Matthew Carlson*

Skepticism and the Digital Information Environment

https://doi.org/10.1515/sats-2021-0008
Published online July 1, 2021

Abstract: Deepfakes are audio, video, or still-image digital artifacts created by the use of artificial intelligence technology, as opposed to traditional means of recording. Because deepfakes can look and sound much like genuine digital recordings, they have entered the popular imagination as sources of serious epistemic problems for us, as we attempt to navigate the increasingly treacherous digital information environment of the internet. In this paper, I attempt to clarify what epistemic problems deepfakes pose and why they pose these problems, by drawing parallels between recordings and our own senses as sources of evidence. I show that deepfakes threaten to undermine the status of digital recordings as evidence. The existence of deepfakes thus encourages a kind of skepticism about digital recordings that bears important similarities to classic philosophical skepticism concerning the senses. However, the skepticism concerning digital recordings that deepfakes motivate is also importantly different from classical skepticism concerning the senses, and I argue that these differences illuminate some possible strategies for solving the epistemic problems posed by deepfakes.

Keywords: deepfakes, evidence, recordings, skepticism, underdetermination

Deepfakes are audio, video, or still-image digital artifacts created by the use of artificial intelligence technology.[1] Deepfakes have entered the popular imagination as sources of serious epistemic problems for us, as we attempt to navigate the increasingly treacherous digital information environment of the internet

[1] Deepfakes are *deep* because the artificial intelligence technology used to create them employs *deep-learning* networks. I will clarify the sense in which they are fake below. One might make the case that *texts*, such as those generated by GPT-3, should also be classified as deepfakes. The idea would be that these texts purport to be written by actual agents (i.e. humans) but are in fact generated by deep-learning networks, in much the same way that deepfakes are. Nevertheless, the word *deepfake* as it is normally used does not refer to texts, and accordingly my main focus here will be on audio, video, and still-image digital artifacts.

***Corresponding author: Matthew Carlson**, Philosophy, Wabash College, 301 W Wabash, 47933 Crawfordsville, IN, USA, E-mail: carlsonm@wabash.edu. https://orcid.org/0000-0002-2193-9952

(Toews 2020). In this paper, I attempt to clarify precisely what epistemic problems deepfakes pose, and why they pose these problems. By drawing parallels between recordings and our senses as sources of evidence, I argue that the existence of deepfakes threatens to undermine the status of digital recordings as reliable sources of evidence. To explain why this is, I draw parallels between deepfakes and skeptical hypotheses that ground classic skeptical arguments concerning the senses. But despite these parallels, I close by highlighting some important differences between skepticism about the senses and skepticism about digital recordings. I believe that these differences are important because they suggest some possible strategies for combating the epistemic threat posed by deepfakes.

1 The Epistemic Problems Posed by Deepfakes

In our normal epistemic practices, we tend to treat recordings as highly reliable sources of evidence concerning their subjects.[2] We are convinced that Nixon was involved in the Watergate conspiracy because there are recordings of him saying so. We are convinced that George Floyd was killed by Derek Chauvin because there are video recordings of this event taking place. In fact, we tend to treat recordings as on a par with our senses themselves. That is, it would not be a stretch to say that we *heard* Nixon saying that he was involved in the Watergate break-in or that we *saw* George Floyd being killed by Derek Chauvin. In Derek Chauvin's trial, prosecution attorney Jerry Blackwell enjoined jurors to employ their "common sense" in assessing the evidence provided by recordings of the event, which were produced by bystanders with their phones. As Blackwell put it, "[y]ou can believe your eyes, ladies and gentlemen. It was what you thought it was. It was what you saw. It was homicide" (Prater and Hartmann 2021). The point of this anecdote is just to draw your attention to the fact that we do tend to treat recordings as reliable sources of evidence, on a par with our own senses.

But why do we treat recordings as being so similar to our own senses? According to Walton's (1984) classic treatment, the answer to this question is that photographs are *transparent* in the sense that we see *through* them and literally see their subjects. As Jerry Blackwell put, in seeing a recording of George Floyd's death, jurors *saw* that event. More recently, Cavedon-Taylor (2013) argues that recordings are sources of perceptual knowledge because, like our senses, they encourage a "default doxastic response" (p. 294). When we see (or hear) an event, regardless of whether it is in a recording or in our direct experience, we tend to

[2] I use the term *recordings* as a generic term to refer to audio, video, and still-image captures made via the use of instruments.

believe that the event occurred. As Cavedon-Taylor points out, this tendency is particularly evident when we watch recordings that are being broadcast live, such as televised sports. When we watch such live events, our default response is simply to believe what we see. When, on the representation on my TV screen, I see Messi score a goal, I simply come to believe that Messi scored a goal. The immediacy of this process is no different than it would have been had I seen the event in person. Whether we are watching on TV or in person, "seeing is believing."

So far, this suggests that we treat recordings as being at least as reliable as our own senses. But in fact, we often treat them as even more reliable. Consider the use of instant replays to help referees determine what actually happened. Did Messi really score a goal, or was he offsides? Because of the speed of the game, this can be very difficult to assess in real time. But we can consult recordings of Messi's run and use these to determine whether he was, in fact, offsides. In cases where such recordings clearly conflict with the judgment of the on-field referees, we typically take the recordings, and not the referees' senses, to be dispositive evidence concerning what actually happened.

As I have explained, in our ordinary epistemic practices, we treat recordings as reliable sources of evidence, on a par with our own senses. But is it reasonable for us to do so? Here is a tempting answer to this question: It is reasonable for us to treat recordings as reliable sources of evidence concerning their subjects because of a certain *causal* relationship that recordings bear to their subjects. The idea that recordings have a *causal* connection to what they represent—like our senses do—is as old as recordings themselves. Concerning the photographic process that he invented in the 1830s, Daguerre 1838/1980 writes that "the daguerreotype is not merely an instrument which serves to draw Nature; on the contrary it is a chemical and physical process which gives her the power to *reproduce* herself" (p. 11, my emphasis). According to Daguerre, daguerreotype images are produced by natural processes, as opposed to processes like painting, which require the intervention of an agent. And this idea, that photographs have a causal connection with their subjects due to the chemical and optical processes involved in their creation, persisted well into the 20th century. In *Camera Lucida*, for example, Barthes (1981) writes: "The photograph is literally an emanation of the referent. From a real body, which was there, proceed radiations which ultimately touch me … like the delayed rays of a star" (p. 80).

Of course, while photographs bear causal relationships to their subjects, this does not imply that they are accurate representations of their subjects in every respect. As Benovsky (2016) puts it, "photographs do not typically depict reality *as it is* (they only depict things from one side, they can involve distortions, blurred background, etc.), but they always at least partly depict something *that was there*, even if they perhaps misdepict [sic] it …" (pp. 77–78, emphasis in original).[3] That is, according to Benovsky, photographs are "always pictures *of something existing*" (p. 77, emphasis in original). Phillips' (2009) account of photography and causation nicely explains this relationship between photographs and their subjects. On her account, a photograph is "a visual image whose relevant causal history necessarily involves a photographic event" (p. 11). And, as she explains:

> A photographic event occurs when a photosensitive surface is exposed to the light and a recording of the light image takes place. The photographic event is *the recording of the light image*. It is important to recognize that in this description 'a recording' is not the same as 'a record'. The record (an object such as a negative) is the result of the recording process (p. 12, emphasis in original).

A photograph is thus a "record" of a photographic event, and by viewing this record, Phillips explains, we can learn about the photographic event. But, as Phillips emphasizes, viewing a photograph does not necessarily give us accurate information concerning the *appearance* of the photographic event in its causal history. Due to the way in which a photograph was captured and processed, it may appear quite different from its photographic event. Still, both Benovsky and Phillips hold that photographs are special because they bear a causal connection to their subjects. According to Benovsky, this explains why photographs are transparent in Kendall Walton's sense, and thus why it is reasonable for us to rely on them as sources of evidence.

The causal connection between a photograph and what it represents may be quite complex indeed. According to Latour (1999), photographs are employed as scientific evidence as part of a complex chain of representations that mediates between the world and our experiences of it. What they represent is a product of what Latour calls "circulating reference," a chain of transformations of representations—such as photographs, diagrams, and equations—that ultimately links our theories to the world. As he puts it:

> We have taken science for realist painting, imagining that it made an exact copy of the world. The sciences do something else entirely—paintings too, for that matter. Through successive stages they link us to an aligned, transformed, constructed world. We forfeit resemblance, in this model, but there is compensation: by pointing with our index fingers to features of an

[3] Benovsky's (2016) article contains very helpful discussion of numerous ways in which photographs, while depicting something real, can depict it in inaccurate ways. For example, telephoto lenses create a "compression" effect that causes objects in the photograph appear to be closer together than they actually are, use of long exposures can create motion blur which is not visible in real life, etc.

> entry printed in an atlas, we can, through a series of uniformly discontinuous transformations, link ourselves to [the world]. … I can never verify the resemblance between my mind and the world, but I can, if I pay the price, *extend* the chain of transformations wherever verified reference circulates through constant substitutions (p. 79, emphasis in original).

Part of Latour's point is that we often do not engage directly with the world itself in investigating it, but with various representations of it. As Latour emphasizes, the world is often far too complex for us to take on directly, so we make it more manageable by working with representations of it. As Latour puts it, a representation "does *more* than resemble. It *takes the place of the original situation* …" (p. 67, emphasis in original).

This idea, that recordings *replace* more than *resemble,* is especially helpful in clarifying the extent to which we rely on recordings as sources of evidence. We expect recordings to give us evidence of things that would otherwise be beyond our ken. For example, we take ourselves to know, in vivid detail, what the South Pole looks like, despite the fact that almost none of us will ever actually go there. In addition to expecting recordings to be reliable, that is, we expect them to be *ubiquitous*. What I mean by this is that recordings populate our current information environment to the extent that we *expect* recordings as evidence concerning events that we were not directly privy to.

The reliability and ubiquity of recordings, taken together, explains why they function as what Rini (2020) calls an "epistemic backstop." According to Rini, an important function of recordings is to ground out the reliability of testimony. We can now see that this is the case because, on the one hand, reliable recordings that corroborate a piece of testimony give us strong evidence that the testimony is reliable. But on the other hand, the ubiquity of recordings leads us to expect that most significant events are being recorded. This constant expectation of surveillance, Rini points out, gives us an incentive to provide accurate testimony, lest we be shown to be unreliable by recordings that contradict what we have said.

Let me sum up some of the points made in this section in order to draw some preliminary conclusions. We do in fact rely on recordings as sources of evidence. We treat them as providing us with evidence, at least on a par with that provided by our own senses, concerning events that we did not observe in person. Moreover, the causal connection between recordings and their subjects gives us *pro tanto* reason to take them to be reliable sources of evidence about their subjects. Even though recordings do not depict their subjects accurately in every detail, their causal link to their subjects makes it *pro tanto* reasonable for us to treat them as evidence. That is, the causal link that recordings bear to their subjects is a central reason why we treat them as evidence, and why it is *pro tanto* reasonable for us to do so.

As noted above, it is possible for even analog photographs to misrepresent their subjects, either through post-production editing, or through techniques employed in the original capture of the photographic event.[4] But due to their digital nature, the problem with deepfakes is not just that they can misrepresent their subjects. Unlike even heavily edited analog photographs, deepfakes need not have any causal connection to their subjects.[5] As I noted above, to create a photograph, there *must have been* a photographic event in the first place. In the case of analog photography, the photograph produced is a physical object, whether a negative, a print on paper, a silver daguerreotype plate, etc. But in the case of digital photography, the photograph produced is just data; it is a file that must decoded by software in order to display an image. This is important because such a file can be created by techniques, such as the algorithms employed in making deepfakes, that can operate in the absence of a photographic event. That is, a deepfake is an artifact that may be indistinguishable from a digital photograph, but which is not a photograph at all, due to the fact that there need not have been a photographic event in the causal history of its creation. By contrast, there is no analogous case for analog photographs. Even a heavily altered analog photograph is still a photograph.[6]

Having clarified the sense in which deepfakes are fake digital recordings, we can start to articulate more clearly what epistemic problems they pose. As Goldman's (1976) famous "fake barn" case shows, the presence of fakes in an environment threatens to undermine otherwise reliable sources of evidence in that environment if we cannot distinguish fakes from genuine items. That is, if we are unable to distinguish between deepfakes and veridical recordings, this threatens to undermine the status of digital recordings as sources of evidence. Because such recordings are ubiquitous—and because we do in fact rely on them to such a great extent—this is a serious problem indeed.

But why does the existence of deepfakes threaten to undermine the status of digital recordings as sources of evidence? In the following two sections, I develop an answer to this question. I proceed by considering and criticizing an account recently offered by Fallis (2020). Then, in Section 3, I articulate a new account on which the epistemic threat posed by deepfakes is analogous to that posed by classic skeptical hypotheses concerning the senses.

[4] See footnote 2.

[5] My discussion in this paragraph concerns still-image deepfakes specifically. But I think it clearly generalizes to other kinds of recordings as well.

[6] See Benovsky (2016).

2 Fallis' Account: Deepfakes and Information

According to Fallis (2020), deepfakes undermine the status of recordings as sources of evidence because they decrease the information content of recordings. Fallis understands information content in the sense of Skyrms' (2010) account, according to which information is to be understood in terms of changes in probabilities. On this account, as Fallis employs it, signals in an agent's environment can carry some quantity of information about states of affairs in that environment. More specifically, signal R carries the information that S iff P(R|S) > P(R|~S).[7]

Applying this specifically to recordings, let R be a recording and S be the state of affairs depicted by that recording. P(R|S) represents the probability of a *true positive*, whereas P(R|~S) represents the probability of a *false positive*. Accordingly, we can say that the quantity of information carried by a signal is the ratio of true positives to false positives. Call this ratio, P(R|S)/P(R|~S), the *information quantity ratio* (IQR).[8] On Skyrms' account, which Fallis is employing here, a signal requires a sender and a receiver. Thus, the probabilities in the IQR depend on both the relative numbers of genuine and fake recordings in the receiver's information environment, and on the receiver's ability to distinguish between genuine and fake recordings. For example, if there are very few fake recordings in the environment, then the probability of the receiver accepting a fake recording as genuine (a false positive) is low, even if the receiver is unable to distinguish between genuine and fake recordings. Alternatively, if the environment contains many fake recordings, but the receiver is adept at distinguishing them from genuine recordings, then the false-positive probability will still be low.

Applying the definition of information that Fallis employs, we can see that recordings became increasingly informative signals over the course of the 20th century. Throughout this period, the use of recording technology, and the technology for viewing recordings, such as video cameras, broadcast television, etc., became increasingly common. Thus, during this period, the number of recordings that were created and viewed increased significantly. In terms of Fallis' definition, we can model this situation by noting that, over the course of the 20th century, the value of the P(R|S) term increased significantly. That is, there was a significant

[7] 'P(R|S)' is the conditional probability of R, given S. That is, it is the probability that R is the case given that S is the case. '~S' denotes the circumstance that S is *not* the case.

[8] By the above definition, R carries information that S iff P(R|S)/P(R|~S) > 1 (i.e. IQR > 1). If P(R|S) = P(R|~S), then R carries no information about S, since R is equally likely to be sent whether or not S is the case. If P(R|S)/P(R|~S) < 1, then R carries information that ~S instead. In the unusual case where P(R|~S) = 0, we can say that R carries maximal information about S, since false positives are impossible.

increase in the probability that, given a state of affairs S, a recording R of that state of affairs would be created and viewed. Using terminology from the previous section, we can thus say that the $P(R|S)$ term captures the *ubiquity* of recordings in a given information environment. But despite the tremendous increase in the ubiquity of recordings in the 20th century, their *false-positive* rate, $P(R|{\sim}S)$, remained quite low.[9] Consequently, by the end of the 20th century, the IQR of recordings had become very high. That is, recordings had become extremely informative to us, in the sense of the above definition of information.

According to Fallis, the epistemic problem posed by deepfakes is that they decrease the amount of information that recordings carry. Because deepfakes increase the probability of false positives—i.e. they increase $P(R|{\sim}S)$—they decrease the IQR of recordings. And it is true that the increased prevalence of digital recording and editing technology has led to a tremendous increase in the false-positive term, $P(R|{\sim}S)$. The use of digital tools has made fake recordings more common than they used to be. In addition to deepfakes, which are still relatively rare, consider the fact that “photoshop” is now a verb in our lexicon. We are all familiar with the fact that it is nowadays not particularly difficult to make a convincing photo, video, or audio recording that grossly misrepresents its subject using digital tools.

But is the information quantity of recordings really lower now as a result of all of these fakes in our environment? Of course, it is true that the IQR of recordings will decrease if $P(R|{\sim}S)$ increases while $P(R|S)$ is held constant. But a moment’s reflection should convince you that this is not an accurate model of our current situation. Nowadays, nearly everyone has easy access to a digital recording device (e.g. a smartphone) and platforms on which to share and view such recordings (e.g. YouTube, Instagram). Because of this, the ubiquity of recordings, $P(R|S)$, has increased astronomically in the 21st century. It is difficult to find statistics that capture just how dramatic this increase in ubiquity has been, but here is one that I find particularly staggering: As of May 2019, 500 h of video were being uploaded to YouTube every minute (Clement 2020). And no doubt the rate has only increased since then. While it is true that the false-positive rate has increased significantly over the last 20 years, I conjecture that the true-positive rate has increased *far more*

[9] No doubt it increased somewhat as techniques for video editing became increasingly sophisticated, but I doubt this made a significant difference to the rate overall. Even very sophisticated fakes, such as those created by the KGB for propaganda purposes, were relatively rare. You might also worry that the proliferation of *fictional* recordings, e.g. Hollywood films, would drive up the false-positive rate. But we need to distinguish between *fictional* and *fake*. The key idea is that a fake recording purports to be a genuine recording of a state of affairs, but it is not. For the purposes of this paper, we can treat fictional recording as genuine recordings, too, in that they are genuine depictions of states of affairs involving actors, costumes, scenery, etc.

in that time period. If this conjecture is correct, then Fallis' analysis actually implies that recordings now carry far *more* information than they did before the development of digital recording and editing technology.

Perhaps we can extrapolate from the present on Fallis' behalf. To make a deepfake, a deep-learning network must have access to an enormous library of training materials; more specifically, recordings of the subject that it is going to fake. This is why it is currently possible to make deepfake videos of celebrities and well-known politicians, but not ordinary people.[10] Thus, it is only possible to produce deepfakes because of the extremely high value of the ubiquity term, P(R|S). Furthermore, this suggests that an increase in the ubiquity of recordings will lead to an increase in the quantity and quality of deepfakes that exist. That is, because of how deepfake technology works, an increase in P(R|S) will lead to an increase in P(R|~S). Of course, if these terms increase at the same rate, this will still not reduce the amount of information carried by recordings, since that quantity is captured by the ratio of these terms. Consequently, the epistemic problem posed by deepfakes, on Fallis' account, must be that *eventually* the quantity and quality of deepfakes in our digital information environment will lead to a rate of increase in the P(R|~S) term that overtakes the rate of increase of the P(R|S) term. When that happens, the quantity of information carried by digital recordings will indeed begin to decrease.

But how long will it be until that happens? I am not in a position to answer that question, but the important point is that this is not happening *now*. In our *current* digital information environment, recordings carry far *more* information, in Fallis' sense, than they ever have in the past, and this is the case despite the existence of deepfakes. Nevertheless, the existence of deepfakes seems to pose an epistemic problem for us in our *current* information environment. As I have just argued, Fallis' account does not explain why this is.

3 Underdetermination and the Epistemic Threat of Deepfakes

In order to explain why deepfakes threaten to undermine the value of digital recordings as evidence, it will be helpful to consider another quirk of Fallis' account. On this account, deepfakes undermine the value of recordings as evidence

[10] Importantly, this is not the case for still-image deepfakes. Technology already exists for generating fake images of novel faces. There is no reason to believe that this technology would not be further developed to generate fake video as well.

because they decrease the quantity of information carried by recordings. As I showed in the previous section, this will happen when deepfakes become sufficiently numerous and convincing as to lead an increase in P(R|~S) that swamps any increases in P(R|S). Given this description of the problem, one natural solution would appear to be simply to increase the value of P(R|S); i.e. increase the ubiquity of recordings in our digital information environment. This could be a "grassroots" effort led by individuals heroically flooding the internet with genuine videos (TikTok to the rescue?), but it would be considerably easier and faster to enlist the help of artificial intelligence technologies to control cameras, generate videos of events, and upload these videos to social media platforms. But this sounds nightmarish. More to the point, it is very hard to see how this could possibly improve our epistemic situation. Increasing the number of genuine recordings in the environment will certainly make it more likely that any given recording is genuine. But recall that the false-positive probability also depends on the ability of viewers to distinguish between genuine and fake recordings. Even if odds are good that a given recording is genuine, if we cannot *tell* whether or not it is genuine, its evidential value for us seems to be significantly degraded. This suggests that the real danger posed by deepfakes is not the relative quantities of genuine and fake recordings in the digital information environment, but rather our increasing inability to distinguish between the two.

Why does our inability to distinguish between genuine and fake digital recordings pose an epistemic problem? To address this question, it will be helpful to consider an analogy to an argument for what I will call "classical skepticism" regarding the senses; the idea that our senses cannot justify our beliefs about the world around us.

Following Brueckner (1994), I think it is most instructive to cast this argument in terms of the Underdetermination Principle (UP): Let E be a body of evidence and P and Q be incompatible propositions. Then E provides justification for believing P only if E supports P more strongly than it supports Q.[11] This principle is, I believe, standardly employed in our ordinary practices of epistemic evaluation. For example, suppose I see a medium-sized black bird flying in the sky, and exclaim, "there's a crow!" You, being inquisitive, ask me why I think this, and I respond by informing you of my evidence, namely that I saw a medium-sized black bird. I think

[11] It may be possible to put this argument in information-theoretic terms as well, by employing the analysis developed by Floridi (2011). My point in the previous section was only to argue that Fallis' information-theoretic treatment does not explain why deepfakes threaten to undermine the status of digital recordings as evidence. That argument was not intended as an indictment of information-theoretic approaches generally. That said, I do think the argument in Section 2 shows that the problem posed by the existence of deepfakes cannot *merely* be that they raise the probability that a given digital artifact is a fake recording.

the appropriate thing for you to say here is: “How do you know it wasn’t a raven, then?” In saying this, you are pointing out that my evidence—seeing a medium-sized black bird—does not support the proposition that I saw a crow more strongly than it supports the proposition that I saw a raven. After all, both are medium-sized black birds. And in that circumstance, it seems that my belief that it was a crow is not justified. After all, as you pointed out, for all I know it might be a raven, and not a crow at all. If that is the case, then my evidence really does not tell me that it is a crow rather than a raven, and so does not justify my belief, just as UP says.

Using UP, the argument for classical skepticism regarding the senses runs as follows. Let P be a quotidian proposition about the world around you, e.g., that you are currently in a room, and let E be the total body of information provided to you by your senses. According to UP, the evidence from your senses provides justification for your belief that you are currently in a room only if this evidence supports your belief more strongly than it supports incompatible alternatives. To take a classic skeptical hypothesis, one such alternative is the hypothesis that you are not actually in a room; that what appears to be a room is instead a fake cleverly constructed by a supremely powerful malicious being, Descartes’ “evil demon,” bent on deceiving you (Descartes 1641/1996). Call this hypothesis Q. Now, the crucial thing to observe is that, on the basis of your senses, it is not possible to tell whether you are in an actual room or whether you are faced with a cleverly constructed fake. We are imagining that the evil demon is sufficiently powerful so as to create an illusion that perfectly mimics any sensory experience you would have were you in an actual room. In this case, the evidence from your senses is *neutral* between P and Q; you simply cannot tell which is the case. This means, of course, that your evidence E does not support P more strongly than it supports Q. Consequently, applying UP, it follows that the evidence from your senses does not provide you with justification for believing that you are currently in a room.

It is worth emphasizing that the skeptical hypothesis Q need not be *true* in order for this argument to work. By the above argument, the mere *possibility* that what appears to be a room around you is in a fact a cleverly constructed fake, once it is made salient to you, is sufficient to undermine the evidence that your senses provide you about your whereabouts. Moreover, it should be clear that this argument readily generalizes; it is a *recipe* for doubting any belief purportedly justified by evidence from the senses. The skeptical hypothesis Q exposes a *gap* between your evidence and the beliefs that this evidence purports to justify. As Stroud (1984) puts it:

> [While contemplating a skeptical hypothesis, we are] in the position of someone waking up to find himself locked in a room full of television sets and trying to find out what is going on in the world outside. … The victim might switch on more of the sets in the room to try to get more

> information, and he might find that some of the sets show events exactly similar or coherently related to those already visible on the screens he can see. But all those pictures will be no help to him without some independent information, some knowledge which does not come to him from the pictures themselves, about how the pictures he does see before him are connected with what is going on outside the room (p. 33).

With respect to our beliefs about faraway places, concerning people we do not know, we are locked in Stroud's room full of television sets.[12] Our only source of information about those distant people and events is the screens we have in front of us. We can try looking at different screens, or different content on the same screen, but none of this can tell us that any of those screens accurately represent the goings-on outside our local environment. In parallel with the argument for classical skepticism rehearsed above, this problem is made particularly difficult by the specter of deepfakes, which make salient the possibility that, while we appear to be viewing digital recordings on our screens, they may not be genuine recordings at all.

Accordingly, deepfakes generate a salient skeptical hypothesis that functions, just as the evil demon functions, to undermine a source of evidence. Let P be the hypothesis that a politician said something self-incriminating, for example, and E the body of evidence in support of this claim; specifically, what appear to be recordings of that politician making self-incriminating statements. As before, these apparent recordings justify your belief about what the politician said only if they support the hypothesis that the politician actually said those things more strongly than they support incompatible alternatives. One such alternative is, of course, the skeptical hypothesis raised by deepfakes. Let Q be the hypothesis that what appear to be recordings of the politician making self-incriminating statements are deepfakes and that in fact they said no such thing. As before, the crucial question is whether you can tell that the videos are genuine, as opposed to cleverly constructed fakes. As deepfake technology continues to improve in quality, the answer to this question will increasingly be "no." By UP, then, it follows that the evidence provided by videos does not justify your belief that the politician said something self-incriminating.

As in the case of classical skepticism, this argument for skepticism concerning digital recordings does not depend on the truth of the skeptical hypothesis. What is important is not that the videos *actually are* deepfakes, but rather, that they *might* be fakes and we could not tell. As before, this argument readily generalizes; it is a

[12] Of course, our main access to the digital information environment is not television sets but smartphones and other internet-connected computing devices. But, even considering the fact that such devices are two-way communication devices as opposed to the one-way television set, that makes no difference to the essential point here.

recipe for doubting any belief held on the basis of evidence from a digital recording. For any belief held on the basis of evidence from a digital recording, one could always stop short: “Maybe not. Maybe this is just a deepfake?” The possibility that any digital recording might be an undetectable deepfake thus exposes a *gap* between the evidence provided by digital recordings and the beliefs that this evidence purports to justify.

This underdetermination-based skeptical argument thus explains why deepfakes undermine the status of digital recordings as evidence. The existence of deepfakes in our digital information environment makes salient to us the possibility that what appears to be a genuine recording is in fact a cleverly constructed fake. As deepfake technology improves, it will be increasingly difficult for us to distinguish between deepfakes and genuine recordings. And, as the underdetermination-based skeptical argument shows, our inability to distinguish fake from genuine recordings will prevent even genuine recordings from providing justification for our beliefs.

Thus, the threat posed by deepfakes is the same sort of threat posed by disinformation campaigns. The goal of disinformation is not to propagate false information, but rather, to undermine the credibility of sources of information generally. As a tobacco industry executive famously put it in a 1969 memo: “Doubt is our product since it is the best means of competing with the ‘body of fact’ that exists in the minds of the general public” (Oreskes and Conway 2010, p. 34). Like disinformation campaigns, deepfakes create doubt. If we cannot distinguish between genuine and fake digital recordings, then, by UP, such recordings cannot provide evidence to justify our beliefs. And absent such evidence, how can we determine what to believe?

4 Overcoming Skepticism About Digital Recordings?

Despite the strong structural similarities in the arguments for skepticism about the senses and skepticism about digital recordings, I think there are two significant points of dissimilarity between the two arguments. Reflection on these points of dissimilarity will, I hope, suggest some possible strategies for combating skepticism about digital recordings.

First, classical skepticism about the senses poses an *in-principle* epistemic problem. *No* body of evidence from the senses could be sufficient to support a quotidian proposition about the world around us more strongly than an incompatible skeptical hypothesis. This is because the skeptical hypothesis (e.g. that what appears to be a room is actually a fake constructed by an evil demon) is constructed in such a way that *any* possible observations one could make—including observations intended to test whether the room is fake—would be consistent with the skeptical hypothesis (Stroud 1984, p. 23). Accordingly, a common response to classical skepticism is: "So what?" As proponents of this response point out, Descartes himself claims that the possibility raised by his skeptical hypothesis is "slight" (Maddy 2017, p. 13). Moreover, practically minded philosophers going back at least to John Locke have pointed out that since *everything*—including our own actions—would seem exactly the same to us whether or not the skeptical hypothesis were true, the truth or falsity of the skeptical hypothesis makes no difference to us whatsoever in the course of our lives, and thus it can safely be ignored.[13]

By contrast, skepticism about digital recordings poses an *in-practice* epistemic problem. Even the best deepfakes are not *impossible* to distinguish from genuine recordings. The problem is that it might be too costly, in terms of time and other resources, to detect deepfakes when it matters. To see this, consider a thought experiment posed by Ewing (2019), writing for NPR:

> Imagine it's the night before a big debate, or Election Day itself. Suddenly a video is everywhere that appears to show a candidate saying something outrageous, or engaged in some kind of inappropriate conduct. If the veracity of that material is unclear for the succeeding 12 or 24 hours or more, that could have an effect on voters' attitudes. Or imagine the mirror image of this scenario: Suppose the clip appears and what it depicts is real—but the candidate involved denies what it contains and says it's a fake, citing all the discussion about fabricated media. Other evidence might emerge proving that the activity in the video or audio was real, but what if all the facts weren't sorted until hours or days later?

Thus, while practically minded philosophers might safely ignore classical skeptical hypotheses, the possibility that a given video is a deepfake presents a clear practical problem. Even if we can determine that a video is fake after some time and effort, by that point, the damage will already be done.[14]

Moreover, while the possibility of deception by a supremely powerful being might be "slight," the possibility of deception by a deepfake is all too real. We know that fake content, including deepfakes, "sells" on social media. For example,

[13] As Locke (1689/1975) puts it, the assurance of our senses is "assurance enough, when no man requires greater certainty to govern his actions by, than what is as certain as his actions themselves" (IV.xi.8).

[14] Consider, for example, the "PizzaGate" affair, in which a man became convinced by fake news reports that Comet Ping Pong, in Washington, DC, was in fact a front for a child sex-trafficking ring and drove there, armed with an assault-style rifle, to investigate (Fisher, Woodrow, and Hermann 2016).

as Madrigal (2017) reports, on Facebook, in the three months leading up the U.S. presidential election in 2016, fake, and in many cases *obviously* fake, election stories generated far more engagement—in the form of likes, comments, and shares—than did genuine stories. And there is a clear, market-driven explanation for this phenomenon. As Madrigal explains, "from [Facebook's] perspective, success is correctly predicting what you'll like, comment on, or share. That's what matters. People call this 'engagement.'" And what do we engage with? Content that is similar to what we have already seen, and best accords with what we already think. As Madrigal puts it, "Facebook's draw is its ability to give you what you want." And this problem is, of course, not unique to Facebook. It is now well known that fake content propagates further and faster on Twitter than does genuine content[15] (Vosoughi, Roy, and Aral 2018; Starbird 2017). Perhaps unsurprisingly, this is not a new problem; sensation, and not the sober reporting of facts, is what sells. As Edward McKernon noted in his 1925 article, "Fake News and the Public," "[t]he very efficiency of the cooperative effort of newspapers in gathering news has caused the Faker to resort to gross exaggeration or absolute fiction in order to make his wares attractive"[16] (p. 534). Nevertheless, the volume of information available to us, and the relative ease with which fakes can be disseminated in our information environment, makes the problem far more pressing now than it was in McKernon's day.

The second important point of dissimilarity between the arguments for skepticism about the senses and skepticism about digital recordings concerns what aspects of the skeptical problem we have control over. As I emphasized above, skeptical hypotheses make salient to us the possibility of a gap between our beliefs and our environment. One lesson of classical skepticism is that, regardless of what we do individually, our environment must also cooperate with us if we are to have knowledge. That is, even if we employ the best techniques of gathering and evaluating evidence, that care will come to naught if we are subject to the whims of a malignant deceiver. This is because we can only influence the doxastic side of the gap between our beliefs and our environment.[17] By contrast, in the case of skepticism about digital recordings, in principle, at least we control both sides of the gap. The digital information environment is as much a product of our own making as are the beliefs that we form in that environment. This suggests that we should

[15] On this point, consider the following elucidating typo from David Ulin's (2015) review of David Shipler's *Freedom of Speech* for the *L.A. Review of Books*. Summing up Shipler's view, Ulin writes that the right of freedom of speech, "enshrined in the first Amendment, serves as a collective superpower, allowing us to diffuse bad ideas." Clearly, *defuse* was intended.

[16] This article, from nearly 100 years ago, contains the earliest use of the phrase "fake news" that I am aware of.

[17] Descartes (1641/1996) is explicit about this. See Meditation IV.

aim to blunt the force of skepticism concerning digital recordings by working on both sides of the gap; by refining both our individual belief-forming methods, and the structure of the digital information environment itself. In closing, I make a few suggestions about how this might be done.

In terms of individual belief-forming methods, part of the problem is that we are facing a 21st-century information environment with 19th-century epistemic norms concerning the reliability of recordings (Rini 2020). As I explained in Section 1, we tend to trust recordings in much the same way that we trust our own senses. And, for much of the 19th and 20th centuries, this was a very sensible strategy because of the way then-current recording technology worked. Recordings had a clear causal link with their subjects, and this causal link rendered them highly reliable as sources of evidence about their subjects. This suggests that, for the reliability of recordings, provenance matters. We could be confident employing recordings as evidence because we could be confident about where they came from; ultimately, they came from a viewpoint on the very scene that they depicted. But since digital recordings are just data, and data can be produced in a variety of ways, we can no longer expect an immediate causal connection between what appears to be a recording and its subject. Fortunately, in the digital information environment, it is possible to track the provenance of data as well. A good place to start is in considering the metadata of recordings, as this can provide information about when and where they were made, and when (if at all) they were edited.[18]

But it is important to note that individual investigative techniques will only get us so far. It is simply too difficult and time-consuming for individuals to track the provenance of information that they find online. Thus, tools for helping to track the provenance of recordings could be built into our digital information environment as well. Perhaps, as Floridi (2018) suggests, blockchain technology could be employed to help track the provenance of digital recordings, and this might help us to ensure that they are genuine; that they have a causal connection with the subjects they depict. Several social media sites, including Facebook, already flag recordings that have been altered in some way. This system could, in principle, be expanded to flag AI-generated content.[19]

Moreover, we should consider structural changes to our online information environment, since certain information structures stymie even the best epistemic efforts of the individuals in them. For example, consider "clumpy" network structures; structures in which nodes of the network are arranged in clusters which

[18] Magnus (2009) makes a similar point about assessing the reliability of information on Wikipedia.

[19] Recently, Microsoft has developed and deployed such technology, which they call the "Microsoft Video Authenticator" (Burt and Horvitz 2020).

are strongly connected to one another, but not strongly connected to nodes outside the cluster. Such structures are typical in filter bubbles or echo chambers.[20] In such structures, you can consistently fail to form true beliefs based on information from the network, even if your individual investigative methods are rational (O’Connor and Weatherall 2019). And, the sharing network of Twitter, for example, is very clumpy. In fact, most fake content on Twitter is shared by a small number of users, from a small number of highly connected sites (Grinberg et al. 2019). This suggests that we might improve our information environment by breaking up such clumps. For example, Twitter’s algorithms could be adjusted so that they demote repeated posts linking to small clusters of sites in short periods of time. This could potentially slow the spread of fake content, including deepfakes, in our information environment.

Lest I come across as unduly optimistic, I should note that a significant downside of any proposal to alter the structure of our digital information environment by the use of algorithms is subject to a serious problem. Any such algorithms can, and will, be “gamed.” For example, there is already an industry devoted to helping businesses and individuals influence their rank in Google searches, and there is no reason to believe that algorithms designed to flag deepfakes and break up clumpy network structures could not be gamed in a similar way.

Thus, my closing suggestions here leave much work to be done on both conceptual and technical levels. But the important point is that, as I have argued, deepfakes pose an epistemic problem because they make salient to us a possible gap between one of our main sources of evidence in the digital information environment—recordings—and reality. This is a significant problem because we have historically treated recordings as very strong sources of evidence, and for much of that history it was reasonable for us to do so. But, to continue to trust recordings in the digital information environment, we must update our own epistemic norms for treating recordings as evidence, and the environment itself. This is not to say that this task will be easy, or even achievable. But it must at least be attempted, since addressing the epistemic problems posed by deepfakes might be the most practically pressing epistemic issue of the 21st century.

[20] See Nguyen (2020) for a discussion of some important epistemic features of these structures.

References

Barthes, R. 1981. *Camera Lucida: Reflections on Photography*. R. Howard, tr. New York: Hill & Wang.

Benovsky, J. 2016. “Depiction and Imagination.” *SATS* 17 (1): 61–80.

Brueckner, A. 1994. “The Structure of the Skeptical Argument.” *Philosophy and Phenomenological Research* 54 (4): 827–35.

Burt, T., and E. Horvitz. 2020. *New Steps to Combat Disinformation*. Also available at https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/.

Cavedon-Taylor, D. 2013. “Photographically Based Knowledge.” *Episteme* 10: 283–97.

Clement, J. 2020. *YouTube—Statistics and Facts*. Also available at https://www.statista.com/topics/2019/youtube/.

Daguerre, L. J. M. 1980. “Daguerreotype.” In *Classic Essays on Photography*, edited by A. Trachtenberg, 11–13. New Haven: Leete’s Island Books. Original work published 1838.

Descartes, R. 1996. *Meditations on First Philosophy*. J. Cottingham, tr. Cambridge: Cambridge University Press. Original work published 1641.

Ewing, P. 2019. *What You Need to Know about Fake Video, Audio and the 2020 Election*. *NPR*. Also available at https://www.npr.org/2019/09/02/754415386/what-you-need-to-know-about-fake-video-audio-and-the-2020-election.

Fallis, D. 2020. “The Epistemic Threat of Deepfakes.” *Philosophy & Technology*: 1–21, https://doi.org/10.1007/s13347-020-00419-2.

Fisher, M., J. C. Woodrow, and P. Hermann. 2016. *Pizzagate: From Rumor, to Hashtag, to Gunfire in D.C. The Washington Post*. Also available at https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html.

Floridi, L. 2011. *The Philosophy of Information*. Oxford: Oxford University Press.

Floridi, L. 2018. “Artificial Intelligence, Deepfakes, and a Future of Ectypes.” *Philosophy & Technology* 31: 317–21.

Goldman, A. I. 1976. “Discrimination and Perceptual Knowledge.” *Journal of Philosophy* 73: 771–91.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. 2019. “Fake News on Twitter During the 2016 U.S. Presidential Election.” *Science* 363 (6425): 374–8.

Latour, B. 1999. “Circulating Reference. Sampling the Soil in the Amazon Forest.” In *Pandora’s Hope. Essays on the Reality of Science Studies*, 24–79. Cambridge: Harvard University Press.

Locke, J. 1975. *An Essay Concerning Human Understanding*, edited by P. H. Nidditch, 1–656. Oxford: Clarendon Press. Original work published 1689.

Maddy, P. 2017. *What Philosophers Do: Skepticism and the Practice of Philosophy*. New York: Oxford University Press.

Madrigal, A. 2017. *What Facebook Did to American Democracy*. *The Atlantic*. Also available at https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/.

Magnus, P. D. 2009. “On Trusting Wikipedia.” *Episteme* 6: 74–90.

McKernon, E. 1925. “Fake News and the Public.” *Harpers Magazine* 151: 528–36.

Nguyen, C. T. 2020. “Echo Chambers and Epistemic Bubbles.” *Episteme* 17 (2): 141–61.

O’Connor, C., and J. O. Weatherall. 2019. *The Misinformation Age*. New Haven: Yale University Press.

Oreskes, N., and E. M. Conway. 2010. *Merchants of Doubt*. New York: Bloomsbury Press.

Phillips, D. M. 2009. “Photography and Causation: Responding to Scruton’s Scepticism.” *British Journal of Aesthetics* 49 (4): 327–40.

Prater, N., and M. Hartmann. 2021. *Chauvin Trial Closing Arguments: ‘You Can Believe Your Eyes’*. *New York Intelligencer*. Also available at https://nymag.com/intelligencer/article/derek-chauvin-trial-everything-you-need-to-know.html.

Rini, R. 2020. “Deepfakes and the Epistemic Backstop.” *Philosopher’s Imprint* 20 (24): 1–16.

Skyrms, B. 2010. *Signals*. New York: Oxford University Press.

Starbird, K. 2017. “Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter.” In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11(1), 1–10. Menlo Park: Association for the Advancement of Artificial Intelligence.

Stroud, B. 1984. *The Significance of Philosophical Scepticism*. Oxford: Oxford University Press.

Toews, R. 2020. *Deepfakes Are Going to Wreak Havoc on Society. We Are Not Prepared*. *Forbes*. Also available at https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/#5e61a3367494.

Ulin, D. 2015. *Review: David K. Shipler’s ‘Freedom of Speech’ Reflects Our Fractured Times*. *Los Angeles Times*. Also available at https://www.latimes.com/books/jacketcopy/la-ca-jc-david-shipler-20150503-story.html.

Vosoughi, S., D. Roy, and S. Aral. 2018. “The Spread of True and False News Online.” *Science* 359 (6380): 1146–51.

Walton, K. 1984. “Transparent Pictures: On the Nature of Photographic Realism.” *Critical Inquiry* 11 (2): 246–77.