Deepfakes, Public Announcements, and Political Mobilization (forthcoming in *Oxford Studies in Epistemology*)

Megan Hyska
megan.hyska@u.northwestern.edu

1 Introduction

This article describes an underattended to way in which synthetic dynamic media generated using deep learning models, now commonly called *deepfakes*, stand to interrupt the practices by which people mobilize for collective action. At the core of this account is an understanding of how this new medium alters the technosocial context for an older medium— that of videography.

The technology of moving images has never been free of manipulations designed to allow it to depict things that haven't actually happened. True videography's emergence in the 1880s was preceded by other forms of dynamic visual media like the magic lantern and zoetrope which, by spinning or shifting images, had shown the viewer many still illustrations or photographs in succession. But such moving images were no faithful report on real events; until the 1870s, it wasn't possible to take still photographs of objects in motion without blurring, and so insofar as the zoetrope presented an object in motion it *had* to present at least a partial falsification of how its model had really moved. While later videography too consists essentially of multiple still images presented very quickly, it differs with respect to its method of capture—a single machine capturing movement as it takes place. Video then *could* present moving images that reflected the world more or less as it had really been in the interval of their capture.

But this is not to say video always did so. It didn't take long for various special effects to find their way into film; in 1896, fillmaker and magician Georges Méliès shocked audiences by making a woman "disappear" on screen— simply by cutting frames in which she was present together with frames in which she was not [Kittler, 1999, 115]. Such tactical cuts, alongside other tricks including reversing or altering the order of frames, and the inclusion of practical effects (e.g. props, prostheses, makeup) were well established in cinematic videography by the end of the first decade of the 20th century [Dixon and Foster, 2008, 13].

But the number of frames involved in even short videographic samples remained a practical barrier to certain in-frame forms of videographic falsification. While sophisticated methods of photographic manipulation were well established by the time videography appeared [Fineman, 2012], even silent films were typically captured at 16 frames per second [Cook and Sklar, 2023] which meant doctoring 16 separate photographs to present even a second of doctored video, and doing so in a way that was consistent with fluid movement. Naturally, as the frame rate increased, the difficulty of fabricating film did too.

But while in cinema certain forms of falsification have always been regarded as acceptable, and indeed as professional achievements, videography has had another life alongside its cinematic one: as a tool of documentary. The practice of documentary, conceived of as the attempt to recruit communications technologies to the task of showing one another what the world is really like, is much older than videography. As Charles Musser has put it, "Documentary practices offered a method of communication that incorporated new media forms as they became available. Projected celluloid-based motion pictures was but one of these" [Glick and Musser, 2018]. But strikingly, in videography's role as a tool of documentary, it has often been treated as the most trustworthy and reliable medium we have, so much so that, as Rini [2020] has put it, video (and audio) recording have functioned as an "epistemic backstop" which "acutely corrects" and "passively regulates" our communication of information by other means. The significance of this epistemic role is particularly notable in the period beginning in the 1960s, when video cameras became widely commercially available, thereby redistributing the currency of epistemic authority that videographic capacity carried. This democratization gathered intensity as video cameras became cheaper, and videos easier to reproduce and distribute. These trends reached a crescendo in the 2000s when the ubiquity of camera phones and the birth of social media unleashed a new era of popular documentary.

We can summarize the unique role of video then as follows: notwithstanding the in-principle manipulability of the medium since its earliest days, the difficulty of producing serious and convincing videographic fakes has meant that we more or less treat video as factive—that is, as a medium which can depict some events only if those events have, under some description, actually taken place. In this respect, it was obviously unlike hand-drawn art, and even unlike something like the zoetrope. In addition, we knew a video when we saw it; video was *contrastively identifiable*— there was no other medium that could be confused with it. These features remained intact even across major technological advances in videography, including the shift from celluloid film to digital storage.

These two features together, factivity and contrastive identifiability, have allowed videography as a medium to play a special role in our epistemic and communicative lives. But these conditions are now being destabilized: deep learning models can now produce partially and wholly synthetic[1] dynamic visual media that are a) non-factive, and b) not contrastively identifiable. They

[1]For an extended discussion of partially synthetic media, see Millière [2022].

are not factive because they can depict a state of affairs without that state of affairs having, under *any* description, taken place. They are not contrastively identifiable because inspection doesn't differentiate them from video. And since contrastive identifiability is symmetric, while true videography remains factive, the advent of this new technology means *video is no longer contrastively identifiable either*. This essay is a contribution to the growing literature concerning the foreseeable epistemic disruptions of this flux in the character of videography and other media which, due to new synthetic media, have lost their contrastive identifiability.

In §2 I review the communicative dimensions of video, and discuss the way that deepfakes stand to disrupt not merely the acquisition of first order knowledge from videographic speech acts, but also the acquisition of higher order knowledge up to and including common knowledge. In §3 I come to the crux of why this matters: common knowledge is implicated, in multiple ways, in people's ability to act collectively. So if, in an environment of ubiquitous deepfakes, videographic speech acts can no longer give rise to common knowledge, they can also no longer function as they once did in political mobilization. §4 closes with a consideration of the possible futures that the ascendance of deepfakes suggests for us.

Before proceeding, a note about vocabulary. As is standard, in what follows we will say that a proposition, p, is *mutual knowledge* among some collection of people when every person in that collection knows that p. And it will generally suffice as a working characterization to say that *common knowledge* of p is present in a group of people just in case every person in the group knows that p, and also knows that every other person knows that p, and also knows that every person knows that every other person knows that p, and so on. This characterization of common knowledge as involving an infinite number of recursively characterized knowledge states, has of course been challenged as psychologically implausible. Insofar as we set out to use the term "common knowledge" to refer to the state that plays a distinctive role in communication and the solution to coordination problems, it has been argued that we should actually concede that the phenomenon in question really involves some finite number of epistemic states, or orthogonally that the recursively specified attitude need only be one of belief, or credence above a certain threshold, rather than knowledge. For this paper, I hope that it's possible to set these issues aside; I suspect that the fundamental point that I'm making would survive substituting any of these notions for the version of common knowledge I suggest in my working characterization.

2 Deepfakes and Common Knowledge

Recent years have seen rapid improvement in deep learning models that can produce samples of text and of still and moving images as well or better than human beings. Among these developments is the emergence of the "deepfake". At the moment, the most common variety of deepfake involves partial local synthetic alteration to a real video, using techniques like face swapping, head

puppetry, or lip syncing techniques [Tolosana et al., 2020, Zakharov et al., 2019, Prajwal et al., 2020]. At the time of this writing, some basic tools for the creation of totally synthetic video based on a text or text-and-still-image prompt are now available to the public, but with outputs that are easily discernible from veridical video. But it is a widespread assumption that tools for the creation of convincing, totally synthetic audio-visual samples will eventually become accessible to even those with very little financing or technical expertise. The technology for the production of deepfakes is meanwhile in an arms race with the technology via which they might be technologically discerned from real video [Farid, 2022]. It is reasonable to contemplate a scenario in which there really is no way, either with the naked eye or with the assistance of other AI tools, to tell the two media apart.

What use have we found for this new technology so far? Some examples are delightful: In February 2023, students at an MIT hackathon used AI graphics tools to create short videos responding to the prompt 'Tell me your dream' [Zhang, 2023]. Other uses, while relatively innocent, have raised issues about the use of people's likeness without their consent; for instance, the 2023 UK television program *Deepfake Neighbor Wars*, uses face swapping to create rudimentary deepfakes depicting celebrities like Idris Elba, Jay-Z, Adele, and Greta Thunberg as neighbors engaged in petty squabbles [Byman Shaw, 2023]. And of course further uses of deepfakes range from the seedy to the abominable: as Ohman [2020] and Rini and Cohen [2022] have pointed out, a primary use of deepfakes to date has been to create pornographic materials which not infrequently function as "revenge porn," that is, as material used to humiliate and discipline the women it depicts [see e.g. Ayyub, 2018]. Such uses also extend to the production of synthetic child sexual abuse material, raising new issues for automated content moderation systems designed to keep this content off the internet and for attempts to identify and help real victims [Harwell, 2023].

What the uses enumerated so far have in common though is that they are not exactly, or not necessarily, designed to trick anyone into thinking that the media they're watching is veridical. But, of course, a natural use of a nonfactive medium that is indiscernible from a factive one is to deceive. And indeed, high profile such uses of deepfakes are now familiar. In March 2022, a deepfake of Ukrainian President Volodymyr Zelenskyy appearing to tell Ukrainian troops to stand down [Simonite, 2022] was widely circulated in Ukraine. In Delhi's 2020 elections, Bharatiya Janata Party official Manoj Tiwari released deepfakes that used lipsync techniques to make it appear that he spoke the minority language Haryanvi in order to woo Haryanvi-speaking voters [Christopher, 2020]. And in May of 2022 a deepfake of Elon Musk promoting a cryptocurrency scam likewise circulated online [Elon Musk [@elonmusk], 2022].

While deception is then a natural use of deepfakes, the epistemology literature on deepfakes to date has been focused on what happens *after* people are deceived. As Rini [2020, 7] has put it, "the most important risk is not that deepfakes will be believed, but instead that increasingly savvy information consumers will come to reflexively distrust all recordings." The proposal of Rini and others [e.g. Fallis, 2021, Matthews, 2023] has been that the presence of

deepfakes in the environment effectively changes the epistemic role that video can play.

While many of these writers note that the the capacity to create totally synthetic dynamic media is not yet widespread enough to make the predicament they consider a technological reality, nor are their worries without existing empirical encouragement. In 2018, suspicion that a video of Gabonese Prime Minister Ali Bongo was a deepfake designed to cover for his death gave rise to an attempted coup [Cahlan, 2020]. More recently, false positives thrown up by AI-detection software created a "second level of disinformation" surrounding the Israeli-Palestinian conflict when images of putative atrocities were falsely adjudicated as AI-generated [Maiberg, 2023]. It is quite realistic then to worry that a general skepticism of video engendered by the ambient threat of deepfakes will have meaningful social and political consequences.

Without any alteration to the intrinsic features of video as a medium, the birth of this new technology, dynamic synthetic media or "deepfakes", threatens to oust video from our epistemic regard. This has far reaching effects for us as would-be knowers. But while the existing literature has emphasized the worry that, in this new information environment, we won't be able to acquire first-order knowledge on the basis of video, I want to draw our attention to a disruptive epistemic effect that goes beyond this. To bring into focus this further disruption, we'll start by considering a toy case:

> **Corruptionville 1:** In the small town of Corruptionville, no one has ever been exposed to a deepfake, nor are they aware of their possibility. In this town, residents all trust their mayor, but otherwise do not like or trust anyone else in town, and this distribution of trust is moreover common knowledge. The houses in Corruptionville all face onto a central square from the same direction in such a way that everyone can see the square but no one can see in anyone else's window. And the town has an unusual approach to the storage of their public funds: they keep them in a set of public coffers in the middle of the square. Now, late one night, the mayor sneaks out and steals some money from these coffers. As it happens, everyone in town was experiencing insomnia that night and saw the mayor's misdeed through their window. However, none of them realizes that anyone else has seen it. But one resident, Betty, had the presence of mind to record the mayor's theft on her phone. And the next day, at a town meeting, she plays the video for all of her neighbors.

When Betty plays this video for her neighbors, she is embedding a piece of technology, a video, in her communicative act— she is, we will say, making a *videographic public announcement* (VPA). The question that should interest us first is what effect her announcement might be expected to have: do the neighbors learn anything new, in the course of being shown the video? One might at first think no: each of them already knew that the mayor had stolen the money, so they learn nothing when they see the recording of this fact. But

this is too quick. A curious thing about public announcements, appreciated since the early days of speech act theory and, since the late 80s, given rigorous formal treatment by work in dynamic epistemic logic [see e.g. Plaza, 2007, van Benthem, 2006] is that they can give rise to further knowledge even in those who already know their content— in this case, even in those who already know that the events in the video took place. Specifically, they can give rise to higher order knowledge, i.e. knowledge about what the speaker and other members of the audience know: when we are in the presence of a public announcement that p we come to know that everyone else in the audience now also knows (or at least has justification for the belief) that p, and we come to know that they know that we know that they have this justification for the belief that p and so on.

What is significant about a public announcement being videographic? In other words why, in Corruptionville, might Betty bother showing her neighbors a video rather than just telling them what she saw? Clearly, it has something to do with the kind of justification or warrant that she thinks a video, as opposed to mere verbal testimony, can offer. A robust literature contends that photography offers perceptual, rather than merely testimonial justification [see e.g. Walton, 1984, Cavedon-Taylor, 2013, Rini, 2020], and accordingly points out that whereas testimonial justification is vulnerable to defeat based on trust of the testifier[2], perceptual justification is not. Because Corruptionville is a low-trust environment, it makes sense that Betty would prefer to provide her audience with a variety of justification that wasn't vulnerable to their lack of trust in her. And what is clear is that, in the pre-deepfake world of Corruptionville 1, announcements that embed video bypass barriers to belief that concern a lack of trust in the announcer. So while all public announcements that p have the capacity to bring about common knowledge that each person has justification for the belief that p, videographic public announcements couple that with common knowledge of the fact that this justification will generally be taken as sufficient for belief that p. Assuming that p is indeed true and that everyone in the given group has indeed seen the VPA, this entails common knowledge that p.

Prior to Betty's videographic public announcement, the residents of Corruptionville had *mutual knowledge* that the mayor stole the money—that is, they each knew this— but after her announcement, they come to have *common knowledge* that the mayor stole the money. This effect was dependent on the publicity of Betty's announcement, and upon its videographic character.

Now, let's consider a variant on our case:

> **Corruptionville 2:** Hold fixed all details of the prior case, except that now the mayor of Corruptionville has (anonymously) been de-

[2]Precisely how to spell out the role trust plays in testimonial justification and its defeat is a matter of some disagreement in the epistemology of testimony: for some theorists, the justification we have for testimonial belief always includes the trustworthyness of the speaker; for others, testimony has a default justification and the (un-)trustworthyness of the speaker becomes relevant only as a possible defeater; for yet others, testimony doesn't rely for its warrant on evidence that the speaker is trustworthy, but instead functions as an invitation to treat the speaker as trustworthy, which confers a kind of non-evidential epistemic warrant.

> liberately circulating deepfakes to the residents of Corruptionville for quite some time before his theft of money from the public coffers. The good people of Corruptionville have been tricked by deepfakes that they took to be veridical videos before, have subsequently realized they'd been tricked, and are now wary of being tricked again. This wariness is now common knowledge.

In Corruptionville 2, what happens when Betty plays her video for the assembled neighbors? Rini [2020] suggests that in an environment like Corruptionville 2, video can perhaps offer only the more easily defeated testimonial justification for a belief, where before it offered perceptual justification. This raises the worry that, in Corruptionville 2, Betty's VPA wouldn't be able to bring about the first order knowledge that the mayor stole the money.

However, this doesn't quite describe what we should imagine to take place in Corruptionville 2; the presence of deepfakes doesn't here endanger first order knowledge of the video's content, because that is already secure— all the residents of Corruptionville already know that the mayor stole the money, and will know that Betty's video is veridical upon seeing it. But the presence of deepfakes does still make a difference: it prevents Betty's videographic public announcement from giving rise to *common* knowledge of the mayor's theft.

Let's walk through the steps to that conclusion: We know that each of the residents of Corruptionville saw the mayor steal the money, but that none of them believe that any of their neighbors saw this. We also know that all the residents will treat video as a trust-vulnerable medium (i.e. as one the justificatory force of which is susceptible to defeat by mistrust of the announcer). Meanwhile, they don't trust Betty, they do trust the mayor, and they know this about one other. So every resident knows that all her fellow residents possess defeaters for the justificatory force of the video. While each resident knows that all their neighbors have seen the video, they have no reason to believe that this brought about belief in the video's contents (so no second order knowledge). They also realize that their neighbors will be reasoning similarly about them, and so have no reason to believe that their neighbors believe that they believe the video's content (so no third order knowledge), and so on. So unlike in Corruptionville 1, here Betty's VPA doesn't bring it about that the neighbors commonly know that the mayor stole the money.

To the existing literature on how the presence of deepfakes might change the communicative dynamics of speech acts in which video is embedded [Pierini [2023], Roberts [2023], Hyska, [forthcoming]], I then contribute the following observation: the presence of deepfakes in the environment alters the character of videographic public announcements.

It is of course worth observing that the residents of Corruptionville 2 are not entirely without a means by which to come to common knowledge of the mayor's theft. After they see the mayor steal the money, they are in a state of what is sometimes called *pluralistic ignorance*: everyone takes themselves to be alone in believing something that is in fact what everyone believes[3]. Often, what holds

[3]Pluralistic ignorance is sometimes characterized more broadly than this as a state in which

pluralistic ignorance in place is that there is a risk associated with revealing oneself to have a minority opinion, and so even though everyone in fact holds the same majority opinion, no one speaks up because they are not in a position to know this— this certainly describes the state of affairs in Corruptionville. Significant work has been done on the question of how pluralistic ignorance is alleviated. As Hendricks [2010] shows, in some cases pluralistic ignorance is resolved by a single public announcement; in Hans Christian Anderson's story of the Emperor's New Clothes, for instance, pluralistic ignorance among the Emperor's subjects is ended when a single child publicly announces that the Emperor is not wearing any clothes. And indeed, in Corruptionville 1, Betty's VPA might be seen to work in just this way. But as Hansen [2012, 74-76] notes, a single public announcement can resolve pluralistic ignorance only where the announcer is a trusted source; where this isn't the case, no one agent's declaration will have the capacity to totally dissolve pluralistic ignorance. Of course, even in Corruptionville 1, it's not that Betty herself is a trusted source, but that videography as a medium is. In Corruptionville 2, where there is no highly trusted medium to compensate for the lack of interpersonal trust, a single public announcement, even a videographic one, will not be enough to resolve the residents' pluralistic ignorance: it can only be dispelled if *many* people speak up. Of course, insofar as what incentivizes residents *not* to speak is the fear of social consequences that will attend their singling themselves out as holding a minority opinion, if some people do speak up despite this risk, this might progressively lower the risk for subsequent actors to do so, giving rise to the cascade of disclosures that actually *could* give rise to common knowledge even in a low trust environment. So certainly the claim here has not been that there is no way for the residents of Corruptionville 2 to reach common knowledge. But we can see that whereas, in the deepfake-free world, overcoming pluralistic ignorance required only one resident to take the personal risk of mking an announcement, in the deepfake-rich world this same effect requires many more people to take some degree of risk. Deepfakes increase the amount of friction encountered in the process of resolving pluralistic ignorance.

The Corruptionville cases feature many stipulated simplifications: we have assumed a uniformly low level of trust among residents, a universal viewership of Betty's video, and a highly homogeneous reasoning process among the video's viewers in both cases 1 and 2. However, I think our general conclusion here holds even for real communities, with all their greater complexity. It is however worth addressing the way that this conclusion is rendered more complex when we drop the simplification concerning lack of trust in particular. Corruptionville's almost universal lack of trust was essential in explaining why, in Corruptionville 1, VPAs had a capacity to bring about common knowledge in a way that mere verbal testimony did not. It is also why, in Corruptionville 2, it wasn't possible for Betty to compensate for the downgraded communicative power of video merely though the assertoric force with which she presented it.

"a number of individuals share the same cognitive error about specific other individuals or categories of individuals" [J. O'Gorman, 1986, 334]. See O'Gorman for a history of the concept of pluralistic ignorance in the social and behavioral sciences.

While it is perfectly realistic to imagine a community in which people do not universally trust one another to provide accurate information— this describes the United States, and most other mass societies—it is surely fairly unusual that the residents of Corruptionville trust almost *no one*. And in an environment where some people do trust one another, perhaps it turns out that the presence of deepfakes shouldn't be expected to make such a radical difference to the capacity of videographic public announcements to give rise to common knowledge. As Harris [2021, 13380] has put it, even in a deepfake rich environment, "insofar as one can be confident that a given source would not share deepfake videos, video footage shared by that source will retain its evidential power." Habgood-Coote [2023] moreover points out that our trust that sources won't deploy deepfakes needn't even be based upon a faith that they have some personal dedication to honesty— it can also arise because we are aware of socially imposed norms that will severely punish them for using this technology.

But we should notice that even these deflationary critics are acknowledging that, in a deepfake rich environment, video's justificatory force comes to be dependent on relations of trust. This is already a significant effect; it changes videography from a trust-indifferent medium to a trust-vulnerable one, which means that first-order knowledge as a result of videographic public announcements will be confined to within networks of trust[4]. And these consequences are even more significant when we consider the production of higher order knowledge. For your public announcement that p to bring about common knowledge that p between me and you, it doesn't suffice that we trust each other; we also have to know that we each trust the other, and know that the other knows this, and so on. In other words, for a public announcement using a trust-vulnerable medium to bring about common knowledge that p within some collection of agents, it already has to be common knowledge among these agents *that they universally trust one another*. When we consider the way that VPAs may be addressed to mass audiences, whose members may not even know one another, the presence of such trust is often far from obvious. The results, I think, are that in an environment of ubiquitous deepfakes, VPAs bring about a smaller and more unevenly textured terrain of higher order knowledge, circumscribed by existing explicit patterns of political or personal affiliation. Having stripped away one of the simplifications present in Corruptionville, we don't discover that deepfakes after all pose no threat to the capacity of videographic public announcements to bring about common knowledge; we merely see the nature of this threat in slightly higher resolution.

[4]Habgood-Coote [2023] disagrees with claims like this on the grounds that the justificatory force of technologies like photography and videography have to some extent been dependent on relations of trust for a long time, even prior to the existence of new synthetic media. But he fails to fully reckon with the ways that these social barriers to falsification have functioned alongside, and indeed been dependent upon, the technical barriers to the falsification of dynamic media that deepfake technologies eliminate. For more extensive critical engagement with this position see Hyska [2023].

3 Deepfakes and Collective Action

We've established that the presence of deepfakes in the environment modifies the potential of videographic public announcements to give rise to common knowledge. And this matters because of the ways that common knowledge enables people to do things together. This section will discuss this connection in detail. All of this is however in the service of making explicit the mechanisms of political mobilization that we then stand to lose because of the way that deepfakes modify our reception of video.

Common knowledge and collective action are entangled in a number of ways, but we might first consider the effects of commonly knowing some *premise of action*— that is, commonly knowing the facts on the basis of which it putatively makes sense for people to act. In Corruptionville, the premise for collective action was the mayor's theft of public funds.

A vast amount of research supports the intuitive point that people will forgo actions if they think they can only be successful in concert with others' actions but are unsure others will act; under such conditions we might say that agent's decision about whether to act is *quorum sensitive*. So consider first a resident of Corruptionville who wants to see the mayor held accountable for his actions (+1), but who believes that any action to bring this about will be unsuccessful unless others act alongside her. Let's say she also believes that others will act to censure the mayor if and only if they know that he stole the money. To make things stark, we'll assume that, in an environment where you were the only one who knew the mayor stole the money, it would be socially costly to take any action to censure the mayor (-1).

	Others know	Others don't know
Act	1	-1
Don't act	0	0

Table 1: Pro-accountability Resident

For such a resident, it is obviously the right thing to act to hold the mayor accountable if you believe that others too know about the mayor's theft, and obviously the wrong thing to do if others don't know. Where this agent doesn't believe that others know that the mayor stole the money, she will make her choice with the second column of the table in mind, and so not act. Common knowledge, which includes knowledge that others know, however, is sufficient to get this resident to make her choice with column 1 of the table in mind, and so to act. This then is one way in which common knowledge of a premise of action can make a difference: it overcomes quorum-sensitivity as a barrier to action.

Consider a different resident of Corruptionville who in fact is not interested in seeing the mayor held accountable; he is happy to see his friend the mayor get away with skimming a bit off the top (+1). However, in an environment where others know that he knows that the mayor stole the money, the mayor won't benefit because he'll certainly be held accountable by others, and if the

resident himself doesn't act he will pay the social cost of being seen as a crony of a corrupt official (-1).

	Others know he knows	Others don't know he knows
Act	0	0
Don't act	-1	1

Table 2: Anti-accountability Resident

Where this resident doesn't believe that others know that he knows what the mayor did, he will act with column 2 of the table in mind and so not act to hold the mayor accountable. But where he comes to believe that others know that he knows, he will make a decision with the first column of the table in mind and so act.

For the pro-accountability resident, it is second order knowledge (i.e. knowing that others know the mayor stole the money) that renders acting to hold the mayor accountable the clearly rational choice. And for the anti-accountability resident it is third order knowledge (i.e. knowing that others know that he knows that the mayor stole the money) which renders action the rational choice. Although I do not sketch all such cases here, I leave it as an exercise for the reader to imagine hypothetical residents such that fourth and higher orders of knowledge about the premise of action are what would be required in order to mobilize them—such cases are eminently constructable. The achievement of common knowledge of some premise of action, in encompassing knowledge at all these orders, then has something to contribute to the mobilization of a wide variety of agents that mere mutual knowledge does not.

So much for hypothetical analyses of how videographic public announcements and their capacity to bring about common knowledge of a premise of action might be mobilizing. But there are many very real examples in which VPAs have, by bringing about common knowledge of a premise of action, mobilized large numbers of people. Vivid among recent such examples are VPAs of police brutality.

It has long been a powerful idea that videographic monitoring of the police would regulate their conduct. But it should be noted that videographic public announcements of police misconduct are not an inevitable result of placing cameras in the vicinity of policing activities. Police body-cam policies, a mainstay of police reform proposals in recent decades, have a mixed record; while some research has found lower use-of-force rates where police body-worn camera (BWCs) were introduced [Michael D. White and Aili Malm, 2020, 31], a large and rigorous study in the DC area for instance found no evidence to reject the null hypothesis that "BWCs *have no effect* on police use of force, citizen complaints, policing activity, or judicial outcomes" [Yokum et al., 2017, 18, italics mine]. Just as footage from body-cams apparently doesn't necessarily discipline police conduct, it is also notable that it quite rarely yields videos that results in significant public mobilization. This is at least in part because these cameras often do not result in footage at all— officers leave their cameras be-

hind or switch them off—and existing footage is often unavailable to the public. As Aschoff [2020] reminds us, there is "no body-camera footage, for example, from March 13 [2020], when Louisville police used a battering ram to bust into Breonna Taylor's apartment in the middle of the night, spraying her apartment with bullets, killing her in her bed," even though the Lousiville police performing the no-knock raid on her house had been issued body cameras [Bella, 2021]. One way of putting this point is that police body cam policies do not translate neatly into more frequent videographic public announcements concerning police violence.

But civilian videography of police misconduct is a different matter. Such videos are not totally novel to the era in which most people carry around a cell phone with videographic capacities— George Holliday's video of Rodney King being beaten by Los Angeles Police was taken on a Sony camcorder in 1991—but there is no doubt that there are now many more of these videos because of the citizen journalism enabled by camera phones [Richardson, 2020, Lawrence, 2022]. And the virality-enabling diffusion capacities of social media have made the posting of these videos function as videographic public announcements with very large publics indeed. Moreover, it is pretty clear that these VPAs have had a massive capacity to mobilize. In summer 2020, George Floyd's murder by Minneapolis Police is estimated to have brought between 15 and 26 million Americans into the streets in protest, vastly more than any other protest movement in US history, and in the middle of a pandemic to boot [Buchanan et al., 2020]. Even if we are skeptical of the capacity of such videographic "sousveillance" (i.e. citizens' surveillance-from-below of the state's activities) to directly incentivize better policing or immediately trigger legal remedy for bad policing, "there's one thing images of police brutality seem to have the power to do: shock, outrage, and mobilize people to demand systemic change. That alone is the reason to keep filming" [Zuckerman, 2020].

While it is impossible to really ascertain the degree to which it was the widely distributed video of, say, George Floyd's murder, as opposed to the mere reporting of it, that brought about the summer 2020 uprising, there is good reason to think that the video was pretty important. That these videos have significant mobilizing potential is suggested by the mere fact that police have often tried to confiscate the phones of those who've recorded them [5]. Lawrence [2022] offers an analysis on which the steady increase in videos enabled by cell phones and social media had, in the decade prior to the 2020 uprisings, been used by Black activists to construct counterpublics that could effectively challenge the erasure of Black victims of police violence in the mainstream media. On this account, by the time 2020 came around, the mainstream media had been disciplined into covering police violence, and covering it as a systemic problem linked to race.

Each video of a human being being killed or maimed by the police is on the one hand documentation of a singular event: of a particular human being struggling to breathe, to protect their one, unique body from taser, baton, or

[5] See e.g. Antony and Thomas [2010] regarding the 2009 shooting of Oscar Grant.

gunfire; to return to their lives peopled with particular friends and families. But even as practices of witnessing these videos have acknowledged this wrenching, too-intimate singularity[6], it is crucial to how these videos have worked upon the American consciousness that each personal tragedy is also treated as a data point to be collated with others. As Richardson [2020] has put it, while no one video of police brutality instantaneously mobilized a massive multiracial swath of American society, their accumulation has over time cast police killings "not as isolated incidents captured serendipitously on camera, but as episodic proof of a pattern of abuse that is decades old" (139).

If, alongside the body of documentation that precedes them and the savvy work of movement communicators, these videos do mobilize, it is of course a further question how they do this. One answer emphasizes their ability simply to bring about first order knowledge. As one writer put it in a reflection on the saga following Rodney King's beating,

> when George Holliday's video surfaced, it signaled to a lot of citizens just how bad police violence visited upon marginalized communities actually was. People either didn't know what was happening or were willfully ignorant of it. They needed to wake up [Smith, 2015].

But implicitly acknowledged even here is that, while some part of the American public may view any given video of police brutality with the shock of learning, for the first time, that the police sometimes visit unjustifiable violence on the citizens they're sworn to protect, this doesn't describe everyone's experience with these videos. There are also the "marginalized communities" who, as the longtime victims of this violence, have always been aware of it. While a new video may bring about new first order knowledge that a particular person was abused thus-and-so, when it comes to the more general mobilizing proposition of police abuse, for these communities "the only thing new is the cameras" [Campbell and Valera, 2020]. Because police brutality videos mobilize these populations as well—indeed, the Black communities who experience disproportionate levels of police violence have been at the forefront of major anti-police-brutality mobilizations—we then need an explanation of how videos of police brutality mobilize that doesn't lean on fresh acquisition of first-order knowledge about the premise of action.

And I think the explanations that movement scholars have offered to fill this gap implicitly invoke a role for something like common knowledge. Keeanga-Yamahtta Taylor tells us that publicized police brutality functions as an "'event that captures people's experiences and draws them out from their isolation into a collective force with the power to transform social conditions" [2016, 153]. It does this because its publicity consists precisely in everyone coming to know that others know of what's just happened, just as they themselves do—in other words, because it gives rise to common knowledge of the event. In this way, it is unlike an abuse suffered personally, news of which is never widely circulated. Where one's willingness to take a particular action is quorum sensitive, as in the

[6] See Richardson [2020] for extensive discussion of what she calls "Black Witnessing".

toy case of the pro-accountability resident above, it makes sense that common knowledge will be mobilizing where mutual knowledge was not.

As in the toy case of the anti-accountability resident, we can also make sense of why videos of police brutality and its fallout might be mobilizing even for people who are indifferent to the existing regime of police abuse. If inaction could previously have been explained by the excuse of ignorance or uncertainty, video footage can take away this excuse. It is one thing to stay home when you can claim not to know about anything that could be a premise for doing otherwise, but it is another, riskier thing entirely to do so when everyone around you *knows that you know* about the prevailing injustice. Indeed, one function of mass protest is to remind the comfortable of the stakes that attend their failing to realize that this excuse has been removed; as Martin Luther King Jr put it in 1969,

> Today's dissenters tell the complacent majority that the time has come when further evasion of social responsibility in a turbulent world will court disaster and death. America has not changed because so many think it need not change, but this is the illusion of the damned. [King, 1986, 328].

So far we have discussed how VPAs depicting police brutality have given rise to common knowledge of a premise for political action, and how this might have played a key role in the uprisings these VPAs preceded. But the role of VPAs in political mobilization do not, I think, end there. Consider the proliferation, during mass protest movements like those in summer 2020, of video documentation *of the protests themselves*. How, we might ask, do these function to feed and maintain the mobilization?

In some cases, videos of protest contribute yet more evidence for the premise of action; in summer 2020, videos showed police pepper-spraying, ramming their SUVs into, and firing rubber bullets at peaceful protestors [Kim, 2020]. These public videographic announcements were then a further source of common knowledge about police brutality.

Documentation of protests themselves also functions to further eliminate quorum sensitivity as a barrier to action. Whereas in the toy case of the pro-accountability resident I simplified by stipulating that this resident believed that other residents would act if and only if they knew that the mayor stole the money, thus collapsing the space between second order knowledge and a confidence that quorum had been reached, in reality we may often doubt that others' knowledge ensures their action. But credible evidence that others are already acting decisively eliminates this barrier.

Finally though, VPAs of protests affect political mobilization not merely by inciting the erstwhile inactive to act, but also by shaping *how* we act. Our foregoing discussion of how, in general, common knowledge is entangled with collective action hasn't yet noted the ways in which common knowledge might be relevant not just to the decision that one will, individually, take action, but to *how* one will act, and relatedly, to an action's being genuinely collective.

Mass political action, including street protest, represents an attempt by very large groups of people to function as a unit. While subsets of these groups will have high degrees of internal organization and coordination, a pervasive feature of movements is that they are composed of "networks of activists, constituent organizations, supporters, and sympathizers whose grasp of plans and intentions is vague or divergent" [Kolers, 2016, 582]. In order for the diverging constituencies of such a network to function, even ephemerally, as a unit, they must choose a common plan of direct action, where multiple plans might be equally good, but where none can be effective if a critical mass of people don't select the same one. Which corporations will we boycott? When and where do we show up to engage in civil disobedience? A key feature of the situations animated by these questions is that there are two or more possible combinations of agents' actions such that no one of the protesters would be better off if only a single agent had acted differently. In other words, these represent *coordination problems*. And the suggestion familiar since Lewis [1969] but tracing its roots back as far as Hume [2000] is that the solution of these problems requires that a certain plan of action come to be common knowledge, whether through explicit communication or the establishment of a convention. Indeed, common knowledge of a shared plan is taken not to be merely helpful for but *constitutive of* collective action, in many of the most influential accounts [e.g. Bratman, 1993, Gilbert, 2009]. The common-knowledge-enabling function of VPAs is then instrumental in, or perhaps even constitutive of, certain forms of coordinated activity.

Beyond enabling the practical task of coordination around a common plan, common knowledge is also what establishes a shared set of symbolic resources, which allow the execution of a plan to resonate with a particular political meaning. A characteristic feature of protest movements for instance is the use of "unity displays" [Tilly et al., 2020] involving matching chants, ribbons, t-shirt colors, or physical movements. A unity display does not in and of itself achieve the movement's end, but it expresses the movement's claims to moral righteousness and to strength: it says that their cause is compelling enough to forge them into one unit, and prefigures, in symbolic terrain, a capacity for coordinated effort in practical matters. The resources deployed in a display of unity may be to some extent arbitrary— everyone wearing a blue shirt might be just as good as everyone wearing a red shirt— but what is crucial to the performance of unity is that a critical mass select the same one. In protest movements that span multiple cities and indeed countries, it is often video of protestors elsewhere in the world that perpetuates the adoption of these symbolic resources, and so the reach of the unity display. Established political symbolism also allows for the performance of continuity across time. Medina [2013, 225] draws our attention to the way in which actions that appear to be taken individually can become coined as symbols which allow others, in repeating them, to invoke their initial context. For these acts to take on their full social meaning as, in Medina's words, "echoing" or "chained to" others', and therefore read to the world as part of a larger protest movement, their symbolic significance must be common knowledge. And VPAs of protests and other political actions are what can allow this to occur.

I have summarized a variety of ways in which arriving at collective action—from overcoming quorum sensitivity and complacency as reasons for individual inaction, to arriving at a coordinated plan, to investing individual political actions with a unified symbolic resonance felt across time and space—makes use of common knowledge. It is no part of my argument of course that video functions as the only way to arrive at such knowledge and so come to function as a collective. Mass political action existed well before the birth of videography and of the internet as a means to make videographic announcements reach a wide public. Even since these technological developments, some notable political mobilizations have not made central use of videographic public announcements. There is for instance an interesting asymmetry in the role that VPAs have had in the Movement for Black Lives and in the roughly contemporary movements around gender-based sexual violence most commonly publicized in anglophone countries as #MeToo[7]. Both of these movements involved levelling accusations of misconduct at powerful people and institutions, and both highlight mistreatment based on identity (i.e. race and gender). But if "Black Lives Matter as a movement originated in images" [Cole, 2016], #MeToo and similar mobilizations based around gender-based violence have mostly proceeded by way of, and indeed thematized the act of, women's verbal testimony. However, even in #MeToo and similar mobilizations around gender-based violence, harassment, and discrimination, it is easy to list off cases where recordings of various kinds have galvanized political mobilization. When #MeToo rose to prominence in 2017, having been initiated a decade earlier by activist Tarana Burke, the investigative reporting that stoked public outrage about predatory men in the Entertainment industry made critical use of an audio recording of Harvey Weinstein admitting to sexual misconduct [Farrow, 2017], taken years earlier by model Ambra Battilana Gutierrez. Moreover, the well of rage that 2017's round of high profile accusations tapped into was certainly primed by the Access Hollywood video, circulated the year before, in which eventual President, Donald Trump described his habit of grabbing women "by the pussy". Meanwhile in 2014, Kenya's #MyDressMyChoice protests had been set off by a video of a woman's public assault in the streets of Nairobi [Igunza, 2014]; in 2019 public outrage about sexual harassment in universities in Nigeria and Ghana was catalyzed when footage of university professors' sexual misconduct was caught on tape by undercover BBC journalists [BBC News Africa, 2019]; and in 2023, a crisis was set off within Spanish womens soccer when a video of Spanish Football Federation President Luis Rubiales forcibly kissing player Jenni Hermoso after the team's world cup victory— and a seperate video of her confirming that it was nonconsensual— circulated on social media [Snape and Kassam, 2023].

While VPAs are not the only means by which people can be mobilized into collective action then, video and the capacity to make videographic public announcements are firmly established as one tool that many of us today use to reach for political collectivity. As Rini [2020] puts it,

[7]For a comparative discussion of the way that digital activists worked to develop counterpublics around gender- and race-based oppression on social media in particular, see e.g. Jackson et al. [2020].

> For better or worse, we have developed a web of epistemic norms assuming reliance upon recordings. In the developed world, there is no one living today who remembers an epistemic environment preceding that reliance. Video and audio recordings, in existence longer than any of us, have always structured our lives. (13)

The observation of this section has been that common knowledge plays an important role, above and beyond mutual knowledge, in moving people into political action and allowing them to function collectively. If, as the previous section contends, deepfakes disrupt the capacity of VPAs to give rise to common knowledge, we then see that this disruption challenges the gestures by which we are drawn into, and consciously position ourselves within, political collectives.

4 Conclusion

Videos and deepfakes are two different kinds of media in the truest sense: they mediate our relationship with the world in fundamentally different ways. But phenomenologically, they form a single category: a moving image that presents the world to us in an idiom that is a minor dialectical deviation from that of our own sensory apparatus. But where this realism and immediacy is divorced from an assumption of factivity, how will we experience the sorts of images that would once have functioned to mobilize us?

Among the most mobilizing videographic media have always been documentations of injustice. It has long been observed though, that photos and videos of violence and suffering are not infallibly linked to remedial political action. Indeed, Susan Sontag [2003] notes that photo- or video-graphic documentation of violence and injustice "may give rise to opposing responses. A call for peace. A cry for revenge. Or simply the bemused awareness, continually restocked by photographic information, that terrible things happen" (13). Likewise registering the non-mobilizing effects of these images, Richardson [2020] notes the ways that having to continually watch police brutality videos has burdened Black Americans. Summarizing this perspective is the Black artist and activist Dread Scott who, in an interview with Richardson, notes that videos of police brutality

> have helped increasing numbers of people see the depth of the problem, but left to its own it's just going to be sort of like lynching photos, where those were used by white people to celebrate a job well done and towards black people to terrorize us [Richardson, 2020, 65].

For all that this essay has emphasized the positive potentials of videographic documentation of injustice then, it is true that such videos have always been politically ambivalent: under the right circumstances they can mobilize, yes, but they can also numb, brutalize, and discipline their audiences[8]. And notably,

[8]For more on the case study that Scott alludes to, lynching photography as both a tool of white supremacist terror and of anti-lynching activism, see Wood [2009], especially section III.

the potential for images of suffering to do all this is seemingly independent of the epistemic features which implicate them in mobilization, like (perceived) veridicality and a perceptual flavor of justification. Sontag [2003, 34] notes her difficulty in looking at the excruciating, but of course fictional, death of a disobedient satyr in Titian's *The Flaying of Marsyas*; particularly violent and bizarre genres of AI-generated pornography will reveal images that many of us will find difficult to look at despite knowing they do not depict a real event of suffering [Cole, 2023]. This phenomenon, of dysphoria elicited by the fictional, opens onto a confusing moral complex in which we are absorbed by the horrors of the merely possible, a domain which is untouchable by action. But so much more confusing is the state of affairs when we encounter an image of suffering and violence not knowing whether it's veridical or not, and perhaps with no serious prospect of finding out. Yet this is the predicament that we can expect to face more and more in coming years: all the affective trappings of witnessing violence and injustice coming increasingly decoupled from a sense that one can do anything about it, or that others will work with us to do so. My worry, in summary, is this: in an environment rich with deepfakes, we can expect both deepfakes and veridical videos that depict violence to continue to be distressing, and indeed to lend themselves to "terrorizing" those who see themselves or their communities depicted, even as they lose their capacity to mobilize. These media will retain the worst functions of images depicting suffering while losing the features that have at other times redeemed them.

This sketch of what I take to be an impending technosocial predicament is of course not intended to suggest that this predicament has no solutions, or that the technology that gives rise to it has no countervailing implications that might redound to the benefit of those who would like to engage in collective political action. A natural solution to the problem would be to, in Rini's expression, find a new backstop that could correct and regulate fake videography, putting the lie to deepfakes that misrepresented the world. And there are a number of suggestions out there for how blockchain technology might serve as this backstop. Blockchain might for instance be used to connect images to their metadata so that their provenance can be tracked across platforms [Koren, 2020], which would at least help determine whether a putative video came from a trusted source. Others have suggested that the blockchain could be used to record people's locations and produce a sort of infallible alibi, were deepfakes produced that depicted them doing something in another location Chesney and Citron [2018, 1814]. It's clear that these ideas wouldn't solve all the problems that I've suggested deepfakes might cause, but they certainly gesture at the fact that such problems are not in principle unsolveable.

As for the politically positive potentials of deepfakes, I have discussed elsewhere the way in which generative AI might be used to help us communicate to one another about the political alternatives that we envision [Hyska, [2023]]. It is also worth mentioning that, whereas videographic surveillance by states and corporations has a chilling effect on protest, generative AI, including deepfakes, suggest new possibilities for the strategy of political resistance that Brunton and Nissenbaum [2015] call *obfuscation*: "the deliberate addition of ambiguous,

confusing, or misleading information to interfere with surveillance and data collection" (1). There is already precedent for using machine learning to produce large quantities of ersatz data in order to mount an obfuscatory defense against would-be IP thieves [Chakraborty et al., 2021]. The capacity to create ersatz videographic data via deepfakes presents the resources for protestors to defend themselves from this surveillance via obfuscation.

What I hope in any case to have shown is that the epistemic implications of deepfake technology are only fully appreciated if we attend to the epistemic lives of collectives, rather than individual would-be knowers. The challenge that such technology poses is not most significantly one for the individual quest to maximize true beliefs, but for our efforts to discern and act upon the world in concert.

References

Mary Grace Antony and Ryan J. Thomas. 'This is citizen journalism at its finest': YouTube and the public sphere in the Oscar Grant shooting incident. *New Media & Society*, 12(8):1280–1296, 2010.

Nicole Aschoff. Smartphones Have Transformed the Fight Against Police Violence. *Jacobin*, June 2020. URL `https://jacobin.com/2020/06/video-recording-police-brutality-george-floyd`.

Rana Ayyub. I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *HuffPost UK*, November 2018. URL `https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316`.

BBC News Africa. Sex for Grades: undercover inside Nigerian and Ghanaian universities. *YouTube*, October 2019. URL `https://www.youtube.com/watch?v=we-F0Gi0Lqs`.

Timothy Bella. Police may have withheld body-cam footage from night of Breonna Taylor's death, lawsuit says. *Washington Post*, July 2021. URL `https://www.washingtonpost.com/nation/2021/07/10/breonna-taylor-body-camera-lawsuit/`.

Michael E. Bratman. Shared Intention. *Ethics*, 104(1):97–113, 1993.

Finn Brunton and Helen Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press, 2015.

Larry Buchanan, Quoctrung Bui, and Jugal K. Patel. Black Lives Matter May Be the Largest Movement in U.S. History. *The New York Times*, July 2020. URL `https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html`.

Dan Byman Shaw. Deep Fake Neighbour Wars, the comedy turning Ariana Grande into a scaffolder. *The Independent*, January 2023. URL `https://www.independent.co.uk/arts-entertainment/tv/news/deep-fake-neighbour-wars-itv-b2268105.html`.

Sarah Cahlan. Analysis | How misinformation helped spark an attempted coup in Gabon. *Washington Post*, February 2020. URL `https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/`.

Felicia Campbell and Pamela Valera. "The Only Thing New is the Cameras": A Study of U.S. College Students' Perceptions of Police Violence on Social Media. *Journal of Black Studies*, 51(7):654–670, 2020. Publisher: SAGE Publications Inc.

Dan Cavedon-Taylor. Photographically Based Knowledge. *Episteme*, 10(3): 283–297, 2013.

Tanmoy Chakraborty, Sushil Jajodia, Jonathan Katz, Antonio Picariello, Giancarlo Sperli, and V. S. Subrahmanian. A Fake Online Repository Generation Engine for Cyber Deception. *IEEE Transactions on Dependable and Secure Computing*, 18(2):518–533, 2021.

Robert Chesney and Danielle Keats Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107: 1753–1820, 2018.

Nilesh Christopher. We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. *Vice*, February 2020. URL `https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp`.

Samantha Cole. The Community Pushing AI-Generated Porn to 'the Edge of Knowledge'. *404 Media*, August 2023. URL `https://www.404media.co/ai-generated-porn-generative-artificial-intelligence/`.

Teju Cole. The Superhero Photographs of the Black Lives Matter Movement. *The New York Times*, July 2016. URL `https://www.nytimes.com/2016/07/31/magazine/the-superhero-photographs-of-the-black-lives-matter-movement.html`.

David A. Cook and Robert Sklar. History of film | Summary, Industry, History, & Facts | Britannica, May 2023. URL `https://www.britannica.com/art/history-of-the-motion-picture`.

Wheeler W. Dixon and Gwendolyn Audrey Foster. *A short history of film*. Rutgers University Press, 2008. URL `https://hdl.handle.net/2027/heb07996.0001.001`.

Elon Musk [@elonmusk]. @cb_doge Yikes. Def not me., May 2022. URL `https://twitter.com/elonmusk/status/1529484675269414912`.

Don Fallis. The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 34 (4):623–643, 2021.

Hany Farid. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, 1(4), September 2022. doi: 10.54501/jots.v1i4.56. Number: 4.

Ronan Farrow. From Aggressive Overtures to Sexual Assault: Harvey Weinstein's Accusers Tell Their Stories. *The New Yorker*, October 2017. URL `https://www.newyorker.com/news/news-desk/from-aggressive-overtures-to-sexual-assault-harvey-weinsteins-accusers-tell-their-stories`.

Mia Fineman. *Faking It: Manipulated Photography Before Photoshop*. Metropolitan Museum of Art, New York, 2012.

Margaret Gilbert. Shared intention and personal intentions. *Philosophical Studies*, 144(1):167–187, 2009.

Joshua Glick and Charles Musser. Documentary's Longue Durée: Reimagining the Documentary Tradition. *World Records*, 2(4), 2018. URL https://worldrecordsjournal.org/documentarys-longue-duree-reimagining-the-documentary-tradition/.

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

Jens Ulrik Hansen. A logic-based approach to pluralistic ignorance. In Jonas De Vuyst and Lorenz Demey, editors, *Future Directions for Logic – Proceedings of PhDs in Logic III*, pages 67–80. College Publications, 2012.

Keith Raymond Harris. Video on demand: what deepfakes do and how they harm. *Synthese*, 199(5-6):13373–13391, 2021.

Drew Harwell. AI-generated child sex images spawn new nightmare for the web. *Washington Post*, June 2023. URL https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/.

Vincent F. Hendricks. Knowledge Transmissibility and Pluralistic Ignorance: A First Stab. *Metaphilosophy*, 41(3):279–291, 2010.

David Hume. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Oxford University Press, Oxford and New York, 2000.

Megan Hyska. The politics of past and future: synthetic media, showing and telling. *Philosophical Studies*. Forthcoming.

Emmanuel Igunza. Kenya's stripping videos cause outrage. *BBC News*, November 2014. URL https://www.bbc.com/news/world-africa-30217462.

Hubert J. O'Gorman. The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences*, 22(4):333–347, 1986. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1520-6696%28198610%2922%3A4%3C333%3A%3AAID-JHBS2300220405%3E3.0.CO%3B2-X.

Sarah J. Jackson, Moya Bailey, and Brooke Foucault Welles. *#Hashtag Activism: Networks of Race and Gender Justice*. The MIT Press, 2020.

Catherine Kim. Images of police using violence against peaceful protesters are going viral. *Vox*, May 2020. URL https://www.vox.com/2020/5/31/21275994/police-violence-peaceful-protesters-images.

Martin Luther King. *A Testament of Hope: the Essential Writings of Martin Luther King, Jr.* Harper & Row, 1986.

Friedrich A. Kittler. *Gramophone, Film, Typewriter*. Stanford University Press, 1999.

Avery Kolers. Social movements. *Philosophy Compass*, 11(10):580–590, 2016.

Sasha Koren. Can Publishers Use Metadata to Regain the Public's Trust in Visual Journalism?, January 2020. URL `https://open.nytimes.com/can-publishers-use-metadata-to-regain-the-publics-trust-in-visual-journalism-ee32707c5662`.

Regina G. Lawrence. *The Politics of Force: Media and the Construction of Police Brutality, Updated Edition*. Oxford University Press, 2022.

David Lewis. *Convention*. Harvard University Press, 1969.

Emanuel Maiberg. AI Images Detectors Are Being Used to Discredit the Real Horrors of War. *404 Media*, October 2023. URL `https://www.404media.co/ai-images-detectors-are-being-used-to-discredit-the-real-horrors-of-war/`.

Taylor Matthews. Deepfakes, Fake Barns, and Knowledge from Videos. *Synthese*, 201(2):41, January 2023.

Jose Medina. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford University Press, 2013.

Michael D. White and Aili Malm. *Cops, Cameras, and Crisis : The Potential and the Perils of Police Body-Worn Cameras*. NYU Press, New York, 2020. ISBN 978-1-4798-2017-7.

Raphaël Millière. Deep learning and synthetic media. *Synthese*, 200(3):231, May 2022.

Carl Ohman. Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, 22(2):133–140, June 2020.

Francesco Pierini. Deepfakes and depiction: from evidence to communication. *Synthese*, 201(3):97, March 2023.

Jan Plaza. Logics of Public Communications. *Synthese*, 158(2):165–179, 2007. Originally published as Plaza, J. A. (1989). Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, & Z.W. Ras (Eds.), Proceedings of the fourth international symposium on methodologies for intelligent systems.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, Seattle WA USA, October 2020. ACM.

Allissa V. Richardson. *Bearing Witness While Black: African Americans, Smartphones, and the New Protest #Journalism*. Oxford University Press, 2020.

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

Regina Rini and Leah Cohen. Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, 22(2), July 2022.

Tom Roberts. How to do things with deepfakes. *Synthese*, 201(2):43, 2023.

Tom Simonite. A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. *Wired*, March 2022. URL `https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/`.

Jamil Smith. Videos of Police Killings Are Numbing Us to the Spectacle of Black Death. *The New Republic*, April 2015. URL `https://newrepublic.com/article/121527/what-does-seeing-black-men-die-do-you`.

Jack Snape and Ashifa Kassam. Spanish football president's kiss sparks outrage after Women's World Cup final. *The Guardian*, August 2023. URL `https://www.theguardian.com/football/2023/aug/21/luis-rubiales-kiss-outrage-spanish-football-fa-president-womens-world-cup-final-spain-jenni-hermoso`.

Susan Sontag. *Regarding the Pain of Others*. Penguin Random House, 2003.

Keeanga-Yamahtta Taylor. *From #BlackLivesMatter to Black Liberation*. Haymarket Books, 2016. URL `https://web-p-ebscohost-com.turing.library.northwestern.edu/ehost/ebookviewer/ebook/bmxlYmtfXzExOTU4MzZfX0FO0?sid=265aff62-068b-4c4c-857a-4c86942da7b9@redis&vid=0&format=EB&rid=1`.

Charles Tilly, Ernesto Castaneda, and Lesley J. Wood. *Social Movements 1768-2018*. Routledge, 4th edition, 2020.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, December 2020.

Johan van Benthem. "One is a lonely number": logic and communication. In Zoe Chatzidakis, Peter Koepke, and Wolfram Pohlers, editors, *Logic Colloquium '02*, number 27, pages 96–129. A K Peters/CRC Press, 2006.

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

Amy Louise Wood. *Lynching and Spectacle: Witnessing Racial Violence in America, 1890-1940*. University of North Carolina Press, Chapel Hill, 2009.

David Yokum, Anita Ravishankar, and Alexander Coppock. Evaluating the Effects of Police Body-Worn Cameras: A randomized controlled trial. Working Paper, The Lab @ DC; Executive Office of the Chief of Police, Metropolitan Police Department, Washington; DC; Office of the City Administrator, Executive Office of the Mayor, Washington, DC, 2017.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9458–9467, October 2019.

Ruihan Zhang. MIT AI for Filmmaking Hackathon 2023 Brings Dreams to Life. *MIT Media Lab*, February 2023. URL `https://www.media.mit.edu/posts/mit-ai-for-filmmaking-hackathon-2023-brings-dreams-to-life/`.

Ethan Zuckerman. Why filming police violence has done nothing to stop it. *MIT Technology Review*, 2020. URL `https://www.technologyreview.com/2020/06/03/1002587/sousveillance-george-floyd-police-body-cams/`.