Small Lies, Big Problems:
The Dangers of Deepfakes Extend Beyond Political Misinformation

By Conor Rowan

Aside from large language models such as ChatGPT and text-to-image services like DALL-E, the latest and perhaps most disconcerting headlines in tech concern AI-generated videos called “deepfakes.” Deepfakes are synthetic images or video generated from artificial intelligence which have become increasingly prevalent in the last year. It appears that photorealistic text-to-video services will soon be available to the average internet user, raising fears about national security in an era of news where seeing is no longer believing.

Articles and papers on deepfakes typically involve thought experiments, such as a fabricated video of President Biden warning American citizens of an incoming intercontinental ballistic missile. In March 2022, a deepfake video circulated online picturing Ukrainian President Volodymyr Zelensky calling for a surrender to Russian troops. Though this video was not totally convincing, deepfakes may soon be indistinguishable from high-resolution camera footage. Concerningly, even deepfake detection tools are resorting to sophisticated metrics such as facial blood flow and blinking patterns in order to keep pace with improvements in the technology. So how worried should we be about deepfake videos? Apart from fears around national security, misrepresentation of public figures, and the subversion of trust which the news tends to emphasize, what other pitfalls can we expect from this technology?

To begin, we should note that disruption to a society’s information environment from technological innovation is not unprecedented. Prior to the widespread availability of the printing press in 18th century England, the Church had a monopoly on information. The eventual ubiquity of the press ushered in a “pamphlet culture” rife with inflammatory political rhetoric and misinformation. English society was radically restructured by this new technology of communication, and in response to the ensuing confusion, citizens slowly learned not to believe everything they read. As individuals adapted to the new information environment, so did institutions–scientific and journalistic enterprises played an increasingly important role in regulating the flow of credible information. Given the obvious ease with which one can distort the truth in writing, a litany of social technologies have been adopted since the introduction of the printing press to help certify credible written content–journalistic standards, citation practices, reputations of institutions, and liability for publishing harmful content. Though the recent era of polarization and dysfunction in the news has demonstrated that the problem of written misinformation has not been solved, our culture has a sufficiently healthy information immune system to view many of the most bombastic written claims with skepticism–advertisements are full of grandiose promises which we view with suspicion, and a text message from a friend providing an update on current events is probably corroborated by a trusted source before being taken seriously. Maybe we can expect a similar trajectory with deepfake video as with writing–that we will rethink how to interpret the medium, and reputable institutions will act as gatekeepers for video that should be trusted, while the rest is seen as

essentially creative. Of course, everyone falls short of this ideal of media literacy, and there is ongoing disagreement about which institutions to trust. Though the problem of managing the prodigious flow of written information unleashed by the printing press and later digital technologies is imperfectly solved, this example still helps illustrate a possible trajectory of our individual and institutional responses to deepfakes.

Consider this: how many times would you need to be fooled by a fake video before you began to shift your understanding of the relationship between video and reality? When the historic film “The Arrival of a Train” was first screened in 1896, audiences responded to footage of an oncoming steam engine by recoiling in fear. Of course, it did not take long for movie-goers to understand the particular relationship that this medium had to the physical world of motion and objects. There are certainly legitimate concerns around political deception and national security with deepfake videos, but I suspect that it will not take long for us to sever the historically solid ties between video and reality, and to understand video as a creative medium which is employed in service of its creator’s ends. Perhaps deepfake images and video will become like paintings or animation–media which depict recognizable places, people, and objects without claiming to represent their nature or behavior in the outside world.

Like writing, which we see as a vehicle for both fact and fiction, video will never be entirely divorced from reality. There is ongoing research into methods to detect AI generated video, which would help users make more informed decisions about what to trust online. The success of these techniques ranges from impressive to pitiful, though as a solution to the problem of deepfake-induced misinformation, relying entirely on detection sets the stage for an arms-race between deepfake detectors and creators. As the detection methods improve, deepfake services will reverse engineer them to erase the detectable fingerprints of AI-origin in their videos. It is unfortunate in some ways that deepfake technology is already sufficiently decentralized so as to prohibit any kind of enforcement of standardized watermarking, which would unambiguously certify the origin of a piece of media. There is another solution, heralded by the Microsoft and Adobe sponsored “Content Authenticity Initiative,” which authenticates the history of an image or video with cryptography. Each picture or video using this service would be stamped indicating that it has been authenticated, and viewers would be able to investigate whether it originated from a device and examine the edits it underwent.

Detection and authentication methods can help people be better informed about information they see online, but some uses of deepfakes are explicitly illegal and should not be tolerated by media platforms. States such as California and Texas have already passed laws criminalizing the use of deepfakes to manipulate elections. Another common application of deepfake technology is to swap women’s faces onto existing pornography. This is primarily done without consent, and the resulting videos can be used to humiliate or discredit the victim. Current privacy and “revenge porn” laws are flexible enough to apply to these uses of deepfakes, and new laws around political misinformation, along with detection and authentication methods, will hopefully deter the creation of chaos and political disorientation in the wake of high-quality deepfake video. There is, however, another class of harm not

addressed by law or the technological tools of media literacy–these are the more subtle problems which will persist even when solutions to handle crime and political misinformation are settled upon.

Speaking on an American Bar Association panel in December, law professor Andrew Woods argued that “the small lies around the social presentation of self” are a bigger part of the online misinformation problem than is typically understood, especially for teens. In part, he is referring to the edited, filtered, and posed photos which define the experience of most social media platforms. Professor Woods, like many others, sees the prevalence of socially dishonest and emotionally manipulative online content as being intimately related to the bleak statistics on the state of teen mental health. NYU business professor Johnathan Haidt has created a database showing the rise of numerous indicators of mental illness around the advent of social media in the early 2010’s. One 2020 government data base found that fully 25% of teenage girls had a major depressive episode in the previous year. A 2018 study linked social media use to ADHD symptoms and sleep deprivation. Tellingly, a 2021 paper found that 40% of social media users often regret their entire session online, and in particular the recommended content. Though the precise causes of these outcomes are complex and multifaceted, they suggest a truth which many feel intuitively–that there are pitfalls to conducting an online social life within our current media environment which manifest as real-world harm. Even though we are aware that much of the content we see online is dishonest and unrealistic, it imprints itself on our mental models of how the social world functions and what other people’s lives are like. This fact, that we can know on some level that online content is dishonest but still be persuaded and influenced by it, is important for considering the potential harms of deepfakes outside the scope of the explicitly criminal or political.

In the years since media platforms such as Facebook, YouTube, and Instagram became ubiquitous, thinkers such as Tristan Harris of the Center for Humane Technology have made progress in understanding the underlying causes of the political and psychological issues associated with digital media. To address these questions, we might first ask: why are services like Google Search, YouTube, Facebook, and Instagram free? The business model of these media companies is to sell data on users' online behavior to third-party advertisers, with the aim of helping these advertisers to tailor their ads to an individual’s preferences. Consequently, in order to produce a higher quantity of useful data, platforms are incentivized to compete to keep users on their site for as long as possible. The way in which time on site is maximized is through curating content, whether through YouTube’s recommended videos or Instagram’s news feed, with complex algorithms trained on an individual’s browsing history to optimize engagement. It might be argued that maximizing time on site is equivalent to providing an enjoyable online experience, but we have seen unequivocally in the last decade that in the world of online media, engagement is a terrible proxy for individual or societal well-being. These sophisticated algorithms, trained to predict and cater to our preferences of what to see online, have unwittingly demonstrated that optimally capturing attention has less to do with producing durable value and more to do with appealing to impoverished notions of group membership and primitive emotions such as anger, disgust, and envy. We see reverberations

of these dynamics in the ongoing problems with both teen mental health and political polarization.

If we survive the near-term turmoil of deliberately provocative political deepfakes, and video comes to be understood as a creative medium, what can we expect from deepfakes in this media landscape of engagement-optimizing recommendation systems? I suspect the most insidious and persistent harms from deepfakes will simply come from exacerbating current problems with social media: increasingly compelling and increasingly dishonest content around the presentation of self; furthering the sense that reality is bizarre, arbitrary, and disorderly; a deepening conviction that people not like you are unreasonable and dangerous; increasing the speed and ease with which content creators can game your psyche to compete for attention; a staggering multiplication of the ability to personalize entertainment and advertising. As media scholar Neil Postman said in his classic book *Amusing Ourselves to Death*, “what the advertiser needs to know is not what is right about the product, but what is wrong about the buyer.” Video is an extremely potent tool for communication and persuasion, and Postman wrote this before the age of personalization. By using these platforms, we have unknowingly agreed to be the subjects of a new type of advertising–not in the sense of seeing advertisements for consumer goods constantly (though there is a healthy amount of this online), rather that the whole platform is a constant advertisement for itself; an advertisement created to optimally exploit *what is wrong with you, the buyer*. It seems like an oversight to locate the harm of deepfakes solely in the realm of news and politics when media platforms have already weaponized the contents of our social lives. Deepfakes unlock a new suite of tools for media platforms to experiment with what captures and holds attention, in spite of the real and potential consequences for users. What fraction of our online existence has to do with sorting out facts anyway?

Deepfakes certainly pose problems for national security and the political process, but given how influenced we are by the storytelling and social signals of our friends and online communities, there is also potential for harm outside the arena of politics. This version of the deepfake problem is less frequently discussed–perhaps in the short term it is less pressing than avoiding mass panic, but the long term effects of deepfake technologies in the hands of the attention economy may prove to be both destructive and stubborn. If we can agree that an underlying driver of our current problems is the paradigm of engagement-based recommendations, then perhaps this is a leverage point to ameliorate the harms of the deepfake dystopia outlined above. One way to approach this is allowing users to choose the objective of their recommendation algorithms, thus fostering a more agentive, critical, and self-conscious online experience. Though the current recommendation paradigm has shown that in some sense, people prefer compelling but negative emotions, I suspect many would not choose anger or jealousy consciously. If state-of-the-art machine learning models can learn to optimize content curation for engagement, we must imagine it is possible to optimize for education, relaxation, or other worthy goals. Of course, it may not be easy to sort out the collateral consequences of optimizing for certain objectives, as we have seen with the competition for attention. Furthermore, it may be difficult to characterize what measurable

online behaviors constitute relaxation or real learning. But we are in a situation where we are knowingly optimizing for the wrong objective–is it far-fetched to claim that improvements might be expected from reformulating recommendation objectives to reflect more noble goals? Though recommending content on the basis of engagement is a way to stay economically competitive, there is increasing cultural pushback on the societal harms perpetuated by media platforms. Giving users this kind of autonomy could be a worthwhile investment in a platform's credibility and a more genuinely satisfying online experience. Furthermore, an early adopter could have the benefit of setting precedents for other tech companies. With the growing disillusionment around social media and the concerning new possibilities that deepfakes unlock, a paradigm shift of this sort is not impossible to imagine. Though updating online recommendation systems may not directly address concerns about trust in the news and misrepresentation of public figures, it may act to curb the virality of this provocative content while explicitly reducing the psychological harms originating in the small lies around how we present ourselves online.

Early media scholar Marshall McLuhan taught that the clearest way to understand a culture is to study its tool for conversation, and with his famous dictum "the medium is the message," that each new communication technology makes possible new types of discourse. What kinds of conversations will we, as a culture, be having with AI-generated video? Though this technology certainly opens up new frontiers of creativity, do we trust that engagement-based media platforms will steward it for anything other than extractive commercial purposes? Ubiquitous deepfake video will entrench and intensify our current problems with social media. A change in how content is recommended is one path forward to address this very fundamental problem. But in the meantime, we might be wise to safeguard our attention and treat our entertainment with caution.