a degree that they become undetectable and so inadmissible as evidence,[5] security footage will remain as a mechanism for documenting crimes and identifying the individuals involved. This would provide the police with the same transparent photographic leads that they have had since the invention of camera surveillance, but it would force them to seek additional corroborating evidence.

On the other hand, digitally altered images are now widespread in other areas. While digitally enhanced images are often regarded as photographs, as in the case of profile pictures on job websites, they have received a plethora of (often morally motivated) criticism for perpetuating false depictions of the human body or promoting unrealistic body ideals. But even with this in mind, some extreme photomanipulations remain uncontroversial, such as the middle-aged Japanese man using FaceApp—a deepfake platform—to turn photographs of himself into images of a young female motorcycle enthusiast. Although most would no longer call the resultant images "photographs," the deception did not generate much ethical backlash from the internet community, and instead was met with good humor, to the extent that the man—Yasuo Nakajima—gained many followers after the deception was revealed (Harwell and Okazaki, 2021).

These examples, combined with earlier ones, suggest an amended understanding of what photographs are, along with images derived from manipulating them. The previously mentioned doctored Mount Everest photo and the ones from the reporting outlets were all distributed by news media as documentary evidence of events that purportedly occurred. So these instances—along with the 1982 *National Geographic* cover[6]—share a common trait insofar as counterfeit images were not just made to look photographic, but presented in contexts in which it is reasonable to assume the veracity of what is displayed. The underlying feature is that they all involve the explicit

[5] See Maras and Alexandrous (2019).
[6] Even though the manipulation was carried out through physical cutting and pasting. See DeVoss (2011).

or tacit violation of social contracts. And, yet, the fact that these fabrications looked photographic was still at issue, for hardly anyone would object if an artist painted an idealized or fictionalized version of a Mount Everest ascent or if a technician rendered the Pyramids using stylized (unrealistic) computer graphics. The example of depictions of Jesus likewise suggests that expectations of transparency remain in photography but not in painting. The discussion, however, also demonstrates that moral responses vary according to context and use. Thus, there are higher expectations attached to media agencies than to private citizens posting doctored images or photos to job websites. In this connection, internet influencers occupy an interesting middle-ground, lodged somewhere between these poles in so far as they regularly use digital manipulation as well as receive ample criticism for disseminating false information and promoting unrealistic ideals, although this censure is not sufficiently serious to thwart the practice.

Where deepfakes are concerned, the moral expectations vary in comparable ways to the examples explored above. Combining Nicolas Cage's face with Julie Andrews' is unobjectionable from a utilitarian perspective insofar as little harm ensues, even if complaints might be raised on rights-based grounds inasmuch as the photographic information is manipulated without permission.[7] When deepfake platforms are used without consent to make pornography, however, the possibility of harm dramatically increases. While perhaps less harmful where celebrities are concerned because there may be an already strong leaning toward viewing the video as fake, non-famous people have suffered harassment and some have committed suicide owing to digital fakery (Young 2021; BBC 2022). Moreover, because secretly filmed sex tapes are associated with at least one celebrity suicide (McCurry 2019), deepfakes can cause irreparable harm to even the famous. These ideas will be examined in what follows, partly to justify the view I stated at the beginning

[7] See Kerner and Risse (2021).

that a primary reason why deepfakes are morally and ethically problematic is because of their technological, historical, and conceptual relationship with photography and, hence, with transparency.

Deepfakes and Intermediary Meanings

A prominent challenge facing any analysis of deepfakes is that they occupy a sort of in-between, or, better, intermediary position. This is in the sense that they use photographically gathered information to create a sheen of transparency, while not being so. Moreover, and despite the known overabundance of digital manipulations in the public sphere, this is exacerbated by the current historical situation in which images that appear photographic are more often than not judged to be transparent; the harm of deepfakes largely comes from this factor. The question of rights even emerges with respect to the most innocuous deepfakes, such as those playfully synthesizing different celebrities faces. I shall explore some of the resonances of this situation, examining the conceptual ambiguity of deepfakes, the shifting role of transparency, and the moral consequences stemming from all of this.

To review, the term "deepfakes" typically refers to realistic looking images of events that did not occur, which are usually generated by web-based counterfeiting platforms. This technology has an historically unprecedented power. Just by using a limited number of still images or videos as source material, it can swap facial or bodily physiognomies, predictively generating outputs that realistically portray what someone would look like from other, unrecorded angles or what they would look like performing movements and making expressions not captured on camera. Deepfake applications can also produce similar effects with audio samples, which can be combined with manipulated video to constitute impressively realistic "footage" of people uttering words they

never said, using mannerisms never captured on camera. This capacity to predict and generate audiovisual presentations beyond a simple aggregate of inputs situates deepfake platforms in the realm of artificial intelligence (AI). Because the technology is open to non-technical dilettantes, who can now produce convincing deepfakes, it may soon be the case that predicating something "photographic" does not carry the same air of truth that it once did. When viewing images that look photographic (rendered broadly to include both still and moving images captured by cameras) we may stop assuming that we are encountering windows onto things that exist or once existed. However, to the extent that most classify deepfakes as photographic forgeries (i.e., a fake of something genuine), the situation remains complicated by the fact that deepfakes are conceptually premised on the presupposition of transparency. The complexities are amplified by the fact that deepfakes are most damaging precisely when they are mistaken as transparent.

This state of affairs differentiates deepfakes from arts such as painting and sculpture. To see, consider the sculpture of President Donald Trump by Ginger and the paintings of Prime Minister Stephen Harper by Margaret Sutherland and President Barack Obama by Alexandra Rubinstein. Trump and Harper are nude in the Ginger and Sutherland pieces, and Obama is about to perform cunnilingus in the Rubinstein depiction. Interestingly, aside from the foreseeable conservative and liberal complaints about impropriety, body shaming, and ageism, the general consensus was that the pieces were simply innocuous political commentary. But now let us picture a different situation: imagine that instead of painting or sculpting those people, the artists had used deepfake platforms to construct photoreal videos or still images of them in the same poses, or engaging in the same activities. It appears likely that protestation would be more aggressive. Why would this be the case?

One potential answer is that the deepfakes would use photographic material to generate new images without the targets' consent. However, the painters and sculptor likewise were only able to execute their works by consulting photographs (in the absence of which they would not have been able to render their subjects). Many photographs of public figures, moreover, are taken without expressed consent, and irrespective of that, were surely used by the aforementioned painters and sculptor without permission. And in fact, had a digital artist used deepfake technology to transform publicly available photographic information into something that looks like a painting or like 1990s computer graphics, one suspects the criticism would not be any more than it was for the nudes that were actually produced. So consent—at least taken in isolation—does not appear to be the main moral stake. But nor either does photorealism alone. After all, while explicit photoreal paintings (e.g., in Chuck Close's style) of world leaders would probably garner fairly vocal complaints from the public, this would likely not match the backlash generated by similarly graphic deepfakes.

If the objection to sexualized deepfakes is not *per se* attributable to creators using photographic information without permission, nor simply a response to the creation of photoreal images, nor solely a reaction to depicting non-consenting parties in tawdry ways, then what is its source? One possibility that encompasses the three just mentioned is that the objection emerges from the fact that many would encounter the images as partially transparent, that is, as occupying an intermediary space where they are neither wholly transparent nor non-transparent. Assuming for a moment that deepfakes of non-consenting parties in sexual acts are more problematic than cartoon animations of the same thing, a case can be made that this is because the deepfake contains a higher degree of photorealism, for it was, after all, constructed from digitally manipulated and amalgamated recordings of events that actually took place. As one commentator

puts it, "deepfakes do not show a celebrity […] actually engaging in sexual activity. Instead, they give the appearance of showing the celebrity engaging in what is, in fact, sexual activity. In other words, they depict someone who was not part of the original sex act taking part in it" (Young, 2021, p. 178). By virtue of creating a transparent-appearing depiction in this vein, such a deepfake involves the non-consensual appropriation of photographic information to portray people in the midst of sexual activity, again without their permission.

That deepfakes have an intermediary status explains the view that a voyeur secretly recording an intimate moment—for instance, someone sitting on the toilet—would cause more distress and be more ethically reprehensible than the construction of a deepfake depicting the same. While this is used as an example, there are real-world occurrences of this, such as the photograph of Alex Rodriguez on the toilet, which was taken without his consent in 2019.[8] Many pundits raised ethical objections to this. By contrast, there was no significant moral censure when Bill Maher's show, *Real Time*, faked an image of Donald Trump sitting on the toilet in 2015. The reason why most find faked images less offensive than actual photographs is likely attributable to the role of transparency: the public takes genuine photographs to be transparent views of an actual situation, whereas they recognize the fictional side of faked images, which was obvious in the Maher episode. This makes the former, transparent images, more invasive than the latter, although it must be said that both sets of images raise significant ethical concerns.

The upshot is that common conceptions of deepfakes and other forms of digital manipulation are mixed: they are partially transparent and this situates them in ambiguous moral territory. For instance, individuals would likely find a deepfake of themselves masturbating more threatening than an animation of the same, although both would probably be less detrimental than

[8] See Helmore (2019).

a real video of the same event. That is, the moral weighting of the violation correlates with how much the material traces to camera recordings of things that actually happened. Furthermore, as a stylistic choice in fine art, photorealism plays an important role when we consider that the non-consensual dissemination of a Chuck Close painting of oneself masturbating might be more problematic than, for example, a Picasso painting of the same, even if the images were based on photographs. The irony is that scholars have often used the possibility of manipulating digital images as a rebuttal to the transparency thesis, but the ethical objections set out above are evidently premised upon popular conceptualizations of photography, which involves transparency, as well as the expectation for it to be transparent, both of which inform the peculiar territory of deepfakes.

Deepfakes and the Destruction of Truth?

In the previous section, I discussed the conceptual, historical, technological, and experiential stakes of deepfakes in their relation to the transparency of photography, and the ethical-moral implications arising from this. In this regard, in the following, while I acknowledge that deepfakes introduce numerous possibilities of harm, I also explore reasons why deepfakes may not lead to the post-truth apocalypse that scholars often predict.[9]

We have already comparatively analyzed videos of sex acts and deepfakes of the same. Granting that most would find it upsetting to be a non-consensual target of such a deepfake, it would be difficult to deny that it would be even more distressing to be a victim of a voyeur's hidden camera, though this may vary between people. It has been stated that counterfeit images did not first arise with the advent of deepfakes, for example, with it long being possible to manufacture damaging videos using body doubles, editing and creative use of angles. But it would

[9] E.g., in Chesney and Citron (2019); Fallis (2020); Rini (2020); Schick (2020), and Kerner and Risse (2021).

have been infeasible for amateurs to convincingly produce such content in the past. Another problem raised by commentators[10] is the concern that deepfakes open wider possibilities for refuting real events, such as a politician denying actual recordings of them on the basis that it is simply a deepfake. A commonly cited example of this is Donald Trump's comment that his enemies faked the Access Hollywood recording of him claiming that he gropes women without consent.[11] Similarly, the tactic was used by a Cameroonian authority to decry a video of soldiers from that country executing civilians.[12] But this situation again existed before the emergence of deepfakes, as is evident from the debates about the authenticity of footage showing the US official Kyle Hatcher with a prostitute.[13] Indeed, this was almost certainly staged as part of a Russian smear campaign. Yet interpretations significantly vary depending on whether the perspective is from the Western or Russian side, just as one's political sympathies may inflect whether or not one accepts the Access Hollywood recording as real or fake.

At the same time, the contemporary age remains one in which materials appearing to be photographic are usually treated as capturing actual occurrences. This is not to deny that the conceptual landscape is shifting. Indeed, these days the term "photograph" is often used even when images are "transparent enough," as when bodies and faces have been filtered to meet conventional standards, or the background is digitally altered to make a vacation scene look more ideal. That said, the continued presumption tied up with the age—namely, that things appearing photographic show events that actually happened—is *itself* exactly the leading edge to what makes deepfakes so threatening. This is largely the reason why sexually explicit deepfakes or even

[10] See Rini (2020) and Schick (2020).
[11] See Rini (2020), Schick (2020), and Kerner and Risse (2021).
[12] See Kerner and Risse (2021).
[13] See Cole and Ross (2009), Dougherty (2009), and Martinez (2009).

photoshopped images have had such harmful, life-altering effects on non-celebrities (Young, 2021; BBC, 2022).

In spite of all the appearance of novelty in these problems, however, it is erroneous to view deepfakes as introducing a radically *new* turn in the capacity to deceive the masses, for deepfakes imply a decidedly *old* turn. That is, the outcome of the proliferation of deepfakes is perhaps more corrective of our presuppositional judgments: the increasing abundance of deepfakes will most likely lead to a situation in which we do not immediately presume that photographs show events that actually occurred or things that actually exist. Airport security video, antiquated photographs of historic events, or images distributed by highly trusted news sources may perhaps act as limited exceptions and would therefore continue to be judged as transparent, however. Moreover, security footage might be rendered inadmissible in court but still used in certain circumstances, for example, aiding police officers in identifying someone so they can take a DNA sample.

Whatever the specifics of these future outcomes may be, however, the general proposition is that as the level of presuppositional trust in audio and visual recordings decreases the presupposition of them as fake increases. Thus, we would be compelled to weigh images with the same critical skepticism at work in testimonial domains before the advent of photography. Instances of claims warranting increased skeptical analysis before deepfakes are obviously numerous, but it would be nonetheless helpful to lay out a few. For example, US leaders lied about an attack in the Gulf of Tonkin to justify escalating the war effort in Vietnam (Hanyok 2000–2001). Another case comes from George H. Bush's time in office, whose administration, alongside the Western media, spread false atrocity propaganda to increase public support for the first Gulf War (Regan 2002). Yet another comes from a subsequent presidential administration, this time of George W. Bush, which misled the public about the presence of

weapons of mass destruction and support for Al-Qaeda in Iraq (Draper 2020). Then, more recently, Russia spread disinformation to influence the 2016 US elections, in a move that follows the older Soviet trope of planting false news stories around the world (Schick 2020). In fact, we can draw from a rich stock of examples throughout human history, wherein rulers and colonialists have repeatedly imprinted messages onto those they control premised on questionable or false information.

The internet is the main vehicle through which deepfakes and the platforms used to make them are disseminated. Indeed, the internet has served to magnify the reach of deepfakes to an astonishing extent compared to what would have been practical via older media avenues. But while the internet is itself often charged with amplifying the spread of disinformation,[14] things are actually more complicated. One view that plays into how the internet does this is that it decentralizes information; and, yet, decentralized networks—for example, the power to start and spread rumors—predate the internet. On the assumption that (dis)information is decentralized to a greater extent in the contemporary internet age, the result follows that what would have otherwise been a fringe actor, such as the online QAnon community, is given a greater voice. This is to the extent that the unwarranted and dangerous conspiracy theories of QAnon are now mainstream enough to have some lawmakers championing them. More hopefully, however, the situation may also work in the other direction, for instance, by thwarting the capacity of leaders to engage in mass manipulation to support unpopular wars.

The broader point to take from this is that the increasing presence of deepfakes compels the more critical audience to approach visual and audio recordings with a comparable level of care as when approaching written or spoken accounts. Before the advent of deepfakes, powerbrokers

[14] E.g., see Schick (2020).

have framed recordings of actual events to mislead the public, which led to the increased scrutiny of images. During the Gulf War, for instance, US leaders and news pundits presented genuine footage of Patriot missiles launching followed by explosions, which they falsely stated to be videos of the destruction of Soviet-made Scud missiles. Another example is that of the statue of Saddam Hussein in Iraq. Its toppling was a genuine, recorded event.[15] But an array of framing techniques were employed, which were not made clear to the viewers of this footage. This included US military personnel assisting in knocking the statue down, also supplying certain props to the people. The camera operators additionally used tight-focused shots to make the scene look more densely packed with celebrating Iraqis than it actually was. So, again, whether in a deepfake era or not, circumspection about still and moving images is warranted.

All that said, these examples fail to address the *personal* threat of deepfakes, such as fabricated sex videos of non-celebrities. Öhman (2019) provides an array of hypothetical examples that can help clarify this point. He starts by considering a sexual fantasy and then a lucid dream with the same content, both kept to oneself. He next ponders a realistic and explicit painting that is used privately and disposed of after one's sexual fantasy comes to an end. He lastly describes a deepfake that is again not shared and destroyed immediately after gratifying one's sexual fantasy. Although there are probably numerous ways in which an ethicist may object to every scenario, for the sake of the following I shall presume that most will agree that the first and second are the least objectionable of the four and the last is the worst.

An interesting consideration to take into account is that it would be creepier to seek consent before engaging in a masturbatory fantasy or lucid dream about somebody than not to. Why? The reason for this is clear: the first two scenarios are necessarily *private*. This also goes some way in

[15] See Fisher (2011) and Lewis (2018).

explaining why one's intuition may veer toward viewing the painting as ethically worse than the fantasy or dream: even if we intend to keep it private, the very fact that it exists as a painting makes it at least possible for it to enter the public domain. The reason why many would probably agree that the deepfake is the most problematic is twofold: first, it is constructed from photographically gathered information, and as such is partially transparent; second, even if there is no intention to share it, should others see it, they may mistake it for a video of actual events.

It is difficult to imagine a future in which a deepfake of the sort just described would cease being problematic, that is, unless attitudes about sex and privacy radically and cross-culturally change. And, yet, it is not too much of a stretch to assert that deepfakes of this sort may cause far less damage *if* our presupposition is that audio and visual footage is not automatically transparent, such that it is always a case of first questioning whether or not the footage shows things that actually happened. An example from my own experience in middle school speaks to this view, in which my classmates put together an amateur newspaper suggesting my involvement with prostitution. Although, as expected, I was a little hurt by this, the harm was minimal precisely because nobody actually believed the story; it was obviously untrue. Deepfakes, however, are different kinds of media from written communication, such that the way we "read" them must be tailored to a more image-based landscape. This is not only insofar as they use lens-gathered information (though CGI could develop in ways as to make this part of deepfaking unnecessary). It is also because pictures hit people in ways that text usually cannot; they have an immediate and palpable impact.[16] Moreover, deepfakes also open a range of other, extremely troubling possibilities, such as portraying minors in sex acts or exaggerating people's appearance to racially mock them. As I have argued above, however, much of the

[16] See Fong et al. (2009).

potential harm caused by deepfakes arises from the fact that people take them to depict things that actually happened, and that as this assumption wanes, the potential damage that deepfakes can do likewise decreases.

Conclusion

My account began by exploring how popular culture has historically promulgated the concept of transparency, examining ethical imperatives implied in this way of thinking about photography. I showed that the advent of digital photography—and the ease of manipulation that went along with it—did not initially change things that much. This is so to the extent that even today, an expectation of transparency often still applies.

I next discussed how deepfakes, as widely encountered and understood in current contexts, occupy an intermediary position with respect to transparency. Accordingly, we have what might be called "in-between" conceptions of deepfakes, which speaks to why we often assess their use on an axis somewhere between paintings (including cartoon animations) and photographs, whether still or moving. I argued that deepfakes do not entail a shift into a post-truth nightmare because deception has long been a part of linguistic and photographic discourses. The call for circumspection about the uses and abuses of media is not new, and the current situation calls for us to assume the same perspective with respect to photoreal images.

Although the potential harm that deepfakes could inflict on individuals is grave, I asserted that the future could see a return to that of the pre-photographic past, but with key differences. Some differences mentioned include new forms of bullying arising from the ease of creating meanspirited depictions, amplified because images can communicate on immediate emotional levels, especially when photoreal, thereby hurting more than words. One similarity with the past, however, may be that more critical viewers will weigh images without any presuppositions of

transparency, just as etchings in old newspapers were viewed: they may or may not have reflected actual events, but they were never confoundable with the actual event itself. In the event that everyone takes it for granted that images might be fake, the situation would be akin to Kant's practical examples where maxims premised on individual objects are incorrectly universalized such that they cancel out the original maxim.[17] This informs perhaps one of Kant's most famous examples in which the maxim to be universalized is lying, which results in a mass falsification that ultimately undermines the heart of the original maxim, which was the intention to deceive. But where Kant used this as an argument *against* lying, we are called upon to adopt a slightly different perspective. For as much as I may desire a future without deepfakes, we occupy an era in which they exist, but the upshot is that widespread belief in deception can serve to take the bite out of malicious deepfakes because the starting presupposition would be skepticism as regards the veracity of all images.

There are also some meta-lessons to be drawn from this article. A corollary to the position that we ought to look at popular meanings when examining artifacts is that asking what things are need not imply essentialism, as some insinuate.[18] After all, popular meanings evolve, which excludes a criterion that seeks for unchanging essential identifying marks. Additionally, popular meanings—if Wittgenstein (1953) is right—do not class things according to essential features in the first place, but instead according to family resemblance. Within a literal family, this can be such that some members share similar noses. A subset of these may share eyes with other members who do not, however, have the same nose. The idea is that we can recognize all as belonging in one family even though no single feature pervades the group. Most definitions of deepfakes involve the manipulation of photographic or audio recordings, but I have suggested a future where

[17] See the *Critique of Practical Reason* 5:27–28 on the example of deposits.
[18] E.g., Jarvie 1987

they might be pure CGI creations. Another author defines deepfakes as malicious disinformation (Schick 2020). However, deepfakes need not be deceptive, as in those involving Nicolas Cage or a future with filmmakers deploying such platforms to make a factually-based movie about Leon Trotsky. Another lesson hinted at but not explored is the idea of pursuing something analogous to a natural law framework in exploring deepfakes. However, rather than drawing ethical implications from a theory of human nature, the approach would extract moral imperatives from the everyday ontologies (natures) of technologies.

All of this reiterates a core point advanced throughout this article: that different procedures are involved in ascertaining the meanings of artifacts and things like elementary particles. Scientific specialists define particles. By contrast, adequate research of an artifact requires a grasp of its everyday use, so that it is, as stated earlier, nonsensical to conclude that an instrument used by a past civilization to write is in fact a weapon because its physical properties make it suitable for killing. In the domain of digital photography, scholars who argue against transparency have significantly neglected the fact that most people continue to treat (use) camera recordings as if they show things that actually happened, meaning that these scholars operate on the premise that popular conceptions are irrelevant to what the multiple instantiations of photography are. But, in so doing, they miss the unique orientation deepfakes call on us to take; they miss the basic axis required when analyzing why it is that deepfakes have the drastic impact they do. A basic irony is that many of the moral problems raised in connection with deepfakes decidedly rest on the common presumption that photographs are transparent: deepfakes are dangerous precisely because they are mistaken as having the same veridical status traditionally attributed to photographs.

References

Agence France-Presse (2017). Indian police sack couple for faking climb to Everest summit. *The Guardian*, 30 August.

Alcaraz, A. (2015). Epistemic function and ontology of analog and digital images. *Contemporary Aesthetics* 13: 1-14.

Atencia-Linares, P. (2012). Fiction, nonfiction, and deceptive photographic representation. *Journal of Aesthetics and Art Criticism* 70: 19-30.

Bazin, A. (1951/1967). *What is Cinema?*, trans. Hugh Gray, 95-124. Berkeley: UC Press.

BBC (2022). Two arrested in Egypt after teenage girl's suicide sparks outrage. *BBC News* 4 January.

Brunner, J. (2020). Fox News runs digitally altered images in coverage of Seattle's protests, Capitol Hill Autonomous Zone. The Seattle Times 14 June.

Cameron, E. (2004). From Plato to Socrates: Wittgenstein's journey on Collingwood's map. *AE: Canadian Aesthetics Journal* 10: 1-30.

Carroll, N. (1996). *Theorizing the Moving Image.* Cambridge: Cambridge University Press.

Cavell, S. (1979/1972). *The World Viewed*, enlarged edition [1979]. Cambridge: Harvard University Press.

Chesney, R. and Citron, D.K. (2019). Deep Fakes: A looming challenge for privacy, democracy, and national security. *California Law Review* 107: 1753-1820.

Cooper, S. (2007). A concise history of the fauxtography blogstorm in the 2006 Lebanon War. *The American Communication Journal* 9: 1-34.

Cole, M. and Ross, B. (2009). U.S. protests Russian 'sex tape' used to smear American diplomat. *ABC*. 23 September.

DeVoss, D.N. (2011) When images "lie": Traveling to the Pyramids with *National Geographic* and thinking about photo illustrations with Martha Stewart. *The Current Educator Innovator* 9 March.

Dougherty, J. (2009). U.S. calls purported sex tape 'doctored' and 'smear campaign.' CNN. 24 September.

Draper, R. (2020). *To start a war*. New York: Penguine.

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology* 34: 623-643.

Fisher, M. (2011). The truth about iconic 2003 Saddam statue-toppling. *The Atlantic*. 3 January.

Fong, G.T., Hammond, D. and Hitchman, S.C. (2009). The impact of pictures on the effectiveness of tobacco warnings. *Bulletin of the World Health Organization* 87: 640-643.

Gaut, B. (2010). *A Philosophy of cinematic art.* Cambridge: Cambridge University Press, 2010.

Goldberg, S., Editor in Chief. (2016). How we check what you see. *National Geographic* 230, n.p. [editorial precedes pagination].

Harwell, D. and Okazaki, S. (2021). A 'beautiful' female biker was actually a 50-year-old man using FaceApp. After he confessed, his followers liked him even more. *The Washington Post* 11 May.

Helmore, E. (2019). Alex Rodriguez bathroom photo highlights permissive privacy laws.

The Guardian 18 May.

Jarvie, I. (1987). *Philosophy of the film: Epistemology, ontology, aesthetics.* London: Routledge.

Kant, I. (1788/2015). *Critique of Practical Reason*. Trans. M. Gregor. Cambridge: Cambridge University Press.

Kerner, C., & Risse, M. (2020). Beyond porn and discreditation: Epistemic promises and perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics* 8: 81-108.

Kracauer, S. (1960) Theory of film: The redemption of physical reality. Oxford: Oxford University Press.

Lewis, J. (2018). Patriot Missiles Are Made in America and Fail Everywhere. *Foreign Policy* 28 March.

Maras, M.-H. and 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: 255-262.

Martinez, E. (2009). From Russia with love: Sex tape burns American diplomat, but is it real? *CBS*. September 24.

McCurry, J. (2019). K-pop singer Goo Hara found dead aged 28. The Guardian 24 November.

Mitchell, W.J. (1992). *The reconfigured eye: Visual truth in the post-photographic era*. Cambridge, MA: MIT.

Moise, E.E. (2019). *Tonkin Gulf and the escalation of the Vietnam War*, revised edition. Annapolis: Naval Institute Press.

Mullarkey, J. (2009). *Philosophy and the moving image: Refractions of reality.* London: Palgrave Macmillan.

Öhman, C. (2019). Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology* 22: 133-140.

Regan, T. (2002). When contemplating war, beware of babies in incubators. *The Christian Science Monitor* 6 September.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers Imprint* 20: 1-16.

Ritchin, F. (2008). *After photography*. W.W. Norton & Company.

Safi, M. (2016). Indian couple banned from climbing after faking ascent of Everest. *The Guardian*, 30 August.

Santayana, G. (c. 1900–1907/1967). The Photograph and the mental image. In *Animal Faith and Spiritual Life: Previously Unpublished and Uncollected Writings of George Santayana with Critical Essays on his Thought*, ed. J. Lachs, 391-402. New York: Appleton-Century-Crofts.

Schick, N. (2020). *Deepfakes: The coming infocalypse*. New York: Hachette.

Sontag, S. (1973/2005). *On Photography*. RosettaBooks LLC.

Strutt, D. (2019). *The Digital Image and reality: Affect, metaphysics and post-cinema*. Amsterdam University Press.

Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6, 1-12.

Walton, K. (1984). Transparent pictures: On the nature of photographic realism. *Critical Inquiry*

11: 246–247.

Wittgenstein L (1953). *Philosophical investigations*. Trans. GEM. Anscombe. Oxford: Basil Blackwell, 1953.

Young, G. (2021). *Fictional immorality and immoral fiction.* London: Lexington.