WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/171005

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

Deepfakes and Political Misinformation in U.S. Elections

Abstract

Audio and video footage produced with the help of AI can show politicians doing discreditable things that they have not actually done. This is deepfaked material. Philosophers and lawyers have recently claimed that deepfakes along these lines have *special* powers to harm the people depicted and their audiences –powers that more traditional forms of faked imagery and sound footage lack. According to some philosophers, deepfakes are particularly "believable", and the technology that can produce them is or will soon be widely available, so that deepfakes will proliferate. I first give reasons why deepfake technology is *not* particularly well suited to producing "believable" political misinformation in a sense to be defined. Next, I challenge the two most prominent philosophical claims –from Don Fallis and Regina Rini--about the consequences of the wide availability of deep fakes.[i] My argument is not that deepfakes are harmless, but that their power to do major harm in liberal party political environments that contain sophisticated mass-media is highly conditional.

Keywords: Deepfakes, misinformation; perception; testimony

1. Introduction

We live at a time when political misinformation is not only widely circulated but widely believed. For example, the misinformation that voter fraud was committed on a scale big enough to compromise the US election in 2020 –commonly known as the "Big Lie"-- is supposed to be believed even now (early 2022) by a majority of Republicans

(https:/www.msnbc.com/rachel-maddowshow/year-later-gop-support-big-lie-remarkably-stable-n1286939). It is accepted by Republicans despite the fact that Republican state election officials, Trump-appointed judges, and some Republican members of Congress rejected it after investigation or court-hearings. The misinformation that various extremists on the left rather than Trump supporters mounted an insurrection in Washington on 6 January 2021 is also believed by large numbers of Republicans.

How could things get any worse, epistemically, among the politically committed but polarised in the United States? This paper considers the conjecture from certain philosophers that things *could* get worse, epistemically, if deepfakes were to become a more common medium of political misinformation.[1] Deepfakes are typically video-films in which an artificially generated image of someone is knowingly used to depict them doing things that they have not actually done, often discreditable things. What is supposed to make them particularly effective tools of misinformation is their "believability" and the fact that the technology that produces them will soon be widely available, allowing deepfakes to proliferate.

These two features of deep fakes will be discussed in turn in what follows. I argue, first, that the believability of visual misinformation is complicated, and not necessarily a matter of the heightened technical quality of deep-faked images or audio content, as is often claimed. The potential wide availability of the technology and its products raise further and perhaps deeper issues. According to the two main arguments from a still very small philosophical literature, deepfakes pose a big threat to the acquisition of knowledge by visual perception and testimony, respectively. The first argument is from Don Fallis:[ii] he claims that the proliferation of deepfaked video makes videos in general less reliable evidence for belief. A second and more substantial suggestion comes from Regina Rini.[iii] Rini's view is that deepfakes matter because video and audio recordings act as

"epistemic backstops" for the transmission of knowledge by testimony, and deepfakes undermine the role of these backstops.

Contrary to Fallis, I argue that deepfakes do not necessarily taint a video pool. I also contest Rini's claim that recordings have a special role with respect to testimony *in general*, and that they uniquely play the role of backstops. My own view is that deepfakes are only one source of misinformation among others, and that in circumstances in which they gain attention, they do not trump all other sources of evidence. Besides, the success of deepfakes as a tool of mass political manipulation is highly contingent, depending on, among other things, being properly timed, and capturing enough of the quickly shifting attention of a highly fragmented online audience, to make the desired kind of impression on those to whom deepfakes are directed.

I focus primarily on the use of deepfakes in liberal party politics in the West. Such political environments permit the wide production of deepfakes, but the obstacles to their reaching susceptible audiences at the right times and with persuasive content are very substantial. I do not deny that widely producible deepfakes could readily be adopted by individuals to settle personal scores in narrower online networks unrelated to an electoral apparatus and party political organizations. But harm in these circles can also be produced, and is regularly produced, by less sophisticated technologies. Deepfakes can also add power to party political dirty tricks departments. But whether they are game-changing additions is an open question.

2. What is special about deepfakes: believability and wide producibility

Visual images are known to have more influence on people than audio (Glasford, 2013; Prior, 2014; Wittenberg et al, 2021). In deepfakes of politicians, the artificial generation of an image involves modelling visual data derived from an actual person. The model is trained using "deep learning" techniques associated with artificial intelligence (AI); hence the term "deep" in "deepfake". A deepfake is thus quite different from a doctored image made with a software like Photoshop.

In some cases, images *and* sound are combined in a deepfake to show someone doing and saying things that they have not said or done. The images and audio material they combine are often based on large quantities of visual and audio data derived from the real people whose exploits deepfakes seek to falsify. Deepfakes do not typically work from material provided by celebrity look-alikes or doubles or celebrity mimics.[iv]

Some deepfakes publicise in comic form the dangers of deepfakes. For example, a deepfake by Jordan Peele shows a deepfaked Obama saying things, including that Donald Trump is a dipshit, that he has never said (at least in public) (https://www.youtube.com/watch?v=cQ54GDm1eL). Other deepfakes are invasive and objectifying. Thus, pornographic deepfakes distributed online falsely depict some celebrity engaged in a sex act –perhaps to make vivid one person's or many people's fantasy of having sex with that celebrity.[v] Still other deepfakes might be used with the intention of ruining the reputation of a public figure, or to invite their prosecution. For example, politically motivated deepfakes might show some elected official apparently doing something unlawful – accepting an envelope apparently full of cash from a known criminal, say.

One question about deepfakes is what harm they do or what rights they violate *in general*.[vi] In the example just given, a deepfake is used to deceive an audience and to create wholly undeserved contempt for a particular elected official.[vii] But deepfakes are not always used in this way. Sometimes they are used as content in vivid educational or dramatic reconstructions of events that

have actually occurred. Thus deepfakes of Stalin, Roosevelt and Churchill might help to bring to life some crucial event, like the meeting at Yalta in the closing stages of World War II. Or again, a deepfaked Obama might be created to act out some imagined gaffe in a film clearly labelled as satirical. In these cases there is no intention to deceive and probably no false belief created in the audience. Nor need there be any reputational harm.

I shall mainly consider whether deepfakes that *are* intended to deceive should be weightier threats to political communication than doctored photographs or audio-visual imagery produced in old-fashioned ways e.g. with body doubles and humanly mimicked speech. The literature on deepfakes concedes that there is nothing new about doctored images, including videos, and doctored sound recordings, intended to mislead or deceive.[8] The same literature concedes that old fashioned fakes of these kinds succeeded in misleading or deceiving past audiences. Presumably, old fashioned images and recordings could mislead some of us today, e.g. concerning events in the pre-digital past of senior politicians.

We know, for example, that undoctored, predigitally printed photographs were taken of a young Justin Trudeau wearing "brownface" at an "Arabian Nights" themed party at a school where he taught.[9] These proved highly embarrassing to the mature Justin Trudeau, Canadian Prime Minister: perhaps a doctored, old fashioned, i.e. low-tech, photograph of some other liberal politician in youthful brownface would prove just as discreditable now to a general audience even in an era where deepfakes are available. Isn't convincingness to the target audience –whatever the technology-- what matters to the deceiver? How does the fact that images or sound recordings are blended with the help of "deep learning", so that they are deepfakes and not mere fakes, make any epistemic difference, if the content of both are believed? It is not as if *only* images produced by deep learning are ever convincing.

The answer to this question seems to be that is (a) algorithms trained by "deep learning" produce images and sound that are increasingly believable;[viii] and (b) the technology of deepfakes can in principle be widely distributed, enabling a large number of partisan users[ix] to create many widely-held false beliefs. These two characteristics are now considered at length in turn. I start with believability.

2.1 Believability in Deepfakes: Seamlessness and High-Quality

We can distinguish between at least two dimensions of believability in deepfakes. One dimension might be called seamlessness: when one simultaneously looks at and listens to the deepfake, there is no impression of discontinuity in the image or the audio taken separately. For example, the face is highly defined and does not look as if it is simply superimposed on a body; and speech both makes sense and does not suffer from sudden changes in the register or timbre of the voice, or obvious gaps in sound. Another dimension of believability concerns the way the audio and visual go together. For example, does the speech that is heard correspond with the facial movements of the person apparently producing the speech, or is it merely approximate, as in an animated cartoon? A deepfake that met these conditions might be "believable" in the sense that it contained clear images and sound, was credibly a recording in real time of an actual event and not a hybrid of faked image and edited-together audio. We might call this *a high-quality* deep fake.

Must a deepfake be high-quality to be convincing? To appreciate the point of this question, we can ask whether an *un*faked recording must be believable along these lines in order to serve a persuasive purpose. One of the most compelling pieces of evidence against the disgraced film producer, Harvey Weinstein, came from an audio recording made by the actress Ambra Battilana Gutierrez, in which he appears to admit to touching her breast the preceding day, and in which he is putting pressure on her to come with him to his hotel room (https://www.newyorker.com/video/watch/harvey-weinstein-caught-on-tape).

The quality of the audio is what one would expect from a hidden microphone. It is not seamless, and there is no visual image. This detracts from its believability in the sense of the previous paragraph, but *adds* to its believability as evidence because the experience of poor audio coheres with the knowledge that it comes from a hidden microphone, and the fact that the microphone is hidden and that Weinstein's side of the conversation contains an admission, shows that, on Weinstein's side, the conversation was very probably unguarded and candid. Now suppose that a deepfaked recording of another person in this style was created to back up the claim that this other person was guilty of sexual harassment: would that have to meet a higher standard of seamlessness to count as convincing?

No. Believability in the sense of being high-quality is not necessary for a deepfake to be convincing as evidence of wrongdoing.[x] Deepfakes are often depictions of discreditable acts, and discreditable acts are not usually carried out brazenly in public, in full view of many witnesses. They are often performed privately, in environments that the wrongdoer controls, and that are out of the range of impartial witnesses, or witnesses hostile to wrongdoers. This is why recordings of wrongdoing so often have to be made secretly, and at the expense of readily intelligible audio or undistorted images. People who are accustomed to the genre of "undercover" reporting are prepared for these departures from high-quality, and even receive them as indicators of the genuineness of what is depicted. By the same token, high-quality may sometimes detract from believability, since producing a high-quality image may create suspicions in an audience that what is shown has been staged rather than caught in an unguarded moment.

Not only is high-quality not necessary for believability: it is not sufficient, either. This is the lesson of a recent experiment (Vaccari and Chadwick, 2020). A representative sample of 2005 respondents were exposed to three versions of the high quality Obama/Peele deepfake, in which a deep-faked Obama is shown saying "Trump is a total and complete dipshit". One version was a 4-second clip in which Obama utters those words. In the second version, the quotation about Trump is shown in a

26-second clip. Finally, there is an educational version in which it is revealed that a deepfaked image of Obama has been used. The first two versions are considered "deceptive" by the experimenters in that nothing is done to correct the impression that the clips show the real Obama asserting what he is shown as saying about Trump.

Participants were asked on the basis of the clips whether it was true that Obama had called Trump a dipshit. The striking finding of the experiment (Vaccari and Chadwick, 2020, p.2) is that

> subjects exposed to either the 4-second or the 26-second deceptive deepfakes were *not* more likely to be deceived than those exposed to the full video with the educational reveal. The 4-second deceptive video was least likely (14.9%) to deceive participants, followed by the 26-second deceptive video (16.4%) and the full video with educational reveal (pp.6-7).

That is, the 4-second clip was least likely to return the answer 'Yes' to the question, "Did Barack Obama ever call Donald Trump a 'dipshit'?", followed by the 26-second clip, followed by the educational video. For the 4-minute video, just over half the respondents answered "No", while just over 35 % answered "Don't know," which was taken by the experimenters to express uncertainty. For the 26-second clip the Nos were 46.7%; the Don't knows 36.9% and 16.4 answered "Yes" (=were deceived). The corresponding percentages for the educational video with reveal were 55.6% (No); 27.5% (Don't know) and 16.9% (Yes).

3. <u>Wide Producibility</u>

We have been considering two claims to distinctiveness of deepfake technology. One is heightened believability, notably through high quality in image and sound. I have been arguing that this is not always a help in producing credible and damaging impressions of people, whether the impressions are well-founded or not. What about the second claim to distinctiveness: namely, that the technology can in principle be deployed by a wide number of users, making it potentially a new

kind of mainstream software, such as Photoshop, but with special dangers in the hands of partisan political activists.

According to Paris and Donovan (2019, p.10; see also Harris (2021) p. 13375), deepfake technology is currently *not* generally available. It is at the exotic end of the spectrum of current and traditional technologies for audio-visual deception. The traditional technologies involve cutting, speeding and slowing, and face swaps. A crude kind of deepfake can be made by a widely available DeepFake app. But this is not always believable.[xi] Let us suppose nevertheless that effective deepfake technology *will* soon be very widely available, so that lots of people can use it. How obvious is it that wide availability lends itself straightforwardly to *political* misinformation?

It is easy to see how someone wanting to tarnish the image of a *personal* enemy online might want to resort to an easy-to-use, widely available deepfake software. This could produce a bad impression among a relatively small circle of friends and acquaintances who know the target. Someone in the friendship group would know whom to include in the audience, and what the relevant email addresses were, or which social media site to post on. The software could be useful to many independent personal projects of getting back at personal enemies.

But the usefulness of a software that could support *mass* deepfaking by political activists against its opponents is less obvious. Effective political persuasion by activists targeting undecided voters is not straightforward. Consider the documented pitfalls of a political campaign that used activists to approach swing voters by telephone. Enos and Hersh (2015, p. 252) show that in the case of the very well-organised Obama campaign in 2012, activists "who were interacting with swing voters [by telephone] on the campaign's behalf were demographically unrepresentative, ideologically extreme, cared about atypical issues, and misunderstood the voters' priorities." They add that they "find little evidence that the campaign was able to use strategies of agent control to mitigate its principal-agent problem (ibid)." It is difficult to see why the attempt by social media activists to

reach undecided voters on social media with the aid of misinformation, including deepfakes, should be free of these problems.

More generally, there are the well-known effects of confirmation bias. People seem more receptive to information that is consistent with and confirms their pre-existing beliefs, than to information which challenges those beliefs (see e.g. Nickerson 1998; Taber and Lodge 2006; Modgil et al. 2021). In the highly polarised and partisan politics of the US (and to a lesser extent the UK and Western Europe in 2020s), this bias has obvious relevance. Many in the potential audience for political persuasion or even for factual reporting, choose their sources in ways that confirmation bias would predict. This means that deepfaked defamation of the politicians they support would probably not easily penetrate the information pools that they use, because those information pools are selected for partisanship and contain stories favourable to the parties the user of those pools supports. Recently, some less partisan potential audience members choose to insulate themselves from news sources altogether, to avoid lowering their mood (https://www.reuters.com/business/media-telecom/more-people-are-avoiding-news-trusting-it-less-report-says-2022-06-14/). That may still leave a group of open-minded and willing consumers of a variety of information sources. But it is unclear what strategy for reaching them (with *or* without deepfakes) would look like.

Political *microtargeting* uses fine-grained big data about media audiences to identify individuals who might be receptive to particular kind of message and vote in a desired way (Zuiderveen et al, 2018). The now defunct Cambridge Analytica was one notorious commercial provider of this kind of service. Might political micro-targeting messages use deepfakes effectively? Very recent empirical work (Dobber et al 2020) shows that political attitudes toward subjects of deepfakes and their parties are affected adversely to a noticeable extent if (small) audiences are microtargeted, but no-one claims that the data required for micro-targeting is readily available to political activists, even those who are able to use deepfake technology.

In short, image-making and un-making in mass liberal democratic party politics are not easy.[xii] They seem to require all of the following: (a) detailed knowledge of which sections of a relevant public with potentially decisive votes are undecided; (b) current information about what concerns them at different times; (c) imaginative judgements about what events in an electoral timetable are opportunities to persuade them to take one's own side, including through misinformation; (d) access to media sources used by the target audience at any one time; and (e) expertise in placing messages on those media sources. Even if deepfakes are widely available, (a) to (e) are not. More obviously still, engaging in the right way with topics that are likely to mobilise the attention and emotions of a large online audience requires uncommon specialist skills sometimes exhibited in professionally produced political advertising. These are probably not the skills of Reddit enthusiasts or politically motivated trolls, still less rank and file members of a political base.

It is true that a one-off, individually-produced deepfake of a particular candidate apparently appearing to accept a large amount of cash, or meeting a well-known deepfaked criminal, might be no harder to produce than a deepfake of a personal enemy, but making it a "believable" image, releasing it at a time when it will do political damage, getting it to trend, and reaching the relevant persuadable audience rather than settled opponents; these are different matters. The possibility of many enthusiasts producing a swarm of deepfakes at will might also undercut a campaign of online influence: it would presumably detract from the credibility of any one deep fake of a particular candidate that many other deepfakes were released of the same person at around the same time. That might be less credible than a single deepfake produced by a political dirty tricks department, and released at an expertly chosen moment of a campaign to sections of the public revealed by prior research to be receptive. But a deepfake deployed in this way would not be exploiting the distinguishing feature of deepfakes that I have been discussing, namely wide producibility.

It is plausible to claim that a political dirty tricks department could fruitfully direct a large number of Reddit enthusiasts to release deepfakes in different political contests. But in that case the wide

availability of deepfake technology is less significant than the directing intelligence who decides when and how it will be used. After all, a political dirty tricks department could itself make locally-effective deepfakes and pass them round to volunteers on the ground for posting at the right time. But quite what can be achieved by many uncoordinated political amateurs with access to deepfake technology is unclear to me. Yet this, presumably, is what distinguishes a widely available deepfake technology from the use of specialist technology by a political dirty tricks department.

3.1 Wide producibility: Poisoning the Video Pool

Fallis has seized on the wide producibility of deepfakes to suggest a distinctive problem: a proliferation of deepfaked videos in a communal video pool reduces the power of videos *in general* to enlarge knowledge beyond what we have by direct visual perception in the relevant community. An undoctored photograph or video arguably makes available a vantage point of an eyewitness to an event or an arrangement of objects at a time. Such a photograph or video can therefore sometimes provide vicarious perceptual knowledge of events or situations that it depicts. But if there are many deepfakes in the pool of videos and a subject cannot tell which video to take as evidence for belief, then that pool is tainted and videos in general will not be a source of knowledge, though they may at times (when they are not deepfakes) produce true beliefs.

So one kind of damage done by the prevalence of deepfakes, according to Fallis, is a reduction in knowledge. Further, if subjects come to believe that there are many deepfakes among videos readily available to them, they may *suspend judgement* about the contents of (available) videos. This is a second way in which deep fakes reduce the amount of knowledge: they reduce the credibility of videos in general, increase suspended judgement or non-belief, and therefore (if knowledge is a kind of belief) increase non-knowledge.

Fallis uses a theory adapted from Brian Skyrms to describe the effect of a large infusion of deepfakes into the pool of videos that a subject consults. The effect, he says, would be a reduction in the

information carried by videos in general in the communal video pool. The idea is not exactly that videos *tell* us less in the presence of deepfakes, but that they are *less reliable as evidence* for what they depict. Skyrms' theory[xiii] of information was developed in part to capture what happens when some signal that is relevant for an animal to its survival and reproduction is transmitted by another animal. In a broadly similar way, according to Fallis, videos are a source of signals to human subjects. Other things being equal, that is, in the absence of deepfakes, a video or photograph will convey the information that Obama is speaking when it is more likely that Obama is speaking than that Obama is *not* speaking. But in the presence of lots of deepfaked videos of Obama speaking, the probability that the information is an indicator of Obama speaking is, if not low, then lower than the corresponding probability if deepfakes were absent. Symmetrically, the chances of forming false beliefs about Obama, or of not knowing what to believe about who is speaking in a video, are higher, and this is epistemically damaging, even if few people are actually deceived.

Fallis relies on the supposed easy producibility and reproducibility of deep fakes to formulate this problem. But the formulation is questionable, because it homogenises deepfakes. Recall that deepfakes sometimes have educational purposes. For example, suppose that a deepfaked image of Franklin Roosevelt is used in a video reconstruction of Roosevelt's exploits as a wartime Allied leader. In one scene of the reconstruction the deepfaked Roosevelt makes imagined small talk with other Allied leaders. This detail is invented but mere background for other events that did take place and are central to the video. Suppose that the video is a great commercial success, and that many versions of it are reproduced –*so* many that they begin to compete in numbers with original photographs and copies of original photographs of Roosevelt. Does the existence of *those* deepfakes do the same damage as the deepfakes of Obama are supposed do to subjects according to Fallis?

I think the answer must be 'No'. The reason is that the informational foreground of the videos –its main message-- depicts exploits that are correctly ascribable to Roosevelt, even if the source of the image of Roosevelt is a lookalike or a big collection of pixels in photos of the actual Roosevelt mined

by an algorithm. Like a crime-scene reconstruction that uses actors, the educational video about Roosevelt is not intended to deceive: it shows "Roosevelt" –deepfaked Roosevelt-- doing things Roosevelt actually did, and, at least in its original version, it is explicitly used as educational material --with a claim to broad historical accuracy notwithstanding its use of deepfakes. In this case, contrary to Fallis, the multiplication of deepfaked images would be epistemically *un*damaging.

Another problem with Fallis's account can be posed by asking what exactly makes a deepfake potentially damaging epistemically to a potential receiver of the deepfake. Does the mere *existence* of a deepfake that I *would* take as evidence, but in fact never encounter, count as damaging? Or must the deepfake be somehow actually present or local to be relevant? This question, which is utterly routine in philosophical literature on defeated justification, has a clear bearing on Fallis's account. It makes a difference to the damagingness of deepfakes, presumably, whether the deepfakes are circulating locally or at least in places that come to the attention of my information sources. Presumably deepfakes circulating only among astronauts on the moon or on servers in Antarctica matter less to me in London than deepfakes circulating on social media sites I regularly visit, or on my favourite television network.

Not only must deepfakes be in circulation in my information-gathering ambit, so to speak; in the political case, it must be a vehicle for the communication of a false message to the discredit of the communicator's political enemies. But for me to form a false belief or a belief with defeated justification as a result of contact with relevantly accessible deepfakes, I have in some sense to be susceptible to them; and I need not be. To take the extreme case, if I am blind and deaf, I may be insusceptible, other things being equal, to misleading audio-visual material of any kind. Cultural and linguistic barriers may be another source of insusceptibility. If there is a video in circulation showing a dance considered disgraceful in Thailand performed by someone whose identity is deepfaked, with a voice-over in Thai, I am insulated by linguistic and cultural barriers from feeling the disapproval for the dancer that, let us stipulate, the deepfake-maker intends their audience to feel. And this is to say

nothing of the kind of insusceptibility that is created by extreme personal partisanship and the choice of news sources that echo that partisanship.

Other examples point in the same direction. Take a deepfake of someone I will probably never encounter or think about, say some minor actor in a Moldovan TV drama series. Deepfakes of that actor might be accessible to me –in that they exist on the surface Web, are reachable by a Google search using the actor's name, and are subtitled in English-- but the actor's name is unknown to me and the actor is irrelevant to my interests. Will any number of deepfakes of that actor circulating on the internet damage me as an epistemic subject? Again, and contrary to Fallis, I think the answer is 'No'. By the same token, a malicious deepfake of someone in a friendship group or circle of acquaintances that does not include me or close friends might be epistemically inert, even if it falsely depicts that person engaged in a racy piece of wrongdoing. In other words, the potential for a deepfake to be damaging to a believer depends on the subject of the deepfake having a place in the interests and pre-existing beliefs of that believer. This does not mean that deepfakes are bound to be harmless or unlikely to be harmful; it means that their taking effect is more contingent than is suggested by those who think that deepfakes will soon vastly increase the effects of misinformation.

Now *celebrities* within a local culture or international celebrities are precisely people who command the interest of large numbers of people in the same community, and across communities, and about whom large numbers of people have beliefs and readily form new beliefs. So, other things being equal, misinformation about celebrities is more likely to be epistemically active and potentially damaging epistemically to large numbers of people than misinformation about non-celebrities. Political campaigning in liberal democracies raises some candidates to the level of celebrities or recruits candidates from celebrity circles; so misinformation about them is more likely to be epistemically damaging to large numbers of those politicians' followers, hostile or friendly, and to the general public, than misinformation about non-celebrities.

Even when Fallis's claim is limited to celebrities, however, it may overstate the damagingness of deepfakes. First, there is the sheer quantity of information, true and false, about celebrities: information-overload may drown out some damaging messages. The widely known fact that celebrities are the subject of lots of gossip, false journalism and even faked (but not deepfaked) images may make everyone hesitate to take *any* material about them, not just deepfaked material, at face-value. Again, as already emphasised repeatedly, the loyalties and rationality of the relevant followers may create barriers that make them insusceptible to damaging information. These barriers may include the acceptance of conspiracy theories that explain away damaging information as the product of malice alone. But if both taking in and accepting the content of deepfakes is highly contingent, then so is the infliction of epistemic damage by deepfakes.

3.2 Wide producibility: undermining epistemic backstops

Regina Rini has identified another kind of epistemic damage done by proliferating deepfakes, this time in relation to knowledge and true belief derived from testimony. According to Rini, the more routinely deepfakes are in circulation, the more the capacity of *any* recording to act as what she calls an "epistemic backstop" is undermined. As things currently are, according to Rini, undoctored recordings act as a check on insincerity and deception in testimony in general, and in the testimony of public figures in particular. But the more that deepfakes proliferate, the more likely "backstop crises" will occur. In backstop crises, the use of epistemic backstops to expose insincerity and deception could be undermined by the claim that the epistemic backstops themselves are mere deepfakes.

According to Rini, the acquisition of knowledge by testimony depends on testimonial norms being observed in conversation. She concentrates on two such norms: speakers should be sincere in what they say –avoid deception-- and express competence or authority in relation to the topic –know

what they are talking about. Of course, compliance with these norms does not ensure that those at the receiving end of information believe no falsehoods, since a sincere and competent speaker can be mistaken. Hearers, for their part, have to be prepared for evidence contrary to what they are told and, normally, they will believe that their informants are sometimes wrong. But when no counterevidence *does* come to light and informants are sincere and competent, receivers of testimony are justified in believing what they are told.

"Acute correction" is Rini's term for what happens when a deceptive claim made in conversation is refuted by the production of either a video or audio recording. The production of the White House tapes in 1974 refuted Richard Nixon's denials of any knowledge of the cover-up of the Watergate break-in. The Gutierrez tape mentioned earlier demonstrated that Harvey Weinstein's interactions with actresses in private spaces were not always by mutual consent. Acute correction shows what can be done to enforce testimonial norms against those who break them --by bringing violations to public attention. But what makes acute corrections possible, according to Rini, is "the passive regulation role of recordings": the existence through recordings of authoritative corroboration or disproof of what informants say.

> This is the idea: Part of the reason our ordinary testimonial practice allows us to trust one another to be sincere and competent is that we all know that, at any time, we *might* be within the range of an audio or video recorder, or might be testifying about an event that occurred near such a device (p. 3).

In still other words, recordings act as an "epistemic backstop" –a settler of disagreements over what was said by informants and over whether testimonial norms have been upheld or violated.

According to Rini, a proliferation of deepfakes potentially undermines the effect of epistemic backstops: it makes authentic and inauthentic recordings harder to distinguish. People will start to

treat all recordings as suspect. When deepfakes undermine the epistemic backstop for politically important claims –such as the claim that Nixon had no role in deflecting the investigation of the Watergate break-in—then large-scale *political* damage is added to large-scale epistemic damage. Although Rini seems to be right to say that audio and visual recordings act as a kind of epistemic backstop, I doubt that what they are a backstop *for* is norms of *general* testimonial practice. It is implausible to claim that awareness of the possibility of being recorded colours or ought to colour *every* speaker's, or even every *typical* adult speaker's, testimonial practice. It is more plausible to claim that this knowledge colours or ought to colour the testimonial practice of only some speakers.

Foremost among the relevant speakers might be public figures, including entertainment celebrities and politicians, since these people know that they are the subject of persistent, intrusive and often false media coverage of the fine detail of their lives. At the margins, this knowledge might also colour the testimonial practice of suspects being sought by police, or people at large who are strongly distrusted by their partners or business associates, and who are the subjects of private investigations by institutions or people who distrust them. People in various kinds of disputes might watch what they say for the same reason –aware that their adversaries might collect evidence for a more formal dispute-resolution process, or as material for an act of public shaming on social media.

Rini denies that *only* the testimonial practice of public figures is backstopped by recordings, and I have just agreed. But she claims, implausibly, according to me, that *everyone's* testimonial practice is backstopped by recordings, given the pervasiveness of mobile phone technologies.

> Is it only public figures who expect to be regularly checked by recordings? Or do ordinary people living day-to-day lives have similar expectations? If it is only public figures, then the epistemic backstop function of recordings will be much narrower than I've suggested, since testimonial norms are determined by the behavior of *all* people, not just the famous. In fact, the answer is a bit more complicated; I think the right distinction is not so much between public figures and ordinary people, but rather between public *events* and private lives (p.4).

By “public events” she means everyday activities involving ordinary people –non-celebrities--that are routinely captured by mobile phone videos and photographs, and CCTV imagery. In the period of the Covid pandemic, we can add events captured by Microsoft Teams or Zoom recording software.

It is undeniable that many everyday events are public in this sense. The question is whether our awareness of the likelihood of being recorded makes us more sincere and authoritative in giving information to others. The answer, surely, is, “Not necessarily”. To illustrate, suppose Adam claims in a voicemail message to be working hard in the office when he is in fact drinking in a bar. It seems to me that the voicemail recording is a at best a weak check on his sincerity if the only audience for the voicemail is a casual acquaintance (as opposed to a partner concerned with Adam’s becoming an alcoholic). Let us now tweak the example. Suppose that at the same bar from which he sends a deceptive voicemail, Adam is recorded by a CCTV camera walking drunkenly out to a carpark and getting into a car. Suppose Adam knows he has been recorded and is shortly afterwards stopped by the police. When asked, will he deny he has been drinking? Here the investigatory powers of the police, the existence of the recording *and* testimonial norms push testimony toward sincerity. But it is probably the investigatory powers of police and the legal consequences of being found to be a drunk driver that are decisive.

Our awareness of the possibility of being recorded, I am suggesting, is not necessarily an influence on our *general* observance of testimonial norms: it operates only in cases where we think that what we say might be actively held against us –either by friends and family or by institutions like the police or the media. Conversational partners do not always receive testimony critically, especially if the testimony concerns the trivial and everyday. If that is right, then the role of recordings as epistemic backstops is more limited than Rini claims.

This does not mean, as already pointed out, that only public figures need to worry about recordings. Ordinary individuals do, too, where they try to deceive people who trust them about something both parties consider important, or when they spread damaging rumours during periods of public

disturbance. Still, the role of recordings is not normally to act as an epistemic backstop for testimonial norms in general. Instead, recordings have a special relation to *disputed* testimony. This is how recordings are used in criminal proceedings, and in some journalistic exposures of wrongdoing, especially where the agents responsible go to great lengths to conceal their wrongdoing. Call this the *formal corroborating role* of recordings.

Whether or not Rini is right about the scope of the backstop function of recordings in relation to testimony, is she nevertheless right to suggest that a future proliferation of deepfaked video and audio recording would undermine the backstop function where it does operate?

> The obvious worry about deepfakes is that they will be used to propagate vivid disinformation…
>
> To see the worry, think ahead to a day when deepfake technology is widely available. The problems will start with events I call *backstop crises* moments when the corrective and regulative functions of recordings are made salient, but then quickly undercut by the spectre of deepfakery. Imagine, for example, that Richard Nixon had said: "Look, that wasn't me on the smoking gun tape. They used that VoCo technology to make it sound like me ordering CIA interference. But it wasn't!" (p.7)

The claim that Rini puts into Nixon's mouth would not have been could not have been formulated in 1974. But now imagine that late in the 2020 US presidential campaign, a recording had emerged in which someone who certainly sounds like Donald Trump is apparently colluding with Russian intelligence operatives to discredit a rival . Suppose Trump insists that it wasn't him, that he has been deepfaked into an entirely fabricated conversation.

It is not necessary for people to accept that the audio is genuine for it to have a subversive effect, according to Rini. Repeated backstop crises can instead have the effect of making people disbelieve

in general that audio or video recordings in general are conclusive. And this is the distinctive harm of deep fakes.

I have offered counter-examples to Rini's general claim that recordings play the role of backstop. I deny that they play this role with respect to norms governing testimony in general. I now want to call attention to a question-begging aspect of her example of an epistemic crisis event. Her example is an updating of the revelation of the tape that led to Nixon's resignation. The tape showed that, contrary to Nixon's denials previously, he was aware of the cover-up of the Watergate break-in. One reason why this revelation was so momentous in 1974 is that there existed at the time what might be called *a common focal point* of media attention. The newspapers and television in America were full of questions about what Nixon knew as well as suspicions that he knew quite a lot. What is more, the newspapers and television sources agreed in their accounts of the facts, for the most part, and these sources of information were intensely consulted by most of the public. No divide existed, as it does now in the US, between the partisan news agendas of Fox on the one hand and CNN on the other. Social media did not yet exist. In short, a few news sources spoke in much more of a single voice to an attentive, mostly trusting, public.

There is no updating the Nixon story to include a deepfake without also updating the state of the American media and its audience. In 2021, there was, notoriously, a highly silo-ed media audience often receiving openly partisan "news" from a few favoured sources, not all of them professed news organizations. It is unclear, therefore, whether there was a common focal point of attention of this audience even by the criterion of what is "trending", and it is not clear how much was taken in for how long, or how much of what was taken in was believed. This means that it is very difficult to recreate today moments of nationwide concentration that were routine in the USA in the last days of the Nixon administration. But in order for an event to trigger a "backstop" *crisis*, it has to unsettle the beliefs of a big audience, and so it has to register with a lot of people at roughly the same time,

all or most of whom react by trading belief for uncertainty. This is quite a feat with respect to an audience with no centre of attention.

There is another aspect of the media landscape of 2021 that distinguishes it from 1974, and this is the widely acknowledged presence of misinformation circulating online and on traditional broadcasting networks. In the USA, even in 2022, there is misinformation on a large-scale about voting irregularities in the 2020 Presidential election. There is misinformation about the origins of the Covid pandemic and the dangers of receiving vaccinations against the Covid virus. And this is to say nothing of the very large numbers of false rumours and other claims routinely spread about celebrity actors. If there is already an epidemic of misinformation available in the electronic and traditional media, how can the additional misinformation that may be created by a future proliferation of deepfakes make the situation any worse epistemically?

Rini suggests that, until the advent of deepfakes, audio and video recordings were a last line of defence against the insincere and the ill-informed among purveyors of testimony. And so her answer to the question of what makes the situation worse is that the existence of deepfakes weakens or removes the authority of video and audio footage, and so removes the last line of defence against misinformation. But this claim seems wrong, for three reasons. First, evidence that is better than recording –namely unedited live television broadcast to a large, attentive audience-- is open to reinterpretation by those whose interests it goes against, as in the case of the insurrection in Washington on 6 January 2021. Second, in periods in which misinformation was far less widely available, photographs and recordings –say of the Loch Ness Monster or supposed UFOs or ghosts— could still be faked, and even if not faked, reinterpreted by the sceptical. This means that they were *not* accepted as authoritative. The fact that these photographs and recordings were not deepfaked, and were not high-quality, does not mean that they were easy to discredit. Third, audio and video can be of action or speech that is *scripted* or *managed* in order to elicit a certain response, as when

people wearing concealed microphones agree to try and capture incriminating actions or admissions.

The events of 6 January 2021 are worth dwelling upon, because they are relatively recent, were widely observed via live television by people around the world, have led to criminal charges by processes meeting due process safeguards, including the use of recordings as formal corroboration, and yet have been denied or explained away by high political office-holders in the USA. The same events have also been denied or explained away by millions of people who belong to the Trump political base. They have been denied and explained away even on one of the rare occasions when, as the insurrection developed, there was enough time for it to become a genuine focal point of media attention. If people like these in the TV audience do not believe their own eyes, or are prepared publicly to reinterpret what they saw with their own eyes, perhaps because they engage in "motivated reasoning" that preserves their pre-existing beliefs (Jennings and Stroud 2021) what more harm could be done by deepfakes?

My own view is that, in a highly partisan, polarized media environment that is full of willingly consumed testimonial misinformation, the addition of deepfaked audio and video is unlikely to make things *much* worse, if worse at all.[xiv] One reason for this, I am claiming, is the lack of a focal point of attention. This means that deepfakes are likely to have effects only in some regions of social media, and only for so long. The regions of social media in question may be big enough or full enough of easily triggered activists to create violence where the activists are concentrated, even in the absence of a focal point of attention. But, in relation to the easily triggered, it is unclear why deepfaked video and audio material --as opposed to something with lower quality, and less dependent on AI--would not have the same effect. Finally, and as already pointed out, it is not easy to make any old piece of misinformation influential among the less easily triggered. The bigger the atrocity portrayed, the more likely, presumably, it is to have an effect, but only if it is believed by

enough people, and only if those people are able and willing to retaliate against the agents of the atrocity. These are big ifs, much bigger ifs than pessimists about deep fakes seem to acknowledge.

4. Conclusion

I have been arguing that the distinctive properties of deep fakes –high-quality and wide reproducibility—do not suit them especially well to political misinformation. High quality in fakes seems neither necessary nor sufficient for persuading an audience that a particular depicted subject has done something discreditable, and this is the central example in the literature of political misinformation. Wide availability of deep-fake technology is, again, neither necessary nor sufficient. If many people create and post their own pictures of discreditable activity, that may make explicit a concerted campaign of misinformation targeting a particular politician and will be self-defeating. Images produced by amateurs will not necessarily be believable. The wider availability of deepfaked technology can also come to the attention of many in the audience to which deepfakes are directed, producing grounds for doubt or uncertainty, rather than conviction that discreditable acts have in fact been done.

It is also disputable whether a proliferation of deepfaked video and audio recordings reduces the number of authoritative sources of perceptual knowledge or undermines the trustworthiness of testimony. The mere existence of deepfaked material does not mean that it exercises an influence on perceptual beliefs. People may even be insusceptible to them, and for mundane reasons. In such cases deepfakes may be epistemically inert. The norms that supposedly govern all testimonial transactions, and that might be undermined by deepfakes, in fact govern many fewer kinds of transactions. Again, a whole spectrum of conditions have to be met for deepfakes to be epistemically damaging. In the case of political misinformation in liberal democracies, deepfakes have to reach and be believed by a wide audience to deceive or produce uncertainty, and this sort of

effect is hard to achieve: for one thing, the audience and its attention are both very fragmented. And the capacity of some of the audience to deny or reinterpret even what they see with their own eyes –when it clashes with their deep convictions—is not to be underestimated.

REFERENCES

Bennett, W.L. and Livingston, S. 2018. "The disinformation order: disruptive communication and the decline of democratic institutions." *European Journal of Communication* 33(2): 122-139.

Citron, D.K. and Chesney, R. 2019. "Deep fakes: A looming challenge for privacy, democracy, and national security." *California Law Review*, 107: 1753-1819.

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

Cole, S. 2018. "Deepfakes Were Created As a Way to Own Women's Bodies – We Can't Forget That". *Vice* https://www.vice.com/en_us/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2

De Ruiter, A. 2021. "The Distinct Wrong of Deepfakes." *Philosophy and Technology* 34: 1311–1332.

Diakopoulos, N and Johnson, D. 2020. "Anticipating and addressing the ethical implications of deepfakes in the context of elections." *New Media and Society* 23: 2072-2098.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., De Vreese, C. 2021. "Do (Microtargeted) Deepfakes have Real Effects on Political Attitudes?" *International Journal of Press/Politics* 26: 69-91.

Enos, R. and Hersh, E. 2015. "Part Activists as Campaign Advertisers: The Ground Campaign as a Principal-Agent Problem." *American Political Science Review* 109: 252-278.

Fallis, D. 2020."The Epistemic Threat of Deepfakes." *Philosophy and Technology* 34: 623–643

Faris, Rob, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. "Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election."2017. cyber.harvard.edu/publications/2017/08/mediacloud

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

Flynn, D. Nyhan, B.and Reifler, J. 2017. "The nature and origin of misperceptions: Understanding false and unsupported beliefs about politics." *Political Psychology* 38 :127-150.

Foer, F. (2018). "The end of reality". *The Atlantic*. https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/

D. E Glasford, D. 2013. "Seeing is believing: Communication modality, anger, and support for action on behalf of out-groups". *Journal of Applied Social Psycholo*gy 43: 2223–2230.

Harris, Keith R. "Video on Demand: What Deepfakes Do and How They Harm". *Synthese* 199:13373-91.

Jafar, M.T., Ababneh,M, Al-Zoube, M and Elhassan,A. "Forensics and Analysis of Deepfake Videos." 2020. *11th International Conference on Information and Communication Systems (ICICS)*: 053-058, doi: 10.1109/ICICS49469.2020.239493.

Jennings, J. and Stroud, N. 2021. "Asymmetric adjustment: Partisanship and correcting misinformation on Facebook" *New Media and Society* published online June 2021.

Johnson, B. 2019. "Deepfakes are solvable - but don't forget that "shallowfakes" are already pervasive." *MIT Technology Review*

https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/.

Kerner, C and Risse, M. 2021. "Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds." *Moral Philosophy and Politics* 8: 81–108.

Modgil, S., Singh, R.K., Gupta, S. *et al.* 2021. "A Confirmation Bias View on Social Media Induced Polarisation During Covid-19." *Information System Frontiers* . https://doi.org/10.1007/s10796-021-10222-9. Published November 2021.

Nickerson, R.S. 1998 "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises" *Review of General Psychology* 2: 175-220.

Paris, B. & Donovan, J. 2019. "Deepfakes and cheap fakes: The manipulation of audio and visual evidence." *Data & Society* (2019) https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf.

Prior,M. 2014. "Visual political knowledge: A different road to competence?" *Journal of Politics* 76: 41–57.

Rini, R. 2021. "Deepfakes and the Epistemic Backstop." *Philosophers' Imprint*.

Rini R and Cohen L, 'Deepfakes, Deep Harms' (n.d. forthcoming)

Rohlinger, Deana A. 2021. "Persuasion and Non-Party Groups in the Digital Age." In *The Oxford Handbook of Electoral Persuasion* ed. E. Suhay. 321. Oxford: Oxford University Press.

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

Taber CS and Lodge M (2006) Motivated skepticism in the evaluation of political beliefs. American Journal of Political Science 50(3): 755–769.

Vaccari, C. & Chadwick, A. 2020. "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news." *Social Media + Society* 6 (2020):1-13.

Wittenberg,C. Tappin,B. Berinsky, A. Rand, D. 2021. "The (minimal) persuasive advantage of political video over text." *Proceedings of the National Academy of Sciences* 118 (47).

Zuiderveen Borgesius, F.J., Moller, J. Kruikemeier, S., O Fathaigh, R. Irion, K. , Dobber, T., Bodo, B and de Vreese, C. 2018. "Online Political Microtargeting: Promises and Threats for Democracy." *Utrecht Law Review* 14(2018): 82-96.

[i] An anonymous referee has pointed out similarities between some of the points I make in what follows and those put independently by Harris (2021), whose work had not previously come to my attention. Like Harris and others (Paris and Donovan (2019), I doubt that deepfakes necessarily constitute a huge departure from earlier, more crude ("cheapfake technologies"). Like Harris and others, I think that the dangers of deepfakes have been overblown. But having doubts about the doomsday scenarios sometimes proposed (see e.g. Foer (2018)) is not the same as denying that deepfakes are potentially capable of producing significant harm. Harris and I agree that harm is possible and even likely in some circumstances. But the considerations we employ are different. For example, Harris has no extended discussion of the mechanics of party political persuasion. Again, he operates with an account of the effects of knowledge of sources of deepfaked information (see his section 3) that I think underestimates the ease of masking sources online.

[ii] Fallis, 2020 .

[iii] Rini, 2020.

[iv] Floridi (2018) has used the term "ectypes" for 3-D printed paintings that result from modelling data concerning the typical subjects, actual brushstrokes, and actual materials in genuine works of famous painters. The ectypes are distinct from, but deeply in the style of, works the painters actually produced, and are distinct again from the works of forgers. They have a claim to be more authentic than mere copies, because they are in some sense the result of a deep distillation of the techniques and materials that produced genuine works. In the same way, the deepfaked Obama is a distillation of the content of unfaked, unstylised images of the real Obama, and so have Obama's imprint in a way that no Obama look-alike or an image of an Obama look-alike could. Or, to take another example of Floridi's, the audio deepfake of the speech Kennedy planned to deliver in Dallas if he had not been shot is a distillation of audio data from the real Kennedy speaking during his life-time. It is an ectype, though the speech was never given by Kennedy.

[v] Cole, 2017, 2018.

[vi] See Kerner and Risse (2021)

[vii] According to De Ruiter (2021),

> The most distinctive aspect that renders deepfakes morally wrong is when they use digital data representing the image and/or voice of persons to portray them in ways in which they would be unwilling to be portrayed. Since our image and voice are closely linked to our identity, protection against the manipulation of hyper-realistic digital representations of our image and voice should be considered a fundamental moral right in the age of deepfakes.

This claim makes sense in the case of deepfaked subjects of pornography and deepfaked depictions of someone apparently accepting cash in secret who has never done any such thing. But it also seems to condemn the use of deepfakes used to satirise a public figure who does not want to be satirised. Again, de Ruiter seems to overstate the importance of the wishes of a subject of a deepfake in relation to that subject's public image. For example, suppose that someone does not want to appear to anyone to be a serial blackmailer or sexual harasser, but is *in fact* a serial blackmailer or sexual harasser. Then, though a deepfake embeds the image of that subject in an episode of blackmail or harassment that never took place, the deepfake need not be unfaithful to the kind of thing the actual subject actually did on many occasions. If the intention of the deepfake manufacturer is to expose this aspect of the subject's life, say to alert the unwary to the dangers of interacting with the actual subject, the fact that the subject does not like the depiction is not decisive in favour of the claim that the deepfake wrongs him. A digital image, to the extent that is attached to a reputation, should not be fully in the control of the subject whose image it is. A reputation is or ought to be based on a person's actual exploits, combined with an impartial moral valuation of those exploits conducted by others. No-one is, or ought to be, fully in charge of their reputation, especially if they engage in very stealthy, serious wrongdoing, or worse, engage in stealthy, serious wrongdoing and publicly proclaim their own virtue. On the other hand, a deepfaked video of what is only characteristic behaviour ought to be circulated as what it is: a depiction of a *kind* of behaviour the subject gets up to, not a depiction of an actual specimen of that behaviour.

[viii] Diakopolous and Johnson, 2020: "One of the earliest systems [of media synthesis technology], presented more than 20 years ago, was already capable of synthesizing speaking faces by splicing together a series of mouth shapes from footage of a person to align to newly input speech. The latest techniques, however, are capable of far greater fidelity and believability due to increased resolution and quality of image sensors, the availability of more data, and advancement in machine learning techniques, such as deep neural networks, that utilize that data. The clear trajectory of the technology is toward more realistic and believable synthesized depictions."

[ix] Rini, p. 5; see also Fallis op.cit., in the sub section entitled, 'The Epistemic Threat of Deep Fakes': "Machine learning can make it possible for almost anyone to create convincing fake videos of *anyone* doing or saying *anything".* This claim is contradicted by Paris and Donavan. Fallis refers to the availability of an app, FakeApp, for creating deepfakes, but the products of this app are not comparable to Peele's products. The product of FakeApp is the substitution of one person's face in a video –almost always a pre-existing video taken from the internet, with another's face. This technology is limited by the content of pre-existing videos, and by the extent to which that content matches the circumstances of e.g. the person being discredited for political or other purposes. If only faces are substituted, other images in the video can be matched to pre-existing video, for example, making a FakeApp deepfake in principle identifiable and challengeable (Jafar et al, 2020).

[x] "[E]xisting recordings of a person's mouth movements and voice can be used to reverse engineer their speech to have them say any sentence. The results can be alarmingly convincing, especially with the low-resolution video that is common online." Vaccari and Chadwick, op.cit. p.2. Convincingness is here tied to a lower standard of resolution being routine on the internet than is afforded by deepfake technology.

[xi] On a website for those wanting to use the DeepFake app, Alan Zucconi says, "Despite what media is claiming, creating deepfakes is not easy. To be more precise, creating deepfakes is very easy, but creating good ones is not." https://www.alanzucconi.com/2018/03/14/how-to-install-fakeapp/

[xii] For a picture of the highly complex left-right media environment at the time of Trump's campaign in 2016, see Faris et al (2017). See also Rohlinger (2019).

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

[xiv] The problem of shallow fakes (videos e.g. of atrocities simply labelled misleadingly) might be more significant. See Johnson, 2019.