Synthese (2024) 204:14
https://doi.org/10.1007/s11229-024-04634-8

INTRODUCTION

Deepfakes: a survey and introduction to the topical collection

Dan Cavedon-Taylor[1]

Abstract
Deepfakes are extremely realistic audio/video media. They are produced via a complex machine-learning process, one that centrally involves training an algorithm on hundreds or thousands of audio/video recordings of an object or person, S, with the aim of either creating entirely new audio/video media of S or else altering existing audio/video media of S. Deepfakes are widely predicted to have deleterious consequences (principally, moral and epistemic ones) for both individuals and various of our social practices and institutions. In this introduction to the Topical Collection, I first survey existing philosophical research on deepfakes (Sects. 2 and 3). I then give an overview of the papers that comprise the Collection (Sect. 4). Finally, I conclude with remarks on a line of argument made in a number of papers in the Topical Collection: that deepfakes may cause their own demise (Sect. 5).

Keywords Deepfakes · Epistemology of technology · Ethics of technology

1 Introduction

Deepfakes are extremely realistic audio/video media. They are produced via a complex machine-learning process, one that centrally involves training an algorithm on hundreds or thousands of audio/video recordings of an object or person, *S*, with the aim of either creating entirely new audio/video media of *S* or else altering existing audio/video media of *S*.

The outputs of this process, the deepfakes themselves, are highly convincing in their appearance. As such, they are widely anticipated to be used for malign purposes. Indeed, malicious use of the technology is already a reality:

✉ Dan Cavedon-Taylor
dan.cavedon-taylor@open.ac.uk

[1] The Open University, Milton Keynes, UK

Published online: 26 June 2024

- In early 2022, a deepfake video of the Ukrainian president Volodymyr Zelenskyy surfaced in which he was depicted 'urging' Ukrainian fighters to lay down arms and surrender to Russia.
- In mid 2023, a deepfake of the English financial journalist Martin Lewis circulated on social media. Lewis was depicted 'promoting' an investment opportunity—one that was, in fact, a cryptocurrency scam—that promised "great investment opportunities for British citizens."
- In early 2020, members of the environment movement Extinction Rebellion created a deepfake of the Belgian Prime Minister, Sophie Wilmès. The Prime Minister was depicted 'admitting' to a link between deforestation and COVID-19.
- In late 2019, an audio deepfake was used to defraud the CEO of a UK-based energy firm of roughly £200,000.
- In late 2020 a US-based anti-corruption nonprofit released two deepfakes, one of Vladimir Putin and another Kim Jong-Un. In the deepfakes, the two leaders praise US politicians for undermining democracy with their recent actions.
- In mid 2023, a presidential campaign advertisement by the Republican Ron DeSantis included deepfake still images of his rival, President Donald Trump, embracing Dr. Anthony Fauci. Fauci is a controversial figure in the eyes of many Republican voters.
- One of the most prevalent uses of deepfakes is to 'face-swap' images of female celebrities (and also private individuals) into pornographic videos, an activity that, when performed non-consensually, is increasingly outlawed in various jurisdictions. At the time of writing, Scarlett Johansson and Taylor Swift are prominent targets for such non-consensually produced pornography.

The above illustrates that deepfakes have the potential to (and have caused actual) financial, political, psychological and moral harms to individuals. This is true both of the individuals depicted in deepfakes as well as viewers, who may (depending on the circumstances) be intimidated, blackmailed, or otherwise put at risk by their content. Moreover, deepfakes are thought to represent an epistemic threat to the accuracy of beliefs that viewers form about the world on the basis of watching or listening to such media. Very often, deepfakes are designed to deceive viewers.

At a more general level, deepfakes are predicted to have deleterious consequences not just for individuals, but for many of our social practices and institutions. In particular, deepfakes are thought to threaten our epistemic reliance on audio and video recordings. In the often-quoted words of political commentator and AI expert Schick (2020, p.9), deepfakes may propel us toward an "information apocalypse"—a digital environment saturated with disinformation. As a result, social practices and institutions that depend upon the veracity of video and audio recordings seem directly threatened by the technology. Democratic processes and procedures are widely considered to be especially vulnerable in this respect. This point has been argued by philosophers (Rini, 2020), legal scholars (Chesney & Citron, 2019), political theorists (Pawelec, 2022) and journalists (Frum, 2020).

However, it is not essential to deepfakes that they are non-veridical or epistemically pernicious. This fact is often overlooked. When deepfakes are defined and introduced in research articles or popular writings, it is not uncommon to see it claimed

that deepfakes show people saying and doing things that they did not. Yet deepfakes may be used to accurately represent events from the past or future. Indeed, as I discuss below, there are examples where they may even be of value in spreading information in the here and now. More generally, the creation of non-consensual pornographic deepfakes is less epistemically pernicious than it is morally so (Harris, 2021; Young, 2021). We should be careful not to tie deepfakes too closely to deception, lest we overlook both positive epistemic uses of the technology, no less than morally problematic ones that have little to do with the intent to deceive.

Many questions remain to be settled about the nature of deepfakes, including the following:

- In what ways might deepfakes have epistemically deleterious consequences?
- Do we have reason to think that the epistemically deleterious consequences of deepfakes will be as severe as many predict?
- In what ways might deepfakes constitute a moral or psychological harm to those depicted?
- Are deepfakes fundamentally different in kind from traditional and digital photographic images?
- Should a liberal democracy control or sanction the production and circulation of deepfakes?
- What positives can we expect from the technology?

In this introduction to the Topical Collection, I first survey existing philosophical research on deepfakes (Sects. 2 and 3). I then give an overview of the papers that comprise the Collection (Sect. 4). Finally, I conclude with remarks on a line of argument made in a number of papers in the Topical Collection: that deepfakes may cause their own demise (Sect. 5).

2 Existing literature: epistemic issues

The existing philosophical literature on deepfakes is small, but growing. Two papers set the background for many of the papers in this Topical Collection and shape the philosophical literature on deepfakes more generally.

First, Rini (2020) has argued that deepfakes will have epistemically deleterious effects for democracy and dependence on audio/video recordings in a highly distinctive way. In making her argument, Rini first claims that recordings realise a unique function in public political discourse; namely, they regulate testimony by functioning as 'epistemic backstops'.

Rini analyses recordings as epistemic backstops in two ways. First, she claims that recordings provide *acute corrections* to false public discourse–they 'set the record straight' when false testimony is given, exposing incompetent and insincere testifying by public figures. Second, Rini argues that recordings have a *passive regulatory effect*, shaping the assertions delivered by public figures in the first place by disincentivising false testimony:

> Our awareness of the possibility of being recorded provides a quasi-independent check on reckless testifying, thereby strengthening the reasonability of relying upon the words of others. (2020, p.2)

But with the introduction of deepfakes, Rini worries that recordings can no longer play these roles, leading to what she calls a 'backstop crisis' (p.8): no more 'smoking gun' tapes (no acute corrections of false or insincere testimony) and significantly less motivation for public individuals to be reliable testifiers (no more passive regulatory effects). Thus, according to Rini, the precise manner in which deepfakes will have epistemically deleterious effects is not that they will cause false beliefs per se, but that the bottom will fall out of our current-day testimonial practices insofar as these are reliant on recordings and their veracity. Worse still, reverting to testimonial norms from the pre-recording era is unlikely to be of much use, Rini thinks, in an epistemic environment "plagued by fake news malefactors and authoritarian deceivers" (p.14).

Second, Fallis (2020) has similarly argued that the epistemic threat posed by deepfakes is not just an increased number of false beliefs—though this is something that he takes seriously—but that they reduce the number of true beliefs that audiences can acquire from videos.

Fallis develops the above idea by first considering the motivations for creating and circulating deepfakes: to fan the flames of scepticism about events reported on and depicted by reliable sources of information; in particular, to sow seeds of doubt in viewers' minds about the legitimacy of reports on local and global events by reputable news agencies. Given the existence and circulation of deepfakes, if an audience is presented with a genuine video of someone they previously believed to be a competent and sincere testifier, Fallis worries that they might now withhold assent to its content because they believe it to be a deepfake. Even if they do believe its content, we might think that the audience is less justified in doing so, given the proliferation of deepfakes.

The precise way in which deepfakes will lead to these epistemic harms, Fallis thinks, is that their proliferation will cause existing, veridical videos to provide less information about what they depict than they did previously. Surveying different accounts of information carrying, Fallis settles on Brian Skyrms's (2010) to express this idea: a signal $\boldsymbol{R}$ carries the information that $\boldsymbol{S}$, if and only if the probability of $\boldsymbol{R}$ given $\boldsymbol{S}$ is greater than the probability of $\boldsymbol{R}$ given not $\boldsymbol{S}$. Unlike other accounts of information carrying, Skyrms's allows that information can be carried by degrees. The more likely it is for $\boldsymbol{R}$ to be sent when $\boldsymbol{S}$ is true than it is for $\boldsymbol{R}$ to be sent when $\boldsymbol{S}$ is false, the more information that $\boldsymbol{R}$ thereby carries about $\boldsymbol{S}$. But with the advent of (and likely increase in) deepfakes, the probability of a veridical video being a false positive has increased (and will likely increase further still). By analogy (2020, p.630), the red, yellow, and black stripes of coral snakes carry the information that they are venomous. But if the number of king snake mimics increases, the coral snake's appearance will not carry as much information about its being a venomous snake as it did previously.

Both Rini's and Fallis's papers rely on the idea that photographic images, or what they call 'recordings' or 'videos', play an important role in our epistemic lives. It is worth pausing to unpack this idea.

Outside of a small literature in aesthetics, the epistemic status of photographic images has received relatively little attention (Maynard, 1997; Currie, 1999; Cohen & Meskin, 2004; Walden, 2005; Abell, 2010; Pettersson, 2011; Hopkins, 2012; Cavedon-Taylor, 2013; Anscomb, 2018; Wilson, 2022). By contrast, one barely knows where to start when referencing the literature on the epistemology of testimony. This asymmetry is surprising. Photographic images, whether still or moving, are key sources of knowledge about how things are in the world. We rely upon them for knowledge about everything from holiday destinations, to potential romantic partners and family history, to the scale of natural and humanitarian disasters across the globe—to say nothing of astronomical events in space billions of miles from Earth.

The neglect of photographic pictures in social epistemology would be justified if our epistemic reliance on such images reduced to, or was simply a species of, our epistemic reliance on testimony. On the face of it, reducing our epistemic dependence upon photographs to epistemic dependence upon testimony is plausible. After all, photographs, whether still or moving, are made *by* people and they are *shown* by people *to* others—much like testimony involves assertions being made by people to others. Arguably, however, the epistemically relevant similarities stop there. Believing something because you are told it and believing something because you see it in a photograph are social practices governed by very different epistemic norms. This point has been underscored by Moran (2005, p.11):

> The status of the photograph as a reason to believe something does not depend on the photographer's own attitude toward it as evidence. It depends only on the camera's ability to record the scene, which need not involve any choice or consciousness on the part of the photographer at all. (The exposure could have been made by a remote timing device.)
>
> […]
>
> By contrast, the *speaker's* choice enters in essentially to the fact that his utterance counts at all as a reason for belief. The point is not that his utterance is voluntarily produced, for that in itself has no epistemic significance and does not distinguish the case from that of the photographer. Rather the point is that the speaker [is] presenting his utterance as an assertion, one with the force of telling the audience something.

This 'freedom' on the part of the speaker matters epistemically, as Moran then makes vivid with the following point:

> [I]f we learn that the photographer is not, in fact, presenting his photograph as a true record of what occurred in the park, the photograph as document retains all the epistemic value for us it ever had. (Ibid.)

If the contrast Moran draws here between testimony and photographs is correct, then our epistemic reliance on the latter cannot be modelled on our epistemic reliance on the former–at least not in terms of its rationality. One way of glossing Moran's point

is that there are defeaters for accepting another's testimony that fail to act as defeaters for accepting what is seen in another's photographs.

Despite the above difference between testimony and photographic images, there are other respects in which our epistemic dependence on photography is akin to our epistemic dependence on testimony. One idea that has been developed here (Cavedon-Taylor, 2013, 2015), and which sets the background for Rini's (2020) discussion, is that the psychology of accepting testimony and the psychology of accepting what one sees in a photograph is of the same kind: it is a matter of spontaneously formed belief. When believing what we are told and when believing what we see in photographic images, we assent by default and only withhold belief if we possess positive reasons for thinking the person or photograph to be *uncreditworthy*—say, if the person speaking is known to be insincere or the place the photograph is published is known to publish propaganda or falsehoods. (This is to assume non-reductionism about testimony—not everyone does, see Fricker (1994).) Just as we generally believe what we are told, absent defeaters, we generally believe what we see in photographic pictures, absent defeaters. One way of thinking about some of Rini's and Fallis's claims is that the proliferation of deepfakes constitutes a normative (and perhaps psychological) defeater here, undermining our ability to rely on photographic pictures in ways we once did.

More recently, Matthews (2022, p.77) has argued that one epistemically deleterious consequence of deepfakes is that they encourage the epistemic vice of intellectual cynicism—the character trait of "habitually distrusting and disengaging from epistemic practices" such as acknowledging video and audio recordings to have significant epistemic value. The remedy, Matthews claims, is for both individuals and institutions to promote the development of viewers' 'digital sensibility', a kind of sensitivity to the credibility of online and video content which Matthews models on Miranda Fricker's (2007) idea of testimonial sensibility to credible speech. Neatly dovetailing with this idea is Keith Raymond Harris's (2024) claim that turning to AI technology to detect deepfakes is undesirable, partly because doing so involves compromising on our autonomy as epistemic agents.

3 Existing literature: other issues

Although epistemic issues are at the forefront of many philosophers' thoughts about deepfakes, worries about moral and psychological harms have been considered too. Although these issues are sometimes intertwined with epistemic ones, some constitute *sui generis* anxieties about deepfakes.

Expanding on the ideas of an influential law article on deepfakes (Chesney & Citron, 2019), Nicholas Diakopoulos and Deborah Johnson (2021) have catalogued a range of potential moral harms that deepfakes may cause to individuals in the context of elections, including: reputational damage (either to candidates or voters), voter intimidation and deliberate misattribution of speech. Like the philosophers mentioned in the previous section, they claim that the harms of deepfakes extend to social practices and go beyond harms to individuals. For instance, Diakopoulos and John-

son claim that deepfakes have the potential to undermine trust in the democratic electoral process itself, leading to scepticism about how individuals are elected to govern.

Adrienne de Ruiter (2021) agrees harms such as the above are a real possibility, but argues that deepfakes may harm the persons portrayed in them even when not disseminated. This reflects the plausible thought that deepfakes that are produced without the consent of the individuals portrayed are morally problematic *tout court*, independent of facts about their circulation (see also Young, 2021, ch.11). According to de Ruiter, the harm caused by non-consensually produced deepfakes is that they represent a threat to an individual's right to 'digital self-representation'—broadly speaking, the right to self-determine one's social identity online. Furthermore, de Ruiter also highlights the psychological harm that may result from seeing and hearing oneself in a deepfake, e.g., doing and saying things one didn't do or say, particularly if the actions depicted are of a pornographic nature (pp.1326–1327).

While agreeing that the creation of non-consensually produced pornographic deepfakes is morally wrong, Öhman (2020) aims to highlight the difficulty in saying *why* producing such images is wrong. First, Öhman claims that if it is morally permissible to have private sexual fantasies about an individual, then it is difficult to see what is morally impermissible about a non-consensually created pornographic deepfake of an individual, so long as it is likewise kept private. However, Öhman claims that, intuitively, there *is* something impermissible about creating deepfakes non-consensually, even if concealment is guaranteed. So the question is: 'what?' Öhman's answer is that private fantasies do not, but private pornographic deepfakes do, play a role in causing and maintaining gender inequality, *by their mere existence*. On Öhman's view, while some individual pornographic deepfakes may seem morally innocuous, by virtue of considering the political context in which pornographic deepfakes, as a phenomenon, exist, their impermissibility is shown in terms of their supporting and perpetuating gender inequality.

Young (2021, ch.12) similarly takes up the question of why deepfakes are problematic yet private fantasies are not, but attempts to solve the dilemma a different way. Young distinguishes private fantasy from pornographic deepfakes by virtue of the latter's being disposed to be publicly accessible, even if it is not. According to Young, this makes the deepfake potentially harmful in ways that idle fantasy is not. Young's view is complex, however, since he believes that fantasy is not totally immune from moral scrutiny and that some forms may be as disrespectful as non-consensually produced deepfakes.

Rini and Cohen (2022) catalogue a number of further moral and psychological harms that may result from the production and circulation of non-consensual deepfakes. First, they observe that deepfake pornography is extremely objectifying in that it can "turn real people into digital toys" (p.147). Second, non-consensually produced deepfakes, pornographic or not, may cause what Rini and Cohen call "illocutionary harm". That is, they may subvert communicative agency and autonomy by compelling assertions from deepfaked individuals that they would not utter, given the choice. For instance, someone portrayed in a salacious deepfake may feel pressured, in order to recover integrity in the eyes of their peers or wider public, to issue a public denial of themselves having committed the embarrassing act that the deepfake depicts them as having performed. Third, non-consensually produced deepfakes may also be used

to gaslight individuals into believing that they *themselves* said or did things that they did not. This is not as far-fetched as it might sound:

> Imagine that one of your friends claims to have heard you say terrible things about your other friend. You certainly do not remember doing that, and you are pretty sure you would never say such a thing out loud. But now your rival pulls out their phone and plays a video: there you are, at your group's favourite pub, looking and sounding just a bit tipsy. And there you go, saying those terrible things. The video is dated from a year or so ago. "I just found it last night," says your so-called friend, "while I was going through old pictures. Don't worry though. I mean, *of course* I'd never show this to you-know-who." (p.153).

In addition, attention has been paid to the ontology of deepfakes and potential distinctions between different kinds of deepfake. Floridi (2018) has suggested we view deepfakes through the lens of 'ectypes'. An ectype is an artefact whose existence causally depends upon a prior artefact (the 'archetype'). Examples include: an impression left by a seal in wax and Lockean ideas, caused by the external objects which are their source. With the ectype/archetype distinction in hand, Floridi claims that among deepfakes we can contrast those which are *authentic* (that is, they wear their identity as deepfakes on their sleeves), but are *not original* in terms of archetypical source (that is, they fail to trace back to the object that they seem to) with deepfakes which are *inauthentic* (that is, they do not wear their identity as deepfakes on their sleeves), but *are original* in terms of archetypal source (that is, they do trace back to the object that they seem to).

As an example of the first kind of deepfake, Floridi mentions a Microsoft-authored, digitally-printed 'Rembrandt' painting, one that was created by Microsoft training an AI on existing Rembrandt paintings and instructing it to analyse these for typical subjects, composition, brushstrokes, and so on. The resulting image created by the AI is authentic in the sense that it is more 'Rembrandt' than any other painting could be. Nevertheless, it is unoriginal in that it fails to trace back to Rembrandt and in that (different) sense also fails to be a 'Rembrandt'.

As an example of the second kind of deepfake, Floridi mentions an audio deepfake of the speech John F. Kennedy was due to give on the day he was assassinated and which was created by training an AI on actual audio recordings of Kennedy's voice. The resulting deepfake of Kennedy's speech is inauthentic in that it is not an actual recording of Kennedy but is, by Floridi's lights, original insofar as it traces back to text Kennedy actually wrote. The deepfake is thus what Kennedy would have delivered on that day, had he not been assassinated. Microsoft's 'Rembrandt', by contrast, is unlikely to be a picture that Rembrandt would ever have painted.

Floridi's discussion is a helpful reminder that deepfakes offer many epistemic benefits, not just pitfalls. Catherine Kerner and Matthias Risse (2021), while fully aware of the dangers of deepfakes—they flag how deepfakes could be presented as 'alternative', long-suppressed documents of disputed events—discuss some potential benefits of the technology. In particular, they point to a number of examples that illustrate how deepfakes can empower the marginalised, be used as effective training tools and (perhaps surprisingly) even help better spread information.

In terms of empowering the marginalised, Kerner and Risse discuss how deepfake technology can allow people to expose and share stories of abuse or injustice on social media platforms anonymously in a more 'human' way than is typical. Often, when a speaker wishes to remain anonymous, their audio testimony is accompanied by a blurred, pixelated or silhouetted image of them. Kerner and Risse note that insofar as deepfakes allow testifiers to 'wear' a virtual mask with human features, facial expressions crucial for expressing and communicating the testifier's emotions can be preserved by the technology, but without compromising the testifier's anonymity.

With respect to being used as training tools, Kerner and Risse discuss contexts in which deepfakes can be an educational aid. For instance, the technology can allow for the creation of medical images, e.g., fake fMRIs, that can be used to train doctors where use of actual medical images might otherwise be costly or an invasion of patient privacy. In terms of being a tool to help spread information, Kerner and Risse discuss how deepfake technology was used in 2019 by a charity to enable David Beckham to deliver an anti-malaria message in nine languages. Crucially, Kerner and Risse (pp.102–105) also examine potential creative and aesthetic gains to be had from deepfakes. One prevalent use of deepfake technology is to swap actors in and out of film clips. While widely used to create non-consensual pornographic images, the method has also been used to create remixed, alternative versions of well-known movies. For instance, searching YouTube.com, you can find Jerry Seinfeld acting in a scene from *Pulp Fiction*; clips from *Terminator 2* starring not Schwarzenegger, but his rival Stallone; and a myriad of excerpts from films starring Nicholas Cage that he didn't, in fact, star in.

Such examples may seem frivolous and have little to do with serious aesthetic value or cinematic artistry. But consider the following hypothetical example suggested by Kerner and Risse: the television series *The Crown*, but featuring the faces of actual members of the British royal family. This counterfactual television series would doubtless be more psychologically impactful, more compelling and more emotionally engaging than the actual series (which is not to say that the actual series lacks in these respects).

Similarly, through deepfake technology, beloved actors and actresses could continue to grace the silver screen long after their death—this might prove particularly desirable in the case of actors and actresses who suffered untimely deaths and were unable, during their lifetime, to achieve the artistic greatness that they seemed destined for. Deepfakes would seem to open the window to new aesthetic and artistic possibilities in film.

Many of the above epistemic and moral issues are likely to resurface when philosophy gets around to having serious discussions about the creative capacities of deepfakes. Moreover, the creative use of deepfakes has many pitfalls of its own. For instance, by using deepfakes to 'revive' stars of the past and in contemporary films, we risk depriving ourselves of new on-screen talent and, as a result, risk depriving others of employment for mere aesthetic or artistic gains. This seems an unfair trade, not to mention concerns one may have about the revived star's inability to consent to their image re-appearing in contemporary films. Relatedly, there is currently widespread anxiety in the music and publishing industries about how the technology will negatively impact song-writers and literary authors. For all that, it is intuitive that

positive uses of deepfakes, creative ones in particular, represents fertile ground for philosophical research on deepfakes.

4 This topical collection

This Topical Collection comprises of 12 papers. Three of these seek to directly challenge Rini's (2020) and Fallis's (2020) claims about the seriousness or uniqueness of the epistemically deleterious consequences deepfakes might bring about.

First, Habgood-Coote (2023), in his contribution to the collection argues that deepfakes are not as unprecedented as they might seem, with similar effects having been achieved even in the early days of photography. Since we do, despite such images, acquire knowledge from photographic pictures, the idea that deepfakes are harbingers of an 'epistemic apocalypse' must be overstated he thinks. Habgood-Coote further argues that close attention to the social-cum-professional contexts in which photographs are not only made, circulated and displayed, but in which fakes are decried, denounced or discredited, should lead us to reject the idea, sketched in Sect. 2, that acquiring knowledge from photographs and videos does not involve epistemic dependence on other people. However, Habgood-Coote rejects the idea that this makes acquiring knowledge from photographs a form of testimony, since viewers of a photograph are not relying on a particular person when acquiring knowledge the picture. Instead, he argues that acquiring knowledge from photographic images involves reliance on a *group* of people participating in a norm-governed practice.

Second, Paloma Atencia-Linares and Marc Artiga (2021) likewise challenge pessimistic accounts of the consequences of deepfakes for our epistemic reliance on photographic images, focusing centrally on claims made by Fallis (2020). According to Atencia-Linares and Artiga, Fallis (2020) is right that deepfake technology can be conceptualised as potentially increasing false positives, but they deny that this is sufficient to make deepfakes epistemically pernicious. To make their case, Atencia-Linares and Artiga analogise producing and viewing photographs and videos to animal signalling models of communication, with deepfakes akin to animal mimicry. Consider: the characteristic bright colours of a coral snake signal its venomosity and some non-venomous milk snakes mimic those colours, producing misleading signals (just as deepfakes mimic the look of veridical photographic images and videos). But while a complete proliferation of milk snakes *would* cause the signalling system to break down, this would be self-defeating for the mimics, as bright colours would no longer be advantageous.

Similarly, Atencia-Linares and Artiga claim that a complete proliferation of deepfakes would similarly result in people ignoring them, undermining their ability to spread false information. Crucially, Atencia-Linares and Artiga identify a number of similar mechanisms that they believe are likely to ensure the reliability of photographic images in the face of deepfakes and which are directly analogous to mechanisms in nature that ensure the reliability of an animal signalling system in the face of mimics. On this view, and like Habgood-Coote's, the ideas of Sect. 2 above are again rejected and the reliability of photographs is pictured as a contingent matter. Photographs are reliable, on this view, not in themselves, but only insofar as fakes

are costly and resource-intensive to produce; certain kinds of fakery are punished but 'harmless' ones are permitted; there happens to be a common interest in photographs being informative; and so on. According to Atencia-Linares and Artiga, a system with mechanisms like this is robust enough to tolerate mimics like deepfakes, possibly even in scenarios where deepfakes become fairly frequent in number.

Third, Keith Raymond Harris (2021), like Atencia-Linares and Artiga (2022) and Habgood-Coote (2023), suggests that attention to factors external to videos and photographs can help allay concerns about deepfakes being an "epistemic maelstrom" (Rini, 2020, p.8). Harris concedes that the epistemic value of 'bare' photographic and video footage may be threatened by deepfakes. However, he claims that the epistemic value of photographic and video footage *as presented by appropriate sources* is unlikely to be threatened much, where sources include particular television channels, news corporations or social media accounts.

Harris's idea is intuitive. Whether a photograph or video is presented by one news corporation rather than another, or by a particular influencer rather than another, seems relevant for whether one ought to believe what one sees in the photograph or video, or whether one should instead suspend judgment. Harris admits that identifying appropriate sources is not easy for viewers and that many trust sources that they shouldn't. But Harris denies that this is a new epistemic problem, let alone one uniquely caused by deepfakes—hence deepfakes are not particularly epistemically catastrophic. Where deepfakes may create a new epistemic burden, Harris claims, is on sources themselves. In the face of deepfakes, life is more difficult for news corporations, television channels, social media influencers, etc. in terms of which videos they may responsibly broadcast to viewers and which they should not. Nevertheless, the problem is not insurmountable, Harris claims, though it may involve sources relying upon external content less frequently.

Near the end of his paper, Harris shifts focus to the moral harms that deepfakes may bring; in particular, harms that deepfakes may cause to a person due to the circulation of non-consensually produced deepfake pornography of that individual. Harris plausibly argues that public individuals who are made to feature in widely-circulated deepfake pornography may come to be associated with sexual activity. Importantly, this association may be non-doxastic and not rise to the level of viewers' beliefs in order to nonetheless harm. For instance, even if the deepfake is not convincing, or if viewers do not believe that the depicted public individual performed the depicted sexual acts, exposure to deepfakes that depict the individual engaging in sexual acts is likely to create an association in the minds of viewers between that individual (and perhaps a group to which they belong) and sex. This association may be unwanted by the individual and may have negative consequences for the individual's career and wellbeing.

Matthews (2023) discusses the epistemology of photographic and video recordings from another perspective, arguing that there are lessons here for mainstream epistemology; in particular, he argues that deepfakes pose a unique challenge to anti-risk analyses of knowledge. For instance, according to the anti-risk theory of Pritchard (2017), knowledge involves true belief that is attributable to (a) cognitive ability and (b) is safe from veritic epistemic risk. In order for a belief to be safe from veritic epistemic risk, there must not be a close possible world in which that belief

is formed via the cognitive ability it was actually formed by, but in which that belief turns out to be false. In fake barn scenarios, where a subject sees the only real barn for miles in an environment densely populated by barn façades, the subject forms the true belief that there is a barn in front of them that meets (a) but not (b); hence, on Pritchard's view, why the belief is not knowledge. Matthews's contribution to this topic is to first analyse deepfakes as fake barn cases ('digital fake barns'). In doing so, he discusses a number of compelling hypothetical examples. Consider the following, inspired by real events:

> **SCAN**: Derek is a radiologist tasked with identifying a batch of scans for lung cancer tumours. He meticulously looks over one of his patient's scans on a computer screen in his well-lit office. Unknown to Derek, a computer hacker has intercepted the other scans in the batch and added deepfake tumours to them. As it turns out, Derek looks at one of the genuine videos and correctly concludes it is tumour-free.

As mentioned in Sect. 3 above, medical images created by deepfakes may prove a useful training tool. So Matthews's idea that such imagery might get mixed in with real medical images is not far-fetched. Indeed, Matthews argues that this fact makes cases like **SCAN** a relevant alternative, while typical fake barn cases involve abnormal environments. Thus, deepfakes constitute a particularly pressing epistemic risk to perceptual beliefs, more so at least than the possibility of finding oneself in fake barn county.

Second, Matthews argues that while Pritchard's view yields the right verdict in cases like **SCAN**, it yields the wrong one in altered cases like **SCAN***. In **SCAN*** there is only one deepfake image in a set of *bona fide* scans. Intuitively, the presence of a single deepfake scan in Derek's local environment does not prevent him from knowing, say, that a particular patient's lungs are cancer-free when looking at a veridical scan. Matthews claims that Pritchard's view yields the opposite, incorrect verdict. After all, a world in which Derek looks at the one deepfake scan, and so forms a false belief, is modally close to Derek's actual one (it is also closer than **SCAN**). Yet, Matthews claims, Pritchard's view must treat **SCAN** and **SCAN*** equivalently. The problem generalises, he argues, to all perceptually-based beliefs acquired from seeing photographs and watching videos, and extends even to testimony delivered through audio-recordings.

Marco Viola and Cristina Voto (2023) tackle the question of how deepfakes will affect our epistemic reliance on photographs in tandem with the question of how they harm. They address the issue of whether the widespread circulation of non-consensually produced pornographic deepfakes will dramatically increase the amount of harm such abusive images cause arguing for a qualified form of optimism that it will not. Their reasoning is as follows:

1. The allure of non-consensually produced pornographic deepfakes, as well as their potential to harm, heavily relies on the special epistemic and affective status that we currently associate with photographic images and videos.

2. Increased (awareness of) the spread of deepfakes will progressively erode the special epistemic status and possibly the affective status of photographs.
3. In the long run, the very diffusion of non-consensually produced pornographic deepfakes will hamper their allure and their potential to harm.

Regarding the affective status of photographs, this is their capacity to seemingly put us in intimate contact with their objects (Walton, 1984) and refers to the way that photographs are psychologically impactful, e.g., they shock, arouse, embarrass, etc. in ways that drawings and paintings don't (which is not to say that paintings and drawings are devoid of all psychological impact). According to Viola and Voto, the epistemically deleterious effects that Rini (2020) and Fallis (2020) claim deepfakes will bring for our reliance on photographic images has at least one positive outcome: non-consensually produced pornographic deepfakes will cease to be so impactful. The proliferation of deepfakes, they believe, is likely to cause viewers to adopt a more sceptical and detached attitude toward photographic media, including videos, one that will rob non-consensually produced pornographic deepfakes of their capacity to harm in the first place. This relatively optimistic stance on deepfakes is said by the authors to be 'qualified' in the sense that Viola and Voto do not deny that deepfakes may morally harm, that deepfakes arise in a gender-biased context (Öhman, 2020) or that deepfakes may create the kinds of associations identified above by Harris (2021). They also admit that it may take significant time for our epistemic-cum-affective responses to photographs to shift in the relevant ways, though they suspect that younger users of technology are already en route toward holding this more sceptical attitude.

Carl Öhman (2022), in his contribution, addresses the question of when we can say of an individual that their likeness, i.e. face or voice, appears in a deepfake. Individuals are increasingly seen to have a moral and legal right over their likeness or 'identity'. So knowing when a deepfake depicts one individual rather than another is a pressing, practical matter. Öhman rejects two tempting answers: (i) that an individual is depicted in a deepfake if and only if the relevant AI or application that generated it was trained on photographs of that individual; and (ii) that an individual is depicted by a deepfake if and only if facial recognition software identifies that individual as depicted in the deepfake. Öhman rejects both of these technologically-grounded answers as unsatisfactory and proposes one that is more viewer-relative. In particular, Öhman devises a heuristic, modelled on Alan Turing's imitation game, that may produce different answers depending on the purpose for which the question of identity of likeness is raised, just as the question of whether a converted school is the same building as the hospital it was converted from may get different purpose-relative answers ('yes', if the purpose is to locate the building; but arguably 'no' if the purpose is to state the building's function). Öhman focusses in particular on purposes for which the question of likeness is raised that relate to personal humiliation, personal reputational damage and damage to the political ecosystem.

Francesco Pierini (2023) likewise addresses the issue of what determines who is depicted in a deepfake. Pierini addresses this question via a different methodology than Öhman's; namely, by inquiring into what standard of correctness is appropriate for deepfakes. The idea that pictures have a standard of correctness is the idea

that there are right and wrong things to see in a picture. A widely affirmed view in philosophy of art, originally due to Richard Wollheim (1980), is that drawings and paintings have an *intentional* standard of correctness insofar as what it is right to see in these pictures is whatever was intended by their makers, whereas photographic pictures have a *causal* standard of correctness insofar as what it is right to see in them is whatever object was before the camera when the photograph was taken. Wollheim's claim goes hand in hand with the ideas from Sect. 2 about how learning from photographs is non-testimonial in nature. Indeed, Pierini links it to the idea that photographs serve documentary and evidential functions whereas paintings and drawings serve communicative functions that are mediated by inference.

Pierini examines what standard of correctness is appropriate for deepfakes. The answer he arrives at is complex. Examples of deepfakes that involve face-swapping he calls 'local deepfakes' since only a region of the image is altered (but see comments below on Millière, 2022). Here, Pierini claims we get a mixed standard of correctness: the parts of the image not altered retain their causal standard of correctness whereas the parts altered acquire an intentional standard of correctness; crucially, the latter do so despite being automated by the image-making process. Examples of deepfakes that are created completely from scratch he calls 'global fakes'. Here, Pierini claims that, despite the image again being created via an automated process, only an intentional standard of correctness applies. The upshot is that deepfakes should excel at communicative functions, which is what Pierini claims is the case via an examination of several deepfakes whose point would be missed if they were viewed as photographic evidence. Pierini leverages this analysis to then explain, on his own terms, what is epistemically troubling about deepfakes: because of their photorealism, they share an appearance with pictures that have a causal standard of correctness and which thereby have evidential functions; this, he claims, is what tempts viewers to wrongly treat deepfakes as evidence. Crucially, Pierini denies Rini's and Fallis's claims regarding the epistemically deleterious effects that deepfakes will bring about for our reliance upon photographs.

Matthew Crippen's (2023) contribution focuses on the related idea that deepfakes are somewhere 'in-between' photographs and non-photographs and what this means for the thesis of photographic transparency (Walton, 1984). The idea that photographs are transparent is that seeing photographs is a way of literally (albeit indirectly) seeing the photographed object, akin to seeing objects through mirrors, microscopes and telescopes. Crippen argues that deepfakes have a 'sheen of transparency' and are thought to be morally problematic (when they are) precisely because they are experienced as giving direct, perception-like access to what they depict. Some have challenged the transparency thesis on the grounds that it is outdated when digital photo-manipulation is now rife. Crippen argues that our finding certain deepfakes to be morally problematic shows that attitudes of transparency are alive and well—we still treat photographs, and photographic-looking media, as windows on to the world. Crippen makes this vivid in the following way: a pornographic or nude depiction of the current or past US President in marble or paint would likely be considered political commentary or satire. But a deepfake with the same content would likely prompt a more visceral reaction or complaint (as was indeed the case when deepfakes of Taylor Swift surfaced in early 2024). Much like Viola and Voto (2023), Crippen

argues that as the number of deepfakes increases, the harm they can do will decrease, since the more familiar we become with deepfakes, the less tempted we will be to treat them as transparent.

Alex Barber (2023), in his contribution, takes up a provocative question; namely, when do deepfakes deserve protection? While it is intuitive that some deepfakes should be criminalized, it is also intuitive, though arguably overlooked, that some should be defended on freedom of speech or freedom of expression grounds. For instance, many deepfakes are satirical or constitute political commentary. Barber defends the claim that deepfakes such as these deserve protection against two views: the 'knee-jerk' view that all deepfakes *qua* fakes deserve no such protections and the more moderate 'dual-policy' view that only deepfakes which overtly signal their identity as deepfakes deserve protecting. Rejecting both views, Barber changes tack and asks whether any new freedom of expression issues are raised by the advent of deepfakes in their various manifestations. Barber's view is that none are. Like many of the authors discussed above, Barber's view is that deepfakes fail to be as unique and unprecedented as their critics claim. For instance, Barber argues that racist deepfakes can be treated in the same manner as racist sketches. Moreover, insofar as non-consensually produced deepfake pornography is a form of defamation or harassment, he claims that it can likewise be countered by existing defamation or harassment laws and does not require new ones. In sum, Barber argues that the measures that will best constrain problematic deepfakes have correlates already in place in order to constrain more traditional, problematic media and actions. Thus, Barber claims that deepfakes do not call for a novel response on our part, either in terms of their protection or restriction.

Tom Roberts (2023), in his paper, examines deepfakes through a novel lens: the philosophy of language, distinguishing two different ways that deepfakes may relate to speech acts. What Roberts calls 'closed deepfakes' are those that purport to record speech acts (such as a deepfake of a wedding ceremony) while 'open deepfakes' are those that purport to address the viewer directly with an illocutionary act, e.g., a deepfake of a threat, promise or apology, seemingly made to the camera/viewers. Roberts argues that the distinction is of value in expanding our understanding of deepfakes; in particular, how deepfakes may deceive and be put to immoral use. Roberts claims that closed deepfakes are the kind of deepfake that the literature has primarily focussed on to date. Such deepfakes constitute a kind of misleading evidence about a person's behaviour. But open deepfakes, with their distinctive perlocutionary properties, risk being more pernicious. Roberts argues that they constitute a kind of direct attempted manipulation of behaviour—they are akin to a fake court summons, where closed deepfakes are akin to a fake courtroom sketch. Like Rini and Fallis, Roberts's view is a pessimistic one about the effects of deepfakes: what is distinctly problematic about deepfakes, Roberts aims to highlight, is that they can be technological artefacts for manipulating others' behaviour.

Raphaël Millière (2022), in his contribution, takes a close look at different types of deepfake and how best to situate them in a taxonomy of audio-visual media. Several of the papers listed above imply that deepfakes are broadly continuous with existing audio-visual media and rely on this claim to challenge Rini- and Fallis-style pessimism about the epistemic consequences of this new technology. But Millière aims

to challenge the idea that deepfake technology is merely a new way of producing already-existent audio-visual outputs. He makes his case by claiming that deepfakes do not readily fit within existing frameworks for analysing audio-visual media. Millière argues for this by taking a close look at the etiology of deepfakes, examining their causal origins in deep-learning algorithms and existing datasets of audio-visual images. According to Millière, this affects the ontology of deepfakes, what type of medium they fundamentally are. (This claim is plausible, since it is often on etiological factors that drawing and painting is distinguished from, say, photography.) One striking example is deepfakes putatively created from scratch, rather than by, e.g., face-swapping or by manipulating one aspect of an already existing image (see also the above discussion of Pierini's paper). As Millière highlights (p.213), because of the way that these deepfakes are created from existing datasets of audio-visual images, they are never wholly fabricated. No deepfake, that is, is ever created from scratch entirely.

The reverse type of deepfake, one which involves a putative local alteration to an existing image, is likewise argued by Millière to be misunderstood. The process of creating such an image involves a deep-learning-based reconstruction of the entire image prior to the algorithm's then making the putatively 'local' adjustment. Two other examples that Millière (p.231) discusses when aiming to highlight the novelty of deepfakes are: (i) the ability of the technology to seamlessly combine fake audio and fake moving visual images; and (ii) the ability to control the outputs of the technology to achieve on-the-fly, real-time effects. On Millière's view, it follows that deepfakes represent a genuine departure from existing digital image creation and editing techniques, no less than a departure from analogue ones. According to Millière, we misunderstand deepfakes if we think of them as simply a more sophisticated form of Photoshop.

Oliver Laas (2023), in his contribution, agrees with Rini and Fallis that deepfakes are epistemically pernicious. But he argues that their accounts fail to get right the fundaments here. Laas argues that the key reason why deepfakes are epistemically harmful is that they will upend our default doxastic attitude of believing that *P* on the basis of a recording that supports the truth of *P*. This is the account of our attitude toward recordings that I outlined in Sect. 2 and which has its roots in Cavedon-Taylor (2013). The uniqueness of Lass's approach is that he analyses this attitude as one of trust, defending the claim that trust is not merely an attitude that we may take toward agents, but is also an attitude that we may take toward technology. Thus, according to Laas, trust can be non-anthropocentric: in trusting a piece of machinery to work, e.g., by unreflectively pressing a button to turn it on, I trust *it*, Laas argues, rather than its makers or designers. This, he claims, is shown by distinctive forms of artefact-directed anger or disappointment (see also Nickel, 2013). But with deepfakes, Laas argues that trust in recordings will no longer be possible—our default doxastic attitude toward them will not be to accept their contents. This is not the end of Laas's analysis, however. He argues that trust in recording technologies can be revived by means of a source-based authentication method. In closing, he sketches a decentralised, blockchain-based vision for how the history of a recording could be monitored in a way that would document access to the recording, including monitoring for third-party interference. With such a tamper-proof, blockchain-based ledger

of information on recordings in place, the veracity of a recording might be confirmed with confidence and our trust in veridical recordings restored.

5 Conclusion

Deepfake technology is currently in its infancy. What the future brings, both for the technology itself and those whose lives will be affected by it, is uncertain. Judging by the current literature, this Topical Collection included, the central question about deepfakes for philosophers is whether they represent anything new under the sun or call for a unique response on our part. Although deepfakes are a novel technology, it is an open question whether they are different *in kind* from what's gone before (be that existing types of media, existing forms of communication, existing forms of disinformation, etc.), or different only *in degree*.

Those in the Topical Collection who suggest that deepfakes fail to be wholly distinctive, or who think that they do not call for a novel response on our part, include, broadly, Atencia-Linares and Artiga (2022), Barber (2023), Habgood-Coote (2023) and Viola and Voto (2023). Those advocating here for the relative uniqueness of the technology or unique consequences of the technology, include, broadly, Harris (2021), Millière (2022), Öhman (2022), Crippen (2023), Laas (2023), Matthews (2023), Pierini (2023) and Roberts (2023). However these last authors differ in their claims regarding *how* deepfakes may be unique, or *what* novel response they call for on our part. Certainly not all in this second list agree with the sceptical views defended by Rini and Fallis.

One important new thread of argument that can be recovered from some papers in this Topical Collection is that deepfakes may only succeed in causing their disruptions, whether moral or political, in the short term (see Atencia-Linares & Artiga, 2022; Crippen, 2023; Viola & Voto, 2023). On the face of it, deepfakes depend for their capacity to harm on either (i) viewers taking the kind of credulous doxastic attitude toward photographic media outlined in Sect. 2, which may already be waning; or (ii) the relatively few number of deepfakes currently in circulation, which is only ever increasing. That is, if our attitude of credulity toward photographic media should cease, or if deepfakes should become more widespread, then we might stop taking them at face-value. If so, then their potential to cause moral or political upset might be quite limited. This suggest that the best line of defence against deepfakes might be to let them proliferate, i.e. in the hope that when 'normalised' they won't be taken seriously anymore and, as a result, will do little damage.

In closing, I want to push back against the above line of argument. First, as Harris (2021) rightly argues, deepfakes can harm the people portrayed in them by creating non-doxastic associations in the minds of audiences, i.e. between the individuals depicted in the deepfakes and the actions that they are falsely portrayed as performing (sexual ones, in the case of non-consensual pornographic deepfakes). That is, deepfakes do not harm, whether morally or politically, only insofar as they cause audiences to form false beliefs. Rather, they may harm insofar as they lead us to associate certain persons with certain actions that they desire to not be associated with. So focussing on our doxastic attitudes toward deepfakes and, more generally, what

deepfakes cause us to believe, may be a red herring when it comes to understanding their potential to harm.

Second, one can imagine a similar claim having been made about trolling in the early days of the internet: trolling is a disruption to communication on the internet, but one that may be short lived since it relies for its success upon (i) people having earnest communicative intentions when engaging with others online; or (ii) the relatively few number of trolls online. It seems to me questionable that (i) and (especially) (ii) hold true as of 2024. Still, trolling hasn't gone away. Indeed, it has become more sophisticated. Trolling now includes acts like 'doxxing' (releasing private information about one's interlocutors to others) and 'swatting' (falsely alleging that one's interlocutor is involved in a serious crime in order to prompt an armed response from law enforcement), not to mention malicious use of deepfakes themselves. Trolling has become relatively 'normalised' in the sense that it is part and parcel of engaging with others on, e.g., a public forum, social media platforms, comment threads, etc. Still, as examples like doxxing and swatting illustrate, despite its normalisation, trolling has not ceased to be a harm and has become only more sophisticated in *how* it harms. The same may prove true of deepfakes.

Acknowledgements My thanks to all authors and referees who made this Topical Collection possible. I would also like to thank the editors and staff of *Synthese* for their guidance, especially Kristie Miller.

References

Abell, C. (2010). The Epistemic Value of Photographs. In Catharine Abell, and Katerina Bantinaki (Eds.), *Philosophical Perspectives on Depiction*. OUP.

Anscomb, C. (2018). The epistemic value of photographs in the age of new theory. *Proceedings of the European Society for Aesthetics*, *10*, 1–18.

Atencia-Linares, P., & Artiga, M. (2022). Deepfakes, shallow epistemic graves: On the epistemic robustness of photography and videos in the era of deepfakes. *Synthese*, *200*, 518.

Barber, A. (2023). Freedom of expression meets deepfakes. *Synthese*, *202*, 40.

Cavedon-Taylor, D. (2013). Photographically based knowledge. *Episteme*, *10*(3), 283–297.

Cavedon-Taylor, D. (2015). Photographic phenomenology as cognitive phenomenology. *British Journal of Aesthetics*, *55*(1), 71–89.

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

Cohen, J., & Meskin, A. (2004). On the epistemic value of photographs. *The Journal of Aesthetics and Art Criticism*, *62*(2), 197–210.

Crippen, M. (2023). Conceptual and moral ambiguities of deepfakes: A decidedly old turn. *Synthese*, *202*, 26.

Currie, G. (1999). Visible traces: Documentary and the contents of photographs. *The Journal of Aesthetics and Art Criticism*, *57*(3), 285–297.

de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy and Technology*, *34*, 1311–1332.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072–2098.

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy and Technology*, *34*(4), 623–643.

Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology*, *31*(3), 317–321.

Fricker, E. (1994). Against gullibility. Bimal Krishna Matilal, and Arindam Chakrabarti (Eds.), *Knowing from words*. Springer.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Frum, D. (2020). The very real threat of Trump's deepfake. The president's first use of a manipulated video of his opponent is a test of the boundaries. *The Atlantic*, 27 April.
Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese*, *201*, 103.
Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese*, *199*(5–6), 13373–13391.
Harris, K. R. (2024). AI or your lying eyes: Some shortcomings of artificially intelligent deepfake detectors. *Philosophy and Technology*, *37*(7), 1–19.
Hopkins, R. (2012). Factive pictorial experience: What's special about photographs? *Nous*, *46*(4), 709–731.
Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81–108.
Laas, O. (2023). Deepfakes and trust in technology. *Synthese*, *202*, 132.
Matthews, T. (2022). Deepfakes, intellectual cynics, and the cultivation of digital sensibility. *Royal Institute of Philosophy Supplement*, *92*, 67–85.
Matthews, T. (2023). Fake barns, and knowledge from videos. *Synthese*, *201*, 41.
Maynard, P. (1997). *The engine of visualization: Thinking through photography*. Cornell University Press.
Millière, R. (2022). Deep learning and synthetic media. *Synthese*, *200*, 231.
Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint*, *5*, 1–29.
Nickel, P. (2013). Trust in technological systems. In Marc J. de Vries, Sven Ove Hansson, and Anthonie W.M. Meijers (Eds.), *Philosophy of engineering and technology, Vol. 9: Norms in technology*. Springer.
Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of deepfake pornography. *Ethics and Information Technology*, *22*(2), 133–140.
Öhman, C. (2022). The identification game: Deepfakes and the epistemic limits of identity. *Synthese*, *200*, 319.
Pawelec, M. (2022). Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society*, *1*, 19.
Pettersson, M. (2011). Depictive traces: On the phenomenology of photography. *The Journal of Aesthetics and Art Criticism*, *69*(2), 185–196.
Pierini, F. (2023). Deepfakes and depiction: From evidence to communication. *Synthese*, *201*, 97.
Pritchard, D. (2017). Anti-risk virtue epistemology and negative epistemic dependence. *Synthese*, *197*(7), 2879–2894.
Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint*, *20*(24), 1–16.
Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy*, *22*(2), 143–161.
Roberts, T. (2023). How to do things with deepfakes. *Synthese*, *201*, 43.
Schick, N. (2020). *Deep fakes and the infocalypse: What you urgently need to know*. Hachette.
Skyrms, B. (2010). *Signals: Evolution, learning and information*. Oxford University Press.
Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese*, *201*, 30.
Walden, S. (2005). Objectivity in photography. *The British Journal of Aesthetics*, *45*(3), 258–272.
Walton, K. L. (1984). Transparent pictures: On the nature of photographic realism. *Critical Inquiry*, *11*(2), 246–277.
Wilson, D. (2022). Reflecting, registering, recording and representing: From light image to photographic picture. *Proceedings of the Aristotelian Society*, *122*(2), 141–164.
Young, G. (2021). *Fictional immorality and immorality in fiction*. Lexington Books.
Wollheim, R. (1980). Art and its Objects. 2nd Edition. Cambridge University Press.