Philosophy & Technology (2021) 34:623–643
https://doi.org/10.1007/s13347-020-00419-2

RESEARCH ARTICLE

The Epistemic Threat of Deepfakes

Don Fallis[1]

Received: 20 October 2019 / Accepted: 28 July 2020 / Published online: 6 August 2020

Abstract
Deepfakes are realistic videos created using new machine learning techniques rather than traditional photographic means. They tend to depict people saying and doing things that they did not actually say or do. In the news media and the blogosphere, the worry has been raised that, as a result of deepfakes, we are heading toward an "infopocalypse" where we cannot tell what is real from what is not. Several philosophers (e.g., Deborah Johnson, Luciano Floridi, Regina Rini) have now issued similar warnings. In this paper, I offer an analysis of why deepfakes are such a serious threat to knowledge. Utilizing the account of information carrying recently developed by Brian Skyrms (2010), I argue that deepfakes reduce the amount of information that videos carry to viewers. I conclude by drawing some implications of this analysis for addressing the epistemic threat of deepfakes.

Keywords Deception · Deepfakes · Epistemic value · Fake news · Information theory · Videos

1 Introduction

Deepfakes are realistic videos created using new machine learning (specifically, *deep learning*) techniques (see Floridi 2018). They are not produced by traditional photographic means where the light reflected from a physical object is directed by lenses and mirrors onto a photosensitive surface. Deepfakes tend to depict people saying and doing things that they did not actually say or do. A high profile example is "face-swap

✉ Don Fallis
d.fallis@northeastern.edu

[1] Department of Philosophy and Religion, Northeastern University, 360 Huntington Ave, 413 Renaissance Park, Boston, MA 02115, USA

porn" in which the faces in pornographic videos are seamlessly replaced with the faces of celebrities (see Cole 2018).[1] But for almost any event, these techniques can be used to create fake videos that are extremely difficult to distinguish from genuine videos. Notably, the statements or actions of politicians, such as former President Obama, can be, and have been, fabricated (see Chesney and Citron 2019; Toews 2020).

In the news media and the blogosphere, the worry has been raised that, as a result of deepfakes, we are heading toward an "infopocalypse" where we cannot tell what is real from what is not (see Rothman 2018; Schwartz 2018; Warzel 2018; Toews 2020). Philosophers, such as Deborah Johnson, Luciano Floridi, and Regina Rini (2019) and Michael LaBossiere (2019), have now issued similar warnings.[2] As Floridi puts it, "do we really know what we're watching is real? … Is that really the President of the United States saying what he's saying?"

In this paper, I offer an analysis of why deepfakes are such a serious threat to knowledge. Utilizing the account of *information carrying* recently developed by Brian Skyrms (2010), I argue that deepfakes reduce the *amount of information* that videos carry to viewers. I also draw some implications of this analysis for what can be done to address the epistemic threat of deepfakes.

2 The Epistemic Value of Videos

In order to survive and flourish, people need to constantly acquire knowledge about the world. And since we do not have unlimited time and energy to do this, it is useful to have sources of information that we can simply trust without a lot of verifying. Direct visual perception is one such source. But we cannot always be at the right place, at the right time, to see things for ourselves. In such cases, videos are often the next best thing. For example, we can find out what is going on at great distances from us by watching videos on the evening news.

Moreover, we make significant decisions based on the knowledge that we acquire from videos. For example, videos recorded by smart phones have led to politicians losing elections (see Konstantinides 2013), to police officers being fired and even prosecuted (see Almukhtar et al. 2018), and, most recently, to mass protests around the world (see Stern 2020). And we are significantly more likely to accept video evidence than other sources of information, such as testimony. Thus, videos are extremely useful when collective agreement on a topic is needed (see Rini 2019).

Many videos that we watch simply show people (reporters, politicians, teachers, friends, etc.) speaking. We do not learn much directly about the world from such videos, other than that a particular person said a particular thing. In such cases, videos are just another way of receiving testimony along with face-to-face conversations, phone calls, and e-mails. But even then, videos can still provide an important epistemic benefit. If we know (a) that **X** is trustworthy and (b) that **X** said that **S** is the case, we

[1] The component parts used to construct a deepfake may have been produced by traditional photographic means. But the video as a whole is not genuine. Cavedon-Taylor (2013, 288) makes a similar point about "Photoshopped" images.

[2] Johnson is quoted in Cole 2018. Floridi appeared on France's *Tech24* in 2018. See https://www.france24.com/en/tech-24/20180706-Deepfake-blurs-lines-between-reality-fiction. Rini has also written an opinion piece for the *New York Times*. See https://www.nytimes.com/2019/06/10/opinion/deepfake-pelosi-video.html

can be justified in believing that **S** is the case (see Fallis 2018, 57). And videos can provide almost as good evidence as a face-to-face conversation that **X** actually said that **S** is the case. For example, watching Lester Holt on the evening news gives me good evidence that he actually said certain things. Thus, if I already know that Lester Holt is a reliable testifier, I am justified in believing that these things are true.

As Floridi suggests, deepfakes seem to be interfering with our ability to acquire knowledge about the world by watching videos. But exactly how do deepfakes harm us epistemically?

3 The Epistemic Threat of Deepfakes

The main epistemic threat is that deepfakes can easily lead people to acquire false beliefs. That is, people might take deepfakes to be genuine videos and believe that what they depict actually occurred. And this epistemic cost could easily have dire practical consequences. For example, Chesney and Citron (2019, 147) ask us to imagine "a video showing an American general in Afghanistan burning a Koran. In a world already primed for violence, such recordings would have a powerful potential for incitement." In addition, "a convincing video in which [a well-known politician] appeared to admit to corruption, released on social media only 24 hours before the election, could have spread like wildfire and proved impossible to debunk in time" (Chesney and Citron 2019, 151).[3]

However, in addition to causing false beliefs, there are other ways that deepfakes can prevent people from acquiring knowledge. For example, even after watching a genuine video and acquiring true beliefs, one might not end up with knowledge because one's process of forming beliefs is not sufficiently reliable. As deepfakes become more prevalent, it may be epistemically irresponsible to simply believe that what is depicted in a video actually occurred. Thus, even if one watches a genuine video of a well-known politician taking a bribe and comes to believe that she is corrupt, one might not *know* that she is.

Moreover, in addition to causing false beliefs and undermining justification for true beliefs, deepfakes can simply prevent people from acquiring true beliefs (see Fallis 2004, 465). When fake videos are widespread, people are less likely to believe that what is depicted in a video actually occurred.[4] Thus, as a result of deepfakes, people may not trust genuine videos from the legitimate news media (see Chesney and Citron 2019, 152; Toews 2020). Indeed, a principal goal of media manipulation is to create uncertainty by sowing doubt about reliable sources (see Oreskes and Conway 2010; Coppins 2019).[5]

[3] Frighteningly, such deepfakes might be disseminated by legitimate political organizations as well as rogue actors. For instance, the Democratic National Committee would not pledge not to use deepfakes (see Coppins 2019).

[4] Schauer and Zeckhauser (2009, 47–48) make a similar point with respect to misleading testimony. As Kant (1996 [1788], 196) points out, in the extreme case, no one will believe anyone if everyone deceives whenever it is to her advantage.

[5] As a result of deepfakes, people may also end up in a worse epistemic state with respect to *what other people believe* about the events depicted in a video. And this can also have dire practical consequences. For example, even if you do not think that the video is genuine, if you believe that your friends are going to be taken in, there may be pressure for you to act as if you believe that the politician is corrupt (see Mathiesen and Fallis 2017, 47).

Of course, deepfakes can only prevent people from acquiring true beliefs if no other reliable source for the same information is available. However, there is often no feasible alternative to video evidence that is equally reliable. Direct visual perception certainly provides reliable evidence that an event has occurred. But people can only make such observations with respect to events in close physical proximity. Photography and testimony allow us to learn about a much wider range of far-flung events. However, they are not nearly as reliable as direct visual perception. Even before deepfake technology, still photographs could be faked with Photoshop (see Rini 2019). And it has always been possible for people to testify convincingly about events that did not actually occur. Psychological studies consistently find that people are only slightly better than chance at detecting lies (see Bond and DePaulo 2006).

Finally, deepfakes can also interfere with our ability to acquire knowledge from video testimony in the same three ways. Even if the person speaking to us in a video looks like someone that we know to be a trustworthy source (such as Lester Holt), it could be a deepfake designed to fool us. If it *is* a deepfake and we believe what the person says, we could easily end up with a false belief.[6] And even if it is a genuine video of a trustworthy source saying something true, we might not be justified in believing what she says, or we might fail to believe what she says, because the video could have been a deepfake.

Now, realistic fake videos of events that did not actually occur are nothing new. For example, during World War Two, the Nazis created propaganda films depicting how well Jews were treated under Nazi rule (see Margry 1992). Also, well before the advent of deepfakes, people have been worried about videos being fake. For example, many people have thought that the videos of the Apollo moon landings were faked (see Villard 2004). So, there is certainly a sense in which deepfakes do not pose a *brand new* epistemic threat.

Nevertheless, deepfake technology threatens to drastically increase the number of realistic fake videos in circulation. Thus, it threatens to drastically increase the associated epistemic costs. Machine learning can make it possible for *almost anyone* to create convincing fake videos of *anyone* doing or saying *anything*. In fact, there are literally apps for that, such as FakeApp and Zao. User-friendly software can be downloaded from the Internet along with online tutorials on how to use that software. One no longer has to be an expert in machine learning with a lot of time and computing resources in order to create a deepfake. As Johnson puts it, “we’re getting to the point where we can’t distinguish what’s real—but then, we didn’t before. What is new is the fact that it’s now available to everybody, or will be ... It’s destabilizing. The whole business of trust and reliability is undermined by this stuff.”

Even though deepfakes can interfere with people acquiring knowledge, deepfake technology also has potential epistemic benefits. Most notably, it might be used for educational purposes. For example, in addition to realistic videos of events that never happened, deepfake technology can also be used to create realistic videos of events that happened, but that were not actually recorded. Thus, extremely accurate reenactments

[6] This risk has arguably increased in recent days. During the COVID-19 pandemic, many interpersonal interactions are now taking place via video conferencing software. And researchers are getting closer to being able to create deepfakes *in real time* (see Thies et al. 2019, Rini 2019). Thus, the person that you seem to be talking to on Zoom might actually be someone else entirely.

of historical events can be created more easily (see Chesney and Citron 2019, 148).[7] Also, content that was originally recorded in one language can be more seamlessly dubbed into another language (see Lu 2019). Face-swapping can be used to allow vulnerable people to speak the truth while preserving their anonymity (see Heilweil 2020). Finally, realistic video lectures can be created just using the sound recording of the lecture and a few still photographs of the lecturer (see Griffin 2019).[8]

However, it seems unlikely that such epistemic benefits will outweigh the epistemic costs of deepfake technology. In any event, my goal in this paper is not to argue that this technology will necessary lead to less knowledge overall or to establish exactly how epistemically bad it is (e.g., as compared with other things such as fake news). My goal is just to explain the epistemic threat that deepfakes pose.

4 Videos Now Carry Less Information

How do deepfakes lead to the epistemic harms described above? My thesis is that, as a result of deepfakes, videos now *carry less information* about the events that they depict. This claim requires some unpacking.

As Jonathan Cohen and Aaron Meskin (2004, 204–205) point out, recordings (photographs, videos, and sound recordings) carry information about the events that they depict.[9] They do so in the much same way that thermometers carry information about the temperature and that tree rings carry information about the age of a tree. And it is because these things carry information that we can use them to learn about the world (see Stegmann 2015, 878–888). Just as we can find out exactly how hot it is by reading a thermometer or how old a tree is by counting its rings, we can also find out that a particular well-known politician took a bribe by watching a video of the event. Moreover, it is an *objective* matter whether something carries information. For instance, tree rings carry information about the age of a tree regardless of whether anyone recognizes this fact. Also, a video carries information about the event that it depicts even if no one ever watches it or forms a belief about that event.

So, videos carry information. But what does it mean for a video to carry *less* information than it used to? One sense in which a video might carry less information is in virtue of providing *less detail*. That is, a video might carry information about fewer states of affairs. For example, unlike color videos, black-and-white videos do not carry information about the colors of the objects depicted. However, this is not the sense of *less information* that I have in mind. Deepfakes can provide just as much detail as genuine videos. Indeed, they would be much easier to detect if they could not.

Alternatively, a video might carry less information in virtue of providing *less evidence* about a particular state of affairs. That is, we cannot be as confident that a

[7] Basically, we can have a much more realistic version of the television show “You Are There” where the venerable reporter Walter Cronkite covered the Salem witch trials and the assassination of Julius Caesar.

[8] In addition, the arrival of deepfakes may spur us to engage in certain epistemically beneficial activities, such as improving our critical thinking and our information environment (see Silbey and Hartzog 2019). But this would not make deepfake technology *itself* epistemically beneficial since we could do these things anyway.

[9] Videos can also carry information about things beyond the events that they depict, such as the skill of the videographer. Cavedon-Taylor (2013, 285–286) makes the same point about photographs. However, my thesis is just that videos carry less information *about the events that they depict* as a result of deepfakes.

state of affairs actually obtains (such as that a particular well-known politician took a bribe) as a result of watching the video depicting that state of affairs. This is the sense of *less information* that I have in mind. But in order to explain how videos carrying less information leads to epistemic harm, we need a formal account of carrying information.

5 A Formal Account of Carrying Information

Several formal accounts of carrying information have been proposed by philosophers. On Fred Dretske's (1981) extremely influential account, **R** carries the information that **S** if and only if **R** provides a guarantee that **S** is true. More formally, **R** carries the information that **S** if and only if $P(\mathbf{S} \mid \mathbf{R}) = 1$ and $P(\mathbf{S}) < 1$. For example, the thermometer reading 72° carries the information that the temperature is 72° because the probability that the temperature is 72° given that the thermometer reads 72° is 1.

Another popular account cashes out carrying information in terms of *counterfactuals* rather than in terms of *objective probabilities*. According to Cohen and Meskin (2004), **R** carries the information that **S** if and only if **R** would not have occurred if **S** were not true. For example, the thermometer reading 72° carries the information that the temperature is 72° because the thermometer would not read 72° if the temperature were not 72°.

These two accounts of information carrying can also be applied to videos. Consider, for example, a video that convincingly depicts a particular well-known politician taking a bribe. On Dretske's account, the video carries the information that the politician took a bribe if and only if the existence of the video guarantees that the politician actually took a bribe. On Cohen and Meskin's account, the video carries the information that the politician took a bribe if and only if the video would not exist if the politician had not actually taken a bribe.

Unfortunately, Dretske's account and Cohen and Meskin's account rule out the possibility of information carrying *coming in degrees*. On their accounts, something only carries the information that **S** if it provides a *guarantee* that **S** is true. In other words, something only carries the information that **S** if it is not possible for this thing to occur when **S** is false. Thus, on their accounts, carrying information is an all-or-nothing affair.

Admittedly, Dretske and Cohen and Meskin can say that a member of one *class* of things (such as videos) is *less likely* to carry information about some state of affairs **S** than a member of another class of things (such as handmade drawings).[10] But the only way that *one particular thing* can carry less information about **S** than any other thing is for it to carry no information about **S** at all. And if we are trying to decide whether the politician took a bribe, what matters is how much information the particular video that we are watching carries about that state of affairs.

Fortunately, more recent accounts of carrying information weaken Dretske's stringent requirement of providing an absolute guarantee (see Stegmann 2015, 870). The specific account of carrying information that I endorse comes from Skyrms (2010). It

[10] Dretske can also measure the degree to which something reduces someone's uncertainty about a state of affairs being true. But that depends on how uncertain she was to begin with. It is not a measure of how much information is *carried* by the object.

was developed in an attempt to understand how animal signals carry information to other animals. For example, the elaborate tails of peacocks carry information about their quality as potential mates, the alarm calls of prairie dogs carry the information that predators are in the vicinity, and the red, yellow, and black stripes of coral snakes carry the information that they are venomous.

On Skyrms's account, a signal **R** carries information about a state of affairs **S** whenever it distinguishes between the state of affairs where **S** is true and the state where **S** is false. That is, **R** carries the information that **S** when the *likelihood* of **R** being sent when **S** is true is greater than the likelihood of **R** being sent when **S** is false. More formally, **R** carries the information that **S** if and only if $P(\mathbf{R} \mid \mathbf{S}) > P(\mathbf{R} \mid \text{not-}\mathbf{S})$.[11] For instance, the prairie dog's alarm call carries the information that there is a predator in the vicinity because it is more likely to occur when there is a predator in the vicinity than when there is not.

Unlike Dretske and Cohen and Meskin, Skyrms thinks that a signal **R** can carry the information that **S** even if **R** sometimes occurs when **S** is false. In other words, Skyrms allows for the possibility of *false positives*. For instance, a prairie dog might mistake a seagull for a hawk and, thus, give the alarm call when there is no predator in the vicinity. In addition, one species might mimic a signal that is commonly sent by another species. For instance, after the venomous coral snake evolved its distinctive appearance to warn potential predators to stay away, the non-venomous scarlet king snake subsequently evolved to resemble the coral snake in order to free ride on this warning system (see Forbes 2009, 241). Thus, seeing a snake with red, yellow, and black stripes does not guarantee that one is dealing with a venomous snake.

As a result, on Skyrms's account, information carrying comes in degrees. Basically, the more likely it is for a signal **R** to be sent in the state where **S** is true than it is for **R** to be sent in the state where **S** is false, the *more* information that **R** carries about **S**. In other words, the higher the probability of a *true* positive relative to the probability of a false positive, the more information that a signal carries. Conversely, the higher the probability of a *false* positive relative to the probability of a true positive is, the *less* information that a signal carries. We can formalize this idea using *likelihood ratios*. A signal **R** carries more information than a signal **Q** about a state of affairs **S** if and only if $P(\mathbf{R} \mid \mathbf{S}) / P(\mathbf{R} \mid \text{not-}\mathbf{S}) > P(\mathbf{Q} \mid \mathbf{S}) / P(\mathbf{Q} \mid \text{not-}\mathbf{S})$.[12] For example, if it is less likely to be deployed by "low-quality" individuals, the peacock's tail carries more information about his reproductive fitness than the sage grouse's strutting behavior carries about his.

It should be noted that talk about how much information a signal carries can be translated into talk about the *reliability of the evidence* that the signal provides. Namely, **R** carries information about **S** if and only if **R** is evidence that **S** is the case. Also, **R** carries more information about **S** than **Q** if and only if **R** is more reliable evidence that **S** is the case than **Q** is. But for purposes of this paper, and following Cohen and Meskin

[11] Skyrms (2010, 35) gives a slightly different formulation. He says that a signal **R** carries the information that **S** if and only if $P(\mathbf{S} \mid \mathbf{R}) > P(\mathbf{S})$. However, my formulation is formally equivalent and avoids certain infelicities. For instance, since the signal **R** occurs after the state of affairs **S**, it makes more sense to talk about the conditional probability of **R** given **S** than to talk about the conditional probability of **S** given **R**.

[12] I assume here that the ratios are both greater than and equal to 1. If $P(\mathbf{R} \mid \mathbf{S}) / P(\mathbf{R} \mid \text{not-}\mathbf{S})$ is less than 1, **R** carries information about not-**S**. I also assume here that the ratio is infinite when $P(\mathbf{R} \mid \text{not-}\mathbf{S}) = 0$. **R** carries the maximum amount of information about **S** if **R** never occurs when **S** is false.

(2004), I use the language of information theory to cash out the intuitive idea that photographs and videos carry information.

The fact that information carrying comes in degrees means that different signals can carry different amounts of information. The probability of a false positive for one signal can be higher or lower than the probability of a false positive for another signal. For example, prairies dogs actually have distinct alarm calls for different predators (see Slobodchikoff et al. 2009, 67). And the alarm call for a coyote in the vicinity carries more information than the alarm call for a hawk in the vicinity if the prairie dog is less likely to mistake something innocuous for a coyote than she is to mistake something innocuous for a hawk. Basically, how much information is carried by a signal can vary with the specific *content* of the signal.

The fact that information carrying comes in degrees also means that the amount of information carried by a particular signal can change over time. As the environmental situation changes, the probability of a false positive may increase or decrease. So, for example, if the number of king snake mimics increases in a particular region, the coral snake's appearance will not carry as much information about its being a venomous snake as it once did. And it is important to note that these probabilities should not simply be interpreted as observed frequencies. For example, if there is an influx of king snake mimics into a particular region, the probability of a false positive will increase even if none of the new mimics have yet been observed.[13]

Another extremely important aspect of Skyrms's account is that he is not measuring the amount of information carried by a signal *from God's eye view*. Instead, Skyrms is concerned with how much information is carried *to the animal receiving the signal*. Consider, for example, the signals of the coral snake and the king snake.

As with most animal mimics, the king snake's appearance is not a perfect copy of the coral snake's appearance. As is pointed out in the handy rhyme ("Red next to yellow, kill a fellow. Red next to black, venom lack"), the stripes are in a different order in the two species. So, from God's eye view, the coral snake's appearance *does* guarantee that you are dealing with a venomous snake.

However, potential predators, such as hawks and coyotes, are not able to distinguish between the red, yellow, and black stripes of the coral snake and the red, yellow, and black stripes of the king snake.[14] Thus, from their perspective, the coral snake's appearance does not guarantee that they are dealing with a venomous coral snake. As far as potential predators can tell, there is a possibility that they are dealing with a non-venomous king snake. Their inability to distinguish the coral snake's appearance from the king snake's appearance is part of what determines the probability of a false positive.[15] As a result, the coral snake's appearance carries much less information

[13] In other words, even though there is currently a high correlation between having red, yellow, and black stripes and being a venomous snake, it is not a *robust* correlation (see Fallis 2004, 474-476). So, the signal does not carry as much information as the observed frequencies might suggest. See Hájek (2019) for other possible interpretations of objective probabilities, such as propensities, infinite long-run frequencies, and Lewis's "best-system" chances.

[14] These potential predators might not have the perceptual ability to distinguish between the two patterns. Or they might simply not have the time to safely make use of this ability while in the presence of a potentially deadly serpent.

[15] Of course, the relative numbers of king snakes and coral snakes in the region is also part of what determines the probability of a false positive.

about what kind of snake it is to potential predators than it does to a human who is familiar with the aforementioned rhyme.

Even though how much information the coral snake's appearance carries is relative to the observer, it is still an objective matter. The coral snake's appearance carries a certain amount of information to anyone in a given epistemic position regardless of whether anyone in that position actually observes the snake or forms a belief about what kind of snake it is. Consider an analogy. In a deterministic world, a fair coin is definitely going to come up heads or it is definitely going to come up tails. So, from God's eye view, the probability of heads is either 1 or 0. However, from the limited epistemic position of humans, the probability is 1/2. And this is an objective matter (see Beebee and Papineau 1997). The probability is 1/2 regardless of whether anyone observes the coin or forms a belief about the chances that it will land heads.

Finally, it is important to keep in mind that Skyrms's account of carrying information is a *mathematical model* of a real-world phenomenon. And as Edwin Mansfield (1994, 13) points out with respect to microeconomic theory, "to be useful, a model must in general simplify and abstract from the real situation." Skyrms's model is intended to capture the important structural features of the phenomenon of signals (and, as I describe in the following section, it can do the same for the phenomenon of videos). The goal is not to calculate precisely how much information is carried by particular signals, such as the prairie dog's alarm call or the peacock's tail.

6 Applying Skyrms's Account to Videos

Skyrms's account of information carrying can be applied to videos as well as signals. Consider, for example, a video that convincingly depicts a particular well-known politician taking a bribe. The video carries the information that the politician took a bribe if and only if the probability of this video existing if the politician actually took a bribe is higher than the probability of this video existing if she did not.[16,17] Moreover, *how much* information the video carries depends on how much greater the former probability is relative to the latter probability. In other words, the video carries more information about the politician taking a bribe when the probability of a false positive is low compared with the probability of a true positive. Thus, with Skyrms's account of information carrying, we can now talk about a particular video carrying less information than another (or than it once did).

In addition, it is important to emphasize that Skyrms's account tells us how much information is carried *to a particular viewer* of the video rather than how much information is carried from God's eye view. How much information is carried to a

[16] A video could carry the information that the politician took a bribe even if it did not *convincingly* depict the event. For example, if the person in the video is clearly wearing a cardboard mask with a photograph of the politician on it, it is not a convincing depiction. Even so, it is possible that such a video is somewhat more likely to be produced if the politician took a bribe than if she did not. However, since deepfakes make it difficult for people to distinguish between genuine videos and *realistic* fake videos, it is not clear that they have much impact on the amount of information carried by such unrealistic videos.

[17] Strictly speaking, what matters is not the probability that the video exists, but the probability that the video is available to be seen. For example, even if all sorts of deepfakes have been produced, they would not decrease the amount of information videos carry to viewers if they never see the light of day (e.g., as a result of effective government censorship).

viewer depends on the probability of a false positive, and the probability of a false positive depends on the viewer's ability to distinguish between genuine videos and fake videos.

To sum up, deepfake technology now makes it easier to create convincing fake videos of anyone doing or saying anything. Thus, even when a video appears to be genuine, there is now a significant probability that the depicted event did not actually occur. Admittedly, the observed frequency of deepfakes, especially in the political realm, is still fairly low (see Rini 2019). Nevertheless, the *probability* of a false positive has increased as a result of deepfake technology. Moreover, this probability will continue to increase as the technology improves and becomes more widely available. So, what appears to be a genuine video now carries less information than it once did and will likely carry even less in the future.

Even after the advent of deepfakes, the amount of information that a video carries is going to depend on the specific content. For example, the probability of a false positive is probably higher when it comes to a politician taking a bribe than when it comes to a politician touring a factory. Even if the technology exists to easily create realistic fake videos of both, fewer people are going to be motivated to fake the latter content. But almost all videos are going to carry less information as a result of deepfake technology.

7 The Epistemic Threat of Deepfakes Revisited

We can now say precisely how deepfakes lead to the epistemic harms discussed above in Section 3. Deepfake technology increases the probability of a false positive. That is, realistic fake videos that depict events that never occurred are more likely to be produced. As a result, videos carry less information than they once did. And this can lead to the significant epistemic harms described above. Consider, for example, a video that convincingly depicts a particular well-known politician taking a bribe.

If videos carry less information than they used to, that means that the probability that a video with this particular content is genuine is lower than it would have been in the past. In particular, suppose that, prior to deepfakes, the probability that a video of this politician taking a bribe is genuine would have been 0.9. But after the advent of deepfakes, the probability that such a video is genuine is only 0.5.[18] If you believe that this politician took a bribe on the basis of this video, the belief that you acquire is five times more likely to be false than it was before deepfakes. And even if your belief happens to be true, it probably does not qualify as knowledge. Your belief is *much less safe* (see Greco 2012, 196). After all, there was only a 50% chance that your belief would turn out to be true.

In addition, you can suffer epistemic harm by failing to acquire true beliefs as a result of deepfake technology. Even after the advent of deepfakes, videos from certain trustworthy sources can still carry a lot of information. For example, suppose that the video of the politician taking a bribe comes from a legitimate news source, such as the evening news or the *New York Times*. Thus, the video is extremely likely to be genuine. However, suppose that, even though you can tell the video comes from this source, you

[18] These specific numbers are just for purposes of illustration. Even if we chose different numbers, the same types of epistemic harm that I describe would still arise.

suspend judgment on whether the politician took a bribe because you are extremely worried about deepfakes. By suspending judgment, you avoid the risk of acquiring a false belief. But as I explain below, you still may be epistemically worse off.

If the odds are sufficiently in your favor of acquiring a true belief, it can be epistemically preferable to believe even though you thereby run a small risk of acquiring a false belief (see Levi 1962; Riggs 2003).[19] So, for example, when it comes to the politician taking a bribe, we might suppose that the epistemic benefit of having a true belief eight times out of ten outweighs the epistemic cost of having a false belief two times out of ten. That is, while you should suspend judgment on whether the politician took a bribe if the probability that she did is less than 0.8, you should believe that the politician took a bribe if the probability that she did is greater than 0.8. Thus, if the probability that a video from this legitimate news source is genuine exceeds this threshold for belief, you would have been epistemically better off believing that the politician took a bribe instead of suspending judgment.

Of course, the epistemic harms that I have described so far are not just the result of deepfake technology. These harms only occurred because you also failed to *proportion your belief* to the evidence (see Hume 1977 [1748], 73, Locke 1975 [1690], 663). In the aforementioned cases, when you acquire a false belief (or a true belief that is not knowledge) from the video, you should have suspended judgment rather than believing what you saw. And when you failed to acquire a true belief from the video from the legitimate news source, you should have believed what you saw rather than suspending judgment. But as I explain below, even if you do proportion your belief to the evidence, there can still be epistemic harm from deepfakes.

Basically, we cannot learn as much about the world if less information is carried by videos. Even if you proportion your belief to the evidence, you are in a less hospitable epistemic environment when the amount of information carried by videos goes down. As a result, as I explain below, you will end up in a worse epistemic state than you would have been in without the decrease in the amount of information carried. In particular, you will end up with fewer true beliefs than you would have had otherwise.

Let us continue to assume (a) that deepfake technology has reduced the probability that the video of the politician taking a bribe is genuine from 0.9 to 0.5 and (b) that your threshold for belief is 0.8. Thus, if you proportion your belief to the evidence, you will suspend judgment on whether the politician took a bribe. Suspending judgment is epistemically preferable to believing that the politician took a bribe when there is a 50% chance that the video is a deepfake. But you are epistemically worse off than you *would have been* without deepfakes. Prior to deepfakes, you would have believed that the politician took a bribe, and you would have been right nine times out of ten. And per our assumptions about your threshold for belief, believing that the politician took a bribe is, under these circumstances, epistemically preferable to suspending judgment.

Of course, it is possible that the advent of deepfakes does not decrease in the amount of information carried by a video past the threshold for rational belief. For example, it might be that enough information is still carried by the video that you will believe that the politician took a bribe even after the advent of deepfakes. Alternatively, it might be that you would have suspended judgment on whether the politician took a bribe even

[19] Only philosophers, such as Descartes (1996 [1641]) in the *Meditations*, who demand absolute certainty would disagree with this claim.

prior to deepfakes. However, even if the advent of deepfakes does not alter what you *fully* believe after watching this video, there can still be epistemic harm. It is possible to measure the distance of someone's credences from the whole truth (see Pettigrew 2016). And if the amount of information carried by a video decreases, you can expect your credences to be further from the whole truth than they would have been.

8 Some Possible Objections

It might be objected that deepfake technology does not *significantly* reduce the amount of information that videos carry. Indeed, it is true that deepfake technology may not *yet* significantly reduce the amount of information that videos carry. Also, it is true that deepfake technology may not significantly reduce the amount of information that *all* videos carry. For example, some videos do not carry that much information to begin with. But as I argue in this section, deepfake technology will soon significantly reduce the amount of information that a large number of videos carry.

8.1 Are Deepfakes Really *Indistinguishable* from Genuine Videos?

In order to significantly reduce the amount of information that videos carry, deepfakes must be indistinguishable from genuine videos. But it might be suggested that it is not all that difficult to distinguish deepfakes from genuine videos. For example, in order to appeal to a certain constituency, an Indian politician recently appeared in a deepfake speaking a language (Haryanvi, a Hindi dialect) that he does not actually speak, but many viewers noticed "a brief anomaly in the mouth movement" (Christopher 2020). However, even though existing deepfakes are not perfectly realistic, the technology is rapidly improving (see Rini 2019; Toews 2020).

Of course, even if a deepfake is indistinguishable from a genuine video based purely on a visual inspection of the image, we might still be able to distinguish it from a genuine video just because the content is *implausible*.[20] For example, despite the high quality of the deepfake, very few people would think that Bill Hader actually morphed into Tom Cruise during an interview on the David Letterman show (see Hunt 2019). However, that still leaves a lot of room for convincing deepfakes. For example, it would not be all that unusual for an Indian politician to speak Haryanvi or for a politician to have taken a bribe.

8.2 Do Videos *Necessarily* Carry Less Information as a Result of Deepfakes?

But even if it is difficult to distinguish deepfakes from genuine videos, it might be suggested that deepfake technology does not *necessarily* decrease the amount of information that videos carry. As noted above, deepfake technology can be used to create realistic videos of events that happened, but that were not actually recorded. In other words, it can be used to increase the number of true positives as well as to increase the number of false positives. And just as the amount of information carried by videos goes down as the probability of a false positive increases, it goes *up* as the

[20] Hume (1977 [1748], 75) and Locke (1975 [1690], 663) make a similar point about testimony.

probability of a true positive increases. However, while it is a conceptual possibility, it seems very unlikely that deepfake technology will lead to a net increase in the amount of information carried by videos. More people are likely to be motivated to create videos of events that did not actually occur than to simply create accurate reenactments. Moreover, empirical studies on fake news (e.g., Vosoughi et al. 2018) suggest that false information spreads faster, and to more people, than true information. Thus, even if accurate reenactments were just as prevalent as videos of events that did not actually occur, we might expect more people to come across the latter.

8.3 Do Videos Carry *Significantly* Less Information as a Result of Deepfakes?

Of course, even if deepfake technology decreases the amount of information that videos carry, it might be suggested that it does not do so to a *significant* degree. It is not yet true that almost anyone can create a deepfake with any content whatsoever. The deepfake apps that are currently available on the Internet often have fairly limited functionality (such as only allowing face-swapping into existing videos). In addition, some of the existing apps have been removed from the Internet (see Cole 2019). But just as the technology is rapidly improving in terms of quality, it is improving in terms of flexibility and is becoming more readily available.

But even once deepfake technology is widely available, it will not significantly reduce the amount of information that a particular video carries if that particular content was already easy to fake. For example, even before deepfake technology, it would not have been difficult to create a convincing fake video of American soldiers burning a Koran. One would just need to rent some uniforms and drive out to the desert with a video camera. In addition, even without deepfake technology, it was not difficult to create convincing fake videos (aka *shallowfakes*) of even *well-known* individuals doing things that they had not actually done. For example, one can speed up a video to make someone seem violent, one can slow down a video to make someone appear to be drunk, or one can cut out parts of a video to make someone sound racist (see Dupuy and Ortutay 2019).

However, without deepfake technology, it would be difficult to fake all sorts of content. For example, what can be depicted in shallowfakes of readily identifiable people or places is significantly constrained by what is already depicted in the genuine source material. Prior to deepfake technology, it *would have been* difficult to create a convincing fake video of a well-known American soldier, such as General Colin Powell, burning a Koran. But this is just the sort of video that one can produce with machine learning. Thus, there are plenty of videos that will carry significantly less information as a result of deepfake technology. Moreover, this is precisely the sort of content that Chesney and Citron worry could have dire practical consequences. It is dangerous if people are misled (when the video is a deepfake) or if people fail to acquire knowledge (when the video is genuine).

9 Alternative Explanations of the Epistemic Threat

I have argued that deepfake technology is reducing the amount of information that videos carry about the events that they depict. In addition, I have argued that, as a result, deepfake technology is interfering with our ability to acquire knowledge from videos by causing false beliefs, causing unjustified beliefs, and preventing true beliefs. My contention is that this explains why deepfakes are such a serious threat to knowledge.

However, even if I am correct that deepfakes interfere with our ability to acquire knowledge from videos as a result of reducing the amount of information that they carry, it might be suggested that this is not the best explanation of why deepfakes are such a serious threat to knowledge. In this section, I consider two alternative explanations. The first appeals to the epistemic difference between photographs and handmade drawings. The second appeals to how recordings underpin our trust in testimony. I argue that these explanations can only account for the seriousness of the threat of deepfakes by appealing (at least tacitly) to the "carrying less information" explanation.

9.1 The Epistemology of Photographs

Not much philosophical work has been done on the epistemology of videos. However, there is a substantial literature on the epistemology of *photographs*. And it might be suggested that this work is applicable to the issue of deepfakes (see Rini 2019). After all, a video is literally a sequence of still photographs taken in quick succession (typically synchronized with a sound recording).

Several philosophers (e.g., Walton 1984; Cohen and Meskin 2004; Walden 2012; Cavedon-Taylor 2013) have tried to identify the property of photographs that makes them epistemically superior to handmade drawings. It might be suggested (a) that videos are epistemic valuable because they have the same property and (b) that they are losing this property as a result of deepfakes.

For example, Kendall Walton (1984, 251) famously claims that photographs are epistemically superior to handmade drawings because they are "transparent. We see the world *through* them." As technology advances, our vision is mediated in ever more complicated ways. We see things through telescopes, microscopes, corrective lenses, mirrors, etc. Walton argues that we also literally *see* the objects and events depicted in photographs. If this is correct, the same would seem to apply to videos (see Cohen and Meskin 2004, 207). And it might be suggested that, as a result of deepfakes, we will no longer be able to see through videos.

However, it is controversial whether seeing is a property that distinguishes photographs from handmade drawings. For example, Cohen and Meskin (2004, 201) have argued that we do not literally see through photographs. Also, Helen Yetter-Chappell (2018, 2032–2038) has recently argued that we can literally see through handmade drawings. But all of these philosophers do agree that, unlike handmade drawings, photographs provide us with a connection to objects and events that is not mediated by the cognitive processes of another human being.

Admittedly, a photographer does get to make a lot of decisions. She decides what to aim her camera at, she decides how to frame the shot, etc. But once the photographer opens the shutter, the features of an object are captured whether or not the photographer notices them. And providing this sort of unmediated connection to objects and events may be the property of photographs that makes them epistemically superior to handmade drawings (see Walden 2012; Cavedon-Taylor 2013).

Deepfakes certainly do not provide us with a connection to objects and events that is not mediated by the cognitive processes of another human being. But the existence of deepfakes does not prevent *genuine* videos from providing an unmediated connection to objects and events. For example, my rearview camera still connects me to the objects behind my car in a way that is not mediated by the cognitive processes of another human being. Even so, it is true that, as a result of deepfakes, videos are *much less likely* to provide us with an unmediated connection to objects and events.

But exactly why is a connection to objects and events that is not mediated by the cognitive processes of another human being so valuable epistemically? After all, we can still acquire knowledge about an event even without an unmediated connection to it. For example, we can hear an accurate report about the event on the evening news or we can watch an accurate reenactment of it. But even though we can acquire knowledge about an event through a mediated connection, it might be suggested that such knowledge is *epistemically inferior* to knowledge based on an unmediated connection (see Cavedon-Taylor 2013; Rini 2019).[21]

One possible reason for the supposed inferiority is that forming beliefs on the basis of a connection that is mediated by the cognitive processes of another human being is *less reliable*. As Scott Walden (2012, 145) notes, "painters or sketchers, their mentation primarily involved in the formative process, can easily add features to their pictures that have no analogues in the depicted scene … Photographers may wish to do the same, but have a much harder time doing so, as their mentation is only secondarily involved in the formative process." But if that is why an unmediated connection is epistemically valuable, we are essentially appealing to the "carrying less information" explanation of why deepfakes are a serious epistemic threat. As discussed above in Section 5, **R** carries less information about **S** than **Q** if and only if **R** is less reliable evidence that **S** is the case than **Q** is. In order to provide an *independent* explanation, an unmediated connection to objects and events must be epistemically valuable *even if* it does not lead to greater reliability.

Another possible reason for the supposed inferiority is that forming beliefs on the basis of a connection that is mediated by the cognitive processes of another human being is *less epistemically autonomous* (see, e.g., Locke 1975 [1690], 101, Fricker 2006). However, it is controversial whether epistemic autonomy has much value beyond leading to greater reliability in certain circumstances (see, e.g., Zagzebski 2007; Dellsén 2020). Given that, it is not clear how the mere loss of epistemic autonomy could be more significant than the loss of knowledge. Thus, the "carrying less information" explanation seems to identify a more serious epistemic threat from deepfakes than the "loss of epistemic autonomy" explanation.[22]

[21] With testimony and reenactments, but not typically with deepfakes, the human being whose cognitive processes mediate our connection to an event *invites us to trust* them. So, it might be suggested that, as a result of deepfakes, videos are much less likely to provide us with *either* an unmediated connection *or* a mediated connection that comes with an invitation to trust. However, much the same line could be taken with respect to this more complicated suggestion as I take with respect to the simpler suggestion in the text. It can only account for the seriousness of the threat of deepfakes by appealing to the "carrying less information" explanation.

[22] Cavedon-Taylor (2013, 295) suggests some other possible reasons. For example, unlike testimonial knowledge, "photographically based knowledge" is a "generative source of knowledge." That is, it allows us to discover new knowledge and not just transmit existing knowledge. But deepfakes only prevent us from using videos to discover new knowledge by preventing us from acquiring knowledge from videos.

9.2 Rini on the Epistemic Backstop

Rini (2019) offers a different sort of explanation for why deepfakes pose an epistemic threat. She points out that, over the past two centuries, recordings (photographs, videos, and sound recordings) have played an important role in underpinning our trust in testimony. First, a recording of an event can be used to check the accuracy of testimony regarding that event (Think of the Watergate tapes). Second, the possibility that an event was recorded can motivate people to testify truthfully regarding that event. (Think of Trump warning Comey to tell the truth since their conversations might have been recorded.) Rini argues that deepfakes threaten to undermine the "epistemic backstop" to testimony that recordings provide.[23] Thus, while Rini agrees with me that deepfake technology interferes with our ability to acquire knowledge, she is concerned about the knowledge that we fail to acquire *from testimony* (because videos no longer provide an epistemic backstop) rather than about the knowledge that we fail to acquire *from watching videos*.

However, while recordings certainly play such a role in underpinning our trust in testimony, it is not clear to me how significant an epistemic harm it would be to lose this epistemic backstop to testimony. It is not as if we do not have other techniques for evaluating testimony. For example, even without recordings, we can consider whether the person testifying has a potential bias or whether other people corroborate what she says (see Hume 1977 [1748], 75). And it is not as if we do not continue to regularly deploy these techniques. Admittedly, given their extensive surveillance capabilities, governments and large corporations may often have access to recordings that bear on testimony that they want to evaluate. But most people, most of the time, do not.

But whether or not losing the epistemic backstop that recordings provide is a significant epistemic harm, this explanation for why deepfakes pose a serious epistemic threat must appeal to the "carrying less information" explanation. It is precisely because deepfakes interfere with our ability to acquire knowledge from the recordings themselves that they provide less of an epistemic backstop to testimony.

10 Implications of Skyrms's Account for Addressing the Epistemic Threat

So, the main epistemic threat of deepfakes is that they interfere with our ability to acquire knowledge, and information theory can explain how deepfake technology is having this effect. But in addition, Skyrms's account of information carrying can suggest strategies for addressing this epistemic threat (and identify some of their limitations). In this section, I discuss three such strategies. The first involves changing our information environment so that it is epistemically safer, the second involves changing *us* so that we are at less epistemic risk, and the third involves identifying parts of our information environment that are already epistemically safe.

[23] In addition to recordings of the event, recordings of the testimony can also be epistemically beneficial since people are less able to deny that they testified in a certain way. Of course, there are also epistemic costs associated with recordings that have to be weighed against these benefits. For example, the possibility of recordings can have a "chilling effect" on the sharing of information (see Fallis 2013, 154).

First, deepfake technology decreases the amount of information that videos carry by increasing the probability that realistic fake videos depicting events that never occurred will be produced. Thus, an obvious strategy for increasing the amount of information that videos carry is to decrease the probability of realistic fake videos being produced. Although there are fewer and fewer technical constraints on the production of realistic fake videos, it is still possible to impose *normative* constraints.[24] For example, laws restricting the creation and dissemination of deepfakes have been proposed (see Brown 2019; Chesney and Citron 2019, 154; Toews 2020). Also, informal sanctions can be applied.[25] Indeed, some apps for creating deepfakes have been removed from the Internet due to public outcry (see Cole 2019). Videos will carry more information as long as the probability of deepfakes being produced is low. It does not matter *why* this probability is low.

Of course, there are some worries about the strategy of banning deepfakes. First, sanctions are not going to deter all purveyors of deepfakes (see Brown 2019). But they are likely to make deepfakes less prevalent. And to paraphrase Voltaire, we do not want to make perfection the enemy of the good. Second, as John Stuart Mill (1978 [1859], 16) points out, access to information can have epistemic benefits even when the information is false or misleading. For example, we often gain “the clearer perception and livelier impression of truth, produced by its collision with error.” But it is not clear that access to information has the epistemic benefits that Mill envisioned when the information is *intentionally* misleading as it is with deepfakes (see Mathiesen 2019, 174).

Furthermore, it should be noted that it is not necessary to ban deepfakes per se. It is only necessary to ban deepfakes of events that did not actually occur that viewers are *unable to distinguish* from genuine videos. For example, the DEEPFAKES Accountability Act only requires “clear labeling on all deepfakes” (Brown 2019). Deepfakes that are labeled as deepfakes need not decrease the amount of information that videos carry.[26] Also, deepfakes that are labeled as deepfakes can still provide the epistemic benefits discussed in Section 3 above.[27] For example, a reenactment can certainly still serve its educational function even if it is labeled as a reenactment.

Second, the amount of information that videos carry does not just depend on the probability that fake videos will be produced. It also depends on the ability of

[24] This seems to be what Walden (2012, 148) has in mind when he asks, “if digital imaging techniques make it easy to undermine the objectivity-based epistemic advantage, will the difference between photographic and handmade images dissipate, or will *institutional factors* limit the extent to which this takes place?” (emphasis added).

[25] Although such informal sanctions have typically been motivated by moral, rather than epistemic, worries about deepfake technology, they can have epistemically beneficial consequences.

[26] Of course, even deepfakes that are labeled as deepfakes can decrease the amount of information that videos carry if people ignore the labels. Researchers are trying to develop ways to label online misinformation that will actually convince internet users that it is misinformation (see, e.g., Clayton et al. forthcoming).

[27] The labeling of deepfakes could get in the way of some epistemic benefits. In just the right situations, we might be able to get people to believe something true by getting them to believe that a deepfake is a genuine video. For example, if the well-known politician actually took a bribe, but the event was not captured on video, we might create a deepfake of the politician taking a bribe. Also, if people will only believe a particular (accurate) message if it comes from a particular (trusted) individual, we might create a deepfake of that individual delivering that message. But situations where deceptive deepfakes have epistemic benefits would seem to be rare.

viewers to distinguish fakes videos from genuine videos. Thus, another possible strategy for increasing the amount of information that videos carry is for us to get better (individually and/or collectively) at identifying deepfakes. As with animal mimics, deepfakes are not perfect counterfeits. So, even if laypeople cannot identify deepfakes with the naked eye, it is still possible for experts in digital forensics to identify them and to tell the rest of us about them. And it is possible for the rest of us to identify such experts as trustworthy sources (see Fallis 2018).

Of course, while this strategy can reduce the epistemic threat of deepfakes, there may be limits to its effectiveness. Consider, once again, the signals of the coral snake and the king snake. Suppose that hawks and coyotes learn how to distinguish the coral snake's appearance from the king snake's appearance. In that case, the king snake would likely evolve to resemble the coral snake more precisely. And in a similar vein, as people develop techniques for detecting deepfakes, other people may have an incentive to create deepfakes that are not detectable by these techniques (see Chesney and Citron 2019, 152; LaBossiere 2019; Toews 2020). In addition, much like with the coral snake and the king snake, even if people have the ability to detect deepfakes, they might not have the time to safely make use of this ability. When deepfakes are used to deceive, the intended victim is often put in a pressure situation (see Harwell 2019).

Finally, in order to address the epistemic threat of deepfakes, we do not have to increase the amount of information carried by *all* videos. Different videos carry different amounts of information. For example, as noted above, a video shown on the evening news is much more likely to be genuine than a random video posted on the Internet. After all, the evening news is a source that has "such credit and reputation in the eyes of mankind, as to have a great deal to lose in case of their being detected in any falsehood" (Hume 1977 [1748], 78). In other words, even without laws against deepfakes, the evening news is subject to normative constraints. Thus, we can try to identify those videos that still carry a lot of information.[28]

But again, while this strategy can reduce the epistemic threat of deepfakes, there may be limits to its effectiveness. Purveyors of deepfakes can try to make it difficult for people to determine whether a video comes from a source that is subject to normative constraints. Indeed, this is precisely what has happened with the phenomenon of *fake news*. With text-based news, there have never been any technical constraints preventing someone from writing a story about something that never happened. Thus, text-based news is only trustworthy if we can tell that the source is subject to normative constraints. So, sources of fake news are typically "designed to look like legitimate news media" (Mathiesen 2019, 166). And as a result, text-based news does not carry as much information as it once did.

[28] Cryptographic techniques can be used to digitally sign (or *watermark*) videos as well as text. If a video is digitally signed by a trustworthy source, we can be reasonably sure that it is genuine. Thus, it carries a lot of information. It is as if a coral snake could make its appearance so distinctive that no other snakes could free ride on its warning system. Admittedly, this strategy only *eliminates* the epistemic threat of deepfakes if the vast majority of genuine videos are digitally signed by a trustworthy source (see Rini 2019). But again, perfection should not be the enemy of the good.

Acknowledgements For extremely helpful feedback on this project, I would like to thank Tony Doyle, Moti Gorin, Peter Lewis, Kay Mathiesen, Neri Marsili, Eliot Michaelson, Kevin Mills, Regina Rini, Ron Sandler, Juan Samuel Santos Castro, Barry Smith, Valerie Soon, Alex von Stein, Mark Wells, Åsa Wikforss, two anonymous referees, and audiences at "Under Pressure: Truth, Trust and Democracy" at the University of London, the CanadianSociety for Epistemology Conference in Montreal, Canada, the International Symposium on Political Lying and Deception in Bogotá, Colombia, the Bodaken Philosophy Symposium Workshop at Colorado State University, and the AI and Data Ethics group at Northeastern University.

References

Almukhtar, S., Benzaquen, M., Cave, D., Chinoy, S., Davis, K., Josh, K., Lai, K. K. R., Lee, J. C., Oliver, R., Park, H., & Royal, D.-C. (2018). Black lives upended by policing: the raw videos sparking outrage. *New York Times*. https://www.nytimes.com/interactive/2017/08/19/us/police-videos-race.html.

Beebee, H., & Papineau, D. (1997). Probability as a guide to life. *Journal of Philosophy, 94*, 217–243.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214–234.

Brown, Nina I. (2019). Congress wants to solve deepfakes by 2020. *Slate*. https://slate.com/technology/2019/07/congress-deepfake-regulation-230-2020.html

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

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. *Foreign Affairs, 98*, 147–155.

Christopher, Nilesh. 2020. We've just seen the first use of deepfakes in an Indian election campaign. *Vice*. https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (forthcoming). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*. https://doi.org/10.1007/s11109-019-09533-0.

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

Cole, Samantha. (2018). We are truly fucked: everyone is making AI-generated fake porn now. *Motherboard*. https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley

Cole, Samantha. (2019). Creator of DeepNude, app that undresses photos of women, takes it offline. *Motherboard*. https://www.vice.com/en_us/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline

Coppins, McKay. (2019). The billion-dollar disinformation campaign to reelect the president. The Atlantic. https://www.theatlantic.com/magazine/archive/2020/03/the-2020-disinformation-war/605530/

Dellsén, F. (2020). The epistemic value of expert autonomy. *Philosophy and Phenomenological Research, 100*, 344–361.

Descartes, René. 1996 [1641]. *Meditations on first philosophy*. John Cottingham, tr. Cambridge: Cambridge University Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge: MIT Press.

Dupuy, Beatrice and Barbara Ortutay. (2019). Deepfake videos pose a threat, but 'Dumbfakes' may be worse. Associated Press. https://www.apnews.com/e810e38894bf4686ad9d0839b6cef93d

Fallis, D. (2004). On verifying the accuracy of information: Philosophical perspectives. *Library Trends, 52*, 463–487.

Fallis, D. (2013). Privacy and lack of knowledge. *Episteme, 10*, 153–166.

Fallis, D. (2018). *Adversarial epistemology on the internet*. In D. Coady and J. Chase (Eds.), Routledge handbook of applied epistemology (pp. 54-68). New York, Routledge.

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

Forbes, P. (2009). *Dazzled and deceived*. New Haven: Yale University Press.

Fricker, E. (2006). Testimony and epistemic autonomy. In J. Lackey & E. Sosa (Eds.), *The epistemology of testimony* (pp. 225–250). Oxford: Oxford University Press.

Greco, J. (2012). Better safe than sensitive. In K. Becker & T. Black (Eds.), *The sensitivity principle in epistemology*. New York: Cambridge University Press.

Griffin, Matthew. (2019). Edtech company Udacity uses Deepfake tech to create educational videos automatically. *Fanatical Futurist*. https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/

Hájek, Alan. (2019). Interpretations of probability. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/probability-interpret/

Harwell, Drew. (2019). An artificial-intelligence first: voice-mimicking software reportedly used in a major theft. Washington Post. https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

Heilweil, Rebecca. (2020). How deepfakes could actually do some good. *Vox*. https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya

Hume, David. 1977 [1748]. An enquiry concerning human understanding. Indianapolis: Hackett.

Hunt, Elle. (2019). Deepfake danger: what a viral clip of Bill Hader morphing into Tom Cruise tells us. *Guardian*. https://www.theguardian.com/news/shortcuts/2019/aug/13/danger-deepfakes-viral-video-bill-hader-tom-cruise

Kant, Immanuel. 1996 [1788]. Practical philosophy. Cambridge: Cambridge University Press.

Konstantinides, Anneta. (2013). Viral videos that derailed political careers. *ABC News*. https://abcnews.go.com/Politics/viral-videos-derailed-political-careers/story?id=21182969

LaBossiere, Michael. (2019). Deep fakes. *Philosophical Percolations*. https://www.philpercs.com/2019/05/deep-fakes.html

Levi, I. (1962). On the seriousness of mistakes. *Philosophy of Science, 29*, 47–65.

Locke, John. 1975 [1690]. An essay concerning human understanding. Oxford: Clarendon Press.

Lu, D. (2019). Dubbing with deepfakes. *New Scientist, 244*(3253), 8–8.

Mansfield, E. (1994). *Microeconomics* (8th ed.). New York: W. W. Norton & Company.

Mathiesen, K. & Fallis, D. (2017). The greatest liar has his believers: The social epistemology of political lying. In E. Crookston, D. Killoren, & Jonathan Trerise (Eds.), Ethics in Politics (pp. 35-53). New York, Routledge.

Margry, K. (1992). Theresienstadt (1944–1945): the Nazi propaganda film depicting the concentration camp as paradise. *Historical Journal of Film, Radio and Television., 12*, 145–162.

Mathiesen, K. (2019). Fake news and the limits of freedom of speech. In C. Fox & J. Saunders (Eds.), *Media ethics, free speech, and the requirements of democracy* (pp. 161–179). New York: Routledge.

Mill, John S. 1978 [1859]. On liberty. Indianapolis: Hackett.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt*. New York: Bloomsbury Press.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Cambridge: Oxford University Press.

Riggs, W. D. (2003). Balancing our epistemic ends. *Nous, 37*, 342–352.

Rini, Regina. (2019). Deepfakes and the epistemic backstop. https://philpapers.org/rec/RINDAT

Rothman, Joshua. (2018). In the age of A.I., is seeing still believing? New Yorker. https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing

Schauer, F., & Zeckhauser, R. (2009). Paltering. In B. Harrington (Ed.), *Deception* (pp. 38–54). Stanford: Stanford University Press.

Schwartz, Oscar. (2018). You thought fake news was bad? Deep fakes are where truth goes to die. *Guardian*. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

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

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

Slobodchikoff, C. N., Perla, B. S., & Verdolin, J. L. (2009). *Prairie dogs*. Cambridge: Harvard University Press.

Stegmann, U. E. (2015). Prospects for probabilistic theories of natural information. *Erkenntnis, 80*, 869–893.

Stern, Joanna. (2020). They used smartphone cameras to record police brutality—and change history. *Wall Street Journal*. https://www.wsj.com/articles/they-used-smartphone-cameras-to-record-police-brutalityand-change-history-11592020827

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Fner, M. N. (2019). Face2Face: real-time face capture and reenactment of RGB videos. *Communications of the ACM, 62*, 96–104.

Toews, Rob. (2020). Deepfakes are going to wreak havoc on society. We are not prepared. *Forbes*. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/

Villard, R. (2004). Did NASA fake the moon landing? *Astronomy, 32*(7), 48–53.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*, 1146–1151.

Walden, S. (2012). Photography and knowledge. *Journal of Aesthetics and Art Criticism, 70*, 139–149.

Walton, K. L. (1984). Transparent pictures: on the nature of photographic realism. *Critical Inquiry, 11*, 246–277.

Warzel, Charlie. (2018). He predicted the 2016 fake news crisis. Now he’s worried about an information apocalypse. Buzzfeed News. https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news

Yetter-Chappell, H. (2018). Seeing through eyes, mirrors, shadows and pictures. *Philosophical Studies, 175*, 2017–2042.

Zagzebski, L. (2007). Ethical and epistemic egoism and the ideal of autonomy. *Episteme, 4*, 252–263.