Deepfakes, Deep Harms

Regina Rini and Leah Cohen

York University, Toronto

This paper has been accepted to run in the *Journal of Ethics and Social Philosophy* (as of July 2021). It has not yet been through copy-editing. Please cite the official version once it appears.

Abstract Deepfakes are algorithmically modified video and audio recordings that project one person's appearance on to that of another, creating an apparent recording of an event that never took place. Many scholars and journalists have begun attending to the political risks of deepfake deception. Here we investigate other ways in which deepfakes have the potential to cause deeper harms than have been appreciated. First, we consider a form of objectification, *virtual domination*, that occurs when deepfaked 'frankenporn' digitally fuses the parts of different women to create pliable characters incapable of giving consent to their depiction. Next, we develop the idea of *illocutionary harm*, in which an individual is forced to engage in speech acts they would prefer to avoid in order to deny or correct the misleading evidence of a publicized deepfake. Finally, we consider the risk that deepfakes may facilitate campaigns of *panoptic gaslighting*, where a myriad of systematically altered recordings of a single person's life undermine their memory, eroding their sense of self and ability to engage with others. Taken together, these harms illustrate the roles that social epistemology and technological vulnerabilities play in human ethical life.

Deepfakes are digitally altered audio or video recordings in which one person's face and/or voice are mapped onto the body of another person, creating misleading evidence of events that never took place. Deepfake videos can be created with open-source software based on machine learning algorithms. Currently this technology is a niche interest, mostly isolated to internet pornography communities, but its use by other actors – some with still more malicious intent – is a very near technological possibility. The aim of this paper is to get out in front of that future, to recognize some of the emergent harms the technology may bring about.

We think that the arrival of cheap and easy to use deepfake technology will have a number of significantly harmful effects, at both personal and social levels. For simplicity, this paper will focus on the *personal* harms of deepfakes. (We discuss the social and political consequences elsewhere.[1]) Put simply: what happens *to* a person who has been deepfaked? What sort of harm is done by being falsely represented in a deceptive recording?

[1] See Rini (2020).

The structure of the paper is as follows. First, we provide a bit more background on the history and technology behind deepfakes. Then we marshal a parade of horribles: several distinct ways in which deepfakes may harm their targets. These include: a new form of objectifying harm that we call *virtual domination*, in which a person's autonomy is invaded by their being represented as engaging in unconsented (and fabricated) sexual encounters; *illocutionary harm*, where a person is forced to engage in involuntary speech acts in order to dispute the content of deepfakes; and, most speculatively, existential trauma caused by *panoptic gaslighting*, when a person's memory and identity are undermined by a myriad of systemically targeted fabrications.

Discount digital deception

The deepfake technology that is available to consumers superficially resembles Hollywood special effects, the sort where dead actors are revived anew onscreen. But these professional techniques are extremely expensive and time-consuming. What makes deepfakes remarkable is that they provide an approximation of the same effect, much more cheaply and quickly. The knowledge, resources and time needed to create deepfakes are substantially lower than for other kinds of video manipulation, and can be done as anonymously as the internet will allow.[2]

Another difference, of course, is that Hollywood stars usually *consent* to being digitally dopplegangered. That's rarely true of the targets of deepfakes. Pornography, always a central causal factor on the internet, drove the early use of deepfakes, including the origin of the term itself. In autumn 2017, an anonymous user going by the handle "deepfakes" posted multiple pornographic videos to the website Reddit. These videos featured the faces of famous actresses mapped onto the bodies of pornographic performers engaging in explicit sexual acts. By February 2018, Reddit and other sites - even PornHub - had banned the videos.[3] But policing the internet is an unending task and the videos abound. A recent study by the digital security firm DeepTrace found that 96% of online deepfakes are pornographic.[4]

By late 2018 public attention began to focus on the potential political risks of deepfakes. In May of that year, the Flemish Socialist Party released a deepfake of Donald Trump appearing to urge Belgium to

[2] Journalist Samantha Cole has done extensive work investigating deepfake technology and the internet communities that favor it. See Cole (2017).

[3] Kelion (2018).

[4] Simonite (2019).

withdraw from the Paris Climate Accords. The video is noticeably unreal. The mouth of the speaker is out of sync with the rest of Trump's facial expressions. If that weren't enough, the final line of the video clearly states, "we all know climate change is fake, just like this video." Yet still, according to *Politico*, "some commenters on the party's Facebook page had apparently not realized the video was a fake".[5]

By mid-2019, government and corporate policymakers had begun debating solutions. In June, the US House Intelligence Committee held hearings on the risks of deepfakes, while the state of Virginia banned the use of deepfakes in "revenge porn".[6] In January 2020, Facebook announced that it would ban deepfakes from its platform, though it granted a somewhat amorphous exception for "parody and satire".[7] Experts continue to debate whether technical or legal solutions are feasible.[8]

We will assume that a solution is not immediately forthcoming. Our goal instead is to illustrate why a solution is urgently *needed*. What are the harms that deepfakes might cause if left unchecked? Some are already recognized in journalism and legal scholarship. Law professors Bobby Chesney and Danielle Citron (2019), for example, enumerate risks of election interference, corporate malfeasance, psychological espionage, and personal blackmail. We agree that these are serious concerns, but we think that the philosopher's lens may help us see more subtle dangers. We turn now to these.

Frankenporn and Virtual Domination

Deepfakes offer their creators a disturbing form of power over other people, one that seems inevitably to lend itself to pornographic misuse. The first faceswapped videos on Reddit featured actresses such as *Wonderwoman* star Gal Gadot engaged in simulated incest. The actresses were not consulted and did not consent to having their images used in such a manner, nor did the pornographic performers. This fact is what eventually led Reddit to shut down its deepfake forum, under its policy against "involuntary pornography". Yet the videos still circulate voluminously in dingier corners of the internet.

[5] Ven der Burchard (2018).

[6] Cox (2019). See https://www.youtube.com/watch?v=tdLS9MlIWOk for the House Intelligence meeting. Rep. Yvette Jones introduced a bill targeting deepfakes at the US federal level, though (as of writing of this paper in late 2020) it has not been acted upon. See https://www.congress.gov/bill/116th-congress/house-bill/3230

[7] Shead (2020).

[8] For discussion of proposed solutions (and their shortcomings), see Farid (2018), Chesney and Citron (2019), Harris (2019), Li and Lyu (2019), Rini (2020).

Deepfake technology crashes into long-running debates about pornography and the objectification of women. In the 1970s and 80s, feminist critics like Andrea Dworkin and Catharine MacKinnon argued that porn functions to affirm perceptions of women as playthings for male viewers, mere objects for gratification rather than full persons with autonomous wills. Other theorists have pointed to the ways in which pornography silences the viewpoints of women. These positions are controversial though, even within feminist communities. Some feminists argue that pornography, when executed carefully and respectfully, can be compatible with or even empowering of women's liberation. Such 'porn-positive' feminists emphasize the agency of individual performers – not to mention female directors and distributors – in designing pornography that expresses women's sexuality without shame.[9]

This is an unresolved debate, one with a good deal of subtlety to it. Even if Dworkin and others are correct that the *general* social function of pornography is to objectify women, it still might be true that the *local* function of some feminist-created pornography is empowerment and de-stigmatization. Women-in-general do not consent to how they are represented in pornography, even if some women consent to their personal presentation in specific pornographic works. Weighing these two points is extremely difficult. But deepfakes obliterate any subtlety or nuance because *no one* consents to deepfake porn.

Journalist Samantha Cole interviewed women who have worked as pornographic performers to get their views about the emergence of deepfake porn. Retired performer Alia James told Cole: "It's really disturbing… It kind of shows how some men basically only see women as objects that they can manipulate and be forced to do anything they want... It just shows a complete lack of respect for the porn performers in the movie, and also the female actresses."[10]

There's something painfully literal in the sort of objectification at work in deepfake porn. Reddit forum users requested the creation of custom videos, with particular actresses swapped into particular sex acts, as casually as specifying the paint job at a car dealership or ordering toppings on a pizza. And as the technology improves, the ability to treat women's images as playthings will only grow. One emerging technique, developed by computer scientists without bad intentions, uses AI to simulate the movements of a real person's entire body by mapping it onto an actor's poses.[11] Once a similar technique is available to deepfakers, they'll no longer be limited to superimposing famous faces onto existing porn clips. Instead

[9] This is an extremely large debate. For important contributions, see Dworkin (1985), MacKinnon (1987), Nussbaum (1995), Strossen (1995), Langton and Hornsby (1998), Maitra (2009), and Bauer (2015). For a recent overview and reorientation, see Cawston (2019).

[10] Cole (2017)

[11] Liu et al. (2019).

they will generate novel simulacra of their targeted celebrities – poseable, pliable representations ordered to do whatever the user desires.

For now, deepfakes are limited to what Cole calls "frankenporn", with the digitally-manipulated face of one woman stitched onto the body of another. Once again, this seems to be an unsubtle manifestation of the worst sort of objectification that feminist critics have always charged to pornography. In deepfake frankenporn, women really are reduced to body parts: a face from here, a torso from there, interchangeable and commodified. What's absent is any sort of independent mind or will.

In an important sense, the entity depicted in frankenporn *cannot* have a determinate will, since it is a composite of the parts of two different people, unified only in digital artifice. This entity is a mereological sum, constituted from the body parts of multiple humans. Its apparent intentions belong to neither of the women whose body parts appear. Instead, it seems to depict a "derived intentionality", like a fictional character, specified by the deepfake's creator.[12]

Yet it does seem to matter to deepfake porn consumers that they are viewing the faces of *particular* women. Their requests target specific celebrities, or in some cases their own ex-girlfriends or acquaintances. This apparent need - to externalize a fantasy of some specific woman doing whatever the user demands - supports the familiar feminist claim that, for at least some men, sexual domination of women is as much about *power* as it is about physical gratification. There certainly appears to be an anti-feminist politics in the communities where deepfake porn is traded. As Cole puts it:

> In these online spaces, men's sense of entitlement over women's bodies tends to go entirely unchecked. Users feed off one another to create a sense that they are the kings of the universe, that they answer to no one. This logic is how you get incels and pickup artists, and it's how you get deepfakes: a group of men who see no harm in treating women as mere images, and view making and spreading algorithmically weaponized revenge porn as a hobby as innocent and timeless as trading baseball cards.[13]

Deepfaked frankenporn, then, is *virtual domination*, an extreme expression of sexual objectification aimed against specific women. As Dworkin puts it, "Objectification occurs when a human being, through social means, is made less than human, turned into a thing or commodity, bought and sold. When objectification occurs, a person is depersonalized, so that no individuality or integrity is available socially…"[14] Frankenporn turns real people into digital toys. Even those unpersuaded by feminist objections to traditional pornography ought to recognize the moral wrong here.

[12] "Derived intentionality" comes from Searle (1992).

[13] Cole (2018).

[14] Dworkin (1985), 15.

Illocutionary harm

We turn now to another potential harm of deepfakes, one more closely tied to their epistemic effects on public discourse. Deepfakes do not have to trick anyone in order to be harmful. Even if a deepfake is ultimately debunked, or never believed at all, it can still hurt the person it falsely depicts by changing the discursive context around them. This point is most clear for public figures like politicians or celebrities. When deepfakes illegitimately force a public figure to react with undesired speech acts, they cause what we will refer to as *illocutionary harm.* In this section we will explicate what makes this a distinctive sort of harm, then catalog several forms it might take.[15]

There's a legend about the American politician Lyndon Baines Johnson. Facing loss in a Texas congressional race, Johnson instructed an aide to spread rumors that his opponent engaged in sex with pigs. "We can't get away with calling him a pig-f****r," said the campaign manager. "No one's going to believe a thing like that." Johnson replied: "I know. But let's make the son-of-a-bitch deny it."[16]

Imagine a 2020s version of the LBJ legend. Now the porcine indecency is no longer mere rumor; instead it has been deepfaked, with the opponent's head digitally inserted into a video of the alleged act. (Best to imagine this case only schematically.) The video quickly goes viral online. On cable news, experts debate the video's veracity while blurred-for-TV excerpts play in the background. Late night comics quickly join in. Most people realize that it's probably fake, but they still laugh along. At first, the politician tries to simply ignore the video, but soon it is everywhere. It becomes hard to do any interview, as even respectable journalists start asking thinly coded questions. Opposing party operatives turn up at rallies dressed in pig costumes. Finally, the politician's aides say: this is only going to stop if you address it directly, once and for all. The press conference is called, the podium prepared. And so there, on live TV, is the son-of-a-bitch denying he f****d a pig.

This is bad. All else equal, people who aspire to public office should not have to call press conferences to deny false allegations of unnatural congress with livestock. In fact, we think that a person placed in this

[15] As this paper went to final editing, we became aware of a very recent paper by Henry Schiller exploring the same term "illocutionary harm" (Schiller 2021). Our use of the term is not the same as Schiller's, though there is some interesting overlap.

[16] That's the frequently told legend, anyway. It's almost certainly not true. The most plausible source we've found is Joseph Califano, an LBJ aide in the 1960s. In Califano's version, LBJ was actually the protesting young staffer in this story! It was LBJ's mentor, Richard Kleberg, who played the 'let him deny it' card against an opponent. Also, the barnyard consort was a sheep, not a pig. See Califano (1991), 118. But we've kept the legend in our main text since it's what frequently appears in political journalism.

position has been harmed, *even if their denial is effective*. That is, even in the unlikely event that everyone immediately accepts the denial and ceases to believe that the video is veridical, the denier has still been harmed simply by having to issue the denial.

The key idea of illocutionary harm is this: a person can be harmed by being illegitimately compelled to perform an undesired speech act. Setting deepfakes aside for the moment, think of simpler examples. Totalitarian regimes frequently force their citizens to engage in compelled speech. Under Mao, the Chinese Communist Party ordered comrades to write 'self-criticism statements', confessions of their complicity in capitalist villainy. Many were also pressured to falsely testify against friends and family. Similar things happened in the Soviet Union and Nazi Germany.

These are all examples of illocutionary harm. Importantly – and perhaps controversially – we hold that this is a distinctive *type* of harm, in that it does not wholly reduce to other types of harm or wronging. Against our view, one might insist that illocutionary harm reduces to some combination of psychological anguish or reputational effects. But we think this misses a key feature of compelled speech. When a person is illegitimately compelled to speak, they are abused *specifically in their capacity as a speaker*.

What does that mean? We have in mind here something akin to Miranda Fricker's account of testimonial injustice. According to Fricker, when a person's testimony is unfairly dismissed on the basis of their membership in a derogated social category, that person has been "wronged in one's capacity as a knower". In addition to whatever material harms might result from not being believed, the target is undermined "in a capacity essential to human value" and so "suffers a great injustice".[17]

Similarly, we think that illegitimately compelled speech involves a distinctive type of harm, a harm to one's capacity as a user of information. In this we follow Rachel McKinney's work on "extracted speech" (McKinney 2016). McKinney's primary examples concern coercion, such as when psychological pressure drives innocent people to confess to crimes. Such pressure "amounts to wrongly undermining, bypassing, or overriding an agent's ability to speak voluntarily" (266), and wrongs victims "as communicative agents" (259).

McKinney distinguishes two ways that extracted speech can be wrongful: first, in a forward-looking way, it can *license* future wrongs against victims (as when an extracted confession licenses the unjust conviction of an innocent defendant); and, second, the mere act of extracting involuntary speech can *itself* be wrongful (regardless of further consequences) when it comes about through subverting a person's communicative agency.

[17] Fricker (2007), 44.

We think that deepfakes pose similar risks. To start off simply, take McKinney's first form of wrongfulness: licensing future wrongs against victims. Suppose a deepfake succeeds in tricking some part of an audience into believing that the target said words they never actually used. This will often license illegitimate treatment.

This sort of harm has already been caused by much simpler manipulations than deepfakes. In 2016, the then-governor of Jakarta, Basuki Tjahaja Purnama, widely known as Ahok, gave a public address decrying his opponents' partisan use of religion. An edited video soon appeared online, in which a word had been clipped from Ahok's remark, causing it to sound as if he were criticizing the Koran itself and not his opponents' appropriation thereof. An enormous public outcry followed, resulting in Ahok losing his governorship and being imprisoned on charges of blasphemy.[18]

Similarly: in 2015, the American anti-abortion pressure group Center for Medical Progress released surreptitious recordings of a 'sting' meeting with representative of Planned Parenthood, maliciously edited to make it appear the latter admitted to profiting from the sale of fetal body parts. As a result, several state governments cut Medicare funding to Planned Parenthood. Then-candidate Donald Trump cited the recordings as grounds for ending federal funding, an ambition he fulfilled in 2019.[19]

In both cases, not everyone believed that the edited videotapes were veridical. But to those who did, the videotapes appeared to license punishment that was in fact unjust. Ahok and Planned Parenthood both suffered wrongs by being portrayed as saying something other than what they actually said.[20]

Yet even when the faked video is widely disbelieved, deepfakes could still impose illocutionary harm, along the lines of McKinney's second type. As we've already stressed, being forced to publicly *deny* an embarrassing rumor can itself be harmful, partly for the reasons McKinney identifies: it subverts the victim's communicative agency. There are many understandable reasons that one may not wish to publicly speak on a topic, such as tact, embarrassment, privacy, and safety. Being compelled to do so by a fabricated recording unjustly compels speech.[21]

[18] Soeriaatmadja (2017). The person who edited the video, a university lecturer, was *also* sent to prison, separately, on hate crime charges.

[19] Kliff (2015), Diamond (2015), Armstrong (2019).

[20] Technically this is probably not "extracted speech" in McKinney's sense, since in these cases the depicted speech act never even happened (at least not as portrayed). One might call this phenomenon "imposter speech".

[21] Some philosophers argue that a person's moral interest in privacy is precisely about being able to control their own social self-presentation; see Nagel (1998), Velleman (2001), Marmor (2015). Thanks to an anonymous referee for suggesting this connection.

Illocutionary harm may happen even when the fabricated speech is truthful and consistent with the (apparent) speaker's beliefs. A deepfake might make a public figure appear to say something that they *do* believe, yet for whatever reason did not wish to express publicly. In other words, the *occurrence of the speech act* might be faked, though the content it expresses might be truthful. The target may then be forced to publicly address the circulating fake, either falsely denying they believe what had been attributed to them, or openly admitting what they'd rather have left unspoken.

An obvious example of this kind of situation happens when a public figure is forced out of the closet. In 2001, a tabloid published (genuine) photos of the Australian-American actress Portia de Rossi with her then-girlfriend. De Rossi did not publicly identify as a lesbian and struggled with how to respond. She later told *The Advocate:* "The most important thing for me was to never, ever, ever deny it. But I didn't really have the courage to talk about it."[22] For the next several years de Rossi avoided talking to reporters about the topic, until she officially came out when she began dating her future wife, Ellen DeGeneres.

In this case, de Rossi's sexuality became a sort of open secret. Hollywood people knew about the tabloid images, of course, but so long as she didn't address them, respectable publications avoided bringing it up. Imagine. however, that deepfake technology had been available in 2001. Imagine someone made a deepfake video seeming to depict Portia de Rossi saying to the camera: "I am a proud lesbian and I want the world to know." In that case, she could not have simply ignored it and counted on respectable media to cooperate; without an explicit denial, respectable media would take it as legitimate news. De Rossi would have been forced to denounce the video – and in doing so, forced to either deny or confirm the rumor, neither of which would be voluntary.

We can see similar risks already in existing technology. If a person gets their hands on your mobile phone, they can send messages to your loved ones, posing as you. A malicious or merely paternalistic acquaintance might say things you think are true but for whatever reason do not wish to say. There's a striking example in Kristen Roupenian's short story 'Cat Person'. Margot is a college student in an unhappy relationship with Robert. She begins ignoring his texts, hoping he'll simply disappear from her life. But he keeps trying to contact her. Finally, her friend Tamara takes her phone and sends Robert the following message: ""Hi im not interested in you stop textng me." Margot is horrified by this, she images Robert "picking up his phone, reading that message, turning to glass, and shattering to pieces."[23]

Yet Margot does not send a followup message disclaiming authorship or denying the content of the first text. After all, it does express her genuine feelings about Robert. She'd not have chosen to say it in quite

[22] Kort (2005).
[23] Roupenian (2017).

that way, bluntly and heartlessly, but it is an accurate representation of her thinking. Unable to bring herself to deny the text, she simply allows Robert to believe this is how she ended things.

This case is fictional, but surely many real people have sent imposter texts on behalf of friends (or enemies). And the rise of deepfakes will make these situations both more frequent and more compelling. A text is one thing; a voicemail or video message is much more affecting. As deepfake technology becomes powerful enough to operate in real-time, it may be possible to fake a live videocall.[24] The more lifelike and compelling the fabrication, the more pressure there will be for the victim to say *something* about it. And when one is illegitimately forced to say something about a topic one would rather not address at all, one has been harmed as a speaker.

These last examples bring to the front an important objection: is there really anything new about deepfakes?[25] Can't the same sorts of harms be caused by already existing technology, such as imposter texts, edited videos, or even forged letters? Is there any cause for *particular* ethical concern about deepfakes?

In a strictly logical sense, the answer seems to be no. The *types* of harm we've considered so far are clearly possible without deepfakes. Rumors alone can damage a person's reputation, producing material harms. Illocutionary harm is possible through low tech means.

But in a more practical sense, deepfakes are a distinct ethical problem. They make the possibility of these harms much easier to bring about, and therefore a much more realistic threat to ordinary lives. With deepfakes, one needn't be a skilled forger, master phone thief, or expert in the dark arts of political rumormongering in order to successfully compel undesired speech. All it takes is a decently powerful computer and reliable wifi connection.

Our worry is that this moral problem, while not theoretically unprecedented, will be practically unfamiliar. Ordinary people – rivalrous coworkers, jilted lovers, bored teenagers – will suddenly have the ability to generate compellingly fabricated evidence of anyone doing or saying anything. The most spectacular consequences will involve public figures, but the most morally troubling may happen on the intimate scale of ordinary enmity. How will our day-to-day relationships, the bonds of routine civility, fare when subversion of recorded reality is a realistic temptation? We have no idea, and we think that is a serious problem.

[24] See Matthias Niessner et al, 'Face2Face: Real-time Face Capture of Reenactment of RGB Videos'. March 17, 2016. YouTube video available at https://www.youtube.com/watch?v=ohmajJTcpNk . See also the project page at https://web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/page.html

[25] We consider related objections in more detail elsewhere; see Rini (2020).

Panoptic gaslighting and existential trauma

There is at least one more way in which deepfakes may generate harms, one which may go beyond facilitating a newly efficient way to do ancient harms. We turn finally to the potential for deepfakes to threaten memory and the existential bases of personhood.

In most deepfake scenarios, there are at least three different participants: a creator who generates fabricated recordings, a target who is falsely represented in the fake, and an audience whose response to the fake causes difficulty for the target. But deepfakes can be troubling even when the target and the audience are the *same person* – that is, when someone views a deepfake falsely depicting *their own* past actions. A fabricated recording could be used to destabilize or even overwrite first-personal, autobiographical memories.

To see the point, imagine you are in one of those complicated triangular friendships where everyone is a bit of a rival for everyone else's time and attention. (Maybe you are a high school student, or just someone whose life continues to feature a lot of drama.) Imagine that one of your friends claims to have heard you say terrible things about your other friend. You certainly don't remember doing that, and you are pretty sure you'd never say such a thing out loud. But now your rival pulls out their phone and plays a video: there you are, at your group's favorite pub, looking and sounding just a bit tipsy. And there you go, saying those terrible things. The video is dated from a year or so ago. "I just found it last night," says your so-called friend, "while I was going through old pictures. Don't worry though. I mean, of *course* I'd never show this to you-know-who…"

You'd be confused, of course. You'd suspect *some* sort of trick. But suppose you'd never heard of deepfakes. Suppose you thought this technology was only possible for wealthy Hollywood studios, not something your petty friend could do on their smartphone. Then what? It's hard to resist video evidence: there you are, you said it. Which should you trust more: your fallible memory, or independent recordings?

Psychologists have shown that fake photographic and video evidence can be used to manipulate autobiographical memories. In one study, participants chose their favorite brands of various consumer products and were photographed with their selections. Later, they were shown fake photos (edited by the experimenters) depicting them with different brands, and many were willing to unselfconsciously claim that *these* really were their favorites. In other words, participants were more willing to spontaneously re-assign their preferences than to challenge fake photographic evidence about their own choices.[26]

Consumer brand preferences are perhaps not that big a deal. But the same techniques can be used to trick people into accepting they may have *done* things they did not. In one study, participants shown faked photos of themselves with broken pencils or unsealed envelopes were more likely to later falsely remember having made those things happen.[27] Worst of all: participants who were shown faked video of themselves cheating in a gambling game were willing to sign false confessions, with many confabulating plausible stories to explain to themselves why they might have cheated.[28]

So, if you were shown a video like the one in our story of friend group rivalry, you very well might believe it. You might fall for it *even if* you do suspect some form of trickery. After all, maybe it did happen. Maybe you did forget. Perhaps the video shows you saying things that, yes, you do sometimes think about your friend – though you try never to say them aloud! Maybe you got a bit drunk, vino did its verifying, and you had forgotten by the next morning. It was more than a year ago, after all. Are you sure you know exactly what you said in every pub conversation of years past?

So even if you know that the recording *might* be fake, you can't be sure. And that is the core of the existential danger of deepfakes: they could be used to create effective skepticism about one's own first-personal memories.

This possibility has several serious consequences which roughly parallel the harms we have already discussed. Most obvious are material harms caused by being tricked. A fake video showing you making a disadvantageous promise or bet might induce you to give up something you shouldn't. Highly honest people would be the most vulnerable to this sort of abuse; even if they know the video might be fake, they may err on the side of honoring even unremembered, uncertain obligations.

A more ominous possibility concerns *gaslighting*, which Kate Abramson defines as "a form of emotional manipulation in which the gaslighter tries (consciously or not) to induce in someone the sense that her reactions, perceptions, memories and/or beliefs are not just mistaken, but utterly without grounds—paradigmatically, so unfounded as to qualify as crazy"[29]. Gaslighting typically involves *telling or implying to* people that things are other than as they perceive or remember. Doing so regularly and persistently can wear down their resistance.

[26] Hellenthal, Howe and Knott (2016).
[27] Henkel (2011).
[28] Nash and Wade (2009).
[29] Abramson (2014), 2.

Motives for gaslighting can be complicated. In the 1944 film that gave us the term, Charles Boyer's character torments Ingrid Bergman's character to get access to her wealth. Driving her mad is only a means to this end. But in the real world, casual gaslighting can be motivated by social jockeying, domestic abuse, retribution, or even internet trolling. It may not even be deliberate; some manipulators are so skilled that they can gaslight without even realizing what they are doing.

The creator of a deepfake may not set out to gaslight their target. If the goal is to trick a third-party audience in order to cause reputational or illocutionary harm, then deceiving the target of the recording may be a mere side-effect. Intentionally or otherwise, in at least some cases deepfakers are likely to make their targets begin to question their own memories.

And it is scarily easily to imagine the extreme case, which we will call *panoptic gaslighting*, where a vicious person sets out to *deliberately* ruin another's grip on reality through systemic use of deepfakes. The abundance of casually recorded and shared videos on social media makes this a very real possibility. Imagine a concerted campaign of just-slightly-changed videos on Facebook, showing a person doing things at last night's party or last week's dinner that are just a bit different than what the victim remembers. Each individual change is fairly unobtrusive by itself, *except* for its just barely noticeable (to the victim) discordance with memory. Done shrewdly and consistently, this sort of abuse could lead the victim to begin to doubt not just particular memories, but the reliability of their memory altogether.

Panoptic gaslighting would be existentially harmful. On some prominent theories of personal identity, what makes someone the same person over time is their ability to veridically recall earlier experiences.[30] A person who begins to systematically doubt their memories loses this connection to past selves. Worse still, a person who *accepts* the contents of deepfakes and develops new false memories could experience a form of identity fracture.

This isn't just a point of abstruse metaphysics. The first-personal experience of being panoptically gaslit - of coming to doubt one's memories generally - would surely be terrible. It would mean helplessness, dislocation, disintegration. Losing faith in your own memories would gradually undo the foundations of self-respect and the ability to withstand pressure from others. As Trudy Govier puts it:

> To discriminate between apt and ill-founded challenges from others, one needs to trust one's own memory, judgment, and conscience. A person who has no resources to preserve her ideas, values, and goals against

[30] See, for instance, Sydney Shoemaker's part of Shoemaker and Swinburne (1984). For further discussion of neo-Lockean 'psychological relation' theories of personal identity, see Parfit (1984), Johnston (1987), Baker (2000).

> criticism and attack from others will be too malleable to preserve her sense that she is a person in her own right, and will therefore be unable to maintain her self-respect.[31]

This sort of vulnerability to manipulation isn't incidental, in the way you can be tricked by scam email. Rather, the tension between our own internal self-concept (chiefly through memory) and the ways others perceive us is essential to how we function as social agents. Bernard Williams makes this point while arguing that, in an important sense, individual people may not even *have* determinant beliefs or desires before engaging with others. On his view, our need to make ourselves trustworthy and responsible to others is crucial to bringing our multifarious mental states into coherent order:

> [W]e must leave behind the assumption that we first and immediately have a transparent self-understanding, and then go on either to give other people a sincere revelation of our belief… or else dissimulate in a way that will mislead them. At a more basic level, we are all together in the social activity of mutually stabilizing our declarations and moods and impulses into becoming such things as beliefs and relatively steady attitudes.[32]

Williams' point here is *not* a postmodern denial of objective truth. Rather, he is highlighting the fact that our social attitudes toward the truth – our reliance on and expectations about one another's sincerity and competence – play an essential role in determining not only what we do, but also what we end up believing, and in an important sense who we *are*. As Williams suggests, this mutual dependence ("we are all together") is a shared predicament, one that demands solidarity and cooperation from people of good faith.

Deepfaked attacks on personal memory are a potent weapon for malefactors who seek to exploit that mutual dependence. A person who has been panoptically gaslit by systemic manipulated video depictions of their past is no longer in a symmetrically dependent relationship with their tormenter. They are instead placed at another's mercy, with not only the contents of their beliefs but also their basic capacity to stabilize their mind upon any determinate belief state, held hostage to the dubitable good will of a deceiver. Deepfakes are more than just dishonest; they hold the potential to truly destroy individuals.

Conclusion

Deepfakes may have valuable commercial and artistic applications. They might permit new sorts of harmless fun. Related technology has already been used to protect the identities of victims testifying

[31] Govier (1993), 111. For related points, see Jones (2012).

[32] Williams (2002), 193. For an important development of this line of thought, see Fricker (2007), 52-55.

about atrocities.[33] But they also might lead to new harms, and not just the obvious practical consequences of epistemic malfeasance.

We have surveyed three categories of distinctive harm in this paper: frankenporn objectification, illocutionary harm, and existentially traumatic panoptic gaslighting. We are sure that more devious minds than ours are already at work on others.

This may seem like a grab-bag of distinct ethical risks, only loosely clustered around the technological vector of deepfakes. But we believe there is a more fundamental commonality to the worries we have raised. It is not an accident that all involve the use of epistemic malfeasance to achieve illicit social manipulation. Whether in objectifying frankenporn, cruel illocutionary harm, or identity-sapping turbo-gaslighting, these uses of deepfakes show the extent that human ethical life is dependent on our epistemic relations.

In recent decades, ethicists and epistemologists have recognized a growing overlap between their concerns and even their methods.[34] We think this a valuable and timely development, as current events make increasingly apparent the social implications of epistemic contention (such as doubt in scientific experts, fake news, deep disagreement, ‘post-truth’). We believe deepfakes are yet another facet of this worrisome convergence, and we hope thoughtful minds turn to forestalling their deep harms.[35]

Works Cited

Kate Abramson (2014). “Turning up the lights on gaslighting”. *Philosophical Perspectives* 28(1): 1-30.

Drew Armstrong (2019). “Planned Parenthood Cut Off From Federal Funding Under Trump Rule”. *Bloomberg* February 22, 2019. https://www.bnnbloomberg.ca/planned-parenthood-cut-off-from-federal-funding-under-trump-rule-1.1218733

Lynne Rudder Baker (2000). *Persons and Bodies: A Constitution View*. Cambridge University Press.

Rima Basu and Mark Schroeder (2019). “Doxastic Wronging”. In *Pragmatic Encroachment in Epistemology* (eds. Brian Kim and Matthew McGrath). Routledge. 181-205.

[33] The documentary film ‘Welcome to Chechnya’ uses a deepfake-like technique to project the faces of volunteer actors over those of real victims of homophobic persecution. As the filmmakers explain, this allowed them to preserve the emotional intensity of their informants’ testimony without placing them in greater danger. See Thomson (2020).

[34] See, for example, Cuneo (2007), Marušić (2015), Basu and Schroeder (2019), Srinivasan (2020).

[35] We owe thanks to the editors and referees for the journal, as well as to audiences at the University of Toronto Centre for Ethics and the Department of Philosophy at the University of Waterloo. This research was supported by York University’s Vision: Science to Applications (VISTA) project.

Nancy Bauer (2015). *How to Do Things With Pornography*. Cambridge: Harvard University Press.

Joseph A. Califano (1991). *The Triumph and Tragedy of Lyndon Johnson: The White House Years*. New York: Simon and Schuster.

Amanda Cawston (2019). “The feminist case against pornography: A review and re-evaluation”. *Inquiry* 62(6): 624-658.

Robert Chesney and Danielle Keats Citron (2019). “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security”. *California Law Review* 107: 175.

Samantha Cole (2017). “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

Samantha Cole (2018). “Deepfakes Were Created As a Way to Own Women’s Bodies – We Can’t Forget That”. *Vice* June 18 2018. https://www.vice.com/en_us/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2

Kate Cox (2019). “Deepfake revenge porn distribution now a crime in Virginia”. *Ars Technica* July 1 2019. https://arstechnica.com/tech-policy/2019/07/deepfake-revenge-porn-distribution-now-a-crime-in-virginia/

Terence Cuneo (2007). *The Normative Web: An Argument for Moral Realism*. Oxford University Press.

Jeremy Diamond (2015). “Trump: I would shut down government over Planned Parenthood”. *CNN* August 4, 2015. https://www.cnn.com/2015/08/04/politics/donald-trump-government-shutdown-planned-parenthood/index.html

Andrea Dworkin (1985). “Against the male flood: Censorship, pornography, and equality”. *Harvard Women’s Law Journal* 8(1): 1-29.

Hany Farid (2018). “Digital Forensics in a Post-Truth Age”. *Forensic Science International* 289: 268−269.

Miranda Fricker (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Trudy Govier (1993). “Self-Trust, Autonomy, and Self-Esteem”. *Hypatia* 8(1): 99-120.

Douglas Harris (2019). “Deepfakes: False Pornography is Here and the Law Cannot Protect You”. *Duke Law and Technology Review* 17: 99-127.

Maria V. Hellenthal et al. (2016). “It Must Be My Favourite Brand: Using Retroactive Brand Replacements in Doctored Photographs to Influence Brand Preferences”. *Applied Cognitive Psychology* 30(6): 863-870.

Linda A. Henkel (2011). “Photograph-induced memory errors: When photographs make people claim they have done things they have not”. *Applied Cognitive Psychology* 25(1): 78-86.

Mark Johnston (1987). “Human Beings”. *Journal of Philosophy* 84(2): 59-83.

Karen Jones (2012). “The Politics of Intellectual Self-Trust”. *Social Epistemology* 26(2): 237-251.

Leo Kelion (2018). “Reddit bans deepfake porns videos”. *BBC News* 7 February 2018. https://www.bbc.com/news/technology-42984127

Sarah Kliff (2015). “I watched 12 hours of the Planned Parenthood sting videos. Here’s what I learned.” *Vox* September 9 2015. https://www.vox.com/2015/8/13/9140849/planned-parenthood-videos-unedited

Michele Kort (2005). “Portia heart & soul”. *The Advocate* August 29 2005. https://www.advocate.com/politics/commentary/2005/08/29/portia-heart-amp-soul

Rae Langton and Jennifer Hornsby (1998). “Free speech and illocution”. *Legal Theory* 4(1): 21-37.

Yuezun Li and Siwei Lyu (2019). “Exposing DeepFake Videos By Detecting Face Warping Artifacts”. *arXiv* preprint https://arxiv.org/pdf/1811.00656v3.pdf

Lingjie Liu et al. (2019). “Neural Rendering and Reenactment of Human Actor Videos”. *ACM Transactions on Graphics* 38(5). https://doi.org/10.1145/3333002

Catharine MacKinnon (1987). *Feminism Unmodified: Discourses on Life and Law*. Cambridge: Harvard University Press.

Ishani Maitra (2009). "Silencing Speech". *Canadian Journal of Philosophy* 39(2): 309-338.

Andrei Marmor (2015). "What Is the Right to Privacy?" *Philosophy and Publics Affairs* 43(1): 3-26.

Berislav Marušić (2015). *Evidence and Agency: Norms of Belief for Promising and Resolving*. Oxford University Press.

Rachel Ann McKinney (2016). "Extracted Speech". *Social Theory and Practice* 42(2): 258-284.

Thomas Nagel (1998). "Concealment and Exposure". *Philosophy and Public Affairs* 27(1): 3-30.

Robert A. Nash and Kimberley A. Wade (2009). "Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence". *Applied Cognitive Psychology* 23(5): 624-637.

Martha C. Nussbaum (1995). "Objectification". *Philosophy and Public Affairs* 24(4): 249-291.

Derek Parfit (1984). *Reasons and Persons*. Oxford University Press.

Regina Rini (2020). "Deepfakes and the Epistemic Backstop". *Philosophers' Imprint* 20(24): 1-16.

Kristen Roupenian (2017). "Cat Person". *The New Yorker* December 4 2017. https://www.newyorker.com/magazine/2017/12/11/cat-person

Henry Ian Schiller (2021). "Illocutionary harm". *Philosophical Studies* 178: 1631-1646.

John Searle (1992). *The Rediscovery of the Mind*. Cambridge MA: MIT Press.

Sam Shead (2020). "Facebook to Ban "Deepfakes"". *BBC News* January 7, 2020. https://www.bbc.com/news/technology-51018758.

Sydney Shoemaker and Richard Swinburne (1984). *Personal Identity: Great Debates in Philosophy*. Blackwell.

Tom Simonite (2019). "Most Deepfakes Are Porn, and They're Multiplying Fast". *Wired* October 7, 2019. https://www.wired.com/ story/most-deepfakes-porn-multiplying-fast/

Wahyudi Soeriaatmadja (2017). "Man who uploaded controversial video of ex-Jakarta governor Ahok sentenced to jail". *The Straits Times* November 14, 2017. https://www.straitstimes.com/asia/se-asia/man-who-uploaded-controversial-ahok-video-sentenced-to-jail

Amia Srinivasan (2020). "Radical Externalism". *Philosophical Review* 129(3): 395-431.

Nadine Strossen (1995). *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*. New York: Scribner.

Patricia Thomson (2020). "Digital Disguise: 'Welcome to Chechnya''s Face Veil is a Game Changer in Identity Protection". *Documentary Magazine* June 30 2020. https://www.documentary.org/column/digital-disguise-welcome-chechnyas-face-veil-game-changer-identity-protection

J. David Velleman (2001). "The Genesis of Shame". *Philosophy and Public Affairs* 30(1): 27-52.

Hans von der Burchard (2018). "Belgian Socialist Party Circulates "Deep Fake" Donald Trump Video". *Politico* May 21, 2018. https:// www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

Bernard Williams (2002). *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press.