Philosophy & Technology (2024) 37:7
https://doi.org/10.1007/s13347-024-00700-8

RESEARCH ARTICLE

AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors

Keith Raymond Harris[1]

Received: 21 September 2023 / Accepted: 3 January 2024

Abstract
Deepfakes pose a multi-faceted threat to the acquisition of knowledge. It is widely hoped that technological solutions—in the form of artificially intelligent systems for detecting deepfakes—will help to address this threat. I argue that the prospects for purely technological solutions to the problem of deepfakes are dim. Especially given the evolving nature of the threat, technological solutions cannot be expected to prevent deception at the hands of deepfakes, or to preserve the authority of video footage. Moreover, the success of such technologies depends on institutional trust that is in short supply. Finally, outsourcing the discrimination between the real and the fake to automated, largely opaque systems runs the risk of undermining epistemic autonomy.

Keywords Artificial Intelligence · Deepfakes · Epistemic Autonomy · Misinformation · Self-Trust · Trust

1 Introduction

Online misinformation has become a major focus of study for academics, and a major cause of concern among journalists and the broader public. While many forms of misinformation are already prevalent online, some commentators fear that novel technologies—especially those for generating deepfake videos—will supercharge the problem of misinformation (Fallis, 2021; Foer, 2018; Rini, 2020; Warzel, 2018). It is natural to suppose that, insofar as novel technologies for deception are the problem, novel technologies for detection are the solution. Thus, considerable thought and investment has been poured into technologies for detecting deepfakes and other forms of misinformation.

In this paper, I argue that technological solutions to the problems posed by deepfakes are deeply limited. After providing a brief overview of how deepfakes threaten

✉ Keith Raymond Harris
keithraymondharris@gmail.com

[1] Ruhr-Universität Bochum, Bochum, Germany

Published online: 10 January 2024

knowledge, I argue that automated detection technologies cannot fully address these threats. Then, I argue that reliance on deepfake detectors, more so than reliance on other forms of artificial intelligence, threatens epistemic autonomy. This threat to epistemic autonomy is especially dire, as automated deepfake detectors may pick up on signs of inauthenticity that are imperceptible to humans, thus generating ineliminable tensions between the detector's verdicts and what seems perceptually to be the case. In this way, technological solutions to the problem of deepfakes threaten to supplant reliance on individuals' own senses, while also undermining trust in those senses.

2 Seeing for Oneself and the Three-Pronged Threat of Deepfakes

Before considering the prospects for technological solutions to the problem of deepfakes, it is necessary, first, to establish a better understanding of that problem. I argue in this section that, when considering the effects of deepfakes on the acquisition of knowledge, deepfakes do not just pose a single problem. Rather, deepfakes pose three distinct but related threats to the acquisition of knowledge. These threats are perhaps joined by threats to understanding, wisdom, and other valuable epistemic states that may or may not be reducible to knowledge. Threats to epistemic states are by no means the only problems associated with deepfakes. As has been argued elsewhere, deepfakes raise a range of moral concerns (Öhman, 2020; Rini & Cohen, forthcoming; Young, 2021, Chapter 11) and might impact the audience's sub-doxastic states even if they do not alter the audience's beliefs (Harris, 2021). I focus here on the effects of deepfakes on knowledge in order to simplify the discussion and because, on the face of things, these effects might seem relatively easily addressed by algorithms for the automatic detection of deepfakes.

The term 'deepfake' is sometimes used to refer to media items in general—including photos, audio, and video—that are generated using deep learning techniques. More commonly, the term is used to refer specifically to video footage that uses deep learning techniques to superimpose the likenesses of persons over existing recorded videos or live video streams. I focus on deepfakes, defined in this relatively narrow way, in what follows. While such deepfakes have to this point overwhelmingly taken the form of pornography (Cox, 2019), the potential of deepfakes to play a deceptive role was recognized early on and is increasingly being realized.

As epistemologists have previously noted, deepfakes are concerning, in part, because they seem to threaten the acquisition of knowledge from video footage—an especially important form of evidence (Rini, 2020). Philosophers have previously argued that to see photographs or videos of a thing is to see the thing itself (Walton, 1984; Yetter-Chappell, 2018) or, more modestly, enables perceptual knowledge of that thing (Cavedon-Taylor, 2013). This point, coupled with the fact that technological developments have led the proliferation of technologies for recording, sharing, and streaming video footage, suggests that individuals can see a much more diverse array of what occurs than was previously possible. Although mass media has long allowed for individuals to see what is occurring in distant corners of the world, and indeed some of what is happening offworld,

the democratization of information sharing through technologies including smartphones and social media platforms radically extends what we can see for ourselves.

This point is epistemically significant not only in terms of the resultant knowledge expansion, but also in terms of the extension of ordinary individuals' epistemic autonomy. Whereas knowing what happened elsewhere in the world might have been impossible, or might have required taking others on their word, we can now, in a radically expanded array of cases, know by seeing for ourselves. This point is epistemically significant, as seeing for oneself—whether literally or in a more inclusive, intellectual sense of "seeing"—is highly valued in both folk and academic epistemology (Pritchard, 2016). Thus, even if it is possible to acquire knowledge through the testimony of others, one is likely to feel some pressure to see for oneself if this option is easily available. Suppose, for example, that one is curious as to the weather outside. One can come to know about the weather by either asking one's partner—who has just returned from outside—or by looking out the window. Even if both options would yield knowledge, most would, I suspect, prefer the former option. In effect, novel technologies have given us access to many small windows through which we can see things for ourselves. In doing so, such technologies have radically expanded what we can know by seeing for ourselves. But this expanded ability to know by seeing for oneself is not secure. As I argue in the remainder of this section, deepfakes pose a three-pronged threat to the acquisition of knowledge by seeing for oneself through video footage. Each of these threats corresponds to a condition that epistemologists tend to think must be satisfied in order to obtain knowledge. Elsewhere, Harris (2023) have discussed the three-pronged threat of social media trolls and bots, and I adopt the terminology of *deceptive, skeptical,* and *epistemic threats* here.

In the early days of Russia's 2022 invasion of Ukraine, a deepfake video appearing to show President Volodymyr Zelenskyy surrendering and calling on Ukrainian soldiers to lay down their arms emerged (Allyn, 2022). The video circulated widely on social media, and was inserted onto a Ukrainian news website by hackers. The apparent intention behind the video was to trick Ukrainian soldiers and civilians into believing that their President had surrendered, thereby facilitating an easier Russian victory. Put differently, the video appears to have been intended to deceive at least some viewers into forming a false belief. Although the deepfake was not well made, and proved ineffective, the video nonetheless illustrates the *deceptive threat* of deepfakes.

In general, deepfakes and other forms of misinformation pose a deceptive threat insofar as they are likely to produce false beliefs in an audience. Deceptive potential is perhaps the most obvious and immediate threat posed by deepfakes. Especially where time is a factor—immediately before an election, for instance—the threat of a deepfake causing deception on a mass scale looms large. Thus, efforts have been made to alert the public to the existence and deceptive potential of deepfakes. For example, in one widely-discussed case, the comedian Jordan Peele partnered with *Buzzfeed* to create a PSA about the deceptive threat of deepfakes (Mack, 2018).

Public awareness of deepfakes raises its own concerns. These concerns are well-put by Hany Farid, in a comment on the effect of the Zelenskyy deepfake:

> The next time the president goes on television, some people might think, 'Wait a minute — is this real?' (quoted in Allyn, 2022)

Farid's comment captures that the harmful effects of deepfakes do not end at the deceptive threat. As awareness of the deceptive potential of deepfakes increases, so too does the *skeptical threat*. In other words, audiences are likely to place less trust in the authority of video footage quite generally, and become less inclined to form beliefs on the basis of video footage, thereby reducing the chances of being deceived, but also limiting opportunities to form true beliefs. Notably, because non-deepfakes vastly outnumber deepfakes, and are likely to continue to do so, an increase in skepticism toward video footage—unless especially well-targeted toward inauthentic videos—would likely reduce true beliefs more than false beliefs.

The skeptical threat of deepfakes is developed most extensively by Regina Rini (2020). According to Rini, deepfakes threaten not only to reduce the perceived authority of video footage but—because video footage has historically been used to verify the authenticity of photos, testimony, and the other forms of evidence—the reduced credibility of video footage will have wider skeptical impacts. In short, if the perceived authority of video footage goes down, it will take the perceived authority of other forms of evidence down with it. Left unchallenged, the skeptical threat of deepfakes has the potential to reverberate widely.

Notably, the skeptical threat of deepfakes does not depend on paranoia or irrationality on the part of ordinary persons. Reducing one's trust in a source of evidence is a reasonable response to actual reductions in the significance of that evidence. Deepfakes threaten to reduce the significance of evidence and may, for this reason, be understood as posing an *epistemic threat*. The epistemic threat of deepfakes is captured neatly by Don Fallis (2021), who argues that deepfakes reduce the amount of information conveyed by video footage. Fallis's account of the threat of deepfakes is based on the following approach to informational signals, adopted from Brian Skyrms (2010):

> [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. (Fallis, 2021, p. 629)

On this approach, deepfakes reduce the amount of information conveyed by video footage by increasing the likelihood that there is video footage depicting an event despite the non-occurrence of that event.

In general, an epistemic threat is one that impedes the ability to satisfy the warrant condition on knowledge. Roughly speaking, knowledge is warranted true belief, where warrant is understood as whatever property distinguishes knowledge from mere true belief. By reducing the amount of information conveyed by video footage, deepfakes can be understood as threatening the acquisition of warrant, and hence knowledge, from even veridical video footage. The threat that deepfakes pose to the acquisition of warrant need not be understood (solely) in terms of information conveyance. Some epistemologists have noted that deepfakes can make the formation

of true beliefs based on video footage lucky (Harris, 2021, 2022; Matthews, 2023). Comparably to how fake barns in one's environment can make it a matter of luck that one forms a true belief that one is in the presence of a barn (Goldman, 1976), deepfakes can make it a matter of luck that one forms true, as opposed to false, beliefs based on video footage. Supposing that knowledge is incompatible with luck, the threat of deepfakes to warrant may be (partly) understood in terms of making it a matter of luck that one's beliefs are true.

We have seen that deepfakes jeopardize the acquisition of warrant from video footage. Additionally, by threatening to cause false beliefs—that is, by posing a deceptive threat—deepfakes threaten the truth condition on knowledge. Finally, as we have seen, deepfakes threaten to reduce the perceived authority of video footage—including authentic video footage. In this way, deepfakes further threaten knowledge by discouraging the formation of true beliefs based on video footage. In short, deepfakes pose a three-pronged threat to knowledge acquisition from video footage. Put differently, deepfakes threaten the ability to acquire knowledge by seeing for oneself through video footage.

3 Technological Solutions for Technological Problems

It is widely hoped that the three-pronged threat resulting from advancements in the development of realistic deepfakes can be countered by advancements in automated deepfake detection technologies (ADDs). ADDs might in principle be used to help remove deepfakes from various online environments, or to label videos as authentic or inauthentic. Additionally, traditional news organizations might rely on ADDs for fact-checking purposes. Researchers' efforts to develop ADDs have been creative and productive. For an excellent overview of these technologies and their shortcomings, from which I draw heavily in this section, see Mirsky and Lee (2022). In general, deepfake detection technologies are based on the principles that automated systems can pick up on subtle artifacts of fakery that are imperceptible, or at least not typically perceived, by human observers. Algorithms can be trained to detect specific anomalies, or to locate, without human direction beyond the provision of training sets, strategies for classifying footage as deepfaked or otherwise. Generally speaking, the former technologies are less promising insofar as any algorithm trained to detect a specific anomaly can be overcome by correcting for that specific anomaly (Mirsky & Lee, 2022).

There are many critiques to be raised against the development of ADDs for addressing the threat of deepfakes. First, algorithms can at best succeed at identifying deepfakes whose flaws resemble those in their training sets. But the detection of deepfakes has the structure of an arms race, pitting the development of deepfakes against the development of ADDs (Mirsky & Lee, 2022). Today's ADDs may be of no use against tomorrow's deepfakes. What is worse, outdated ADDs may provide a false sense of security, and confer an unwarranted air of authenticity on sophisticated deepfakes that manage to evade detection. Relatedly, it is worth noting that, even if a video is not a deepfake, it may nonetheless be manipulated using less sophisticated techniques (Harris, 2021). Technologies designed solely to detect

deepfakes will fail to detect, and potentially promote credulity toward, videos that are misleading in other ways. In other words, ADDs do not fully address the deceptive threat of manipulated videos, and can even be expected to facilitate deception in cases of false negative verdicts concerning undetected deepfakes and true negative verdicts concerning videos that are otherwise misleading.

The preceding concern for the use of ADDs is not merely speculative. Existing empirical research has shown that, where systems for labeling misinformation are in place, unlabeled misinformation enjoys a boost in perceived accuracy. This is the *implied truth effect* (Pennycook et al., 2020). The present concern for ADDs is that awareness of the use of such systems will lend perceived credibility to video content that evades detection—either because of advancements in deepfake technology or because the video in question, while misleading, is not a deepfake.

That ADDs are only one side in what is effectively an arms race also means that such technologies have limited potential for addressing the skeptical threat. Even if detection technologies are on par with the development of deepfakes themselves, skepticism may be grounded in the inability to verify this. To see this, suppose that the detection strategies that exist at a time perfectly identify all extent deepfakes as such. In this case, the fear may yet remain that some deepfakes have evaded deception. Notably, this fear will be most pronounced among those who recognize the limitations of ADDs described above.

I have thus far focused mainly on the potential for certain deepfakes to evade detection. But the potential for false positives is a further serious concern, and one that might severely decrease trust in the authenticity of real video footage. For a simple example, consider an ADD that works by recognizing unrealistic patterns in the lighting of the foreground and background. Such a system might flag certain authentic videos—especially those taken in unusual lighting conditions—as deepfakes. More generally, whatever features typically indicate that a given piece of footage is a deepfake might be present, in certain rare cases, even in authentic footage. In this way, algorithmic deepfake detection tools can encourage undue skepticism toward even authentic videos. Crucially, supposing that non-deepfakes continue to vastly outnumber deepfakes, even a miniscule rate of false positives would lead to the mislabeling of a vast number of non-deepfakes (Dolhansky et al., 2020). Adding to this concern is recent empirical evidence suggesting that warnings about deepfakes themselves reduce willingness to believe based on even non-deepfake video footage—in other words, that such warnings constitute a skeptical threat in their own right (Ternovski et al., 2021). Such findings accord with a wave of more recent work suggesting that efforts to warn about or otherwise reduce receptivity to misinformation can, by making the threat of misinformation more salient, reduce trust in even reliable information (Hameleers, 2023; Modirrousta-Galian & Higham, 2023; Van Duyn & Collier, 2019). These findings suggest that, even when ADDs are accurate, they may exacerbate the skeptical threat by drawing attention to the possibility that video footage is fake.

There are thus straightforward reasons to doubt that ADDs fully address the three-pronged threat of deepfakes. The reliance on such technologies has also been critiqued on more general grounds. According to Joshua Habgood-Coote (2023), reliance on such technologies amounts to a misguided form of *techno-solutionism*.

Habgood-Coote argues that the challenges raised by deepfakes are continuous with the challenges raised by earlier media manipulation techniques.[1] These are social challenges that call for social solutions. In particular, the threat of deepfakes exist against a backdrop of often warranted distrust in mainstream media and other institutions, and this threat cannot be fully addressed without confronting this distrust.

In the remainder of this paper, I begin to develop two further concerns for the use of ADDs to confront the threat of deepfakes. The first of these, discussed in Sect. 3, builds on Habgood-Coote's concerns about attempting to confront this threat without addressing the underlying social conditions under which deepfakes are likely to cause problems. Then, in Sects. 4 and 5, I argue that ADDs present a unique threat to epistemic autonomy.

4 The Trust Problem

In the preceding section, I provided a brief overview of ADDs and some of their shortcomings. It is worth reiterating, at this point, that the promise of such technologies is to pick up on indicators of fakery that are not perceptible, or not typically perceived, by human observers. In other words, where such detectors are successful, they will sometimes render judgments that are in tension with the perceptions of human observers. An ADD might declare a video that appears real to human observers to be inauthentic, and vice-versa. Thus, on the face of things, these technologies will only help humans to better distinguish between deepfakes and non-deepfakes, where humans trust[2] these technologies more than their own senses.

This initial gloss is too quick. One might put greater trust in one's own faculties than the abilities of an automated system, but nonetheless treat unexpected outputs of that system as grounds for examining a given video more closely. Suppose, for example, that one initially regards a given piece of video footage as authentic. Suppose that footage is flagged as fake by an ADD. Even if one does not defer to the detector, one might nonetheless think that the system's judgment is a sufficient reason to scrutinize the video more carefully.[3] In this way, ADDs might serve to prompt more careful consideration of evidence by humans. Additionally, over time, comparing one's judgments to that of ADDs may help one to develop one's own discriminatory abilities.

[1] Britt Paris and Joan Donovan (2019) provide an excellent overview of deepfakes and their continuity with earlier media manipulation techniques.

[2] Philosophers sometimes invoke thick conceptions of trust according to which, for example, trust involves an expectation of goodwill toward the truster on the part of the trustee, and failures on the part of the trustee are appropriately met with feelings of resentment. Here, I understand trust in a thin sense, according to which trust in a person or thing amounts to a willingness to rely on that person or thing for some specified purposes.

[3] The point here resembles the point, common in literature on the social epistemology of disagreement, that initial disagreement is grounds to reconsider one's evidence even if it is not grounds for abandoning one's belief.

Still, even these limited forms of reliance on AI require individuals to place some degree of trust in the outputs of AI. A significant concern for the reliance on AI to combat deepfakes is thus that such trust may be difficult to come by. Much has been made of declining trust in the mainstream media and other institutions, but it is worth noting that trust in the technology sector is also limited. In the United States, for example, trust in the technology sector has declined rapidly since 2017 (PAC, 2022). Insofar as users associate the reliability of ADDs with the trustworthiness of the sector that produces and—as is likely in the case of social media platforms—employs them, there is thus reason to doubt that individuals will place much faith in the outputs of such tools. This challenge is exacerbated by the fact that some footage vetted by ADDs is likely to be politically charged. Consequently, the familiar lack of trust characteristic of political polarization is likely to compromise individuals' willingness to rely on the outputs of AI deepfake detection tools, especially where those outputs challenge existing political outlooks. Even if one accepts AI-generated labels of accuracy that accord with one's own political attitudes—for example those that judge to be false videos that cast members of one's own party in an unfavorably light—one will likely be less inclined to accept politically discordant judgments. Resistance of this sort might reflect motivated reasoning, but might also involve a rational process whereby one's prior beliefs impact the perceived chances of video footage being veridical. In general, it is difficult to determine whether resistance to evidence reflects motivated reasoning or the influence of prior beliefs (Williams, 2023) but, while these explanations differ in their rationality, both can reduce receptivity to good evidence.

As I suggested above, even limited reliance on AI as a prompt to more careful consideration might be compromised by a lack of trust in ADDs and the people and platforms that deploy them. Thus, for example, a lack of trust may render one unlikely to take a second look at a piece of video footage when that video is flagged as inauthentic. In such a case, however, trust in the reliability of AI tools need only be partial and temporary. Such tools may lead individuals to recognize perceptible imperfections—or the perceptible absence of these—to which they have not previously attended.

However, the challenge will be most severe if deepfakes reach a stage of development at which deepfakes are, for human observers, entirely indistinguishable from non-deepfakes. Supposing that ADDs are nonetheless reliable at this stage, correct judgments as to authenticity will require, of human persons, more than the temporary suspension of judgment as to the authenticity of video content. Rather, to assess video footage correctly, human persons will need to largely disregard the evidence of their own senses in favor of the outputs of an automated system whose functioning they do not well understand. Notably, this lack of understanding need not be due to failings on the part of the individuals in question. Given the opacity of AI, individuals may through no fault of their own be unable to understand how deepfakes are detected or what features of video footage ADDs pick up on. Under these conditions, it is to be expected that individuals will be unconvinced by outputs of deepfake detection systems that conflict with their own judgments. This is especially true of those outputs that are misaligned with the deeper convictions of the individuals in question.

In this section, I have argued that the effectiveness of ADDs is likely to be compromised by a lack of trust in these systems and the human individuals and institutions with which they are associated. In the next two sections, I argue that, even if individuals are willing to rely on such systems, reliance on such systems is problematic insofar as it jeopardizes individuals' epistemic autonomy.

5 Reliance on Artificial Intelligence

In Sect. 1, I suggested that recent advances in technology—especially those that facilitate the recording, streaming, and sharing of video footage—have expanded what individuals can see for themselves and, in this way, have expanded ordinary individuals' epistemic autonomy. But I have also argued that deepfakes reduce the efficacy of seeing for oneself as a way of obtaining knowledge. Deepfakes thus reduce the epistemic value of epistemic autonomy or, more modestly, of the realization of epistemic autonomy through seeing for oneself. In this way, deepfakes seem to discourage the exercise of epistemic autonomy. One might think that, at least in principle, ADDs solve this problem by helping to remove or label deepfakes. But this is too quick. I now argue that, when it comes to epistemic autonomy, the cure afforded by ADDs may well be worse than the disease.

Over the past few decades, epistemologists have devoted a great deal of attention to questions concerning epistemic autonomy. Inquiries in this area have considered the nature of epistemic autonomy (Carter, 2020; Elgin, 2021; Encabo, 2008; Grasswick, 2018; King, 2021; Matheson & Lougheed, 2021; Zagzebski, 2012) and whether epistemic autonomy is a worthy ideal (Ahlstrom-Vij, 2013, Chapter 4; Battaly, 2021; Coady, 2002; Dellsén, 2021; Hardwig, 1985; Matheson, 2022; Roberts and Jay Wood, 2007). Very plausibly, whether epistemic autonomy has value depends on how it is defined. To determine whether the impact of ADDs on epistemic autonomy is a cause for concern, it will first be necessary to consider whether there is a valuable form of epistemic autonomy.

To begin, consider some recent proposals as to the nature of epistemic autonomy. According to Elizabeth Fricker, the epistemically autonomous person "takes no one else's word for anything, but accepts only what she has found out for herself, relying only on her own cognitive faculties and investigative and inferential powers" (Fricker, 2006, p. 225). Sandy Goldberg characterizes epistemic autonomy similarly, writing that an epistemically autonomous person "judges and decides for herself, where her judgments and decisions are reached on the basis of reasons which she has in her possession, where she appreciates the significance of these reasons, and where (if queried) she could articulate the bearing of her reasons on the judgment or decision in question" (Goldberg, 2013, p. 169). According to these *austere* conceptions, epistemic autonomy might be equated roughly with epistemic independence. An epistemically autonomous person might be prompted to investigate certain questions by the influence of others, but that person will not treat the judgments of others as reasons for her beliefs. In what follows, I will use the term 'reliance' to refer narrowly to treatment of others' judgments in this way—where *judgment* is understood broadly enough to

include both human assertions and the outputs of artificially intelligent systems. While other forms of reliance arguably undermine autonomy, it is this specific form of reliance that most plausibly threatens epistemic autonomy.

So understood, epistemic autonomy would likely lead to at least one of two problems. On the one hand, refusal to rely on others might leave one with little material on which to base one's beliefs. One might thus be extremely limited in what one believes (Roberts and Jay Wood, 2007, pp. 259–260). More realistically, however, one's reach would exceed one's grasp, and thus one would persist in holding a broad range of beliefs untethered and untutored by the reasons of others (Hardwig, 1985). Arguably, this is the problem that occurs when conspiracy theorists attempt to "think for themselves" or "do their own research" rather than relying on epistemic authorities (Ballantyne et al., 2022; Buzzell & Rini, 2023; Levy, 2022). Thus, on this austere approach to epistemic autonomy, it is unclear that epistemic autonomy is a worthy aim.

In light of these concerns about epistemic autonomy understood in austere terms, some epistemologists have proposed that epistemic autonomy can be reconciled with reliance on others (Roberts and Jay Wood, 2007, p. 260). Nathan King summarizes this approach well, writing that:

> Autonomy requires thinking *for* ourselves, but not *by* ourselves. (2021, p. 88).

One way to reconcile epistemic autonomy with reliance on others is to contrast autonomy with heteronomy, rather than with dependence (Encabo, 2008). On this approach, reliance on others is consistent with the preservation of one's epistemic autonomy insofar as this reliance is expressive of one's identity. Thus, by selecting those to rely upon based on one's beliefs, values, and other features of one's identity, one can exercise epistemic autonomy through one's reliance on others (C. Z. Elgin, 2013; Grasswick, 2018).

To better grasp this point, consider why, intuitively, reliance on others appears to threaten autonomy. Zagzebski makes the point compellingly in the following passage:

> Many people who live for a time in another country, or study their wisdom literature in depth, find that their trust in their own beliefs is undermined. It is common to think, "I would have had different beliefs if I had grown up in a different place, and it is an accident of history that I have the beliefs I have. I could have been Hindu, Muslim, Buddhist, Christian, atheist, or many other things." The same line of thought applies to philosophical positions and attitudes about political arrangements. I am a believer in libertarian free will but I could have been a determinist. I am a believer in Western democracy, but I could have believed in Islamic theocracy. (2012, p. 245)

Once one recognizes the extent of one's epistemic reliance on others, it is very easy to conclude that one's beliefs are less an expression of oneself, and more an accident of one's social epistemic circumstances. However, the tension between reliance on others and epistemic autonomy is lessened when one recognizes

oneself as the author (or co-author) of one's social epistemic circumstances. We are not simply victims of social epistemic circumstance, but have the opportunity to partly construct the social epistemic niches that ultimately shape our beliefs. Epistemic autonomy is thus arguably consistent with reliance on others—and indeed is furthered by such reliance—so long as one's patterns of reliance reflect features of one's identity. Epistemic autonomy of this *broad* kind is consistent with reliance on, and regulation by, the reasons of others. Notably, the relevant others may include epistemic authorities, and thus this sort of epistemic autonomy need not involve thinking for oneself. So understood, epistemic autonomy is far more conducive to epistemic success than epistemic autonomy understood in austere terms, at least so long as one chooses the objects of one's reliance wisely.

One might argue that selectiveness about objects of reliance is better understood in terms of epistemic interdependence than epistemic autonomy (Levy, 2023). While I am sympathetic to this suggestion, I think that our choices about who to associate with—including for exchanges of information—are important ways in which we assert features of ourselves. This is perhaps most clear when one's choices buck expectations in one's immediate social environment. For example, when a teenager in a deeply religious community opts to pursue and to rely upon secular sources of information that are not well respected in his or her community, the teenager plausibly exhibits a form of autonomy. Especially insofar as this process involves investigations into the credibility of various sources, the selection and ultimate reliance on sources reflects the intellectual character and effort of the individual. But, even in cases in which one's selection of sources involves no act of teenage rebellion, or indeed where one's choices are well-aligned with the choices made by other members of one's community, such choices reflect features of oneself. For this reason, I think it is appropriate to think that choices about who to trust are appropriately construed as exercises of epistemic autonomy.

Some approaches to epistemic autonomy considered thus far are stated in terms of reliance on other persons. However, in light of increasing reliance on the outputs of artificially intelligent systems as bases for belief, it is worth considering whether such definitions of epistemic autonomy are too restrictive. Consider a few examples. Within medicine, diagnoses and recommendations are increasingly informed by AI. Many companies use chat bots to communicate information to consumers, and this practice can be expected to expand as chatbots become increasingly sophisticated. More recently, it has been suggested that sophisticated chatbots might replace browser searches as the default mode of retrieving information online (Wong, 2023).

I will argue below that a particular form of reliance on AI—reliance on the output of ADDs—compromises epistemic autonomy. It might be thought that *only* reliance on other human persons presents a *prima facie* challenge to epistemic autonomy, and thus that reliance on AI cannot threaten epistemic autonomy. However, there are at least two reasons to think that reliance on artificial intelligence sometimes threatens epistemic autonomy. First, there is a sense in which dependence on AI *does* involve reliance on other human persons. One form of reliance on others is direct, and involves taking the word of others—or some similar expression on the part of others—as a reason to believe a specific proposition. For example, I directly rely on you when I take your word for it that it that your clock is accurate. However,

reliance on others can also be indirect. When I come to believe that it is 11:03AM based on the reading of your clock and your prior assurance that the clock is generally accurate, I rely on you indirectly with respect to my belief about the time. Such indirect reliance is pervasive, and need not concern the use of material tools and instruments. If I take your assurance that some heuristic or problem-solving method yields correct results, I rely indirectly on you concerning my belief in those results. Consider for example the simple long division algorithm taught to children as a way of breaking down relatively complex calculations into a series of simpler steps. If I accept your assurance that following these steps will yield a correct answer to a particular mathematical problem, I am thereby indirectly reliant upon you concerning my belief in that answer (cf. Levy, 2007, p. 186). Notably, indirect reliance on others may dissipate as one confirms the accuracy of an instrument, method, or the like.

When one relies on AI, one relies indirectly on those who have developed it and who attest to its reliability. Even this indirect form of reliance poses at least a potential challenge to the valuable form of epistemic autonomy described above. I expand on this point below. For the present, it bears noting that, if reliance on artificial intelligence potentially threatens epistemic autonomy for the reason just described, so too does reliance on those tools, instruments, methods, and so on whose accuracy we accept on the word of others.

There is, however, a second and more distinctive reason to suspect that reliance on artificial intelligence sometimes threatens epistemic autonomy. Many interactions with systems powered by artificial intelligence increasingly resemble interactions with human persons. It is possible in principle to interact with a customer service chatbot or online social bot without realizing that one is not in conversation with another human. One way to develop this point is to argue that systems powered by artificial intelligence can assert or issue testimony.[4] However, for present purposes, it is enough to note that, from a human interactant's perspective, interactions with such systems may closely resemble interactions with other human persons. Within these interactions, a person might behave just as on would in interactions with another person. Suppose that one allows such interactions to inform one's beliefs. It would appear arbitrary to think that epistemic autonomy *is not* potentially threatened in such interactions, but *is* potentially threatened by the acquisition of information through comparable interactions with human persons.

The above argument requires some clarification. ADDs are unlikely, in typical cases, to provide information through interactions comparable to interactions with other human persons. Insofar as ADDs inform individuals' judgments about the authenticity of video footage, it is likely that they will do so by generating simple labels or scores for pieces of video footage. The point of the preceding argument is hence not that ADDs present a potential threat to epistemic autonomy because interactions with such systems resemble interactions with human persons.[5] Instead,

[4] Thus far little philosophical attention has been devoted to this issue, but see Ori Freiman and Boaz Miller (2020) and Billy Wheeler (2020).

[5] It is worth noting, however, that some ways of relying on human persons resemble the way in which we are likely to rely on ADDs. In particular, we now sometimes form beliefs on the basis of simple labels and scores—generated by human persons—concerning the accuracy of information.

the argument above serves to show that it is implausible that reliance on AI systems in general cannot compromise epistemic autonomy. Thus, that ADDs are powered by artificial intelligence is not by itself a reason to deny that such systems potentially compromise epistemic autonomy.

The preceding arguments suggest two ways in which we might rely on artificial intelligence. First, we might rely on artificial intelligence as we would other tools and methods, thereby relying indirectly on human designers. Second, we might treat artificial intelligence as functionally similar to ordinary human persons, rather than tools, instruments, or methods. Whether we construe artificial intelligence as more instrument-like or more person-like has implications for how beliefs based on information retrieved from AI are justified (Duede, 2022). For present purposes, however, the most important point is that reliance on AI is at least a candidate for threatening epistemic autonomy.

6 Deepfake Detection, Self-Trust, and Epistemic Autonomy

In the previous section, I argued that relying on AI presents at least a potential threat to epistemic autonomy. In this section, I argue that reliance on ADDs, in particular, would in many cases compromise epistemic autonomy. On an austere construal of epistemic autonomy, the argument that reliance on ADDs and other forms of AI would compromise epistemic autonomy would be a simple matter. One who relies on AI does not exclusively depend on what she has found out for herself, and does not trust solely in her own cognitive faculties and skills. However, as I now argue, some forms of reliance on ADDs would compromise even the broad form of epistemic autonomy.

On the broad construal, epistemic autonomy is compatible with reliance on others insofar as this reliance reflects features of oneself, including one's beliefs, values, and so on. When construed in this way, some ways of relying on ADDs might well reflect oneself in relevant respects. For example, if one identifies closely with a particular ADD—because for example one had a role in its development or is deeply familiar with the process by which it was trained or its track record of success—one's reliance on that ADD may reflect one's identity.

However, such cases would be highly atypical. Suppose, for example, that ADDs come to be widely employed by social media platforms to label video footage. Realistically, even if platforms emphasize transparency about the use of ADDs, most users will have little understanding about how such systems work, how they are developed, and so on. As we have seen, this lack of understanding need not be the fault of ordinary users but will at least in many cases be an inevitable consequence of the opacity of ADDs. In these conditions, the reliance on AI is not an expression of epistemic autonomy. It might be thought that individuals can exercise epistemic autonomy through their indirect reliance on the developers of ADDs, or the platforms that vouch for the reliability of these systems. I return to this point in the concluding section. For the present, it bears recalling that most ordinary individuals

have little trust in the technology sector, and hence would be unlikely to identify with the developers of ADDs in this way.

To further develop the argument that ADDs are likely to compromise the epistemic autonomy of many ordinary persons, it is worth comparing reliance on ADDs to the reliance on AI within a more familiar and more extensively theorized context. Artificial intelligence promises to revolutionize processes of diagnosis and prediction in medicine,[6] and will increasingly assist in making treatment recommendations. Because the detection of deepfakes is most closely analogous to diagnostic classification, I focus here on the diagnostic role of AI. Philosophers and other researchers have devoted considerable attention to whether it is possible (Ferrario et al., 2021; Ryan, 2020), and if so appropriate (Alvarado, 2022; Durán & Jongsma, 2021; Hatherley, 2020), for patients and medical professional to trust the outputs of diagnostic AI. One consideration relevant to the latter question is whether reliance on diagnostic AI would compromise human epistemic autonomy.

In the short term, the main use of artificial intelligence within diagnostic contexts is to analyze large quantities of data for the purposes of early screening. For example, diagnostic algorithms might be used to provide preliminary analyses of vast numbers of medical images, selecting a subset of patients for further review by human physicians (Hunter et al., 2022). This limited diagnostic role for artificial intelligence leaves the epistemic autonomy of human physicians intact, especially as it serves mainly to better focus, rather than supplant, physicians' cognitive labor. However, as diagnostic artificial intelligence advances, AI systems are likely to increasingly generate more reliable diagnoses based on patient information and medical images than human physicians. Indeed, the equivalence or superiority of certain algorithms to human experts has already been demonstrated in some diagnostic contexts (Menzies et al., 2023; Tschandl et al., 2019). Cases are likely to arise in which the independent judgments of human physicians conflict with those of artificially intelligent diagnostic systems. In such cases, reliance by physicians on AI diagnoses presents at least a *prima facie* threat to the epistemic autonomy of physicians. One might thus suspect that, even if artificially intelligent deepfake detectors pose a threat to epistemic autonomy, it is not a *unique* threat.

However, there are several reasons to be especially concerned about the compatibility of reliance on ADDs with epistemic autonomy. First, in a future in which deference to AI in medical diagnostics becomes commonplace, professionalization into the medical community will likely involve education concerning the principles and reliability of diagnostic algorithms. In fact, incorporation of information concerning artificial intelligence into medical curricula is already underway. For this reason, future reliance on diagnostic algorithms within the medical context will likely be aligned with the beliefs and values of members of the medical community, and hence need not conflict with their epistemic autonomy in the broad sense. This contrasts with the reliance on ADDs to judge the authenticity of video footage, which cannot be expected to be accompanied by systematic education concerning the reliability and value of such systems.

[6] It bears noting, though, that some claims as to the revolutionary potential of AI in the medical context appear to be overly optimistic (Smith, 2023, pp. 211–212).

Second, AI will be used within medical diagnostics to make classification judgments concerning non-ideologically-charged matters[7]—the presence or absence of malignant tumors—for example. In contrast, as we have seen, ADDs can be expected to provide labels for politically and otherwise ideologically significant content. Thus, the outputs of ADDs are more likely than the outputs of medical diagnostic algorithms to include judgments that conflict with beliefs significant to individuals' identities.

Finally, the application of medical diagnostic AI will be restricted to specific types of data in specific contexts. In contrast, ADDs can be applied to video footage[8] depicting all sorts of events. As emphasized in Sect. 1, such footage has long been a trusted way of forming beliefs, and hence to affirm the necessity of reliance on ADDs would be to acknowledge the inadequacy of long-standing and general belief-forming practices. In effect, to rely on an ADD instead of one's own faculties is to accept the vulnerability of the process by which many of one's most confidently-held beliefs have been formed. In this way, the reliance on ADDs fosters a general kind of self-doubt that is inimical to epistemic autonomy both within the context of reliance and beyond. The risk is that, in coming to rely on ADDs, we will cease to regard seeing for oneself through video footage, unaided by AI tools, as a way of coming to know. Instead, when it comes to the formation of beliefs through perception of video footage, we will regard the evidence of one's senses as an insufficient basis on which to form beliefs and will thus rely more heavily on external aids and less on our own perceptual judgements. Even on a broad construal of epistemic autonomy, this externalization of reliance will, insofar as it fails to reflect features of ourselves, amount to ceding some epistemic autonomy. Thus, even if ADDs can address the three-pronged threat of deepfakes, reliance on such tools may result in the degradation of epistemic autonomy. This is not a decisive consideration against the use of such technologies, but it is a significant cost that ought to be factored into decisions about how and if to deploy such technologies.

7 Concluding Remarks

In this paper, I have argued that technological solutions offer little hope of satisfactorily addressing the three-pronged threat of deepfakes. In part, this is because the value of ADDs is offset by corresponding developments of deepfake technologies themselves. Consequently, ADDs run the risk of improperly providing stamps of approval for advanced deepfakes. Similarly, ADDs cannot provide assurance that advanced deepfakes have not gone undetected. At the same time, reliance on ADDs has the potential to undermine self-trust and epistemic autonomy.

[7] To say that these matters are not ideologically charged is not to say that they are value-neutral. For example, diagnostic contexts plausibly involve judgments about the relative importance of avoiding different kinds of error (cf. Douglas, 2000).

[8] While I focus on ADDs here, there are comparable technologies for detecting manipulated photos and audio content.

The main upshot of this paper is thus that the prospects for purely technological solutions to the threats of deepfakes are dim. Yet these threats are potentially dire. Some solution is needed. Given the scale of the problem, technological solutions—especially those that help to extend and focus the efforts of human persons—undoubtedly have some role to play. But, as Habgood-Coote suggests (2023), we ought not expect future technology to offer a silver bullet to the challenges posed by deepfakes. Rather, confronting the threats of deepfakes plausibly requires confronting the social and institutional ills that allow for deepfakes to pose potent threats. Even if one cannot trust video footage by itself, one might trust certain sources and channels to deliver authentic footage (Harris, 2021). Discovering and relying upon trustworthy sources is itself an expression of epistemic autonomy in the broad sense. While locating such sources is no easy task and will likely require increased transparency on the part of sources and channels, addressing the threat of deepfakes through allocations of trust that better reflect the trustworthiness of sources holds the promise of promoting epistemic success while simultaneously preserving and enhancing the epistemic autonomy of ordinary persons.

Acknowledgements Thanks to audiences at the 2023 Visual Trust conference on image-making in Barcelona and at the 2023 Changing Minds Online conference in Tilburg for their feedback on material included in this paper.

Authors' contributions KRH is the sole author of the content of this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This is a publication in the context of INTERACT!, funded by the ministry of culture and science of North Rhine Westphalia (PROFILNRW-2020–135).

Availability of data and material N/A.

Declarations

Ethics approval and consent to participate N/A

Consent for publication N/A.

Competing interests The author has no relevant financial or non-financial interests to disclose.

References

Ahlstrom-Vij, K. (2013). Epistemic Paternalism. *Palgrave Macmillan UK*. https://doi.org/10.1057/9781137313171

Allyn, B. (2022). Deepfake video of Zelenskyy could be "tip of the iceberg" in info war, experts warn. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Alvarado, R. (2022). Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics, 36*(2), 121–133. https://doi.org/10.1111/bioe.12959

Ballantyne, N., Celniker, J. B., & Dunning, D. (2022). "Do Your Own Research." *Social Epistemology*, 1–16. https://doi.org/10.1080/02691728.2022.2146469

Battaly, H. (2021). Intellectual Autonomy and Intellectual Interdependence. In J. Matheson & K. Lougheed, *Epistemic Autonomy* (1st ed., pp. 153–172). Routledge. https://doi.org/10.4324/9781003003465-12

Buzzell, A., & Rini, R. (2023). Doing your own research and other impossible acts of epistemic superheroism. *Philosophical Psychology, 36*(5), 906–930. https://doi.org/10.1080/09515089.2022.2138019

Carter, J. A. (2020). Intellectual autonomy, epistemic dependence and cognitive enhancement. *Synthese, 197*(7), 2937–2961. https://doi.org/10.1007/s11229-017-1549-y

Cavedon-Taylor, D. (2013). Photographically based knowledge. *Episteme, 10*(3), 283–297. https://doi.org/10.1017/epi.2013.21

Coady, C. A. J. (2002). Testimony and intellectual autonomy. *Studies in History and Philosophy of Science Part A, 33*(2), 355–372. https://doi.org/10.1016/S0039-3681(02)00004-3

Cox, J. (2019). Most deepfakes are used for creating non-consensual porn, not fake news. *Vice News*. Retrieved March 3, 2022, from https://www.vice.com/en/article/7x57v9/most-deepfakes-are-porn-harassment-not-fake-news

Dellsén, F. (2021). We Owe It to Others to Think for Ourselves. In J. Matheson & K. Lougheed, *Epistemic Autonomy* (1st ed., pp. 306–322). Routledge. https://doi.org/10.4324/9781003003465-21

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv:2006.07397

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science, 67*(4), 559–579.

Duede, E. (2022). Instruments, agents, and artificial intelligence: Novel epistemic categories of reliability. *Synthese, 200*(6), 491. https://doi.org/10.1007/s11229-022-03975-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-2020–106820. https://doi.org/10.1136/medethics-2020-106820

Elgin, C. (2021). The Realm of Epistemic Ends. In J. Matheson & K. Lougheed, *Epistemic Autonomy* (1st ed., pp. 55–70). Routledge. https://doi.org/10.4324/9781003003465-5

Elgin, C. Z. (2013). Epistemic agency. *Theory and Research in Education, 11*(2), 135–152. https://doi.org/10.1177/1477878513485173

Encabo, J. V. (2008). Epistemic merit, autonomy and testimony. *Theoria, 23*(61), 54–56.

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology, 34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics, 47*(6), 437–438. https://doi.org/10.1136/medethics-2020-106922

Foer, F. (2018). The era of fake video begins. *The Atlantic*. Retrieved August 4, 2019, from https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/

Freiman, O., & Miller, B. (2020). Can Artificial Entities Assert? In S. Goldberg (Ed.), *The Oxford Handbook of Assertion* (pp. 413–434). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190675233.013.36

Goldberg, S. (2013). Epistemic Dependence in Testimonial Belief, in the Classroom and Beyond: Epistemic Dependence in Testimonial Belief. *Journal of Philosophy of Education, 47*(2), 168–186. https://doi.org/10.1111/1467-9752.12019

Goldman, A. I. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy, 73*(20), 771. https://doi.org/10.2307/2025679

Grasswick, H. (2018). Epistemic Autonomy in a Social World of Knowing. In H. Battaly (Ed.), *The Routledge Handbook of Virtue Epistemology* (1st ed., pp. 196–208). Routledge. https://doi.org/10.4324/9781315712550-17

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese, 201*(103), 1–23. https://doi.org/10.1007/s11229-023-04097-3
Hameleers, M. (2023). The (Un)Intended Consequences of Emphasizing the Threats of Mis- and Disinformation. *Media and Communication*, 11(2). https://doi.org/10.17645/mac.v11i2.6301
Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy, 82*(7), 335. https://doi.org/10.2307/2026523
Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese, 199*(5–6), 13373–13391. https://doi.org/10.1007/s11229-021-03379-y
Harris, K. R. (2022). Real Fakes: The Epistemology of Online Misinformation. *Philosophy & Technology, 35*(3), 83. https://doi.org/10.1007/s13347-022-00581-9
Harris, K. R. (2023). Liars and Trolls and Bots Online: The Problem of Fake Persons. *Philosophy & Technology, 36*(2), 35. https://doi.org/10.1007/s13347-023-00640-9
Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935
Hunter, B., Hindocha, S., & Lee, R. W. (2022). The Role of Artificial Intelligence in Early Cancer Diagnosis. *Cancers, 14*(6), 1524. https://doi.org/10.3390/cancers14061524
King, N. L. (2021). *The Excellent Mind: Intellectual Virtues for Everyday Life* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190096250.001.0001
Levy, N. (2007). Radically Socialized Knowledge and Conspiracy Theories. *Episteme, 4*(2), 181–192. https://doi.org/10.3366/epi.2007.4.2.181
Levy, N. (2022). Do your own research! *Synthese, 200*(5), 356. https://doi.org/10.1007/s11229-022-03793-w
Levy, N. (2023). Against Intellectual Autonomy: Social Animals Need Social Virtues. *Social Epistemology*, 1–14. https://doi.org/10.1080/02691728.2023.2177521
Mack, D. (2018). This PSA about fake news from barack obama is not what it appears. *BuzzFeed News*. Retrieved March 20, 2022 from https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed
Matheson, J. (2022). Why Think for Yourself? *Episteme*, 1–19. https://doi.org/10.1017/epi.2021.49
Matheson, J., & Lougheed, K. (2021). Introduction. In J. Matheson & K. Lougheed, *Epistemic Autonomy* (1st ed., pp. 1–18). Routledge. https://doi.org/10.4324/9781003003465-1
Matthews, T. (2023). Deepfakes, Fake Barns, and Knowledge from Videos. *Synthese, 201*(2), 41. https://doi.org/10.1007/s11229-022-04033-x
Menzies, S. W., Sinz, C., Menzies, M., Lo, S. N., Yolland, W., Lingohr, J., Razmara, M., Tschandl, P., Guitera, P., Scolyer, R. A., Boltz, F., Borik-Heil, L., Herbert Chan, H., Chromy, D., Coker, D. J., Collgros, H., Eghtedari, M., Corral Forteza, M., Forward, E., … Kittler, H. (2023). Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: A multicentre, prospective, diagnostic, clinical trial. *The Lancet Digital Health*, 5(10), e679–e691. https://doi.org/10.1016/S2589-7500(23)00130-9
Mirsky, Y., & Lee, W. (2022). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys, 54*(1), 1–41. https://doi.org/10.1145/3425780
Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General.* https://doi.org/10.1037/xge0001395
Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology, 22*(2), 133–140. https://doi.org/10.1007/s10676-019-09522-1
PAC. (2022). 2022 Public affairs pulse survey report: what Americans think about business and government [White Paper]. Retrieved October 8, 2022 from https://pac.org/wp-content/uploads/2022/09/Pulse_Survey_Report_2022.pdf
Paris, B., & Donovan, J. D. (2019). *Cheap Fakes* (p. 47). Data & Society Research Institute.
Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science, 66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478
Pritchard, D. (2016). Seeing it for oneself: Perceptual knowledge, understanding, and intellectual autonomy. *Episteme, 13*(1), 29–42. https://doi.org/10.1017/epi.2015.59
Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint, 20*(24), 1–16.

Rini, R., & Cohen, L. (forthcoming). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy, 22*(2), 143–161.

Roberts, R. C., & Wood, W. J. (2007). *Intellectual virtues: An essay in regulative epistemology*. Oxford University Press.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information* (1st ed.). Oxford University Press Oxford. https://doi.org/10.1093/acprof:oso/9780199580828.001.0001

Smith, G. (2023). *Distrust: Big Data, Data-Torturing, and the Assault on Science*. Oxford University Press.

Ternovski, J., Kalla, J., & Aronow, P. M. (2021). Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments [Preprint]. Open Science Framework. 10.31219/osf.io/dta97

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., … Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), 938–947. https://doi.org/10.1016/S1470-2045(19)30333-X

Van Duyn, E., & Collier, J. (2019). Priming and Fake News: The Effects of Elite Discourse on Evaluations of News Media. *Mass Communication and Society, 22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Walton, K. L. (1984). Transparent Pictures: On the Nature of Photographic Realism. *Critical Inquiry, 11*(2), 246–277.

Warzel, C. (2018). Believable: The terrifying future of fake news. *Buzzfeed News.* Retrieved March 6, 2023, from https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news

Wheeler, B. (2020). Reliabilism and the Testimony of Robots. *Techné: Research in Philosophy and Technology*, *24*(3), 332–356. https://doi.org/10.5840/techne202049123

Williams, D. (2023). The case for partisan motivated reasoning. *Synthese, 202*(3), 89. https://doi.org/10.1007/s11229-023-04223-1

Wong, M. (2023). AI search is a disaster. *The Atlantic*. Retrieved March 3, 2023, from https://www.theatlantic.com/technology/archive/2023/02/google-microsoft-search-engine-chatbots-unreliability/673081/

Yetter-Chappell, H. (2018). Seeing through eyes, mirrors, shadows and pictures. *Philosophical Studies, 175*(8), 2017–2042. https://doi.org/10.1007/s11098-017-0948-8

Young, G. (2021). *Fictional immortality and immoral fiction*. Lexington Books.

Zagzebski, L. T. (2012). Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780199936472.001.0001