Synthese (2023) 202:40
https://doi.org/10.1007/s11229-023-04266-4

ORIGINAL RESEARCH

Freedom of expression meets deepfakes

Alex Barber[1]

Received: 31 July 2022 / Accepted: 6 July 2023

Abstract
Would suppressing deepfakes violate freedom of expression norms? The question is pressing because the deepfake phenomenon in its more poisonous manifestations appears to call for a response, and automated targeting of some kind looks to be the most practically viable. Two simple answers are rejected: that deepfakes do not deserve protection under freedom of expression legislation because they are fake by definition; and that deepfakes can be targeted if but only if they are misleadingly presented as authentic. To make progress, following a discussion of why freedom of expression deserves protection in a democracy, the question is reframed. At issue is not whether the arrival of deepfakes brings new and potentially serious dangers (it does), nor whether these dangers call for measures that potentially limit freedom of expression (they do), but whether the need for such measures raises any new and unfamiliar freedom-of-expression challenges. The answer to *that* question, surprisingly, is no. The balancing act needed to cope with the arrival of deepfakes brings plenty of difficulties, certainly, but none of the measures likely to be effective in tackling deepfake harms raises freedom-of-expression concerns that aren't familiar from consideration of non-deepfake harms. In that respect, at least, the arrival of deepfakes makes no difference.

Keywords Deepfakes · Record keeping · Democracy · Freedom of expression · Free speech · Epistemic backstop

1 Introduction

Though often put to benign ends, deepfakes and related synthetic media are also widely seen as posing a potentially serious threat to democracy. All democracies need reliable public information; audio and video recordings have become an increasingly important source of such information; but this source is undermined once one can no longer distinguish authentic instances from fakes. The worry is thus that, as fake and

✉ Alex Barber
alex.barber@open.ac.uk

[1] Faculty of Arts and Social Sciences, The Open University, Milton Keynes MK7 6AA, UK

Published online: 20 July 2023

misleading audio and video images rapidly become realistic, easy to produce, immune to technological detection, and widely disseminated, even authentic audio and video images will become untrustworthy—a variant of the so-called 'liar's dividend' that gives a dangerous new weapon to those hoping to undermine democratic norms. Nina Schick, for example, a political technology advisor, calls deepfakes the 'latest evolving threat' in 'our increasingly dangerous and untrustworthy information ecosystem' and a threat to democratic society.[1] Several philosophers, meanwhile, have written of the dangers deepfakes pose to the epistemic practice of record-keeping, and so by implication to a service vital to modern democracies.[2]

In this article I consider an important but unexamined wrinkle in this debate.[3] If we accept—as I think we should—that the deepfake phenomenon poses a potential threat to the health of democracies, and can harm in other ways too, we might sympathize with measures to target them. Yet deepfakes also have representational properties. That makes them potential candidates for protection under freedom-of-expression principles, principles that *themselves* lie at the heart of any functioning democracy. We therefore face a conundrum: within a democracy, what protections do deepfakes deserve in virtue of their being products of expressive acts?

Here I entertain three possible answers, plumping for the third. The first I call the 'kneejerk' answer (Sect. 3 below). It holds that freedom-of-expression considerations have no purchase: deepfakes, being definitionally fake, are an epistemic pollutant, liable for targeting by whatever measures are needed to deal with the harms they can produce. Property rights of various kinds could be grounds for protection in some cases, perhaps, but freedom of expression is an irrelevance. As my label for it implies, this first answer is beset with difficulties and is at best a stepping-stone towards the more nuanced 'dual-policy' answer considered in Sect. 4. This recommends different treatments for different cases: overt deepfakes, whose inauthenticity is meant to be discernible by their consumers (e.g. most satirical deepfakes), should be protected to the same degree as any other form of expression; covert deepfakes, meanwhile, whose inauthenticity is hidden, are generally liable for targeting through, say, censorship, fines, or other restrictions. Though more credible than the kneejerk answer, this too is found wanting, so I end by recommending what I call a 'makes-no-difference' answer (Sect. 6).

This third answer follows a reframing of the question. Our conundrum is as much about freedom of expression in a democracy as it is about deepfakes, so in Sect. 5 I explore what makes freedom of expression desirable in the first place and suggest we think of the threat posed by deepfakes as a challenge to existing freedom-of-speech infrastructures, though hardly the first such challenge. Our question is really whether the task of recalibrating these infrastructures in response to the arrival of deepfakes poses any challenges *distinctive of* deepfakes. In Sect. 6, I argue that measures most

[1] Schick (2020: p. 2, 10). See also Ovadya (2019) and Chesney and Citron (2019: p. 1786).

[2] Regina Rina gives an especially compelling account of the threat posed to what she calls the 'epistemic backstop' function of recordings (2020: pp. 4–5). See also Floridi (2018: p. 320) and Fallis (2021: p. 638).

[3] Unexamined in philosophical discussions of deepfakes, at least. The discussion by Chesney and Citron (2019: pp. 1789–1792) is situated within the US legal context, with its distinctive First Amendment tradition, but several of their observations are made use of below.

likely to be of use in containing deepfake harms are compatible with freedom-of-expression principles, or at least no less compatible than they are when deployed against non-deepfake harms. The very real anti-democratic dangers in this area—-such as the co-opting of deepfake-panic to justify over-suppression, or, conversely, insincere appeals to freedom of expression to justify disruption of the information ecosystem—are already familiar in other domains.

Before all of this, in the next section I consider what deepfakes are and how best to understand their key properties in the present context.

2 Deepfakes: what they are, their variability, how they represent, and the dangers they pose

While deepfakes emerged only recently in philosophical years, in technological years they are old news. The word itself derives from a 2017 reddit user who went by the name 'Deepfake' and adapted pre-existing technology to integrate images of female celebrities into pornographic videos. The label caught on and is now applied more broadly, though the overwhelming majority of deepfakes still contain coercive pornography.[4] It has no standardized definition, unsurprisingly so given these origins and the pace of change, but I here characterize deepfakes as audio or video recordings that have been digitally altered using machine learning algorithms to create misrepresentations of events that never took place.[5]

Deepfakes already come in many varieties. They can be video or audio. They can be made to appear as a stored or found record or as a live feed, and in the latter guise can be used in scam calls. Like most tools, they can be used benignly or malignly. Benign actual or potential uses include: a "live" concert by rejuvenated band members appearing on stage as they might have done in their prime; a globally recognized footballer being deepfaked into speaking diverse languages as part of a malaria vaccine charity campaign; use of the same underlying technology to tell whether historical artworks are themselves faked; a posthumous deepfake video of a school-shooting victim arguing for gun-control; and grieving relatives finding solace by interacting with deepfakes of the deceased.[6] One further dimension of variety, mentioned already, will be especially important in what follows. In covert deepfakes, the inauthenticity is not intended to be evident to the audience (though a poor-quality covert deepfake may not deceive many); in overt deepfakes, the inauthenticity is either explicitly signalled

[4] Ajder et al., 2019. The persistence of the deepfake label is not ideal: though widely interpreted as a simple portmanteau of 'deep-learning' and 'fake', the name was presumably also a nod to *Deep Throat*, a commercially successful 1972 pornographic film made using coercion and rape (MacKinnon, 2006: pp. 18–19).

[5] Adapted from Rini and Cohen (2022) but broadened to include reality-warping synthetic media more generally and not just face- and body-swapping software. I am sidestepping the question of how or whether to draw a boundary that excludes beautifying filters, etc.

[6] For bands see Cairns (2022); for art see Floridi (2018: pp. 318–319); for malaria charity see 'How we made David Beckham speak 9 languages', 9 Apr 2019, https://www.youtube.com/watch?v=CF_e0kMCW2o; for school shooting see 'Chilling Deep Fake Gun Control Ad', 6 Oct 2020, https://www.youtube.com/watch?v=5HwWkeanfLA. For more on benign uses see Silbey and Hartzog (2019), Chesney and Citron (2019), and Kerner and Risse (2021).

Table 1 Tiers of representation relevant to both deepfakes and authentic perceptual records

Tier 3	Using something's appearance of being a perceptual record to create meaning
Tier 2	Appearing to be a perceptual record
Tier 1	Being a perceptual record

or else is meant to be clear from the context. Therapeutic, satirical, and other overt deepfakes are like fictional literature in that they do not present themselves as a true record of events.

Since we are considering deepfakes in the context of freedom of expression, it will be useful to appreciate the nature of their representational properties. Their core feature is that they *appear* to be a perceptual record but *are not in fact* so. In this respect they are like fake photographs, which appear to be a visual record of a scene at a particular moment but are not. How a photograph (or any other genuine product of a recording device) represents is a controversial matter in the philosophy of art, for reasons that need not detain us here. For our purposes it is enough to say that a photograph (or any other… etc.) is an automated record of a perceptual environment that is readable as such by a typical human perceiver.[7] A deepfake, then, appears to be, but is not, an automated record of a perceptual environment that is readable as such by a typical human perceiver.

Malign agents can exploit this gap between appearance and reality, subject to the appearance being convincing. So too can benign agents, to create imagined but unreal scenarios (for art or satire, say). Even when the false appearance is highly realistic, a deepfake can be presented as a *merely* apparent perceptual record (within a documentary on deepfakes, for example). The difference between covert and overt cases is thus about how the deepfake is packaged and presented, not about degree of realism.

We can impose some order on this complexity by distinguishing between three interacting tiers of representation when thinking about either faked or authentic perceptual records (Table 1). At ground level (Tier 1) we have the kind of representation authentic videos, photographs, etc. have, and which deepfakes merely appear to have: being a perceptual record, the notion glossed above. Above this (Tier 2) we have the fact of their appearing to be a perceptual record. This appearance is possessed by both deepfakes and authentic videos, etc., which differ merely in whether the appearance is veridical. Finally, and again like authentic videos, etc., deepfakes can acquire a further layer of contextual meaning by being used in a particular way (Tier 3). I might use a make-yourself-look-old app, for example, *either* to misrepresent myself as older than I am *or* to predict how I will look in a few decades, accurately or not as it may be. The difference here arises from how the image is used and interpreted rather than from properties intrinsic to the image itself.

[7] Being a 'record' implies it is veridical, i.e. is not introducing significant distortions such as would be introduced by, say, a piece of camera-internal software that made subjects appear to smile. Being 'automated' is what distinguishes it from, say, a realistic sketch. Being 'readable as such by a typical human perceiver' means the record cannot simply be lots of stored 1s and 0s. For more thorough treatments of photographic representation, see Costello and Phillips (2009), Cavedon-Taylor (2013), and Kulvicki (2020).

Why are deepfakes such a concern? We can distinguish two categories of potential harms, call them *individual* and *environmental*. Individual harms are those done directly to specific persons or organizations thanks to the specific content of the relevant deepfake. In covert cases the danger is clear, as when a duplicated voice is used for telephone fraud, to ruin an organization's reputation, or simply to cause someone to have a false belief. Individual harm can also be caused by overt cases, through emotional harassment for example. Deepfaked pornography, however realistic, does not have to be taken as authentic for it to harm the dignity of the depicted individuals. Environmental harm, meanwhile, is the wider damage done by deepfakes, not necessarily to those directly incorporated into specific images. Deepfaked pornography, overt or covert, can be ideologically harmful in what it says about women, for example, and may even be said to subordinate them. This harm is distinct from any harm done to individuals whose images are blended without consent. Sheer ease of production makes deepfakes a likely weapon for use against other oppressed groups (including via the targeting of individuals who are treated as representative of those groups simply because they enter public life).[8]

One of my chief concerns in this article is with environmental harm of a particular sort: the damage done by deepfakes to the epistemic environment—and hence democracy—through their undermining of the trustworthiness of authentic perceptual records. Most commentators are quick to observe that environmental epistemic harms are more serious in the longer term than individual epistemic harms, since the scepticism that will naturally emerge as deepfake deceptions become known about will protect against the latter but feed the former.[9] This *type* of harm is not new. Faked videos have been around for a while, and faked photos are almost as old as authentic ones. Skilled human-voice impersonation means that even deepfaked audio is somewhat familiar in the type of challenge it poses. Deepfakery, however, has five traits that massively amplify the problem: ease of production, ease of dissemination, realism, immunity to technological detection, and the sheer pace of change. In truth it is not yet easy to produce realistic, technically undetectable deepfakes, but it would be naïve to assume this will still be true in a decade or so.

The epistemic threat posed by deepfakes is in this respect like that posed by fake news more generally: traditional propaganda methods have fused with emerging technologies such as social media and machine learning to energise a threat already posed to democratic norms. Blatant lies, for example, can now be disseminated so easily that conspiracy theories have a newly acquired significance as a political tool (Cassam, 2019). The challenge posed by deepfakes is nonetheless distinctive. Authentic perceptual records, if they can be known to be such, have an immediacy to them that avoids the need to trust some intermediary's account of events. Rini (2020) has

[8] See Rini and Cohen (2022) for a rich account of the harms that can be done to those who have been deepfaked. For a review of the pre-deepfake debate on pornography's harms and free speech, see Mikkola (2019).

[9] Floridi (2018: 320); Ovadya (2019, 'Myth 1'); Vaccari and Chadwick (2020: p. 2); Schick (2020: pp. 130–134); Rini (2020: p. 8); Fallis (2021: p. 625); Kerner and Risse (2021: p. 98). Some sense of the danger can be had from the 2017 coup launched against the Gabon president, triggered by suspicion that a genuine video meant to prove he was alive was fake (Ajder, Patrini et al., 2019: p. 10).

called this their 'epistemic backstop' function. Televised debates and leaked recordings of political candidates making compromising statements have become a staple of elections the world over. This potency depends, of course, on its being possible for ordinary citizens to trust in the authenticity of perceptual records, which is exactly what an explosion of deepfakery threatens to make far harder.

What to do? The fact that deepfakes can be used for good or bad already suggests that a policy towards them of uncompromising intolerance is unwise. In addition, deepfakes are at least sometimes expressive in a way that, on the face of it, earns them protection on freedom of expression grounds (their use in satire being a clear example). Let us turn, then, to our main question. Should freedom of expression protections in a democracy be extended to deepfakes and, if so, then when and how?

3 The kneejerk answer

A tempting first answer is: no, they should not. Two considerations speak in favour of this thought, one theoretical and the other practical. The theoretical consideration is that deepfakes are, definitionally, fake. Unlike genuine videos, deepfaked videos merely *appear* to be perceptual records. Their suppression would not therefore warrant the suppression of authentic videos. Nor do deepfakes have those redeeming qualities of falsehoods championed by Mill, such as containing elements of the truth or revivifying our attachment to the truth (Fallis, 2021: p. 639). In short, deepfakes seem to be nothing but an epistemic pollutant, liable for targeting. The practical consideration is that clearing up this pollutant is unlikely to be effective without the use of automated detect-and-target technology, unmediated by a human able to make contextually nuanced distinctions. Having to consider first whether freedom of expression norms have been violated on a particular occasion would make this, to say the least, more difficult.

As my pejorative label for it implies, this first answer is at best a first step towards something better. Whether its apparent practical advantage will persist is unclear at present. Technology alone, as opposed to the slow laborious process of fact-checking, reverse-image searching, etc., may well not be up to the task for much longer (Chesney & Citron, 2019: pp. 1787–1788, though see Lang, 2019). If and when it is not, the theoretical advantage too will disappear. True, *known* deepfakes could be targeted as nothing but pollutants, but we may soon be unable to tell that an image is a deepfake save by inference from what it purports to be a record of, an inference that may well depend on political opinion. A compromising image as of a political leader accepting a bribe, for example, would be deemed genuine or not according to background beliefs about that politician's integrity. Targeting the image could then easily be portrayed as, and could indeed be, politically motivated censorship.

But the strongest reason for rejecting this first answer is that it treats all deepfakes as malign when that is demonstrably not the case. Freedom of expression measures seem important to the protection of at least some benign instances, even if we think only narrowly of their use for the expression of opinion (never mind therapeutic or artistic uses). MIT's *In the Event of Moon Disaster*, which reconstructs President Nixon reading out the news that the 1969 moon landing has failed, and a deepfake of President Obama giving a speech about deepfakes, were both created to warn of the potential

dangers of deepfakes.[10] Deepfakes are also increasingly likely to figure in satire. The comedian Bruno Sartori already uses them to target anti-democratic political figures in Brazil. He is keenly aware that his political enemies would like to gloss over the distinction between acceptable and unacceptable uses of deepfake technology, allowing them to justify shutting down his satirical operation by tarring both it and unquestionably malign uses with the same brush.[11] Undifferentiated suppression in more established democracies would create the same dangers. Cristian Vaccari and Andrew Chadwick are right to warn that deepfakes and the like will 'create new opportunities to campaign on promises to restore "order" and "certainty" through illiberal policies curtailing free speech and other civil rights' (2020: p. 9; see also Chesney and Citron, 2019: p. 1786).

4 The dual-policy answer

These objections to the kneejerk answer might incline us towards invoking the distinction between covert and overt cases and taking different approaches towards each. On such a dual policy, overt cases would mostly deserve protection while covert cases would not. *In the Event of Moon Disaster*, for example, satire, and deepfakes used in the context of grieving, would receive freedom of expression protections when there is no attempt to hide the fakery. Covert cases meanwhile would be ripe for targeting. Those currently using deepfakes benignly but covertly, such as in advertising or charitable promotions like the Malaria vaccine promotion case, could be given a choice: make the fakery explicit or expect to be targeted. The situation would be analogous to the requirement in French advertising legislation that the phrase '*photographie retouchée*' be placed on digitally reshaped images of models' bodies. Overt deepfakes are potentially usable to malign ends, of course. Racist stereotyping made using deepfake technology is one obvious example. But a supporter of the dual-policy approach could insist that such cases be dealt with on the same terms as sketches, so long as they are marked as overt. The fact of their being deepfakes would thus become irrelevant.

Something like this dual policy has been proposed or adopted in various legislative and philosophical contexts. Don Fallis, for example, gestures at support for the view in passing when he writes that 'it is not necessary to ban deepfakes per se. It is only necessary to ban deepfakes of events that did not actually occur *that viewers are unable to distinguish* from genuine videos'.[12] Fallis does not explicitly argue for a dual-policy response to freedom-of-expression concerns since that is not the topic he is addressing, but we can try to fill the gap. One appealing feature of the approach is the apparent simplicity and neutrality of its application. It is not *as* simple or neutral to

[10] 'In the Event of Moon Disaster' (https://moondisaster.org/); 'Fake Obama created using AI video tool—BBC News' (https://www.youtube.com/watch?v=AmUC4m6w1wo).

[11] WITNESS video, 'Bruno Sartori—Deepfakes as satire and parody in Brazil', https://www.youtube.com/watch?v=vhbZtXkhQC0, 24'.

[12] Fallis (2021: p. 639), italics in original. China is pushing ahead with a law requiring labelling (Hine & Floridi, 2022); see also the proposed DEEP FAKES Accountability Act, progressing through the US Congress (https://www.congress.gov/bill/116th-congress/house-bill/3230), which requires labelling or watermarks.

apply as the knee-jerk policy, but determining whether a deepfake is overt or covert is more straightforward than, say, determining whether it is benign. We can also envisage tentative arguments for each of the two sub-policies out of which it is composed. All overt cases express something at Tier 3 (Sect. 2, Table 1), so are at least candidates for protection on freedom of expression grounds. Covert cases, meanwhile, do not express anything, and so do not qualify for protection.

This last claim may seem odd, but a case can be made for it using the three-tiers account of Sect. 2 plus Grice's distinction between natural and non-natural meaning.[13] Consider the difference between wet flagstones and an utterance of the Welsh sentence *Mae hi wedi bod yn bwrw glaw*. Both can mean that it has been raining, but in distinct senses of the word 'mean'. The difference comes out if we suppose that it has not in fact been raining. Perhaps someone wetted the flagstones to fake its having been raining. In that case, we would stop saying that the wet flagstones mean that it has been raining; they would merely appear to mean that. This truth-of-what-is-meant requirement is a mark of what Grice calls natural meaning ('meaning$_n$'). The utterance, in contrast, continues to mean what it does irrespective of whether it has been raining, a mark of non-natural meaning ('meaning$_{nn}$'). A genuine photo is more like the flagstone case whereas a sketch is more like the utterance. A genuine photo of recent rain, for example, means$_n$ it has been raining but does not mean$_{nn}$ it, whereas a sketch of the same would mean$_{nn}$ it but not mean$_n$ it (Grice, 1957: pp. 382–383). A fair extension of this view would be that a faked photo or (to bring things round to our present concern) a deepfaked video of recent rain would *appear* to mean$_n$ that it has been raining but would not in fact mean$_n$ it—just like the fake-rained-on flagstones. But nor would the deepfaked video have meaning$_{nn}$ if it is covert as opposed to overt. Overt deepfakes are somewhat like sketches, with an openness of communicative intention characteristic of meaningful$_{nn}$ acts (Barber, 2019: 155). Covert deepfakes are precisely not open in this way.[14]

The more closely one inspects this dual-policy approach, the less promising it becomes. To apply the dual policy, for example, notice that it is *not* always enough to perform the relatively easy task of telling overt from covert cases, in the hope of then being able to apply the relevant sub-policy. Rather, one must also be able to tell covert cases from authentic recordings. This is because (on the dual policy) a purported recording is liable for targeting if it is a covert deepfake, whereas it is a candidate for protection if it is an overt deepfake *or else a genuine recording*. This means that the practical worry, seen already in connection with the kneejerk answer, carries over. Absent an effective technology for discerning inauthenticity, in the background software for example, this policy will be extremely difficult to implement without the potential intrusion of political judgement.

A second worry is that the difference between overt and covert cases is unstable. Overt cases sometimes escape the lab, so to speak (Floridi, 2018: p. 320). *In the Event of Moon Disaster* has, inevitably, found a home on moon-landing conspiracy sites. A video created in Pakistan for educational purposes showed a child being kidnapped

[13] Grice (1957). Others make similar distinctions (e.g. Dretske, 1988: pp. 51–69), but Grice's is more familiar.

[14] Notice how this helps justify the suggestion above that malign overt deepfakes be treated on equivalent terms to malign sketches (e.g. racist cartoons).

(overt) but was taken out of its context and presented as real (covert), prompting rioting over several weeks that resulted in several deaths (Vaccari & Chadwick, 2020: p. 1). Running in the other direction, a covert case can very easily be turned into something with a claim to protection under freedom of expression norms simply by affixing a comment such as "I believe this is a genuine photograph", "I know they have had to say it is faked in order not to be targeted by the government or sued, but I believe it is genuine", or simply "This may or may not be genuine, what do you think?" Such comments would presumably often be sincere, since publishers and authors could simply not know whether something is a deepfake, yet they would strip of all utility the argument that covert cases do not express anything. Once affixed with a comment of this kind, deepfakes clearly *do* express something at Tier 3 in the earlier hierarchy.

5 Freedom of expression in a democracy

Having rejected two flawed answers to our question ('Should freedom of expression protections in a democracy be extended to deepfakes and, if so, then when and how?'), it is time to change tack. Our question has as much to do with freedom of expression in democracies as it does with deepfakes, yet neither of the two simple answers just rejected was prefaced by reflection on that other topic. I propose reframing our question as: What *new* freedom of expression concerns are raised by the advent of deepfakes in their various manifestations? The answer, a little surprisingly, is almost none. Showing this is for Sect. 6. The present section does the groundwork needed for reframing the question in this way. It considers what makes freedom of expression valuable in the first place.

There are, inevitably, competing views on this matter. Thus we find autonomy views (variants of the thought that freedom of expression should be prized because it is the exercise of personal self-government); truth views (which stress the epistemic benefits of the free exchange of ideas); democracy views (which see free speech as vital to some legitimizing feature of democratic systems, such as ownership of or meaningful consent to laws); and recent individual contenders such as Seana Valentine Shiffrin's thinker-based account (which derives freedom of expression protections from the more fundamental freedom to exercise and develop our rational, emotional, and perceptual capacities) or Matteo Bonotti and Jonathan Seglow's relational account (extrapolated from various ethical properties embedded in the dynamics of communication).[15] This puts us in a predicament. We might select one from among these various accounts and use it to anchor our subsequent discussion, but that would make the conclusion to be derived in the next section unnecessarily controversial, since plenty of readers would have selected otherwise.

I propose instead to sketch a perspective that is open-ended and hybrid enough to avoid being needlessly controversial, yet contentful enough that we can use it to think about deepfakes. To that end, I will consider John Stuart Mill's canonical defence of freedom of opinion in *On Liberty*, alongside its limitations. For our purposes, these

[15] Baker (2011) [autonomy]; *Abrams v. United States*, 250 U.S. 616 (1919) [truth]; Post, 2011, Weinstein, 2011 [democracy]; Shiffrin (2014: pp. 79–115); Bonotti and Seglow (2022), who also criticise the other accounts here mentioned.

limitations are as instructive as those elements of his position that even most non-Millians about freedom of expression would want to get behind. The perspective that emerges from this discussion is not a finished account of the value of freedom of expression so much as an ecumenical take on the topic (it includes aspects of the standard three accounts, for example) that gives us what we need when we return to deepfakes in Sect. 6.

Using Mill's discussion as our vehicle here is useful not only for its familiarity but for its acknowledgement of at least two distinct reasons for prizing freedom of expression. The first of these, resonating with the autonomy view, is that freedom of expression is a freedom like any other, to be cherished like any other. Call this the *general* value of freedom of expression, in view of its being common to all freedoms. For Mill, exercising freedoms, of expression, of movement, of association, of property, etc., is a core component of human wellbeing. By his harm principle (1859: p. 22), it and the other freedoms can legitimately be restricted by the state only if their exercise harms others. It is important to note that the harm principle is an 'only if' principle, directing the state to back off when there is no harm, not instructing it to intervene whenever there is. Indeed, Mill dedicates an entire chapter of *On Liberty* to showing that the expression of unpopular and potentially false opinions deserves protection, up to a point, *even if* this has the potential to cause harm to others. That point was famously exceeded by an imagined orator urging on an angry mob with a speech declaring that corn dealers starve the poor (1859: p. 101) but not by a newspaper editorial using the same anti-corn dealer rhetoric (p. 100), notwithstanding the potential for harm even in the latter case.

Mill rests his justification for this extra helping of toleration, so to speak, on freedom of expression's having a second source of value: epistemic value (resonating with the truth view).[16] Call this its *specific* value, since it is not attributable to freedoms in general. Mill offers four reasons against a policy of prohibiting the expression of opinions deemed false and potentially dangerous. None of these reasons consists in a value for the person exercising a freedom. Instead, Mill looks to the dynamics inherent in open debate, which he regards as the optimal means for eradicating error.[17] First, he says, seemingly false opinions could turn out to be true after all, since no government is infallible. Second, opinions could be genuinely mistaken yet contain important insights missed by official dogma. Third and fourth, even if they are wholly false, having to argue for a truth can have the twin benefits of promoting better understanding of it and stronger attachment to it (1859: p. 95).

This specific value gives us one way of appreciating why freedom of expression is peculiarly important in democracies (resonating with the democracy view, and of interest to us given our interest in the threat deepfakes appear to pose for democracies). No system of government is worth wanting unless the sovereign power in that system

[16] Confusingly, 'Millian' in this context often functions as a shorthand for just one or the other of these two sources, rather than as a hybrid of the two. van Mill (2021) stresses Mill's views on freedom and the harm principle, for example, whereas Bonotti and Seglow associate him with the truth view (2022: pp. 517–518). Since I am merely using Mill's approach and its shortcomings as a vehicle to create a hybrid perspective, I will not dwell on its proper interpretation.

[17] His examples are of public political speech, but he would presumably advocate the epistemic merits of open discussion in, say, scientific, workplace, or domestic contexts, and for the same reasons.

makes competent decisions. Such competence (using Mill's reasoning) requires a deliberative environment in which strengths and weaknesses of competing perspectives can be freely aired. It follows that, when the sovereign power is the populace, the populace should be able to engage in free and open debate. In democracies, then, freedom of expression has the vital epistemic function of delivering an optimally informed, competent, deliberating, rational populace.

Let us turn now to some shortcomings in Mill's view (as here interpreted). One respect in which his arguments are clearly limited is in their range. He had in mind the specific case of the suppression (unwarranted in Mill's view) of printed opinions deemed false and dangerous by the state, but freedom-of-expression questions arise in many other contexts. Very different ranges of issues arise, and different norms apply, according to whether one is considering expression within, say, a school, a religious setting, a public body, a national newspaper, a chat forum, or a science department. Mill's brief but influential discussion is largely silent on all but a relatively narrow range of cases.

In addition, Mill failed to properly take account of the impact of power asymmetries on public debate. This impact can take at least two forms. First, being relatively powerless often entails having less of a voice. What presents itself as a free market of ideas is thus often, in practice, a hollow affirmation of a status quo that effectively permits only powerful groups to advocate loudly for themselves. Second, some groups are in practice far more vulnerable than others to the harms that can arise from the expression of toxic views. Any willingness to overlook the harms that can result from a policy of tolerance will therefore often have a disproportionate impact on such groups. Although distinct, these two considerations tend to reinforce one another, since the same groups tend to be both vulnerable and deprived of a loud voice.[18]

Other more theoretical objections might be raised against Mill's view (as here presented). Some might charge it with misconstruing the significance of free discussion for democracy, for example, on the grounds that free expression is as important for ownership and consent as it is for the truth-tracking potential of open deliberation. But instead of trying to amend or replace Mill's view with something better, let us take stock. This discussion of his view was designed to yield a platform from which to think about what difference deepfake harms, and the need to tackle those harms, might make to how we think about freedom of expression. We do not need a complete and defensible view of freedom of expression to do that. We need a sense of the lay of the land, and we now have it. A plausible account of what makes freedom of expression valuable is going to have to acknowledge both the twin benefits just described (the specific epistemic value of an open deliberative environment, especially in a political system where the public is meant to be sovereign; the general value that comes with exercising a freedom). Equally, though, Mill's account is radically incomplete, is naïve about the conditions needed for public deliberation to be epistemically credible, and is possibly theoretically mistaken. This view and an awareness of its apparent shortcomings give us enough to go forward.

Thus, it is already evident that freedom of expression will require balancings of various kinds. Value to the expresser will need to be balanced against serious

[18] He was perhaps more alert to these dangers than is generally acknowledged (Gordon, 1997).

harms that might arise from particular expressive acts, for example; and deliberative forums will generally be epistemically skewed without measures to counter the effects of social inequalities, lack of a moderator or gatekeeper, or algorithms that reflect the perverse incentives facing social media organizations (van der Linden, 2023: pp. 103–132). I will presuppose that these balancings will take the form of calibrations of an ever-evolving freedom-of-expression infrastructure. By a 'freedom-of-expression infrastructure' I mean, somewhat loosely, such things as legislation, including laws governing copyright, defamation, fraud, as well as media licencing policies and consumer law, journalistic norms, research and teaching policies at universities and schools, internet protocols, counterintelligence organizations, employment law, whistleblowing organizations, policies to ensure a plurality and variety of platforms for expression and the consumption of expression, and (loosest of all) a commitment by society and culture at large to a freedom-of-expression ethos, under some conception of that ethos. The nature of this infrastructure, and so our attitude to its proper calibration, will be closely dependent on the characteristics of the democracy it makes possible and that makes it possible, on the specifics of the given jurisdiction, and on the immediate context (a school is not a parliament; public bodies, media organizations and religious organizations may have distinct responsibilities or entitlements).

This calibration task is already considerable thanks to our rapidly changing information ecosystem. Social media, the dark web, fake news, easily disseminated hate speech and propaganda, are of necessity driving a revolution in thinking across a diverse range of contexts. Our task is more limited, thankfully. We need only to think what difference the advent of deepfakes might make to the calibration challenge. This is taken up in the next section.

6 The makes-no-difference answer

We now have a clear sense of how to proceed. We need to consider the various harms made possible by deepfakes (outlined in Sect. 2), think about various strategies for containing those harms, and ask whether those strategies raise distinctive freedom-of-expression concerns—or, in the language introduced at the end of Sect. 5, raise distinctive challenges to the task of calibrating freedom-of-speech infrastructures. In this final section, I argue that there are no such distinctive challenges.

This is emphatically not to deny that deepfakes constitute a new and distinctive threat that needs to be taken seriously. They clearly do, as Rini and other authors have argued persuasively.[19] Nor is it to say that in addressing these threats we shouldn't think about freedom of expression. Our earlier consideration of the kneejerk answer showed that we should. Deepfake panic, motivated by a concern for democratic norms, could easily be co-opted and used to justify anti-democratic suppression, for example. The claim is simply that measures likely to be effective against deepfake harms do not seem to raise substantially new and unfamiliar freedom-of-expression difficulties (at a conceptual level, anyway—legislative or technological details will be a different matter). In this limited sense, the arrival of deepfakes makes no difference.

[19] See notes 1 and 2 and Sect. 2.

Section 4 already contained an example we can use to illustrate this deflationary approach. Overt deepfakes can be put to malign ends, such as racist stereotyping in images openly made with deepfake technology. Such harms, it was suggested there, might be dealt with on the same terms as racist stereotyping in sketches. So long as they are marked as overt, the fact of their being deepfakes becomes irrelevant.[20] This point can be made without our having to reach a settled view on how the harm done by racist sketches might be addressed within a well-calibrated freedom-of-expression infrastructure; it requires only the thought that racist overt deepfakes be treatable in the same way, whatever that is. Being a deepfake, in this context, makes no difference.

This pattern appears to generalize. In what follows, I run through the types of deepfake harm outlined in Sect. 2 and consider a range of potentially effective measures and reflections directed at containing those harms. These measures and reflections are already familiar in non-deepfake contexts, and their application in a deepfake context does not seem to generate any new freedom-of-expression concerns. This may sound like a low-key result, since I won't be pretending to have shown how to contain the deepfake threat, nor, *a fortiori*, how to contain it in a way that is compatible with freedom-of-expression concerns. Still, it is a more welcome outcome than the alternative, which would be that deepfakes *do* raise sui generis problems for freedom of expression. It also gives us a welcome alternative to the flawed (and potentially dangerous) kneejerk and dual-policy answers.

Let us start with harm done to the epistemic environment by deepfakes, deferring individual harms and non-epistemic environmental harms to later in the section. The undermining of trust in perceptual records is often seen as one of the chief threats posed by deepfakes because it serves the interests of those seeking to undermine trust in basic democratic institutions. I will present five measures or reflections that bear on the task of containing that threat. Each also bears on the task of containing threats posed by fake news more broadly, and there is nothing peculiar to deepfakes that should amplify any legitimate freedom-of-expression worries we may already have.

The first is the consideration that containment of malign deepfakes need not reduce to a simple choice between permitting and prohibiting. It is now possible to flag an image as a possible or suspected deepfake; or, conversely, to award it verified-source status. Graded verdicts of this kind are already familiar in non-deepfake contexts, as with flagging of Twitter posts and quality queries on Wikipedia. Such systems are rarely perfect in their operation and are open to abuse, but they are a welcome tool in the calibration task and—the key point for the present argument—their permissibility from a freedom-of-expression perspective is not affected by their being used to contain the harm done by deepfakes in particular.

Second, we should move away from the notion that tackling malign deepfakes is solely or even largely a technical task. It is instead likely to require co-ordinated action built on integrated stable alliances across different spheres (Schick, 2020: pp. 201–205). Technical fixes are likely to be a necessary part of that alliance's toolkit. Blockchain technologies, for example, better known for their use in creating trustworthy registers of crypto-currency transactions, could be adapted to create reliable

[20] This was presented as a possible move made by a supporter of the dual-policy answer, an answer subsequently rejected on other grounds. For why it is credible to treat overt cases as like sketches, see Sect. 4, note 14.

records of origin for images (Floridi, 2018: p. 321; and see Chesney and Citron, 2019: pp. 1814–1817). But such developments will be ineffective without co-ordinated support in the form of, for example, new legal frameworks, media-watch organizations, social policy responses, counterintelligence, and cultural adaptation.[21] This strategic observation, that we need integrated alliances, applies to disinformation and misinformation generally, not only to deepfakes. The key point in the present context, once again, is that if building such strong alliances into our freedom-of-speech infrastructure is warranted in the case of non-deepfake fake news, it is warranted for deepfakes themselves. Overreach would be overreach either way.

Third, notice that a well-calibrated freedom-of-expression infrastructure will support different forms of expression, including a diversity of opinions, across a range of contexts, platforms, and platform-types. Tolerance policies apt in one context may not be apt in another. Such diversity is vital to the twin goals of allowing individuals the freedom to produce and consume expressive acts as they see fit, while *also* promoting islands of epistemic authority that stand up more robustly to disinformation attacks. Deepfakes do not change the equation: they can be less stringently policed in some contexts without this inevitably leading to mistrust of authentic images accessed via islands of epistemic authority. Or, more carefully, if that is true for non-deepfakes, it is probably true for deepfakes too.

Fourth, several forays have been made into educating the public about deepfakes, with a view to building resilience. *In the Event of Moon Disaster*, mentioned earlier, was exactly that. If done clumsily, such alerts could have the wrong effect, making people wary even of genuine perceptual records and so doing the anti-democrat's job for them. A rounded education campaign would make clear the danger posed to the epistemic environment, explain that creating doubt is often precisely the goal of deepfake makers, and offer practical advice on how to avoid being fooled without simply doubting all images.[22] Information literacy approaches to fake news in general are increasingly familiar. They teach how to be a critical consumer of information and are compatible with freedom of expression principles. When Sander van der Linden describes methods for building 'immunity' to misinformation, he sometimes means this almost literally. He demonstrates the effectiveness of 'inoculation', weak doses of fake news that are subsequently revealed (2023: pp. 169–194, 255–271). *In the Event of Moon Disaster* and the like, he points out, adopt a similar strategy (273–4). Whether they are as effective is unclear, but they seem every bit as compatible with freedom-of-expression principles.

Fifth and finally, perhaps the most effective measures for protection against fake news, including deepfakes, would involve making societies more democratic. Democratic institutions are more likely to be robust against attempts to generate mistrust if they are more trusted in the first place, which would involve further entrenching

[21] Counterintelligence, for example, helped de-tooth a deepfake of President Zelenskiy of Ukraine calling for his compatriots to acquiesce to the 2022 Russian invasion. This was "pre-bunked" in that it was publicly predicted months in advance. ('Deepfakes v pre-bunking: is Russia losing the infowar?', *Guardian*, 19 Mar 2022.)

[22] See, for example, MIT's *Detect Fakes* toolkit (https://www.media.mit.edu/projects/detect-fakes/overview/) and the human rights organization WITNESS's list of preparations for the arrival of deepfakes (https://lab.witness.org/projects/synthetic-media-and-deep-fakes/).

broadly democratic values such as equality, a high-quality education system, inclusivity, openness, accountability, and the rule of law.[23] Here is not the place to sketch a plan to fix democracy, of course, but we do not need such a thing in order to appreciate that if this suggestion would help to defang fake news in general, its application to the problem of deepfakes would not raise any new freedom of expression concerns.

This is hardly a definitive list of sensible measures to diminish the damaging political impact of deepfakes on the epistemic environment, but it hopefully illustrates the approach and makes it provisionally credible. Let us therefore turn to the other kinds of deepfake harms listed in Sect. 2, where we find the same pattern. These include a variety of individual harms to persons or organizations connected to what the deepfake depicts, and non-epistemic environmental harms. Once again, measures to contain such harms seem to replicate measures needed to contain existing non-deepfake harms, and to do so without generating any new freedom-of-expression concerns.

We already saw one illustration of this, the strategy of treating racist overt deepfakes as racist sketches (a non-epistemic environmental harm in both cases). Measures to counter the use of deepfakes in fraud, defamation, or harassment, likewise do not raise any new freedom-of-expression concerns, since we already accept that anti-fraud, anti-defamation, and anti-harassment legislation and powers, under the right circumstances, legitimately place limits on permitted expression. As ever, the need to deal with a real threat could be misused to justify a disproportionate response (libel laws permit unjust gagging in some jurisdictions, for example, and anti-fraud measures could be used to ride roughshod over privacy concerns); but there is nothing about deepfakes that makes achieving this already difficult balance any more difficult. Being incorporated into deepfaked pornography without consent generates novel and concerning forms of personal harm, for example (Rini & Cohen, 2022), but categorising this as harassment or defamation would quickly and easily make suppression compatible with legitimate freedom-of-expression principles. Or consider the commercial exploitation of a well-known individual's voice in advertising without their consent. This is made technologically easier by deepfake technology, but that does not seem to raise any intractable new ethical difficulties for how such cases are dealt with. Finally, some pornography may or may not subordinate women as such (over and above harming the individuals depicted), but debate over whether it does and over what prohibitions this would warrant is not obviously affected by whether the images are deepfakes rather than authentic (if misleading) perceptual records.

Summing up, in this final section I have been highlighting a pattern. The various harms made possible by the advent of deepfakes, and the countermeasures we might adopt to tackle those harms, overlap with harms and countermeasures already known about in non-deepfake contexts and do not seem to generate *extra or novel* freedom-of-expression concerns. Or, to put it in the language of Sect. 5, the advent of deepfakes does not seem to call for any unfamiliar risky moves in the already difficult task of calibrating the freedom-of-expression infrastructure. There may yet turn out to be measures where this is not the case, measures that also turn out to be unavoidable, and currently unforeseen deepfake harms will doubtless come to light, so the conclusion

[23] A recent study (Open Society Institute Sofia, 2019) linked a population's capacity to withstand fake-news attacks to its score on a range of democratic measures, including free press, education levels, trust of others, and perceptions of corruption.

is tentative. Still, even with what we have seen, we have a more credible and less dangerous answer to the paper's overarching question than the kneejerk or dual-policy answers that first come to mind when thinking about the topic.

Acknowledgements I am grateful to Dan Cavedon-Taylor, Azita Chellappoo, Derek Matravers, Mark Pinder, participants at the Philosophy of Digital Images conference (University of Liverpool, May 2022), and, above all, this journal's diligent referees, for advice on earlier versions.

Declarations

Conflict of interest The author has no financial or other interests to declare.

References

Ajder, H., Patrini, G., Cavalli, Francesco, & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace.
Baker, C., & Edwin. (2011). Autonomy and free speech. *Constitutional Commentary, 27*(2), 251–280.
Barber, A. (2019). Lying, misleading, and dishonesty. *The Journal of Ethics, 24*(2), 141–164.
Bonotti, M., & Seglow, J. (2022). Freedom of speech: A relational defence. *Philosophy and Social Criticism, 48*(4), 515–529.
Cassam, Q. (2019). *Conspiracy theories*. Polity.
Cavedon-Taylor, D. (2013). Photographically based knowledge. *Episteme*, *10*, 283–297.
Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.
Costello, D., & Phillips, D. M. (2009). Automatism, causality and realism: Foundational problems in the philosophy of photography. *Philosophy Compass, 4*(1), 1–21.
Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. MIT Press.
Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy and Technology*, *34*, 623–643.
Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology, 31*, 317–321.
Gordon, J. (1997). John Stuart Mill and the "Marketplace of Ideas". *Social Theory and Practice, 23*(2), 235–249.
Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377–388.
Hine, E., & Floridi, L. (2022). New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, *4*(7), 608–610.
Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics, 8*(1), 81–108.
Kulvicki, J. (2020). *Modeling the meanings of pictures: Depiction and the philosophy of language*. Oxford University Press.
Lang, F. (2019). Adobe trains AI to detect deepfakes and photoshopped images." *Interesting Engineering,* 17 Jun 2019. https://interestingengineering.com/innovation/adobe-trains-ai-to-detect-deepfakes-and-photoshopped-images. Retrieved June 2023.
MacKinnon, C. A. (2006). *Are women human? and other international dialogues*. Harvard University Press.
Mikkola, M. (2019). *Pornography: A philosophical introduction*. Oxford University Press.
Mill, J. S. (1859). *On liberty*. John W. Parker and Son.

Open Society Institute, Sofia (2019). Findings of the media literacy index 2019. *Policy Brief* 55. https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf. Retrieved June 2023.
Ovadya, A. (2019). Deepfake myths: Common misconceptions about synthetic media.“ *Alliance for Securing Democracy*, https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/. Retrieved June 2023.
Post, R. (2011). Participatory democracy and free speech. *Virginia Law Review*, *97*(3), 477–489.
Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers Imprint, 20*(24), 1–16.
Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy, 22*(2), 143–161.
Schick, N. (2020). *Deep fakes and the Infocalypse: What you urgently need to know*. Monoray.
Shiffrin, S. V. (2014). *Speech matters: On lying, morality, and the law*. Princeton University Press.
Silbey, J., & Hartzog, W. (2019). The upside of deep fakes. *Maryland Law Review, 78*(4), 960–966.
Cairns, D. (2022). Abba: The inside story of how we transformed into avatars.“ *The Sunday Times (London)* April 30 2022. https://www.thetimes.co.uk/article/abba-the-inside-story-of-how-we-transformed-into-avatars-59ttv3f5f. Retrieved July 2023.
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*(1), 1–13.
van der Linden, S. (2023). *Foolproof: Why we fall for misinformation and how to build immunity*. London: 4th Estate.
van Mill, D. (2021). Freedom of speech. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (spring 2021 edition).* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/info.html
Weinstein, J. (2011). Participatory democracy as the central value of american free speech doctrine. *Virginia Law Review, 97*(3), 491–514.