Deepfakes, Public Announcements, and Political Mobilization

Megan Hyska, Northwestern University, United States
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

This paper considers the possibility that widespread access to deep learning models capable of quickly generating novel, high-quality videos—deep fakes— will represent a disequilibrium between technologies of communication and technologies of forgery. Specifically, I argue that this disequilibrium will leave long-distance communication without a crucial means of *showing*, rather than merely *telling*, that certain facts obtain. And, because a certain sort of political activity depends upon this kind of showing, I make projections concerning the way that advances in deepfake technology can be expected to affect the practice of politics.

The technological state of affairs that this paper is concerned with is not yet a reality; although mostly convincing audio-visual deepfakes of entirely novel events are now possible, their production requires significant time investment and technical expertise, with various technical obstacles to their production by generalized deep-learning models still to be overcome (Bommasani et al., 2022, 32). However, the recent explosive advances in turning generalized "foundation models"[1]to tasks like generating novel text (e.g. GPT-3) or still images (e.g. DALL-E) suggests the plausibility of eventually scaling and democratizing the ability to produce highly convincing novel audio-visual samples via artificial intelligence (Chesney and Citron, 2018, 1772-1773). In such world, faked video can be expected to become ubiquitous. And this ubiquity would, I argue, have profound effects on the overall dynamics of information flow and the way that mass politics proceeds.

Let's bear in mind some general dynamics of information flow, technologies of communication, and politics. Humans, like most organisms, survive by conditioning their behavior on information from their environments. Simplifying a complicated issue, we will say that an information transmission event is *direct* just in case the state of affairs that the information is about is itself the proximal cause of the sensory impression by which the receiver acquires the information, and indirect otherwise. Much of what humans and other animals learn is then indirect: you don't see the predator, but you do see the rustling in the bushes. This willingness to take one stimulus as a proxy for another is the premise that all our technologies of communication are based on, from written language, to the phone call, to the video clip, to the e-mail.

Forgery is the technological domain that exploits our willingness to accept indirect information: where people are willing to accept stimulus$_A$ as an indi cation of stimulus$_B$ and act accordingly, the forger's task is to bring about this behavior by producing stimulus$_A$ without stimulus$_B$. Different communications technologies have had different degrees of forgeability across different moments of history and, in any era, living among others successfully has required a facil ity with identifying which information technologies of the day were forgeable. An equilibrium between paranoia and credulousness has been struck partially on the basis of there being a sphere of near or total unforgeability: some com municatiions technologies that are regarded as nearly unforgeable, whether it be the proprietary wax seal on a letter, the video clip, or the block chain packet.

An innovation in the domain of forgery doesn't necessarily upset this equilibrium, where innovations in

[1] The term "foundation model" is a recently coined one referring to machine learning models that are "trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al., 2022, 3)

communications technology stay one step ahead: the possibility of good faith communication is sensitive to the relative, rather than absolute, sophistication of forgery. But ubiquitous deepfakes represent a deep incursion into the sphere of unforgeability, and advances in detection technologies are hard-pressed to catch up, particularly since these technologies can in fact be used to train less detectable deepfakes going forward (see for discussion Farid, 2022). There is moreover not a clear successor to video technology as a method of communication.

What can we expect the effects of this incursion to be? As Rini (2020) notes, the true danger of deepfakes is "not that they will trick us into believing false content, but that they will gradually eliminate the epistemic credentials of all recordings" (8). Fallis (2021) has pointed out that the advent of deepfakes effectively changes how much information a video contains, and thus perhaps whether it can give rise to knowledge.

The point I want to make is distinct from Fallis'. Whereas his claim is that there is a *degree* of informativity that deepfakes will strip from videography as a technology, mine is that there is a *kind* of informativity that is so stripped. The kind in question is that of showing. Two different uses of communications technology are showing and telling. The distinction as I'm drawing it is inspired by the discussion in Grice (1957), and made explicit in later work in pragmatics like Sperber and Wilson (2015). Cases of telling are instances of what Grice famously dubbed non-natural meaning, where the speaker intends a) that the audience come to bear an attitude toward a piece of information, b) that the audience recognize this intention, and c) that the recognition of this intention is what is in part causally responsible for the audience in fact coming to bear the attitude. Call these the informative, communicative, and causal intentions respectively. Showing is just like telling but without the causal intention. In showing cases, the signaler does intend to get some information across to the receiver (the informative intention) and does intend this intention to be transparent to the receiver (the communicative intention) but they intend that the receiver come to believe the information not because of the signaler's inferred intention but because they are faced with some independent evidence of it. Grice's famous case involves Kind Herod showing Salome St John the Baptist's severed head: even if Salome infers Herod's informative intention, her recognition of his intention is causally otiose when it comes to the formation of the belief that St John the Baptist is dead– the presence of his severed head has already taken care of that.

One crucial difference between a video of an event and verbal testimony concerning the event then is that the video *shows* what happened, whereas the testimony *tells* you about it. It's not precisely that verbal testimony as a communication medium is incapable of showing, but that, unaugmented by ostension toward things in the environment, it can at most engage in second order showing (i.e. showing the audience that the speaker is speaking in such and-such a way). For first-order showing, where we are not in eye or ear shot of the subject matter we want to communicate about, and so cannot resort to ostensive gesture, we rely heavily on photography and videography.

Recall that cases of showing that p require the presentation of evidence that p beyond the communicator's own intention to get the audience to believe that p. If videos cease to be extra-intentional evidence that what they depict has really taken place, then the presentation of a video can no longer count as showing. And it isn't unreasonable to think that deepfakes do threaten the status of videos as extra-intentional evidence. Ina world of ubiquitous deepfakes, I could only regard a video as credible evidence if I trusted its source not to have fabricated it. If the reason that the presentation of a video brings about a certain cognitive response in its audience involves trust that the audience wouldn't mislead them, then what we have is an instance of telling, not showing.

Why should it matter that we lose one of our primary methods of showing? I argue that showing is a key

mechanism by which extra-institutional political collectivity emerges and comes to exercise power. Of course, deep fakes do not threaten the sort of showing that takes place in person. But increasingly since the 90s, a lot of political collectivity formation and mobilization has made indispensible use of the internet (see e.g. Tufekci, 2017). And online, it is videos and photographs that we rely upon for showing.

In making visible this indispensible role for showing in political collectivity formation, I distinguish between *group-addressing* and *group-forming* political discourse[2]: the former addresses an already consolidated group of loyalists, and the latter aims to draw individuals *into* one group or out of another. Showing is particularly important to the latter variety of politics, because it can function in the absence of pre-existing trust. The legacy of political organizers realizing this is long; in the 19th century, European anarcho-socialists endorsed "propaganda by the deed" as a messaging strategy that targeted workers and peasants who "make their way home worn out from fatigue and have little inclination to read socialist pamphlets or newspapers" by "showing them what they cannot read, of teaching them socialism by means of actions and making them see, feel, touch" (Brousse, 1877, 150). And in the 21st century US, the Movement for Black Lives plausibly would not have occurred without our contemporary forms of political showing: video clips of police misconduct, as well as of the responding protests. In summary, I argue that creating the grassroots political formations that regular people can use to collectively wield power has always required a means of offering a not-yet-mobilized public evidence beyond the organizer's testimony. In our own age, video evidence has been the paradigmatic means of doing so. Pervasive deepfakes will strip videography as its capacity to show, rather than tell, and in so doing it presents a challenging to political organizers. This might suggest the approach of an age of ascendant social atomization— at least, it makes vivid the scale of technological creativity that organizers will find need and scope for in the coming decades.

References

Arendt, H. (1951). *The Origins of Totalitarianism*. Harcourt Brace Jovanovich, San Diego, CA.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., R´e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tram`er, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs].

Brousse, P. (1877). Propaganda by the Deed. In Graham, R., editor, *Anarchism: A Documentary History*

[2] The distinction is roughly analogous to Arendt's (1951) distinction between indoctrination and propaganda, or McAlevey's distinction between mobilizing and organizing (2016)

of Libertarian Ideas, Volume One: From Anarchy to Anarchism (300 CE to 1939), pages 150–151. Black Rose Books, Montreal, Canada.

Chesney, R. and Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107:1753– 1820.

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

Farid, H. (2022). Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, 1(4).

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66(3):377–388.

McAlevey, J. F. (2016). *No Shortcuts: Organizing for Power in the New Gilded Age*. Oxford University Press.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. 20(24):16.

Sperber, D. and Wilson, D. (2015). Beyond Speaker's Meaning. *Croatian Journal of Philsoophy*, 15(44):117–149.

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, New Haven and London.