Deepfakes and Democracy: A Catch-22?

Dan Cavedon-Taylor

dan.cavedon.taylor@open.ac.uk

The Open University

Forthcoming in the *Journal of the American Philosophical Association*

Please reference/cite final version only

Abstract: Deepfakes are AI-generated media. When produced competently, they are near-indistinguishable from genuine recordings and may mislead viewers about the actions of the individuals they depict. For this reason, it is thought to be only a matter of time before deepfakes have deleterious consequences for democratic procedures, elections in particular. But this pessimistic view about deepfakes and their relation to democracy is flawed whether it means to pick out current deepfakes or future ones. Rather than advocating for an optimistic view in its place, I outline the opposite: a nihilistic account of deepfakes and their relation to democracy. On the nihilistic view, the harms that deepfakes pose for democracy are significantly more serious than those implied by the pessimistic view. Nihilism says that the real threat that deepfakes pose for democracy is that their existence counts against reforming current politics to be more truth-oriented.

Keywords: deepfakes; democracy; social media; voter behaviour

1. Introduction

Deepfakes are AI-generated media. They include still images, audio files and moving images, either individually or in combination. When produced competently, deepfakes are highly realistic. They mimic, with near-perfect accuracy, the perceptual properties of genuine recordings of actual events involving real people. And they are only getting more convincing.

For instance, still deepfakes mimic photographs: a deepfake still image of you shaking hands with Marine Le Pen would be near-indistinguishable from a photograph of you shaking hands with Marine Le Pen. Audio deepfakes mimic audio recordings: an audio deepfake of you drunkenly shouting racial slurs would be near-indistinguishable from an audio recording of you drunkenly shouting racial slurs. Video deepfakes mimic video recordings: a video deepfake of you assaulting a police officer would be near-indistinguishable from a video recording of you assaulting a police officer. While generating deepfakes like these requires specialist knowledge, sharing them online is child's play.

The production and circulation of deepfakes has the potential to cause numerous moral harms. These have been variously catalogued by philosophers and include reputational harms, harms related to sexual objectification, gaslighting, and more besides (Öhman 2020; de Ruiter 2021; Harris 2021; Young 2021; Rini & Cohen 2022; Benn forthcoming). However, there is another alleged deleterious consequence of deepfakes, one that is conceptually distinct from, though in practice often overlapping with, their capacity to cause moral harm: deepfakes are said to be a direct threat to democracy.

This idea is widely defended, by philosophers (Rini 2020), legal scholars (Chesney & Citron 2020), political theorists (Pawelec 2022) and journalists (Frum 2020). It is also *prima facie* intuitive. Consider the three fictional examples above. If you were a candidate for political leadership somewhere, the production and circulation of those deepfakes might discredit *you* in the eyes of the electorate while simultaneously being an afront to *their* autonomy to make informed voting decisions (e.g., by causing then to hold false beliefs about

your actions and character). Thus, such deepfakes would cause moral harms in election contexts to both the depicted candidate and the electorate (Diakopoulos & Johnson 2021). But, in addition, the creation and dissemination of such media would seem to undermine democratic processes and procedures themselves. For what is also problematic about deepfakes, in the context of elections at least, is something epistemic: the representational contents of deepfakes are fabricated, yet such media look and sound like genuine recordings. Worse still, deepfakes are commonly presented *as* recordings by those who circulate them and so constitute disinformation. This makes deepfakes epistemically pernicious in election contexts, no less than morally so: they risk causing voters to have false beliefs about candidates and their actions, thereby undermining informed voting. Call this view of deepfakes and their relationship to democracy 'the pessimistic view'.

In this paper, I argue against the pessimistic view. I claim that it misunderstands the relationship between deepfakes and current democracy. In contrast with the pessimistic view, some have argued for cautious optimism about the effects of deepfakes, particularly when it comes to epistemic and moral matters (Atencia-Linares & Artiga 2022; Voto & Viola 2023). But the view I develop here is not at all optimistic about deepfakes and their relation to democracy. On the view I shall outline, deepfakes are a threat to democracy, but the pessimistic view both mislocates and understates the nature of that harm.

I suggest that the real challenge that deepfakes pose for democracy may be that of a catch-22: given the advent of deepfakes, democracy may be damned if we reform it to be more veritism-friendly, but also damned if we don't—for truth-reformed political ecosystems may provide the perfect environment for deepfakes to cause significant disruption, much more than they currently do. If that's right, then deepfakes count against making democratic politics more truth-oriented. This view is not pessimistic about the impact of deepfakes for democracy, but nihilistic: deepfakes are a *pro tanto* reason against reforming democratic

politics in truth-oriented ways, because such reforms risk exposing the electorate to greater abuse by deepfake malefactors.

In section 2, I develop and defend my argument against the pessimistic view. In section 3, I offer an explanation for why deepfakes have yet to significantly impact democratic elections. This serves as a segue into section 4, where I outline the nihilistic position on deepfakes and their relationship to democracy. In section 5, I reply to objections. Section 6 concludes by arguing against an optimistic view.

Before proceeding, I want to clarify two issues. First, as outlined above, the nihilistic view is that deepfakes provide a *pro tanto* reason against reforming democratic politics to be more truth-friendly. Whether deepfakes constitute an *all things considered* reason against ameliorating democracies in this way is a matter on which nihilism is neutral (see section 5). So the nihilistic view is not as alarmist as it might sound. Nonetheless, its central claim—that deepfakes provide a reason against reforming democracies in truth-relevant ways—amounts to an extremely negative (and perhaps surprising) account of deepfakes and their relation to democracy. Second, I want to emphasize that my aim here is not to engage in empirical predictions about what the future might bring for democracies, given the advent of deepfakes, so much as to make the nomological possibility of nihilism vivid to readers, i.e. as an underappreciated hazard of improving democracies' orientations toward truth.

2. Against the Pessimistic View

What I call 'the pessimistic view' affirms that deepfakes will have deleterious consequences for democratic procedures, chiefly elections. To date, Regina Rini (2020) has provided the most detailed defence and elaboration of this view. Her version of pessimism includes several auxiliary claims about recordings and their roles in regulating testimony. These have been challenged by others (Harris 2021; Habgood-Coote 2023; Sorell 2023). Here, I object to

pessimism in general, a view not unique to Rini but defended by several other philosophers too (Diakopoulos & Johnson 2021; Barber 2023; Roberts 2023).

2.1 The Objection

My objection to the pessimistic view is straightforward: there is no evidence of the election interference it predicts. If pessimism were true, then it is reasonable to think that deepfakes should have had an impact on elections by now. But there is no evidence that they have. Therefore, we have reason to think that the pessimistic view is on the wrong track.

That is my objection in a nutshell. Allow me to elaborate it by outlining a striking asymmetry: there are many well-documented cases of deepfakes having caused moral harms to individuals, mainly women. This includes both private individuals and celebrities. For instance, in April 2024, the BBC reported on 'Jodie', a woman who found deepfakes of herself online after enduring years of harassment (West 2024). Jodie discovered that these images had been created and circulated by a close male friend from university—he subsequently admitted as much and was handed a suspended prison sentence. A few months earlier, Taylor Swift became the target of a campaign to create pornographic deepfakes of celebrities that originated on the anonymous imageboard website 4chan.org. The images quickly spread to social media networks (Belanger 2024). Swift is not the first female celebrity to be harassed and objectified in this way. But she is arguably the first to have deepfake pornographic images of her go viral. Crucially, these are not isolated cases. With the advent of deepfakes, adult women and teenage girls posting photographs of themselves online face the very real prospect of being 'face-swapped' into pornographic deepfake videos and still images. The dangers that deepfakes pose to women are clear and present. By contrast, the harms that deepfakes pose to elections are conspicuous in the headlines by their absence.

This is not to say that no attempts have been made to use deepfakes to interfere with elections. One notable example is the mid-2023 presidential campaign advertisement by the

U.S. Republican Ron DeSantis. Circulated on social media, the advertisement included deepfake still images of DeSantis's rival at that time for the Republican nomination, President Donald Trump, embracing Anthony Fauci. Fauci, Chief Medical Advisor to the President during the COVID-19 pandemic, is a controversial figure in the eyes of many Republican Party supporters. As such, DeSantis's use of these deepfakes in a campaign advertisement was a clear attempt to sway political opinion by discrediting a political opponent in the eyes of potential voters. However, what is notable is how little effect the image had: half a year later, DeSantis suspended his campaign and endorsed Trump, who went on to win the 2024 U.S. Presidential Election. An analysis of similar attempts to use deepfakes to influence voter behaviour in 10 other countries over the course of 2023 concluded that they had little effect on election outcomes (Łabuz & Nehring 2024).

These are a limited number of cases from which to challenge the pessimistic view. But pessimistic claims about the effects of deepfakes on electoral processes have consistently failed to be vindicated. In 2020, philosopher Nathan Colaner and computer scientist Michael J. Quinn expressed anxiety about the potential role that deepfakes might play in that year's U.S. election:

> The personal harms that can be unleashed by deepfakes are limited only by the imaginations of bad actors, but they are dwarfed by the scale of societal harms we may soon experience... It is chilling to think of the effect they may have on the 2020 US presidential election. (2020, np)

This is a statement in support of the pessimistic view and the potential of deepfakes to upset a recent, major election (see Diakopoulos & Johnson 2021 for similar claims). However, the effect of deepfakes on the 2020 U.S. presidential election did not materialize. By contrast, politicians' speech acts to the effect that the election was rigged caused $30 million dollars'

worth of destruction to the U.S. capitol and nine deaths, four in the riots and five by suicide afterward.

To be sure, deepfakes remain a new technology—what the future brings for both them and democracy is unclear. Nevertheless, the fact that deepfakes have immediately caused significant moral harms while having almost no measurable impact on democratic processes should give us pause about whether the pessimistic view is a credible account of deepfakes and their relation to democracy. Even if there are elections where they have had *some* impact, that would be a far cry from the kind of epistemic "chaos" (Rini 2020 p.7) or "maelstrom" (p.8) that the pessimistic view predicts we would find ourselves in because of deepfake-based election interference. It is a far cry further still from the moral maelstrom that deepfakes have caused for women like 'Jodie' and Swift.

Now, Rini does suggest that the epistemic turmoil deepfakes might bring may be "slow-boiling" (2020, p.8.). That is, the pessimistic view may be read as a claim about the capacity that deepfakes will develop in the future, once the technology becomes more advanced. But why? Current deepfake technology is sufficiently advanced that expertly produced deepfakes are near-indistinguishable from genuine photographs, audio recordings, and video recordings. Do deepfakes need to be *completely* indistinguishable from recordings to cause election interference? I think we should be sceptical, as I shall now argue.

2.2 A Reply

Does it matter for deepfakes' ability to disrupt elections that they are not yet wholly indistinguishable from genuine recordings? If so, then my objection to pessimism is easily rebuked. The pessimist might reply, "Just wait—when deepfakes become indistinguishable from real recordings, then we will see the electoral interference our view predicts."

Crucially, this reply provides pessimists with an explanation for why deepfakes have caused moral harms to women yet so far failed to impact elections: to cause harm to women,

deepfakes do not need to be indistinguishable from genuine recordings, whereas this is required to impact elections. Deepfakes of women are produced for viewers' sexual arousal but do not have to be indistinguishable from recordings to satisfy this function (plausibly, it is not necessary for viewers to believe that the depicted individual performed the depicted sex act to be aroused). But when it comes to deepfakes designed to interfere with elections, these deepfakes are produced to *mislead* viewers, i.e. about the actions of a candidate. So they cannot achieve their purpose unless they induce viewers to falsely believe that the depicted individual performed the depicted act. In contrast with pornographic deepfakes, this requires that they be indistinguishable from real recordings. And although they come close, current deepfakes fall short here. Hence, why we have seen examples of pornographic deepfakes causing moral harms, but not deepfakes of politicians causing election interference.

I think we should reject this defence of pessimism. We should also reject the explanation it offers for why the technology has so far caused moral harms to women but failed to impact elections.

First, audio deepfakes are already indistinguishable from genuine recordings. This is evidenced by cases where audio deepfakes have been used to defraud individuals and companies of significant amounts of money. For example, as far back as 2019, criminals used deepfake technology to mimic via telephone the voice of a CEO of a UK-based energy company. The deception was so convincing that an employee acquainted with the CEO transferred $243,000 to a non-standard supplier, who of course turned out to be the fraudsters (Stupp 2019). The success of this scam, and similar ones, suggests that deepfake audios of politicians have the capacity, in the here and now, to mislead voters and affect election outcomes. But they haven't. Hence my objection that the pessimistic view appears misguided.

Second, deepfakes don't need to be entirely indistinguishable from recordings to disrupt elections. Near-indistinguishability, a feature of deepfakes produced competently in

the present, is enough. All it takes for deepfakes to go viral on social media is for non-experts to be fooled: people who may take only a quick look at the video on a relatively small phone screen and then hastily share it with followers, whether innocently or for the sake of 'clout' (i.e. online social status). Although a widely shared piece of media may ultimately be identified as a deepfake, corrections don't go viral on social media in the way that misinformation does (Vosoughi et al. 2018). Corrections also sometimes fail to be as widely believed as the initial misinformation that they refute, as Rini (2020, p.7) acknowledges. So the idea that deepfakes need to be wholly indistinguishable from recordings in order to be in a position to cause electoral mischief is false.

Third, the idea that deepfakes need to be indistinguishable from recordings to be in a position to cause election interference is not only false, but it also wrongly implies that only doxastic reactions to deepfakes will influence voter behaviour. But consider an extreme, though not unimaginable, example: a deepfake of a candidate engaging in bestiality. This deepfake need not be wholly realistic to harm the candidate's chances of election, and so need not be indistinguishable from a genuine recording to influence a result. As Keith Raymond Harris (2021) has argued, deepfakes may harm by creating non-doxastic associations in the minds of viewers, ones linking the person depicted in a deepfake with the action they are depicted as performing. For example, the pornographic deepfakes of Swift were so extreme that they were unlikely to cause viewers to *believe* that she performed the depicted actions. However, their creation and circulation nonetheless harmed Swift by virtue of creating in the minds of viewers an association between her and sexual activity, one that she understandably found distressing and unwelcome. Similarly, a deepfake of an election candidate engaging in a taboo or immoral act like bestiality need not be indistinguishable from a genuine recording to nonetheless affect the candidate's chances of election. The deepfake may likewise cause voters to associate the candidate with bestiality, irrespective of whether it is *believed*, something

likely to render the candidate an object of ridicule in ways sufficient to impact voter sentiment.

In sum, pessimists cannot defend their view by replying that their account concerns the effects of *future* deepfakes on democracy, i.e. when the technology becomes sufficiently advanced as to be completely indistinguishable from genuine recordings. Complete indistinguishability is a red herring insofar as deepfakes have the potential to disrupt elections in their current state, no less than their negative impact on women. Therefore, if framed in terms of current deepfake technology (Colaner and Quinn 2020; Diakopoulos & Johnson 2021), pessimism is refuted by the lack of significant impact of deepfakes on recent elections. However, if framed in terms of future deepfakes (Rini 2020), the view incorrectly assumes that indistinguishability from recordings is necessary for deepfakes to cause electoral disruption, while also overlooking that some deepfakes, i.e. audio-based ones, already meet this standard. Either way, pessimism is flawed.

At this stage, a question arises: given that deepfakes haven't significantly disrupted any elections, *why* haven't they? One explanation is that pessimism is just alarmist AI 'doom-mongering'. But I suspect that the answer is more nuanced. Deepfakes really do appear to have the potential to cause electoral damage. But that deepens the puzzle: since it is plausible, given the current state of the technology, that deepfakes *can* affect electoral outcomes, why haven't they? In particular, why aren't more political deepfakes being created and circulated?

3. What Explains the Impotence and Absence of Deepfakes?

Here is my answer: in many parts of the world, democracy is in a sufficiently poor state—in terms of its being hostile to matters of truth and accuracy—that deepfakes will make little difference for the reason that they *can't* make things much worse. They are just one more truth-unfriendly factor among a much larger veritism-hostile morass. That, I suggest, is what

explains both the impotence and absence of deepfakes in current election contexts, despite their apparent capacity for disruption. Deepfakes are one more pollutant in an already contaminated environment, one which is so tainted as to effectively 'mask' the disposition that deepfakes have to cause political upset.

What do I mean by saying that democracy is in a poor state, veritism-wise? In the main, I have in mind facts that will be familiar to many readers. Across the world, liberal democracies are saturated with political spin, deception, promise-breaking, bullshit (more politely, what Quassim Cassam (2019) calls 'epistemic insouciance') and outright lies. I am referring here not just to speech acts by politicians, but also those made by their supporters, political activists, the media and even voters themselves. As Adam Gibbons (2024, p.1007) observes "Voters themselves will often bullshit about whether politicians are bullshitting." Another aspect of political life that stands in the way of veritistic speech in politics is moral grandstanding. Not infrequently, politicians are motivated to make their assertions not from a desire to speak truthfully, but from a desire to appear morally superior to their rivals (Tosi & Warmke 2016, p.200).

All this (and more) adds up to a situation where competent and sincere testimony is hard to come by in the political domain. This makes it difficult for voters to maintain true beliefs in that domain about, e.g., what policies elected officials will enact if voted in, which party most closely reflects their values, which political issues are most pressing in their country, which policies will have greatest impact on the relevant issues, and so on. It also makes it costly for voters to expend effort identifying competent and sincere testimony. Indeed, there are well known burdens on being politically informed that may render it rational to remain ignorant on such matters (Downs 1957; Somin 2013).

Moreover, I take it that voters are aware of many of the above facts. Voters may be exasperated when politicians spin, omit crucial details, bullshit, or lie—but they are rarely surprised. This is not to say that voters are a bastion of reasonableness. Far from it. Belief is

commonly said to aim at the truth (Velleman 2000; Boghossian 2003). But political beliefs are often shaped by factors at best only accidentally related to truth, including the beliefs and attitudes of those that they happen to associate with or wish to emulate (Huemer 2016; Hannon & de Ridder 2021). As such, political beliefs are often tools for social bonding, formed because of social rewards and punishments rather than a dispassionate concern for truth (Williams 2021). Moreover, political beliefs are often maintained in truth-insensitive ways, to the extent that there is research on the neural correlates of their persistence in the face of counterevidence (Kaplan et al. 2016). Even when voters' political beliefs do update in rational ways, voting behaviour does not necessarily follow suit (Swire et al. 2017).

I am not offering the above as an especially novel analysis of the truth- or evidence-insensitive nature of political speech and the epistemically irrational or socially adaptive nature of political belief. What I wish to highlight is the relevance of such facts for whether deepfakes are a direct and immediate threat to democracy, as the pessimistic view affirms. For once we bear in mind the evidence-unfriendliness of the existing political landscape, we can thereby appreciate why current deepfakes do so little damage and are rarely even seen: they are one more piece of information in an environment where (i) there is not a high premium on accuracy to begin with; (ii) attempting to identify accurate information is burdensome; and (iii) becoming informed on such matters also incurs a number of costs. This means that, first, many current political environments disincentivise voters from taking deepfakes seriously: if and when they come across suspicious-looking media, voters may simply shrug their shoulders and move on rather than waste time digging deeper. Second, the state of current political environments may also disincentivise bad actors from creating and circulating deepfakes in the first place; it is unlikely to be worth their time and effort when much information in the political domain is taken by voters with a pinch of salt and their voting behaviour not infrequently fails to obey epistemically rational norms, such as updating in the light of (putative) evidence.

Granted, as mentioned in section 2.2, deepfakes can cause political disruption by creating non-doxastic associations in viewers' minds between the depicted politician and the action(s) they are depicted performing. So bad actors might still have *some* incentive for creating deepfakes, even if their deepfakes fail to be believed. However, there are far simpler ways to create such associations in the minds of the electorate, and these act as a countering disincentive against spending time creating deepfakes. For instance, crudely photoshopped images or creatively produced memes of politicians performing various actions are much easier to generate than deepfakes but are just as apt for widespread circulation on social media.

Thus, for all that is undesirable about current political environments around the world and their hostility to truth, that may, ironically, provide protection against deepfakes.

In affirming that deepfakes are a threat to democracy, the pessimistic view assumes that democracies around the world are in a healthy enough state that deepfakes *can* do them harm and that it will be worth the time of meddling actors to create such media. Both assumptions are erroneous. Democracies around the world are not in a good state. Hence why deepfakes have failed to cause them much (if any) harm and hence why we see little effort expended to create deepfakes that are seriously politically disruptive.

4. For the Nihilistic View

Above, I gave an argument against the pessimistic view, in both present- and future-oriented forms. I also offered an explanation for why deepfakes are not more prevalent in current political environments, given their capacity to disrupt elections—their disposition to do so is masked by those environments' truth- or evidence-hostile nature. In this section, I outline a nihilistic account of deepfakes and their relation to democracy. On this view, the real challenge that deepfakes pose for democracy is that they constitute a reason *not* to reform the

current political climate to be more truth- or evidence-friendly, something that has traditionally been considered essential for a well-functioning democracy.

To begin, suppose that we were successful at reforming current politics to be more veritism-friendly. Suppose that we created incentives for politicians to speak credibly and sincerely, while simultaneously creating disincentives for epistemic insouciance, spin and moral grandstanding. Suppose further that we combined this with penalties for outright lies. (Suppose further that these incentives, disincentives and penalties were to be effective.) We will likely also want to create various fact-checking mechanisms to monitor politicians' speech and behaviour. We might also create a legislative-based mechanism to provide the incentives and enforce the penalties. This is no small task, clearly, but introducing such mechanisms and policies is not unprecedented. In 2009, the UK's *Daily Telegraph* exposed widespread exploitation of the allowances and expenses scheme by members of the British Parliament for personal financial gain. The scandal led to resignations, repayments and even imprisonments. It also resulted in the creation of the Recall of MPs Act 2015, a mechanism that allows for the removal of members of the British Parliament suspected of significant wrongdoing or illegality. Similar recall procedures exist in other democracies (Welp & Whitehead 2020). For instance, in 2003 a gubernatorial recall election in the state of California led to the election of Arnold Schwarzenegger.

One might argue that the threat of recall elections—whether in the UK, USA or elsewhere—has had a limited influence on politicians' speech and behaviour. But the point is that more scaled-up and widespread mechanisms *of this kind* might, in principle, provide disincentives against political dishonesty. Either way, my aim is not to defend or assess particular courses of action to combat truth-insensitive political speech.[1] Others have

[1] An anonymous referee suggests that the very incentives that cause politicians to be dishonest and bullshit may incentivize those entrusted with designing and enforcing these

undertaken this task (Amazeen 2015; Cassam 2019; Habgood-Coote 2019; Marsden et al. 2021; Fritts & Cabrera 2022). Instead, my concern is with how the public will likely *react* to deepfakes, should we be successful in reforming politicians' linguistic behaviour, regardless of the means employed. Here, my proposal is that in veritistically- and evidentially-improved political environments, deepfakes could potentially take root and wreak havoc on elections.

Let me unpack this idea further. I argued above that, currently, much information in political ecosystems is taken by voters with a pinch of salt, meaning that they have, at best, a low degree of credence in that information. By contrast, a veritistically-improved political environment is one where voters are likely to place a higher degree of credence in the information they encounter in this domain. Although it is logically possible that an environment where politicians speak more sincerely and competently might still be one where they are distrusted, it is far more likely that voters' attitudes towards politicians would, over time, adjust to become more credulous. That, after all, is part of the goal of reforming politics to be more truth-oriented. We don't simply want to make politicians improved *as individuals*; we want them to be truthful *sources of information*, producing credible assertions that the public can believe, free from charges of epistemic recklessness.

The difficulty is that an increase in electorate credulity creates a situation ripe for abuse by deepfakes. As I argued in section 3, it is plausible that deepfakes are not believed, or even created, because voters don't put sufficient stock in information that they encounter in the political domain. Conversely, a scenario where voters are more credulous is one where this greater willingness to believe can be exploited by deepfake troublemakers. Indeed, increased credulity among voters is likely to incentivise malefactors to create and circulate deepfakes precisely because such media would be more likely to be taken as truth. For

---

mechanisms to likewise be dishonest and bullshit about who should be sanctioned. See especially Gibbons (2024).

example, in a veritistically-reformed political ecosystem—where putative evidence is taken seriously and belief is adjusted accordingly—a deepfake of a candidate engaging in taboo or immoral activity would no longer be just one more potential oddity in a sea of questionable content. Rather, it would be a salient and news-worthy depiction of a candidate, one likely to raise suspicions among voters about the candidate's actual involvement in the depicted activities. Although, as argued in section 2, deepfakes do not need to induce false beliefs to cause electoral interference—they can do so merely by creating non-doxastic associations—misleading voters is the most potent and dangerous way that deepfakes might affect electoral outcomes. That is what we risk exposing ourselves to by improving the democracies in truth- and evidence-related ways.

Thus, on the nihilistic view, the real threat that deepfakes pose for democracy is that they provide a reason *not* to reform the current political climate. For the aim of reforming political speech in general, and politics as a whole, is to remove epistemic pollutants that interfere with the electorate forming and maintaining true political beliefs—something that, as noted, they already struggle with, but which is foundational for democracy to work as it should. But a truth-hostile environment is one where deepfakes can't do much damage, whereas a truth-friendly one is. Therefore, successfully reforming politics risks exposing the electorate to a new and different epistemic pollutant: deepfakes. This means that reforming democracies to be more truth-oriented may fail, either way. Given the advent of deepfakes, making politicians more credible and voters more epistemically rational may simply take us out of the frying pan and into the fire.

In essence, deepfakes may put us in a catch-22: if we *don't* reform current political environments, then the upside is that deepfakes will continue to do little damage, but we will allow spin, deception, bullshit, promise-breaking, and outright lies to continue unabated (not to mention acquiescing to voter epistemic irrationality). If we *do* reform the current political environments, then we may *ipso facto* make them highly susceptible to deepfakes, since the

electorate is then likely to be more trusting of information they come across in the political domain (and may react to putative evidence, like deepfakes, in epistemically appropriate ways). In other words, deepfakes constitute a *pro tanto* reason not to do the very thing that democracy requires of us to function properly; namely, promote a truth- or evidence-friendly political environment. Hence, the consequences of deepfakes for democracy may not be pessimistic but nihilistic. That is the view in outline. I now discuss further aspects of it by considering five objections.

5. Objections and Replies

Objection 1: The nihilistic view is not logically distinct from the pessimistic view. Both agree that the consequences of deepfakes for democracy are negative in general and that deepfakes have the capacity to negatively affect elections in particular.

Reply: The nihilistic view is, like the pessimistic view, a negative one: it claims that deepfakes have negative consequences for democracy and it agrees with the pessimistic view that deepfakes have the capacity to negatively affect elections. The views have this in common. Nevertheless, they are distinct. The two offer different accounts of the conditions under which deepfakes may cause problems for democracies as well as the nature, severity and novelty of that harm.

First, the pessimistic views says that deepfakes have the capacity to affect elections, either in the present or near-future. The nihilistic view agrees, but with a crucial caveat that the pessimistic view overlooks: this is conditional upon facts related to the health of democracies, such as the credibility of politicians, whether the electorate take encountered political information seriously to begin with, etc. Thus, the views disagree about the circumstances in which deepfakes can disrupt elections.

Second, although the two views agree that deepfakes can harm democracies, the nihilistic view sees such harms as indirect, compared to those outlined by the pessimistic view. On the nihilistic view, the harms that deepfakes pose for democracy are indirect insofar as the view says that deepfakes may negatively affect our *motivations* for improving democracy in truth- or evidence-oriented ways. The nihilistic view affirms that deepfakes are a *pro tanto* reason against making such improvements and that deepfakes have the potential to impeded those reforms insofar as they may expose the electorate to greater manipulation by deepfakes, incentivising the production of deepfakes in greater numbers. By contrast, the pessimistic view says that deepfakes directly harm democracy by affecting election outcomes in the immediate or near future. Therefore, the nature of the harms that deepfakes may inflict on democracies differ between the two views.

Finally, in terms of severity and novelty, the harms attributed to deepfakes by the nihilistic view are arguably greater and more unique than those attributed by the pessimistic view. Regarding severity, while it is bad for democracy if deepfakes affect election outcomes, it is far worse if they count against efforts to improve democracy's orientation toward truth- and evidence-sensitivity. For if deepfakes put in question whether we ought to make the political domain more truth-oriented, then they strike at the very foundations of democracy, where informed voting is key. Regarding novelty, the pessimistic view says that deepfakes may affect election outcomes, but many factors already do that. Indeed, this is what truth-insensitive political speech does. But while truth-insensitive speech *is* disruptive to democracy, it does not, itself, challenge democratic reform in the way that, according to nihilism, deepfakes do. Hence the pessimistic and nihilistic views are distinct.

Objection 2: The nihilistic view depends upon a causal claim: that improving the trustworthiness of politicians will, in turn, make voters more credulous of information in the

political domain (including any deepfakes that they come across). But no evidence has been given for this claim.

Reply: The nihilistic view indeed relies on the causal claim that improving the trustworthiness of politicians will, in turn, make voters more credulous of information in the political domain generally. This forms the basis for its claim that reforming political speech might make voters more susceptible to manipulation by deepfakes, potentially rendering such reforms self-defeating.

This aspect of the nihilistic view appears speculative. But I want to reemphasise that I am not advancing the nihilistic view as a prediction of what the future will bring for deepfakes or democracy. Rather, I am outlining the nihilistic view as a nomological possibility, one that we might inadvertently find ourselves in by making politics more truth-oriented and which we might therefore need to prepare ourselves for. The aim of the paper is not to engage in empirical prediction, but to make the possibility of nihilism plausible. With this in mind, it should be noted that the causal connection proposed by the nihilistic view—between increased trust in politicians and increased credulity toward political information generally—is *prima facie* plausible. That is all the view requires. Moreover, I take it that one of the main reasons for voter scepticism about political information is the truth-insensitive nature of the assertions voters encounter in that domain, and which often originate with politicians. As these assertions improve in truth-relevant ways, it is reasonable to expect that the electorate will not only become more credulous of politicians but become less sceptical of political information in general. Again, voters *want* politicians to be credible.

Objection 3: The nihilistic view says that making politics more truth-oriented means suffering greater manipulation by deepfakes. But this is short-sighted: making political environments more truth-oriented *includes* addressing threats from deepfakes. So the nihilistic view rests upon a false dilemma between (i) allowing truth-insensitive political speech to prosper, but

thereby offering a shield against deepfakes, or else (ii) improving political speech, but thereby allowing deepfakes to flourish.

Reply: The nihilistic view does say that we may suffer greater manipulation by deepfakes if we clean up the political environment to be more truth-oriented. In doing so, it does distinguish the threat posed to democracy by deepfakes from those posed by truth-insensitive political speech. However, it is not obviously wrong to separate these two threats. One way to bring this into focus is to reflect on the kinds of solutions proposed for each problem.

Solutions to widespread truth-insensitive assertions by politicians, such as recall elections, focus on politicians themselves. Sometimes the emphasis is not so much on incentives or disincentives, but personal improvement. For instance, Cassam (2019) outlines several 'self-help' techniques for bullshitting politicians, including showing greater respect to others, reading more widely and improving one's listening skills. Cassam's approach to these matters is relatively unique, and a reflection of his vice-epistemological viewpoint. Alternatively, the emphasis might not be on improving politicians *per se*, but on improving regulation of the communication channels via which their truth-insensitive assertions reach the electorate. These solutions have focussed either on better independent fact-checking or better control of social media (Amazeen 2015; Marsden et al. 2021). For instance, Megan Fritts and Frank Cabrera (2022) advocate for making social media companies financially liable for damages caused by misinformation spread through their platforms.

When it comes to deepfakes, the solutions that might curtail their damage look very different. First, there are no solutions akin to Cassam's suggestion that politicians focus on self-improvement. This is not a coincidence. Those overseeing the production of deepfakes in Western democracies are likely to be motivated by financial gain (or perhaps the desire to create political mischief), while those based elsewhere who are looking to interfere with Western elections may, additionally, be motivated by a desire to attack democracy itself. It is

difficult to imagine what strategies for self-improvement might work on people who are so influenced by greed or revulsion at democracy that they would, as a result, attempt to disrupt elections. In practical terms, it is harder still to see how we might incentivise such persons to self-improvement, especially in the case of external actors looking to interfere in Western democracies.

By contrast, solutions to the harms that might be caused by deepfakes are, in the main, identification-based. These come in two forms. There are identification-based solutions that are technological and there are identification-based solutions that are viewer-centric. In the main, technological solutions involve designing computing tools to reliably detect when photorealistic media is created in ways characteristic of deepfakes. For instance, the tool might be trained to identify 'tell-tale' signs of a deepfake, such as certain face-warping artefacts or visual-noise patterns (see Rana et al. 2022 and Yu et al. 2021 for reviews). On the other hand, viewer-centric solutions focus on developing our perceptual and affective sensitivities, either to online content in general or photorealistic media in particular. For instance, Taylor Matthews (2022) has argued that the harms of deepfakes might be mitigated by the cultivation in viewers of a 'digital sensibility'. This is not simply to affirm the platitude that we should be cautious about what political media we interact with online and share with others. Rather, Matthews' idea is virtue-based. Emphasis is placed on 'digital exemplars', i.e. tech-savvy ideals, who would flag potentially untrustworthy content to the folk, thereby slowly fine-tuning the latter's sensitivity to trustworthy content.

We can sum up these contrasting approaches using the analogy of producer and consumer: mitigating the harms posed by deepfakes is thought to involve helping consumers (the electorate) to detect such media, either with their own eye or with technological aids. By contrast, avoiding the harms posed by truth-insensitive political speech is thought to involve either improving producers' (the politicians') dispositions to generate truth-insensitive content or, in addition, better controlling the channels through which such content spreads to

consumers. The difference in these approaches is not unreasonable. Being an elected politician arguably comes with a duty *to the electorate*, one that might be used to incentivise them to more truth-sensitive speech when addressing voters. However, those producing deepfakes have no analogous partiality-constituted duty to which we might appeal in order to stop them creating such media, especially when based in another country and potentially wedded to a different form of political life. Reforming truth-insensitive speech and mitigating the harms of deepfakes are different problems, calling for different solutions. The nihilistic view is correct to distinguish them.

Objection 4: But cleaning up our political environments can be a means to combatting deepfakes. The honest politician will (i) condemn malevolent actors for spreading deepfakes about their opponents; and (ii) be believed when truthfully affirming that a piece of media of themselves is a deepfake that constitutes disinformation about their own actions. So if politicians were more truthful, we might look to them to prevent deepfakes from causing harms to democracy.

Reply: This sets an unrealistic standard for what making politicians more honest should consist in. Reforming politics so that politicians avoid spin, deception, promise-breaking and bullshitting does not include the thought that politicians should, as a result, actively work to correct false beliefs of their electorate about the actions of rival candidates. In general, there seems no duty on politicians to point out false beliefs that the electorate may have about their political opponents This is not to deny that it would be a good thing if politicians did this. Nor is it to deny that it might be effective against deepfakes. The point is that such actions would be supererogatory; they go 'above and beyond' what we can reasonably hope of the honest, reformed politician. As such, it is an unrealistic and overly demanding expectation. It may also be highly impractical, if not self-sabotaging. If politicians are required to correct the electorate's false beliefs about their opponents, or else risk charges

of dishonesty, then they may endanger their own campaigns. There may simply be too many voters with false beliefs for them to spend time correcting these, while also running their own campaigns competently.

Likewise, to think that a politician could affirm that a *prima facie* credible piece of media of themselves is a deepfake, and be widely believed, we must imagine a highly idealised situation. Someone who could rebuff putative audio, photographic or video evidence of wrongdoing or illegality on their part (assuming the contents were not absurd) and, crucially, be believed by both supporters and detractors, would have to possess extraordinary levels of credibility. Again, in conceiving of such a situation, we go well beyond typical ideas about what reforming political speech, and democracy in general, to be more truth-sensitive should amount to. After all, no matter how credible a politician is, if media that looks and sounds like a recording of them engaging in (non-absurd) wrongdoing or illegality were to surface, it would be epistemically rational to investigate its authenticity, irrespective of the depicted politician claiming that it must be a deepfake.

Objection 5: Granted, deepfakes and truth-insensitive political speech might call for different responses. And granted, it may be unrealistic and overly demanding to look to politicians to identify and decry deepfakes. Still, nothing in the nihilistic view shows that deepfakes can't, *in principle*, be combatted. Indeed, identification-based strategies aim to do that. Thus, the nihilistic view rests upon an unsupported scepticism that no strategies to mitigate the political harms of deepfakes will succeed, when there are plausible ones on standby.

Reply: Again, it is worth being clear that the nihilistic view is that deepfakes constitute a *pro tanto* reason against truth- or evidence-oriented reforms to democracy. The view does not say that deepfakes are an all things considered reason against such reforms. So nihilism does not say that deepfakes cannot in principle be combatted. Nevertheless, there is reason to think that attempts to curtail the spread of deepfakes, particularly via social media, will in fact

fail. If so, then deepfakes may end up constituting an all things considered reason not to bring about the relevant changes

This is a bold claim, admittedly, but consider identification-based strategies, like those mentioned above. These suffer from the recurrent mistake of assuming that only doxastic attitudes to deepfakes affect voters' sentiments toward politicians. Yet it is naïve to think that a deepfake of a politician engaged in taboo or immoral activity, even if identified as a deepfake (whether by technological means or a virtuous eye), won't affect some voters' feelings toward that politician. Identifying a piece of media as a deepfake may, at best, mitigate *some* of its effects upon voters, such as their forming the false belief that the depicted politician performed the depicted act. But identifying a piece of media as a deepfake offers no prevention against non-doxastic associations being formed in the minds of voters between the depicted politician and the depicted act (although, as argued in section 3, there are easier ways to achieve this). Worse still, given that voters often fail to be epistemically rational in how they maintain their beliefs in the political domain (see section 3), identifying a piece of media as a deepfake is no guarantee of voters failing to believe it anyway. Thus, a real possibility is that identifying a piece of media as a deepfake may not mitigate *any* of its potential effects upon some voters, doxastic or non-doxastic. Then there is also the fact that, as mentioned in section 2.2, corrections don't go viral on social media in the way that misinformation does, and may fail to be as widely believed as the initial misinformation that they aim to counter. Similarly, although one can imagine politicians attempting to counter any negative non-doxastic associations caused by deepfakes by publishing media of themselves performing positive, morally upstanding actions, this is unlikely to have the desired result. For many, a politician posing for another 'wholesome' photo-opportunity is unlikely to be worth sharing, let alone a second look, whereas a salacious piece of media of them might well be.

All told, we should take seriously the idea that we may be relatively powerless in the face of deepfakes and that steps to mitigate their influence on voting behaviour may be only

minimally or partially effective. Given how compelling they are, and given how easily political belief can be shaped by truth-irrelevant factors, there may be little we can do to diminish the influence of deepfakes on the electorate. If that were to be the case, then deepfakes may end up being an all things considered reason against improving the politics of dishonesty and not merely a *pro tanto* one, especially as they become easier to create. But it is important to be clear that this is something nihilism *per se* leaves open. Whether we ought to upgrade the threat posed by deepfakes in this manner, from *pro tanto* to all things considered, turns upon many empirical variables, not only regarding the effectiveness of attempts to identify and counter deepfakes, but other, less predictable matters too. And, again, my aim is not to speculate about what the future brings for democracies and deepfakes, but only to examine some possibilities, like nihilism, not yet made salient.

In particular, much depends here on the governance of the social media platforms through which deepfakes circulate and whether their owners would want to remove deepfakes from their networks in the first place. A blanket ban seems unlikely, given many positive use cases, including not only artistic ones, but positive epistemic ones too (Kerner & Risse 2021, pp.97-100; Cavedon-Taylor, 2024, pp.12-14; but see Flattery & Miller 2024 for complications). A narrower ban on deepfakes of politicians in particular seems unlikely too. Not only are there overtly satirical uses of deepfakes that feature politicians while being poor targets for censorship, but social media platforms increasingly see themselves as guardians of free speech. As such, those who create deepfakes of politicians may claim that their media should be protected on such grounds (see Barber 2023 for discussion). So even if there were foolproof ways to identify and remove deepfakes from social media platforms—or indeed the people who share them—it does not follow that this is what the owners of such platforms, or indeed its users, will want.

6. Against an Optimistic View too

The pessimistic view that I began this paper with is not the only account of deepfakes in the philosophical literature regarding their potential to harm. Some have argued in defence of cautious optimism instead. The central idea is that, ironically, the more widespread or normalized that deepfakes become, then the *less* harm they can do, since a situation in which deepfakes proliferate is one where all photorealistic media will fail to be taken seriously by viewers, including both genuine recordings and deepfakes.

The optimistic view sketched above has been made in two contexts: first, the moral harms of deepfake pornography and, second, the question of whether deepfakes will erode the epistemic value of recordings, a claim associated with Don Fallis (2021). In relation to the former, Marco Viola and Cristina Voto suggest that the harms of deepfake pornography might be short-lived, should deepfakes proliferate:

> In a world where most intimate images were *known to be* deepfakes, we would be less worried about what images and videos (including real photographic images and videos) could reveal about us, because hardly anyone would assume by default that they were revealing something about us. (2023, p.12)

In relation to Fallis's (2021) claim that deepfakes will erode the reliability of recordings, Paloma Atencia-Linares and Marc Artiga (2022, p.16) offer a similar claim. They analyse photographs and recordings through the lens of animal signalling, analogizing deepfakes to mimics. As part of this analysis, Atencia-Linares and Artiga, like Viola and Voto (2023), argue that the increased production of deepfakes is likely to be self-undermining, insofar as mimics depend for their success on not outnumbering the honest signallers they impersonate. For instance, some non-venomous snakes mimic the stripes of venomous coral snakes; in doing so, they succeed in deceiving and evading predators. But should the mimics become too

numerous, then the system breaks down: copying the stripes of coral snakes would, in such a situation, *fail* to deceive predators. Likewise, Atencia-Linares and Artiga claim that the more deepfakes there are in circulation, the more they and photographic media (including recordings in general) may simply be ignored. But that would undermine the motivations for creating deepfakes in the first place. So deepfakes cannot become too numerous without thereby failing to be taken seriously.

To my knowledge, an optimistic view of deepfakes has not been extended to deepfakes of politicians. But it is not difficult to envisage what such a view might claim: deepfakes will do little to damage democracy, either in general or in relation to elections in particular, since an increased awareness (or circulation) of deepfakes of politicians will cause viewers to suspend judgment in all photorealistic media and recordings of politicians, including deepfakes. In such a situation, deepfakes can cause little election interference.

Applied to deepfakes of politicians, the optimistic view is not one we should welcome.[2] For many of us, recordings are the only epistemic access we have to our elected leaders. As such, they constitute the central means by which we hold politicians accountable, i.e. by reviewing recordings of their past speech acts, promises, etc. So a situation in which viewers cease to form beliefs about politicians on the basis of recordings of them is an undesirable one (see also Rini 2020; Harris 2021; Matthews 2022). Still, might that be what comes to pass? Two lines of argument suggest not.

First, it is worth reemphasising that deepfakes do not need to be believed in order to affect elections (see section 2.2). The optimistic view, like the pessimistic one, fails to recognise this. For instance, say that the optimistic view is right: eventually, deepfakes of politicians fail to be believed because, due to their increased circulation, all photorealistic media of politicians fail to be believed. Still, deepfakes may cause non-doxastic associations to

[2] Thanks to one of the journal's anonymous referees for pressing me to develop these claims.

be formed in the minds of the electorate (although, as mentioned, if that were the aim then far simpler means are available to achieve this). Moreover, they may still function as political satire, akin to political cartoons that mock, 'send up' or caricature politicians and their behaviour. Indeed, this is already a key function of deepfakes. In 2018, comedian Jordan Peele featured in the creation of a deepfake that depicted former U.S. President Barack Obama insulting Donald Trump as a "total and complete dipshit." However, there is evidence that satire in general and political cartoons in particular affect viewers' sentiments toward electoral candidates (Baumgartner 2008; Zurbriggen & Sherman 2010; see Mag Uidhir 2013 for philosophical issues). Indeed, Abraham Lincoln is widely believed to have won the 1860 U.S. presidential election partly due to the influence of a political cartoon, *Compromise with the South* (Vinson 2014). So even supposing that an increased awareness of deepfakes will cause viewers to suspend judgment in the content of all photorealistic media of politicians, such media may still exert some influence on election outcomes in the way political cartoons have.

Second, the optimistic view is unduly naïve. Imagine a similar position on trolling and communication, affirmed in the early days of the internet: trolling has deleterious consequences for communication via the internet, but the more widespread trolling becomes, the less damage it can do, since people will learn to ignore the trolls. I think it is clear that someone affirming this 'optimistic' view of trolling in, say, 1994, would, from our perspective in 2024 onwards, seem short-sighted. Trolling behaviour has not only continued since the early days of the internet, but as the online world has changed, so too has trolling (Sanfilippo et al. 2017): from targeting Facebook memorial pages or Wikipedia entries, to women being sexually harassed in online games or dating apps, not to mention more extreme acts like 'doxxing' (releasing private information about one's interlocutor to others) and 'swatting' (falsely alleging that one's interlocutor is involved in a serious crime in order to prompt an armed response from law enforcement). Trolling is also increasingly linked to online

radicalisation and extremism (Munro forthcoming). So while trolling may or may not have increased in frequency since the early days of the internet, it has certainly become more varied, more extreme and—if you are unfortunate enough to be the target of its more severe manifestations—almost impossible to ignore.

The particular way in which the optimistic view of trolling is naïve is that it assumes that trolling will not alter significantly as online technology evolves. Yet, as the examples above illustrate, new opportunities for online collaboration or communication have offered new opportunities for trolling while new technologies have offered new tools for trolls to exploit. Arguably, the optimistic view about deepfakes and democracy is naïve for the same reason: new opportunities for online political collaboration or communication may similarly offer new opportunities for political deepfakes to cause harm and new technologies may likewise offer new tools to bring about, or change the nature of, those harms. These are possibilities that the optimistic view at best ignores and at worst denies. Finally, just as extreme forms of trolling that affect one's offline life (like doxxing and swatting) are impossible to ignore, it is naïve to assume that voters can easily ignore salacious or offensive photorealistic political media, even when known to be deepfakes, especially in cases when these depict one's most preferred (or loathed) candidate.[3]

References

Amazeen, M. (2015). "Revisiting the epistemology of fact-checking." *Critical Review: A Journal of Politics and Society* 27 (1) 1–22.

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

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

[3] My thanks to colleagues at the Open University for discussion of an earlier version of this paper. Thanks also to the journal's editor and two of its anonymous referees for comments that significantly improved it.

Baumgartner, J. C. (2008). "Polls and elections: editorial cartoons 2.0: The effects of digital political satire on Presidential candidate evaluations." *Presidential Studies Quarterly* 38 (4): 735-758.

Benn, C. (forthcoming). "Deepfakes, Pornography and Consent." *Philosophers' Imprint*. Available at: https://philpapers.org/rec/BENDPA-5

Boghossian, P. A. (2003). "The normativity of content." *Philosophical Issues* 13 (1): 31–45.

Cassam, Q. (2019). *Vices of the Mind: From the Intellectual to the Political*. OUP.

Cavedon-Taylor, D. (2024). "Deepfakes: A survey and introduction to the topical collection." *Synthese* 204, 14.

Colaner, N. & Quinn, M. (2020). "Deepfakes and the value-neutrality thesis." *Seattle University Ethics and Technology; Viewpoints*. Available at: https://www.seattleu.edu/ethics-and-technology/viewpoints/deepfakes-and-the-value-neutrality-thesis.html

Belanger, A. (2024). "4chan daily challenge sparked deluge of explicit AI Taylor Swift images." *Ars Technica*. Available at: https://arstechnica.com/tech-policy/2024/02/4chan-daily-challenge-sparked-deluge-of-explicit-ai-taylor-swift-images/

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

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

Downs, A. (1957). "An Economic Theory of Democracy." Harper.

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

Flattery, T., & Miller, C. (2024). "Deepfakes and dishonesty." Philosophy and Technology 37 (120): 1-24.

Fritts, M., & Cabrera, F. (2022). "Fake news and epistemic vice: Combating a uniquely noxious market." *Journal of the American Philosophical Association* 8 (3): 454-475.

Frum, D. (2020). "The very real threat of Trump's deepfake. The president's first use of a manipulated video of his opponent is a test of the boundaries." *The Atlantic,* 27 April.

Gibbons, A. (2024). "Bullshit in politics pays." *Episteme* 21(3): 1002-1022.

Habgood-Coote, J. (2023). "Deepfakes and the epistemic apocalypse." *Synthese* 201: 103.

Hannon, M., & de Ridder, J. (2021). "The point of political belief." In their *Routledge Handbook of Political Epistemology*. Routledge.

Huemer, M. (2016). "Why people are irrational about politics." In J. Anomaly, G. Brennan, M. Munger, & G Sayre-McCord (eds.) *Philosophy, Politics, and Economics*. OUP. pp. 456–467

Harris, K. R. (2021). "Video on demand: What deepfakes do and how they harm." *Synthese*, *199* (5–6): 13373–13391.

Kaplan, J., Gimbel, S., & Harris, S. (2016). "Neural correlates of maintaining one's political beliefs in the face of counterevidence." *Scientific Reports* 6 (1), 39589.

Kerner, C,. & Risse, M. (2021). "Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds." *Moral Philosophy and Politics* 8 (1): 81–108.

Mag Uidhir, C. (2013). "Epistemic misuse and abuse of pictorial caricature." *American Philosophical Quarterly* 50 (2): 137-151.

Marsden, C., Brown, I., & Veale, M. (2021). "Responding to disinformation: Ten recommendations for regulatory action and forbearance." In Moore & Tambini (eds.) *Regulating big tech: Policy responses to digital dominance.* OUP. pp. 195-230.

Matthews, T. (2022). "Deepfakes, intellectual cynics, and the cultivation of digital sensibility." *Royal Institute of Philosophy Supplement* 92: 67–85.

Munro, D. (forthcoming). "Internet Trolling: Social Exploration and the Epistemic Norms of Assertion." *Philosophers' Imprint*. Available at: https://philpapers.org/rec/MUNITS-2

Öhman, C. (2020). "Introducing the pervert's dilemma: A contribution to the critique of deepfake pornography." *Ethics and Information Technology* 22 (2): 133–140.

Pawelec, M. (2022). "Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions." *Digital Society* 1, 19.

Rana, M., Nobi, M., Murali, B., & Sung, A. (2022). Deepfake detection: A systematic literature review. *IEEE access* 10 25494-25513.

Rini, R. (2020). "Deepfakes and the epistemic backstop." *Philosophers' Imprint* 20 (24): 1–16.

Rini, R. & Cohen, L. (2022). "Deepfakes, deep harms." *Journal of Ethics and Social Philosophy* 22 (2): 143–161.

Roberts, T. (2023). "How to do things with deepfakes." *Synthese* 201: 43.

Sanfilippo, M., Yang, S., & Fichman, P. (2017). "Trolling here, there, and everywhere: Perceptions of trolling behaviors in context." *Journal of the Association for Information Science and Technology* 68 (10): 2313-2327.

Somin, I. (2013). *Democracy and political ignorance: Why smaller government is smarter*. Stanford University Press.

Sorell, T. (2023). "Deepfakes and political misinformation in U.S. elections." *Techné Research in Philosophy and Technology* 27 (3): 363-386.

Stupp, C. (2019). "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case." *Wall Street Journal*. Available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Swire, B., Berinsky, A., Lewandowsky, S., & Ecker, U. (2017). "Processing political misinformation: Comprehending the Trump phenomenon." *Royal Society Open Science* 4 (3) 160802.

Tosi, J. & Warmke, B. (2016). "Moral grandstanding." *Philosophy and Public Affairs* 44 (3):197-217.

Velleman, D. (2000). "On the aim of belief." In his *The possibility of practical reason*. OUP.

Viola, M. & Voto, C. (2023). "Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images." *Synthese* 201: 30.

Vinson, J. (2014). *Thomas Nast: Political Cartoonist*. University of Georgia Press.

Vosoughi, S. Roy, S. & Aral, S. (2018). "The spread of true and false news online." *Science* 359 (6380): 1146−1151.

Welp, Y. & Whitehead, L. (2020). "The politics of recall elections." In their The Politics of Recall Elections. Springer.

West, K. (2024). "I was deepfaked by my best friend." *BBC News*. Available at: https://www.bbc.co.uk/news/uk-68673390

Williams, D. (2021). "Socially adaptive belief." Mind & Language 36 (3): 333-354.

Young, G. (2021). *Fictional immorality and immorality in fiction*. Lexington Books.

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). "A survey on deepfake video detection." *IET Biometrics 10* (6): 607-624.

Zurbriggen, E., & Sherman, A. (2010). "Race and gender in the 2008 US presidential election: A content analysis of editorial cartoons." *Analyses of Social Issues and Public Policy* 10 (1): 223-247.