Philosophers' Imprint

VOLUME 20, NO. 24
AUGUST 2020

Deepfakes and the Epistemic Backstop

Regina Rini
York University

Deepfakes are fabricated video or audio recordings created through machine learning technology. A computer program uses a large data set of real recordings to build a model of the facial/vocal characteristics of a person, then superimposes this onto recordings of another person. The effect is an *apparent* recording of a well-known person doing or saying something they never did. If you haven't seen a deepfake, you should probably stop and watch one before continuing with this paper; it's a phenomenon that's hard to grasp without ostensive demonstration.[1]

So far, the output isn't entirely convincing. But that may change, and soon. Which means we should start asking: What could happen to our collective democratic processes of information-sharing and debate when our leaders can be deepfaked into doing or saying whatever a malicious agent desires? Hints of this worry already appear in journalistic discussions, but outside computer science and legal studies, there has been little published by academics.[2] To my knowledge, there is nothing yet from philosophers on the subject. So consider this essay an exercise in prophylactic political epistemology; I aim to raise worries about deepfakes and the erosion of knowledge in democratic societies *before* it begins to happen.

Testimony, Recordings, and the Epistemic Backstop

I'll come back to deepfakes shortly. First, we need to get clear on the epistemic environment in which they operate. I'm going to argue that audio and video recordings currently serve a distinctive role in public discourse. Specifically, they regulate our testimonial practices, providing what I'll call an *epistemic backstop*.

Our collective epistemic practices are highly reliant on testimony, the transmission of knowledge via say-so. When a credible person tells me that she has seen or heard something, I am typically justified

1. For an especially entertaining example, you can watch "Barack Obama" insult Donald Trump in a demonstration from Buzzfeed and Jordan Peele. See Mack (2018).

2. See Maher (2018). For legal academic perspectives, see Chesney and Citron (2019); Silbey and Hartzog (2019); Pfefferkorn (2019).

in believing the contents of her claim simply on the basis of her testifying to it. Of course, this justification can be defeated by evidence against the proposition, or undercut by evidence that my interlocutor is unreliable. But absent those factors, I seem to have a default justification to accept testimony as evidence, and in many cases to adopt a belief simply on testimonial recommendation. Since any individual knower only has so much time and cognition available, our knowledge of the world would be severely impoverished if we could not rely upon testimony.[3]

Of course, some philosophers — being philosophers, after all — have skeptical worries about our reliance on testimony. But most theorists accept *some* form of justification via testimony, and all socially functional humans rely on it in practice, whatever their theoretical qualms. A major part of the rationality of this practice comes from widespread awareness of enforceable *testimonial norms*. When a person attempts to provide testimony, she is taken to be implying that she is both *sincere* and *competent* about the matter on which she testifies. Put simply: She really means it, and she knows what she's talking about.[4] A person who regularly transmits misinformation violates one or both of these norms and will likely acquire a reputation as a liar or an idiot. Fear of those reputational consequences gives everyone — even those lacking in good will — a reason to cooperate with the norms.

An excerpt from Michael Cohen's February 2019 testimony before the US House Oversight Committee provides a remarkably succinct illustration of *both* the ways that bad testimonial practice can blemish one's reputation. Cohen admitted to transmitting disinformation for his erstwhile client, but claimed a distinction in their respective epistemic failings:

> Rep. Comer: You called Mr. Trump a cheat. What would you call yourself?
>
> Michael Cohen: A fool.[5]

So far, these are just familiar philosophical theses about the nature of testimony. But now I want to draw attention to the role of *recordings* in regulating our testimonial practices, something rarely appreciated by analytic epistemologists. Neglect of this role has perhaps been reasonable until now, but deepfakes will make it dangerous.

Here is the main idea: Video and audio recordings function as an *epistemic backstop*. They regulate our testimonial practices, sometimes bluntly though more often subtly. The availability of recordings undergirds the norms of testimonial practice, increasing the incentive for testifiers to speak with sincerity and competence. 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.

I think that recordings do this in two distinctive ways: actively correcting errors in past testimony and passively regulating ongoing testimonial practices. I will take these in turn.

Acute correction is the active role recordings play in regulating testimony. Think of familiar "let's check the tape" scenarios: Someone has testified that an event transpired in such-and-such a way, but then a surreptitious recording reveals otherwise. Perhaps the most famous

3. The philosophical literature on testimony has grown significantly in the last three decades. For important examples, see Coady (1992); Lackey (2008); and Goldberg (2010). Among philosophers, there is an active debate whether we should think of testimony as a basic source of epistemic justification (a position sometimes called non-reductionism), or as just one among several sources of input to some more basic epistemic capacity (sometimes called reductionism). I don't think we need to settle that debate here; so far as I can tell, it won't matter to most of my points about how people use testimony in ordinary life, whatever we might say theoretically. See Lackey (2006) for an overview of the debate and reasons to think the dichotomy itself may be a mistake.

4. Here, I am synthesizing (and simplifying) convergent points from two different philosophical literatures: the knowledge norm of assertion (Williamson 1996; Lackey 2007) and the nature of interpersonal trust (Baier 1986; Jones 1996).

5. See Prasad and Matza (2019).

example is Richard Nixon's "smoking gun" tape. In 1974, Nixon denied for months that he'd had any foreknowledge of the Watergate coverup. His aides testified to the public that the president was innocent; they became the outward links in a testimonial chain running back to Nixon himself. But it turned out that the president's historical grandiosity had led him to install a secret audio recording system in the Oval Office. After much dissembling and legal evasion, Nixon turned over a June 1972 recording on which he could be heard ordering the CIA to interfere in the FBI's Watergate inquiry. The effect was devastating; Nixon was revealed to have unambiguously flouted the sincerity norm of testimony, and his aides further down the testimonial chain were left out in the cold. This last fact seems to have been especially important, as many of the Nixon true-believers now felt personally betrayed by a president exploiting their testimonial assistance. Even his innermost defenders turned on him. Within weeks, the president resigned.[6]

You might think modern politicians have learned from examples like this. But recordings still provide acute correction in politics. In spring 2018, Junichi Fukuda, a senior official in the Japanese Finance Ministry, was accused of sexually harassing a reporter during an interview. Fukuda denied the allegations — but it then emerged that the reporter had pressed the audio record button on her mobile phone. When she publicly released this recording, Fukuda continued to deny the allegations. He admitted the voice in the recording perhaps *sounded* like him, but claimed: "I cannot tell since I can only hear my voice through my own body." The public was not convinced. Fukuda was quickly forced to resign.[7]

Recording can also provide acute correction in cases of conflicting or confused testimony. An extremely famous example is the Zapruder film of the 1963 assassination of John Kennedy. In this case, there were thousands of independent eyewitnesses, an entire crowd in a large public location. But people were sure they witnessed different things: a gunshot echoing from here, another shot from there. Complexities of speed, emotion, distance, and memory made it hard to tell which eyewitness testimony to trust. The Zapruder film provided an independent check on all the testimonial noise, allowing investigators to calibrate a single narrative (though, of course, not everyone was convinced).[8]

As important as these acute corrections are, I suggest that the passive regulation role of recordings is still more important. This is the idea: Part of the reason our ordinary testimonial practice allows us to trust one another to be sincere and competent is that we all know that, at any time, we *might* be within the range of an audio or video recorder, or might be testifying about an event that occurred near such a device. This is especially true for public figures, who can expect to acquire the interest of smartphone-equipped observers in every room they enter. Your background awareness that the event you are testifying about *might* have been recorded is a good reason to be as sincere and competent as possible, even if you lack characterological honesty or aptitude.

Of course, that's not always enough to keep people from lying or overstepping their competence. People still do it, as the Fukuda case shows. But I suspect that our tendency to mis-testify has declined significantly since the arrival of widespread recording. We can see this passive regulatory effect operating in some complex examples played out in national political life. In May 2017, US President Donald Trump fired FBI Director James Comey, who alleged this was retaliation for his refusal to help the president suppress an investigation of Russian interference in the 2016 presidential election. Responding to the

6. Woodward and Bernstein (1976, 308–341).

7. Yamaguchi (2018).

8. The Zapruder film also provides instruction on the risks of *over*-reliance on recordings. According to a modern theory by journalist Max Holland, the investigating Warren Commission was misled by assuming that Zapruder's tape covered the entire assassination, when in fact the first gunshot (which missed) was fired before Zapruder turned on his camera. Trying to interpret all three shots within the film's short timeframe led to factual inconsistencies that were quickly seized upon by conspiracy theorists. See Holland (2014).

allegation, Trump tweeted: "James Comey better hope that there are no 'tapes' of our conversations before he starts leaking to the press!"[9]

This was a rather insidious exploitation of the testimony-regulating function of recordings. Trump was implicating that, like Nixon, he keeps a secret recording device in the Oval Office, and that he had tapes showing Comey's testimony to be false. But no such tapes ever emerged. The press, as well as Comey himself, called on Trump to release the implied recordings. A month later, the twittering president followed up: "With all the recently reported electronic surveillance, intercepts, unmasking and illegal leaking of information, I have no idea … [second tweet] whether there are 'tapes' or recordings of my conversations with James Comey, but I did not make, and do not have, any such recordings".[10]

This is a remarkable illustration of the multiple roles of recordings in testimonial practice. When Trump was unable to produce the "tapes", many observers took this as a knock against his own credibility and further support for Comey's testimony. After all, the Director of the FBI was surely aware of the Nixon precedent for a self-bugged Oval Office, and presumably experienced passive regulation of his testimony by the thought that tapes might later emerge. Trump's incompetent insinuation of challenging recordings instead reminded the public that Comey had a strong reason to be telling the truth.

As these examples show, recordings (or the possibility thereof) play several important roles in supporting the rationality of our testimonial practices. It is surprising, then, that epistemologists have not much attended to these roles. I suspect this is because analytic philosophers prefer to focus on necessary or conceptual conditions for knowledge transmission, the sort of factors you could find at work in the epistemic practice of any era or any society, unlike the historically contingent effects of recordings. The first known audio recording was made in France in 1857, while the oldest extant visual recording, *Roundhay Garden Scene*, dates to 1888.[11] A phenomenon that began to exist only 150 years ago is unlikely to attract the attention of a perspective attuned to universal logical constants.

But this is a misleading way of thinking. Though recordings are historically contingent, their role in contemporary social epistemic practice is pervasive. Recording technology is older than any currently existing person. Everyone living today (at least in wealthy countries) has *always* lived with the implicit understanding that recordings provide a check on disputed testimony. Our social epistemic practices have adapted to five generations' awareness of this possibility. As a result, the epistemic role of recording can be diaphanous, since it is hard to see a lens put in place before you were born.

It's worth asking about the scope of these epistemic effects. Is it only public figures who expect to be regularly checked by recordings? Or do ordinary people living day-to-day lives have similar expectations? If it is only public figures, then the epistemic backstop function of recordings will be much narrower than I've suggested, since testimonial norms are determined by the behavior of *all* people, not just the famous.[12] In fact, the answer is a bit more complicated; I think the right distinction is not so much between public figures and ordinary people, but rather between public *events* and private lives. It is still true (for now) that most of us have no expectation that our private comments at home are being recorded (unless you live with TikToking teenagers). But when we are in public urban spaces, we know that we're more likely than not covered by CCTV cameras or traipsing through the background of any number of strangers' selfie-directed phones. So we should indeed expect recordings to regulate our testimony *about public events*, even if we are not public figures ourselves. The Zapruder film, again, did focus on some very public figures, but

9. @realDonaldTrump Twitter account, May 12, 2017. <https://twitter.com/realdonaldtrump/status/863007411132649473>.

10. @realDonaldTrump Twitter account, June 22, 2017. Two tweets: <https://twitter.com/realDonaldTrump/status/877932907137966080>; <https://twitter.com/realDonaldTrump/status/877932956458795008>.

11. NPR (2008); Smith (2016).

12. Thanks to anonymous referees for *Philosophers' Imprint* for encouraging me to clarify this point.

the *testimony* it corrected came from ordinary people. So it may be that the backstop function of recordings applies only to our testimonial knowledge of public events. But that category is already a large and politically important one.

I suggest, then, that the epistemic role of recordings has fallen into a characteristic gap in analytic philosophers' attention. Between the logically universal and the suddenly emergent lies the realm of historically entrenched social practice. As a discipline, analytic philosophy has tended to ignore that realm.[13] But this becomes dangerous when social conditions suddenly change. What was diaphanous while intact may become disorientingly opaque once fractured. That is what I fear may soon happen to the epistemic role of recordings. To see this danger, we need to understand deepfakes.

Deepfakes and Public Deception

Deepfakes came to mainstream public attention in December 2017 through an article in *Motherboard* by tech writer Samantha Cole.[14] As with so much on the internet, pornography was the main vector. The basic idea is this: If you feed hundreds of hours of a celebrity's video appearances into a machine learning algorithm, the algorithm will build an adaptable model of their face, which can then be digitally inserted over the face of a different person in a different recording. So if you have a desire to project your favorite actress into your favorite pornographic scene, the technology can make your (invasive and creepy) dream come true. Connoisseurs of the form gathered on the website Reddit, where they swapped source images and fabricated outputs. One user — whose Reddit handle "deepfakes" came to stand for the phenomenon itself — provided a free software tool ("FakeApp") allowing anyone with a decent home computer to make their own videos. Following media attention, Reddit banned the deepfake community for violating its "involuntary pornography" rules.[15] Of course, the internet is a wide and wild place, and deepfake pornography is far from eradicated. In summer 2019, digital security firm Deeptrace tracked 15,000 deepfakes on the web, nearly double the number from earlier that year. 96% of the videos were porn.[16]

Obviously, the use of deepfakes to present people doing and saying things they never did — especially in pornography — leads to serious ethical worries.[17] But in this paper, I will focus on the risks of *epistemic* mischief. The most obviously dangerous applications are in politics. Several journalistic entities have already published political deepfakes, featuring, e.g., Donald Trump's face superimposed on Angela Merkel's body or Barack Obama appearing to call Trump a "total and complete dipshit". In May 2018, the Flemish Socialist Party posted a deepfake video appearing to show Trump urging Belgium to withdraw from environmental treaties. The Party later claimed that it meant the video to be only a provocation to conversation, not to fool anyone into believing its content. Yet some internet commenters appear to have taken the video as real.[18]

In January 2018, technologist John Wiseman documented what may have been the first attempted deepfake with a political motive. The request was posted to the Reddit deepfakes forum (before that forum was banned) and does not appear to have been fulfilled. This is the text as presented in a screencap Wiseman posted to Twitter:

13. Continental philosophers, of course, are much more comfortable in the land of historically entrenched social practice. A huge literature of obvious relevance to the topic of this paper opens out from Walter Benjamin's *The Work of Art in the Age of Mechanical Reproduction* (1935). I am unaware of anything in continental work touching *directly* on my argumentative claims, though my own training is mostly analytic. In genuine social-epistemic fashion, I would be pleased if a reader with better knowledge of other traditions pointed me to a close parallel.

14. Cole (2017).

15. Robertson (2018).

16. Simonite (2019).

17. I have a separate paper, co-authored with Leah Cohen, discussing the ethics of deepfakes. It is provisionally titled "Deepfakes, Deep Harms", and you can email me (rarini@yorku.ca) to request a draft.

18. See von der Burchard (2018); Silverman (2018).;

> In Russia we have a big problem with gay activist rights. The governor of Chechnya (region in Russia) Ramzan Kadyrov supposedly ordered the execution of 200+ gay people in his region. Vladimir Putin (our president) is a big pussy and doesn't want to do anything with him. Can someone please make a gay video with Ramzan and some other random guy? We can get it viral within his region. Whatsapp channels with 100K+ subscribers will go crazy over it. Everyone will know it's fake, but his reputation will be done. (but not Ramzan with Putin together, because it would just be considered stupid)[19]

This was a request to utilize the technology for a political goal, trying to undermine a politician. In this particular case, the politician is rather loathsome, but of course there is nothing confining the use of deepfakes to laudable political objectives. And the example points to something very important. Notice that the Reddit user says, "everyone will know it's fake, but [Kadyrov's] reputation will be done". The thought is that a deepfake video needn't be positively *believed* by viewers in order to be effective. Merely getting a suggestive video into public distribution may be enough, even if many viewers realize it is fake.[20]

Still, we might ask: Just how convincing *is* this technology? If you look at current examples on the web, especially those by Reddit amateurs, you might be dubious. Many deepfake videos have give-away digital artifacts. The edges of the jaw are blurred unnaturally. The eyes don't blink quite as you'd expect. Though you might not notice these things while watching quickly on a low-resolution platform, you can see them if you pay attention. A critical viewer with access to a large digital platform might spread a debunking just behind the deepfake itself.[21] So perhaps deepfakes won't be very successful political tools.

The problem with this hope is that, of course, new technology eventually improves. To understand what future deepfakes will be like, we shouldn't look at Reddit amateurs. We should look at professional computer scientists. In 2016, a group of researchers in Germany and California presented a technique called "Face2Face", which uses an algorithmic process to impose a famous person's visage over an actor's *in real time*. You can watch a demonstration online: An actor sits in front of a camera, moving his face through various contortions. At the same time, the computer screen displays the face of George W. Bush doing the same things.[22]

You might be tempted to think that fabricated video is of little importance without corresponding audio. Even if a malefactor can make Barack Obama's face do whatever they want, we'll be able to tell that it's not speaking with Obama's voice. (The Obama deepfake mentioned earlier employs noted mimic Jordan Peele to provide the voice. Not everyone will have access to such effective vocal talent!) Unfortunately, there are deepfake technologies for voice as well. Researchers at Princeton and Adobe (the makers of Photoshop) debuted a technique called "VoCo" in 2017. It allows the user to alter the content of a spoken audio recording simply by typing new words in the transcript. The algorithm synthesizes what the speaker's voice would

19. /r/deepfakes post by Reddit user "ethan1el", as image-embedded in tweet by John Wiseman: @lemonodor Twitter account, January 25, 2018. <https://twitter.com/lemonodor/status/956652112678551552>.

20. This point is also shown by another example, which seemed to some observers to be the first case of an actual political deepfake. In late 2018, Gabonese president Ali Bongo appeared in a video that some opponents claimed was machine-generated. Bongo had suffered a stroke months earlier and had not been seen in public since; some opponents thought he had in fact died, with the military using his deepfaked image to maintain their power. It now seems unlikely that the video was deepfaked (Bongo has reappeared), but the accusation contributed to a failed coup attempt. See Breland (2019).

21. In early 2020, Facebook announced that it would "ban" deepfakes and direct resources toward automating their detection. However, the social media firm explicitly exempts "parody or satire" from the ban, a distinction whose ambiguity will surely be exploited by malefactors. See Shead (2020).

22. Niessner et al (2016). YouTube video available at <https://www.youtube.com/watch?v=0hmajJTcpNk>. See also the project page at <https://web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/page.html>. More recently, some of the same researchers proposed a 3-D modeling technique doing the same with fabricated videos of a person's entire body. See Liu et al (2018).

have sounded like with the altered phonemes — as if you can make any voice read from any script you'd like.[23]

You might now say: Okay, this technology is scary, but won't it be cancelled out by equally powerful detection technology? Perhaps we can turn machine learning loose on the internet and have it diagnose characteristic patterns in manipulated recordings. Unfortunately, there are two problems with this thought. The first is technical. Even if counter-deepfake technology temporarily overcomes fakery, this may just be the first move in a machine learning arms race, where fakers continually change strategies to stay a bit ahead of detection. As digital forensics expert Hany Farid explains, "Not only can these automatic tools be used to create compelling fakes, they can be turned against our forensic techniques in the form of generative adversarial networks (GANs) that modify fake content to bypass forensic detection".[24]

And remember that we can see only the publicly available deepfake research done by universities and corporations. One must assume that several extremely well-funded national intelligence services are also hard at work on their own in-house versions of deepfakery. For all we know, their technology is already substantially improved, and they are simply waiting for the right moment to deploy it. (Would we know it if they already *had*?) In any case, the point of this paper is to think ahead, perhaps five years or ten, to the effects of an improved version of the technology — to allow us to begin preparing for the epistemic repercussions before they become dangerously real.

The second problem with technical solutions is more fundamental. It is a problem of social epistemology. Even if we did have reliable deepfake detection technology, past experience with fake news suggests that corrections rarely travel as far as initial fakes and are often not as readily believed.[25] That suggests the *best* case scenario is a kind of epistemic chaos, with corrections chasing vivid fakes around the internet, being variously championed by partisan communities. Which returns us to the central concern of this paper: What will happen to the backstop function of recordings once their unreliability becomes a matter of regular public debate?

23. Jin et al (2017). You can see a demonstration at <https://www.youtube.com/watch?v=RB7upq8nzIU>.

24. Farid (2018, 268).

25. See Vosoughi, Roy, and Aral (2018).

Backstop Crises

The obvious worry about deepfakes is that they will be used to propagate vivid disinformation. As legal scholars Bobby Chesney and Danielle Citron (2019) point out in an influential law review discussion, deepfakes are ripe for election interference, corporate malfeasance, psychological espionage, and personal blackmail.

But I think that the most important risk is not that deepfakes will be *believed*, but instead that increasingly savvy information consumers will come to reflexively distrust *all* recordings. In other words, the backstop baby may get thrown out with the deepfake bathwater.

To see the worry, think ahead to a day when deepfake technology is widely available. The problems will start with events I call *backstop crises* — moments when the corrective and regulative functions of recordings are made salient, but then quickly undercut by the spectre of deepfakery. Imagine, for example, that Richard Nixon had said: "Look, that wasn't me on the smoking gun tape. They used that VoCo technology to make it sound like me ordering CIA interference. But it wasn't!"

This would not have been a very plausible claim in 1974. But now imagine that late in the 2020 US presidential campaign, an audio recording emerges which certainly sounds like Donald Trump colluding with Russian intelligence operatives. Suppose Trump insists that it wasn't him, that he has been deepfaked into an entirely fabricated conversation.

I don't know how many people would believe Trump, or how many would believe the recording. I suspect views would break along predictably partisan lines. But I can say this: Though not myself a fan of Donald Trump, I would have serious doubts about the veracity of this tape. Even with strong prior reasons to suspect that Trump *would* collude with Russian intelligence, knowing about deepfake technology

provides reason to doubt this piece of supposed evidence. After all, many political actors, domestic and foreign, would have strong motives to create such a tape, regardless of its underlying truth.

Notice, then, that the point of this example doesn't require that everyone *believe* a deepfake. The point is the public controversy itself. A backstop crisis implies the sudden public realization that there is no longer any such thing as a "smoking gun" tape. Even if we end up disbelieving our first famous deepfakes, we will all come away with a growing sense of displaced epistemic reality. Once everyone has seen that a supposedly authoritative recording can be reasonably challenged, we will all start to wonder about the next recording, and the next, and the next…

Here is another scenario to make the idea more vivid. Imagine a politician video-calling in to a TV news show and, under pressure from interviewers, saying something foolish or offensive. Instantly, the pundits decree this a deadly gaffe, the beginning of the end of the politician's career. But then, the politician's staffers claim that this wasn't the politician at all. They assert that their politician was nowhere near a camera at the time. Rather, it must have been an actor, using real-time deepfake technology to steal the image and voice of the poor misrepresented politician![26]

Imagine the fallout from such a spectacle. Hours and hours of cable news fulmination over the plausibility of a guest's appearance being "hacked". Linguists and computer scientists summoned to provide rival millisecond breakdowns of the recording: "See the lip move like that, there? That can't be real!" Or: "Look, that partial blink is very hard to fake with an algorithm. This is the real thing!" Imagine this going on for days or weeks with no clear resolution. In the end, presumably, public opinion will break along partisan lines. If you like the politician, you'll believe the denial. If you don't, you'll allege that they are crying deepfake wolf and hate them all the more for it.[27]

Regardless of the outcome — whether we believe the politician or the video — this process *itself* is the backstop crisis. We will all confront a suddenly plausible skepticism about the knowledge-bearing potential of video and audio, after lifetimes of relying on them as solid testimonial anchors. This has the makings of an epistemic crisis on the order of beginning to suspect that a hallucinogen has been pumped into the city's water supply.

As backstop crises follow one on another, video and audio recordings may lose their status as acute correctors of the testimonial record. Earlier, I stressed that the acute corrective role of recordings is less important than their passive regulatory role. We can see that most clearly now, as we try to imagine the long-term effects of highly public backstop crises. Once we know that recordings can be deepfaked — and that *veridical* recordings can be convincingly dismissed — our motivation to be responsible testifiers may slowly erode. We no longer need worry that someone might have recorded the public events about which we are testifying. If the recording goes against us, we can always cry, "Deepfake!" And we might even be right.

This is the gravest danger of deepfakes: not that they will trick us into believing false content, but that they will gradually eliminate the epistemic credentials of *all* recordings, to an extent that video and audio no longer serve their passive regulative function in testimonial practice. As that happens, the reasonableness of expecting testifiers to be (usually) sincere and competent will begin to diminish. Within a few years, we may have little reason to trust the testimony of strangers, as the norms securing their anticipated cooperation come gradually undone. Backstop crises triggered by deepfakes may be only the harbingers of a slow-boiling but deeply consequential epistemic maelstrom.

26. This scenario roughly follows what Chesney and Citron call the "liar's dividend" implication of deepfakes: "As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authoritative video and audio as deep fakes [sic]" (Chesney and Citron 2019, 1785).

27. I've previously written about the similarly troubling role of partisanship-in-testimony reception regarding fake news; see Rini (2017).

The Epistemology of Recordings: Perceptual versus Testimonial Knowledge

My claims about the consequences of backstop crises are speculative and somewhat vague. In this section, I'll aim to be a bit more philosophically precise by grounding deepfakes' challenge to the backstop in terms borrowed from philosophical work. Surprisingly, there isn't much philosophical writing on the epistemology of *recordings*, but there is a substantial literature regarding the epistemology of still photographs. When a photograph provides epistemic access to its subject, what sort of epistemic role is it playing?

Much of this literature is framed in response to Kendall Walton's *transparency* thesis, which holds that photographs enable direct, literal perception of their objects.[28] Suppose, for instance, that I want to learn about Queen Victoria's dog Looty, the supposed ancestor of all Pekingese in Britain and North America. (Looty was so named because she — along with many equally priceless objects — had been stolen from the Chinese Imperial Summer Palace by rampaging British troops during the Second Opium War.) First, I look at an oil painting of Looty, executed by Friedrich Wilhelm Keyl in 1861. Then, I look at a photograph of Looty, grumpy on an ornate stiff-backed chair, taken by William Bambridge in 1865.[29] According to Walton's view, in the latter case I am literally *seeing* Looty herself, across time via this photograph, just as surely as if I were seeing her across space via a telescope. By contrast, when I look at the painting, I am not literally seeing Looty but only a pictorial representation of her, mediated through Keyl's perceptions and artistic expression.

Walton's proposal has been extremely controversial, chiefly due to the oddity of implying that right now I am literally seeing a dog who died 140 years ago.[30] But the same point comes through in other ways. Robert Hopkins shifts from the idea that I literally see Looty to the slightly more arcane suggestion that a photograph can place me in a necessarily veridical *seeing-in* relation to Looty. Assuming the photograph was made and reproduced according to standard photographic norms, then my experience of seeing Looty in the picture is roughly as reliable as seeing Looty with my own eyes.[31] I'm not literally seeing Looty, but (assuming there were no shenanigans in the causal pathway from Bambridge's shutter to my screen) I gain the same epistemic access to Looty as if I were.

Dan Cavedon-Taylor makes an important related point, claiming the distinct epistemic advantage of photographs over hand-made pictures is that the former generate *perceptual knowledge*, while the latter can only support *testimonial knowledge*. When I look at Keyl's painting of Looty, I am relying upon the painter's skill and honesty in pictorial representation just as much as I would rely upon his competence and sincerity in a verbal testimonial report. By contrast, Cavedon-Taylor argues, the photo is not mere testimonial evidence; it belongs in the same category as ordinary perceptual experience, with the same epistemic immediacy. He says: "When we see a photograph that depicts x as F, say, our default doxastic response is to believe that x is F — and to only withhold assent if we possess reasons *against* thinking the photograph creditworthy".[32]

Another example might best illustrate this contrast. Leo von Klenze's 1857 painting *The Temple of Concordia at Agrigento* depicts the

28. Walton (1984). As a referee for *Philosophers' Imprint* reminds me, Walton himself was not trying to make an epistemic point. But many of the reactions to his article are framed in epistemic terms.

29. Both the painting and photograph are held by the Royal Collection. You can see them online at, respectively, <https://www.rct.uk/collection/406974/looty> and <https://www.rct.uk/collection/2105644/looty-the-pekingese>. To learn more about Looty herself, see Haven (2010).

30. For discussion, see Cohen and Meskin (2004).

31. Hopkins (2012). This is a slightly compressed version of a nuanced view. One worry about Hopkins' view is that he appears to concede that photographs are only reliable when viewers see in the photo "only facts consistent with *perfect* general knowledge of how things are" (719). Given the difficulty of attaining second-order knowledge that one's judgements cohere with perfect general knowledge, I worry that Hopkins' proposal encourages practical skepticism regarding *all* real-world photographic experiences. But I'll leave that thought to the side here.

32. Cavedon-Taylor (2013, 294).

eponymous Sicilian ruin. When I saw this painting hanging in the Alte Nationalgalerie in Berlin, I was confused. I've been to Agrigento, and something about the painting seemed wrong to me, though I couldn't articulate just what. Later, I looked up a photo of the temple and understood: The viewing angle Klenze depicts is not actually possible; the temple and the city simply are not spatially arranged in that way. Klenze seems to have taken artistic license with perspective in order to achieve a more pleasing composition.[33]

This example shows the different epistemic powers of paintings and photographs. My pictorial experience of the temple in Klenze's painting was at odds (inchoately) with my perceptual memory. I was unable to resolve this tension until I examined a photo of the same ruins, which quickly decided me in favor of my memory and against the painting. This is because, just as Cavedon-Taylor suggests, paintings provide only testimonial evidence, which is typically trumped by perceptual evidence of the sort photographs provide. Because I know that Klenze's aesthetic sensibilities play a crucial role in his depiction of the temple — and this is not true to the same extent in photographs — I treat the painting with the same merely provisional authority as testimony. As Cavedon-Taylor puts it: "The conditions under which it is rational to believe the content of another's testimony are stricter than those under which it is rational to believe the content of another's photograph" (2013, 288–289).

Plausibly, all of the above is just as true for audio and video recordings as it is for still photographs. Recordings provide us with a form of perceptual evidence, which enjoys a stronger presumptive authority than testimonial evidence. And this is exactly why recordings are so well suited to provide the epistemic backstop. Their stronger evidential weight allows them to provide acute correction of deviant testimonial practices and passive regulation of trustworthy testimonial norms.

33. You can see the painting here: <https://www.smb.museum/en/exhibitions/detail/concordia-kunst-und-wissenschaft-in-eintracht.html>. Gallery text (on display in May 2018) confirms that Klenze manipulated the temple's relative positioning for aesthetic reasons.

We can now recast the points of the preceding sections more precisely. The danger posed by deepfakes is that successive backstop crises will gradually transform our attitudes toward the epistemic status of recordings. In effect, recordings will be demoted from sources of perceptual evidence to sources of mere testimonial evidence. And if they are simply just *another* source of testimony, then they cannot be relied upon to *correct or regulate* testimonial practice. Recordings could become no more authoritative than paintings, only as reliable as the reputation of their creator. A recording is then just another node in the testimonial web, equally subject to partisan repudiation or inflation. And there is less reason to be responsible in one's testimonial practices if recordings offer no more than a source of conflicting testimony.

Cavedon-Taylor and Hopkins make similar points about the effects of Photoshop on the epistemology of still photographs.[34] Indeed, this worry was raised by Barbara Savedoff as early as the year 2000, not long after consumer digital photography first appeared:

> If we reach the point where photographs are as commonly digitized and altered as not, our faith in the credibility of photography will inevitably, if slowly and painfully weaken, and one of the major differences in our conceptions of paintings and photographs could all but disappear.[35]

This point raises a natural objection to my worries about deepfakes. If recordings have a perceptual-type epistemic role similar to still photographs, and if the reliability of photographs has been under assault from Photoshop for two decades, why haven't we *already* seen the sort of epistemic catastrophe I described in the last section? It seems like

34. Hopkins (2012, 723). Cavedon-Taylor actually goes further: If our awareness of digital manipulation thoroughly erodes our trust in photos, then we might conclude that photographically based belief, in order to be rationally grounded, must be formed on the basis of positive reasons for thinking the photograph has been reliably produced. This will have the effect of rendering photographically based knowledge exclusively inferential in nature, and not testimonial. (Cavedon-Taylor 2013, 296)

35. Savedoff (2000, 202).

we've adapted well enough to widespread knowledge that photos can be manipulated — so why won't we adapt just as well to deepfakes?

The Distinctive Threat of Deepfakes

In fact, we've known that photographs are vulnerable to manipulation since long before Photoshop. In 1920, Arthur Conan Doyle published an article vouching for photographs of fairies taken by two young girls in West Yorkshire. Of course, the creator of Sherlock Holmes would not have done so without ruling out alternative explanations. First, he sent the negatives for inspection by Kodak technicians, who refused to provide a certificate of authenticity, though they agreed the materials "show no signs of being faked".[36] Conan Doyle then considered that the girls might have simply drawn the fairies on carboard and posed them very carefully. But he concluded that "the girl's own frank nature is, I understand, a sufficient guarantee for those who know her". To be safe, Conan Doyle's investigator

> tested her powers of drawing, and found that, while she could do landscapes cleverly, the fairy figures which she had attempted, in imitation of those she had seen, were entirely uninspired, and bore no possible resemblance to those in the photograph.[37]

The fairy folly was only the start of a century-long trend. Stalin employed entire teams to bring scalpel and airbrush down upon the photographic records of disfavored comrades.[38] In one extraordinary case, the 1926 original poses Stalin alongside three others. A succession of retouched versions shed comrades as decisively as the Politburo. The last image — Stalin all alone — has been so tortured that it is plainly more painting than photograph.[39]

So we have known for decades that photos can be manipulated, yet our testimonial norms seem largely intact. Why, then, should we worry much about deepfakes? In answering this objection, it's important to remember that our question is *not* "Why are deepfakes more likely to mislead than earlier photographic manipulation?" As I've stressed, my main worry is *not* that deepfakes succeed in tricking anyone. Rather, the worry is that backstop crises triggered by contested deepfakes will lead to erosion of the reliability that recordings provide to our testimonial practices. There are at least four reasons to worry that deepfakes are distinctively threatening here.

First, there is a psychological difference between still photographs and audio-video recordings. Recordings are typically more gripping, perhaps because they extend over time and support articulated narratives. People happily spend money to gather in theaters and watch movies, but they don't do so to view still photographs anywhere near as frequently. Perhaps recordings simply *feel* more like reality than photographs. We rarely experience the world in static images, but a well-executed recording mimics our everyday perceptual activities. This suggests we may be more psychologically invested in the reality of recordings than still photographs. When backstop crises force us to confront the growing unreliability of recordings, the skeptical consequences may be amplified by their psychological power.

Second, deepfakes, unlike familiar Photoshop images, are generated by a machine learning process that permits *efficient response* to epistemic challenges. Currently, if we doubt someone's testimony about an alleged public event, we can challenge them to produce a recording. And if we doubt the first recording they provide, their ability to produce a second corroborating recording would be strong evidence in their favor. For example, in 2016, Trump campaign manager Corey

36. Smith (1997, 384).

37. Conan Doyle (1920). Sixty years later, the women admitted this was exactly what they had done — though they still insisted there really had been fairies around.

38. See King (1997).

39. These images are in the public domain and may be viewed (as of March 2019) at <https://commons.wikimedia.org/wiki/File:Soviet_censorship_with_Stalin2.jpg>.

Lewandowski was accused of grabbing the arm of a reporter and pulling her away from questioning his candidate. The campaign quickly dismissed the charge, claiming that "not a single camera or reporter of more than 100 in attendance captured the alleged incident".[40] Reporters took up the challenge, locating two different video recordings, from alternate angles, showing Lewandowski doing just what was alleged.

Part of why the campaign's challenge was compelling — and why successfully meeting the challenge was even more so — is that we currently assume that no one will be able to swiftly produce fake video recordings tailored to novel (purported) facts in the way that we know Photoshop is relatively cheap and easy to use. Though Hollywood can produce extremely compelling videos of unreal things, we know that this takes enormous resources of time, money, and human talent. That's why it is implausible to explain away effectively-met challenges to recordings; currently, the idea of overnight faked recordings sounds more like a conspiracy theory than a plausible debunking. But this thought will become less powerful after we have witnessed a few backstop crises. Once deepfake technology is widely known, it will be possible for public figures in situations like Lewandowski's to simply insist that any purported recording evidence is just more fakery — and who will be able to say they are wrong?

Third, the efficiency of machine learning will enable mass production of deepfakes at a similar rate to textual fake news. This will allow epistemic malefactors to "spam" the epistemic environment with conflicting false videos of prominent events, in much the way they currently do with fake news. This is an established tactic of authoritarians seeking to disrupt public information channels. Neutralizing effective public epistemology doesn't require tricking people into believe any particular falsehood. It just requires convincing them not to trust *any* information source. Once information channels are thoroughly saturated with obvious garbage, people fall into a sort of epistemic learned helplessness, where they give up trying to critically assess information and simply believe whatever conforms to their worldview.[41] There is every reason to expect such malefactors will use deepfakes in the same way.

Once that happens, we will see the vulnerability of recordings demonstrated frequently and shockingly, in a way that hasn't (yet) happened to still photographs. Of course, it's possible that photos will *also* soon be undermined by machine learning mass-production. My point is only that the machine learning technology behind deepfakes leads to a predictable threat distinct from earlier photo manipulation techniques.

Finally, and perhaps most importantly, the difference between Photoshop and deepfakes may be simply this: Audio and video recordings already function as backstops *for* still photographs. A still photo of a single moment can mislead about causal relations among objects in a way that is often quickly resolved by a corresponding video. When a still image appears to show something dubitable, critics typically demand to see a video of the same incident.

For instance, in late 2019, a strange photo circulated on Twitter appearing to show former Vice President Joe Biden sucking on his wife's finger on the stage at a campaign rally. Journalist Luke Darby dug further, noting

> it definitely looks like the former vice president — at a public event and in front of human people — is nibbling his wife's finger like a goldfish at feeding time. But the Internet has produced doctored photos before, and maybe a video of the event will clear things up?[42]

(As it turns out, the video shows something only slightly less strange.)

40. Flores (2016).

41. See Tufekci (2018); Lynch (2016). For a worked example of how the Putin regime has used this technique in the past, see Snyder (2018, chapter 5). A related technique, which Margaret Roberts calls "flooding", is sometimes used by Chinese internet censors; see Roberts (2018).

42. Darby (2019).

As this incident suggests, the epistemic authority of still photographs *has* already been eroded by Photoshop, but our evidential practices around photographs are, like verbal testimony, apparently still undergirded by the epistemic backstop of recordings. Perhaps the evidential status of still photos has *already* shifted from vehicles of perceptual knowledge to mere testimonial knowledge. But so long as recordings still serve their backstop functions, they can acutely correct and passively regulate photographic evidence in the same way as testimonial practice.

However, if this is true, then we already implicitly place more epistemic weight on recordings than we appreciate. The epistemic backstop regulates not only our testimonial practices, but also our use of other sorts of documentary evidence, like still photographs. This suggest that the threat deepfakes pose to the backstop may be even more consequential, exposing neglected decay in our epistemic relations to photos along with undiagnosed vulnerability in testimony.

Toward a Worrisome Epistemic Future

Where does this leave us? If my worries are right, then in the near future, we may face a sudden collapse of the backstop to our testimonial practice, as we can no longer trust that recordings provide authoritative correction. Following highly public backstop crises, we may find that our adherence to testimonial norms of competence and sincerity becomes progressively less reliable as people realize there is rarely an independent check on their testimony. What happens then?

One answer is surprisingly positive. You might think that we never *should* have put so much trust in recordings. After all, recordings have never been strictly invulnerable from manipulation; given extensive time and resources, skilled fakers could produce convincing videos long before the rise of machine learning. So it might be a good thing if deepfakes shake us free from our unreasonable reliance on recordings.[43]

43. I thank audience members at the 2019 Central APA in Denver for convincing me of the merit of this objection. Similar ideas appear in Silbey and Hartzog (2019).

I think this is the wrong attitude to take. It assumes an implausibly Spartan epistemology, one where we restrict our evidence to near-infallible technologies. I think this is an unrealistic epistemic standard that would deprive us of valuable sources of information quite often. To put it another way, I think that our epistemic reliance on recordings *has been* reasonable, even if it will soon cease being so.

But you don't have to agree with me about this to see the danger of deepfakes. Even if abandoning our reliance on recordings might be epistemically admirable, there will be serious *transition costs*, and these ought to worry everyone. For better or worse, we *have* developed a web of epistemic norms assuming reliance upon recordings. In the developed world, there is no one living today who remembers an epistemic environment preceding that reliance. Video and audio recordings, in existence longer than any of us, have always structured our lives. To really appreciate what their discredit would mean for our testimonial practice, we must look back to a time when people relied solely on the testimony of eyewitnesses and newspapers for knowledge of public events. If I am right about the effects of deepfakes, we may be very suddenly plunged into just such an environment, without preparation or training. When we fall into dispute about public events, there will be no recorded backstop to resolve things.

There were, of course, testimonial norms before the 19th-century creation of recording technology. Those norms might be adequate if we could return to them. But how do we get back there? None of us have the benefit of life-long training in an unregulated testimonial environment. We may become dangerously credulous, or perhaps reactively paranoid. I cannot predict the future, but I am confident that we will not quickly rediscover 19th-century testimonial norms without a lot of trouble along the way.[44]

44. Among other problems, we should worry about losing the epistemically equalizing benefits of recordings. 19th-century epistemology left members of marginalized groups especially vulnerable to testimonial injustice (Fricker 2007). Reliable video recordings have sometimes provided marginalized

More fundamentally, I worry that 19th-century testimonial norms simply won't *work* in a modern world of instant global communication. We may have been fortunate to enjoy the testimony-regulating effects of the epistemic backstop during the last 150 years of rapid technological and social change. But we may now be forced to navigate future changes without that protection. In a social epistemic environment already plagued by fake news malefactors and authoritarian deceivers, I am less than fully confident we will weather the transition with our social and political systems intact.[45]

References

Annette Baier (1986). 'Trust and Antitrust'. *Ethics* 96(2): 231–260.

Walter Benjamin (1935[1969]). 'The Work of Art in the Age of Mechanical Reproduction'. In *Illuminations* (ed. H. Arendt). New York: Schocken. 217–251.

Ali Breland (2019). 'The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink". *Mother Jones* March 15, 2019. <https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/>.

Dan Cavedon-Taylor (2013). 'Photographically Based Knowledge'. *Episteme* 10(3): 283–297.

Robert Chesney and Danielle Keats Citron (2019). 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security'. *California Law Review* 107: 175. <https://ssrn.com/abstract=3213954>.

groups with the means to confront disbelieving publics, as with recent smartphone videos of police brutality against Black people in the US. (I owe this important point to Daniel Saunders.)

45. I would like to thank Daniel Saunders for invaluable research assistance on this project. I also owe thanks to Leah Cohen for assistance on a related earlier project — and for first bringing deepfakes to my attention. Thanks also to audiences at the 2018 "Fake Knowledge" conference at the University of Cologne; the 2019 Central APA in Denver; the 2019 "Ignorance in the Age of Information" conference at Scripps College; the University of St. Andrews; and the editors and referees for *Philosophers' Imprint*. This research was supported by the VISTA project at York University.

C. A. J. Coady (1992). *Testimony: A Philosophical Study*. Oxford: Oxford University Press.

Jonathan Cohen and Aaron Meskin (2004). 'On the Epistemic Value of Photographs'. *Journal of Aesthetics and Art Criticism* 62(2): 197–210.

Samantha Cole (2017). 'AI-Assisted Fake Porn is Here and We're All Fucked'. *Motherboard* December 11, 2017. <https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn>.

Arthur Conan Doyle (1920). 'Fairies Photographed'. *Arthur Conan Doyle Encyclopedia*. <https://www.arthur-conan-doyle.com/index.php?title=Fairies_Photographed>.

Luke Darby (2019). 'What Is Biden Doing to His Wife's Hand?' GQ December 1, 2019. <https://www.gq.com/story/what-is-biden-doing>.

Hany Farid (2018). 'Digital Forensics in a Post-Truth Age'. *Forensic Science International* 289: 268–269.

Reena Flores (2016). 'Donald Trump's Campaign Denies Getting Rough with Reporter'. CBS *News* March 10, 2016. <https://www.cbsnews.com/news/trump-campaign-responds-to-charges-of-getting-physical-with-reporter/>.

Miranda Fricker (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Sanford Goldberg (2010). *Relying on Others: An Essay in Epistemology*. Oxford: Oxford University Press.

Cynthia Haven (2010). 'Stanford Historian Tells Why the West Rules — For Now'. *Stanford News* September 14, 2010. https://news.stanford.edu/news/2010/september/morris-west-rules-091410.html.

Max Holland (2014). 'The Truth Behind JFK's Assassination'. *Newsweek* November 20, 2014. <https://www.newsweek.com/2014/11/28/truth-behind-jfks-assassination-285653.html>.

Robert Hopkins (2012). 'Factive Pictorial Experience: What's Special About Photographs?'. *Noûs* 46(4): 709–731.

Zeyu Jin, Gautham J. Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein (2017). 'VoCo: Text-Based Insertion and Replacement in Audio Narration'. ACM *Transactions on Graphics* 36(4): 96.

Karen Jones (1996). 'Trust as an Affective Attitude'. *Ethics* 107(1): 4–25.

David King (1997). *The Commissar Vanishes: The Falsification of Photographs and Art in Stalin's Russia*. New York: Henry Holt and Company.

Jennifer Lackey (2006). 'It Takes Two to Tango: Beyond Reductionism and Non-Reductionism in the Epistemology of Testimony'. In *The Epistemology of Testimony* (eds. J. Lackey and E. Sosa). Oxford: Oxford University Press. 160–182.

--- (2007). 'Norms of Assertion'. *Noûs* 41(4): 594–626.

--- (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, Christian Theobalt (2018). 'Neural Animation and Reenactment of Human Actor Videos'. <*arXiv*:1809.02658v1 https://arxiv.org/pdf/1809.03658.pdf>.

Michael P. Lynch (2016). 'Fake News and the Internet Shell Game'. *New York Times* November 16, 2016. <https://www.nytimes.com/2016/11/28/opinion/fake-news-and-the-internet-shell-game.html>.

David Mack (2018). 'This PSA About Fake News from Barack Obama Is Not What It Appears'. *Buzzfeed* April 17, 2018. <https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed#.noWQR06Em>.

Stephen Maher (2018). 'Fake Video is a Big Problem. And It's Only Going to Get Worse'. *MacLean's* December 28, 2018. <https://www.macleans.ca/opinion/fake-video-is-a-big-problem-in-2019-it-gets-worse/>.

Matthias Niessner, Christian Theobalt, Marc Stamminger, Michael Zollhöfer, and Justus Thies (2016). 'Face2Face: Real-time Face Capture and Reenactment of RGB Videos'. <http://www.graphics.stanford.edu/~niessner/papers/2016/1facetoface/thies2016face.pdf>.

NPR (2008). '1860 "Phonautograph" Is Earliest Known Recording'. *Talk of the Nation* April 4, 2008. <https://www.npr.org/templates/story/story.php?storyId=89380697>.

Riana Pfefferkon (2019). 'Too Good to be True? "Deepfakes" Pose a New Challenge for Trial Courts'. NWL*awyer* September 2019. <https://law.stanford.edu/publications/too-good-to-be-true-deepfakes-pose-a-new-challenge-for-trial-courts/>.

Ritu Prasad and Max Matza (2019). 'Reaction to Cohen testimony'. BBC *News* February 27, 2019. <https://www.bbc.com/news/live/world-us-canada-47390920>.

Regina Rini (2017). 'Fake News and Partisan Epistemology'. *Kennedy Institute of Ethics Journal* 27(S2): 43–64.

Margaret Roberts (2018). *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton: Princeton University Press.

Adi Robertson (2018). 'Reddit Bans "Deepfakes" AI Porn Communities'. *The Verge* February 7, 2018. <https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban>.

Barbara Savedoff (2000). *Transforming images: How Photography Complicates the Picture*. Ithaca, NY: Cornell University Press.

Sam Shead (2020). 'Facebook to Ban "Deepfakes"'. BBC *News* January 7, 2020. <https://www.bbc.com/news/technology-51018758>.

Jessica Silbey and Woodrow Hartzog (2019). 'The Upside of Deep Fakes'. *Maryland Law Review* 78(4): 960–966.

Craig Silverman (2018). 'How to Spot a Deepfake Like the Barack Obama–Jordan Peele Video'. *Buzzfeed* April 17, 2018. <https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed>.

Tom Simonite (2019). 'Most Deepfakes Are Porn, and They're Multiplying Fast'. *Wired* October 7, 2019. <https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/>.

Ian Smith (2016). '"Roundhay Garden Scene" Recorded in 1888, is Believed to be the Oldest Surviving Film in Existence'. *The Vintage News* January 10, 2016. <https://www.thevintagenews.com/2016/01/10/roundhay-garden-scene-is-believed-to-be-the-oldest-known-video-footage/>.

Paul Smith (1997). 'The Cottingley Fairies: The End of a Legend'. In *The Good People: New Fairylore Essays* (ed. P. Narvez). Lexington: University Press of Kentucky. 371–405.

Timothy Snyder (2018). *The Road to Unfreedom: Russia, Europe, and America*. New York: Penguin.

Zeynep Tufekci (2018). 'It's the (Democracy-Poisoning) Golden Age of Free Speech'. *Wired* January 16, 2018. <https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/>.

Hans von der Burchard (2018). 'Belgian Socialist Party Circulates "Deep Fake" Donald Trump Video'. *Politico* May 21, 2018. <https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/>.

Soroush Vosoughi, Deb Roy, and Sinan Aral (2018). 'The Spread of True and False News Online'. *Science* 359(6380): 1146–1151.

Kendall L. Walton (1984). 'Transparent Pictures: On the Nature of Photographic Realism'. *Noûs* 18(1): 67–72.

Timothy Williamson (1996). 'Knowing and Asserting'. *Philosophical Review* 105(4): 489–523.

Bob Woodward and Carl Bernstein (1976). *The Final Days*. New York: Simon & Schuster.

Mari Yamaguchi (2018). 'Japan Finance Official Denies Sexual Misconduct Allegation'. *The Seattle Times* April 18, 2018. <https://www.seattletimes.com/nation-world/japan-finance-official-denies-sexual-misconduct-allegation/>.