Article

new media & society
1–27
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1461444820925811
journals.sagepub.com/home/nms
SAGE

Anticipating and addressing the ethical implications of deepfakes in the context of elections

Nicholas Diakopoulos
Northwestern University, USA

Deborah Johnson
University of Virginia, USA

Abstract

New media synthesis technologies are rapidly advancing and becoming more accessible, allowing users to make video and audio clips (i.e. deepfakes) of individuals doing and saying things they never did or said. Deepfakes have significant implications for the integrity of many social domains including that of elections. Focusing on the 2020 US presidential election and using an anticipatory approach, this article examines the ethical issues raised by deepfakes and discusses strategies for addressing these issues. Eight hypothetical scenarios are developed and used as the basis for this analysis, which identifies harms to voters who view deepfakes, candidates and campaigns that are the subjects of deepfakes, and threats to electoral integrity. Four potential forms of intervention are discussed with respect to multi-stakeholder responsibility for addressing harms, including education and media literacy, subject defense, verification, and publicity moderation.

Keywords

Anticipatory ethics, deepfakes, electoral integrity, synthetic media, technology ethics

Corresponding author:
Nicholas Diakopoulos, Department of Communication Studies, Northwestern University, Evanston, IL 60208, USA.
Email: nad@northwestern.edu

Introduction

New technologies used to synthesize media are rapidly advancing and becoming more accessible, allowing users to make compelling video and audio clips of individuals doing and saying things they never did or said. Users can, for instance, synthesize a particular individual's voice based on a transcript, swap one person's face onto another person's body in a video, or synthesize an entirely new video of someone speaking based on audio that's lip-synced to their face. Recorded audiovisual media is becoming more and more malleable, facilitating an ease of editing almost analogous to text.

The technology offers a host of potential benefits in entertainment and education, from multilingual advertising campaigns and assistive technologies for people who have lost their voice to bringing dead artists "back to life" to engage museum-goers. But it can also challenge aural and visual authenticity and enable the production of disinformation by bad actors. Popularly referred to as "deepfakes" (owing to their reliance on a new form of machine learning known as "deep learning"), these techniques have the potential to destabilize domains such as news reporting, where audio and video may be treated as evidence that something actually happened. Less technology-intensive "cheapfakes," such as the widely circulated clip of House Speaker Nancy Pelosi that was slowed to make her appear intoxicated, have already demonstrated the potential for even low-tech manipulated video to have an impact. As deepfake technology becomes increasingly easy to use and broadly accessible via open source tools and readily available services, a whole new level of realism, speed, scale, and ability to personalize disinformation are enabled (Diakopoulos, 2019). As such, deepfakes contribute to the broader problem of "fake news" by technically enabling both the more widespread fabrication or manipulation of media that may be deliberately used for the purposes of disinformation (Tandoc et al., 2018; Wardle and Derakhshan, 2017) and the introduction of uncertainty which can affect trust in news on social media (Vaccari and Chadwick, 2020).

Because deepfakes can distort the information voters use to make decisions, they have the potential to undermine the integrity of democratic elections. In this article, we take an anticipatory approach (Brey, 2017) in considering how deepfakes could affect the 2020 US presidential elections. We systematically develop eight hypothetical scenarios and use these as the basis for thinking through two fundamental ethical questions: (1) What harms result from the use of deepfakes in elections? and (2) What can be done to limit those harms?

To address the first question, we provide an analysis of the ethical issues prompted by the scenarios, offering nuanced articulations of the harms to voters who view deepfakes (deception and intimidation), candidates and campaigns that are the subjects of deepfakes (reputational harm and misattribution), and the integrity of elections (undermining trust). To address the second question, we draw on our ethical analysis and consider four potential forms of intervention that could limit the harmful effects of deepfakes: education and media literacy, subject defense, verification, and publicity moderation. For each intervention strategy, we provide a responsibility analysis that identifies how various actors—social media platforms, journalists and news media organizations, technologists, political campaigns, policy makers, and individual citizens—could be involved in addressing the harms wrought by deepfakes. This analysis identifies modes of intervention that are relevant to elections as well as other domains in which deepfakes could be problematic. As an interdisciplinary study, our analyses serve to enrich disciplines such as technology ethics, anticipatory ethics, communication studies, and science and technology studies, while at the same time providing a broader, synthesized contribution to the pragmatic discourse needed to address the ethical implications of deepfakes.

A primer on synthetic media technology

To anticipate the social and ethical implications of a new technology, it is important to first understand the state of the art of the technology, including current limitations and any new opportunities it affords (Sarewitz, 2011). Media synthesis is an applied technology that draws on a constellation of underlying basic technologies related to computer graphics, computer vision, and machine learning that have been steadily advancing over the last few decades. One of the earliest systems, presented more than 20 years ago, was already capable of synthesizing speaking faces by splicing together a series of mouth shapes from footage of a person to align to newly input speech (Bregler et al., 1997). The latest techniques, however, are capable of far greater fidelity and believability due to increased resolution and quality of image sensors, the availability of more data, and advancement in machine learning techniques, such as deep neural networks, that utilize that data. The clear trajectory of the technology is toward more realistic and believable synthesized depictions.

Video synthesis techniques largely fall into two main categories: *face-swap* and *facial reenactment*. Face-swap involves taking the facial movements of a person in an original video and mapping a target person's face on top so as to make the target person look like they said or did the things the original person said or did in the video. Face-swaps rely on the voice and shape of the person in the original video, while substituting the facial identity of the target person. In contrast, facial reenactment involves synthesizing a target person's face based on the input expressions from another person's face (puppeteering), or from a different input voice track (lip-sync). Face-swapping technology initially gained notoriety due to its use in the creation of non-consensual pornography, where someone's face is swapped into an existing pornographic video (Cook, 2019). Research on the method has continued to advance robustness to partial facial occlusions and skin tone differences while further reducing the amount of training data needed (Nirkin et al., 2019). But the technique is already widely available through open source tools (e.g. Faceswap, DeepFaceLab) and online services which lower the barriers to use in terms of technical expertise and cost (Lee, 2019).

In comparison with face-swapping, facial reenactment has received more attention in the academic research literature. For instance, the Face2Face system enables real-time facial puppeteering by taking an input video of an actor's face and transferring the mouth shape and expressions onto a synthesized target face (Thies et al., 2016). Despite not including a voice track, output videos drew attention and triggered initial concerns over the misuse of the technology by demonstrating examples of an actor controlling the faces of Donald Trump and Vladimir Putin. Facial reenactment was also achieved in a system using an input voice clip and video footage from President Obama to synthesize new footage of the president saying things he had said in another speech (Suwajanakorn et al., 2017). This technology can be paired with the target person's real voice as input, the voice of an impersonating actor, or a fully synthesized voice mimicking the target person. Some of the limitations include not being able to work with oblique video and needing enough training data to model a diversity of mouth shapes. Recent advances also allow for the interactive control of head positions and expressions in facial reenactments, while more realistically synthesizing matching eye gaze, blinking, and upper body pose (Kim et al., 2018). Another method allows users to directly edit the transcript of a video to insert, remove, or change the words of the person speaking, synthesizing both the necessary speech and lip-synced video of the person to match the new transcript (Fried et al., 2019). Video synthesis techniques also work beyond the face and upper body, allowing for motion transfer from an actor's entire body, such as a dancer, to a target person (Chan et al., 2018).

Voice synthesis has also advanced, driven by new deep learning techniques that are able to render a text transcript as spoken audio. For instance, Google's WaveNet system was trained on about 25 hours of speech data from professional speakers and used to synthesize English-speaking voices that were rated by human evaluators on a "naturalness" scale from 1 to 5. People rated the synthesized voice a 4.21 on average, while real human speech was rated as 4.55, only about 8% better (Van Den Oord et al., 2016). More recent techniques train a generic model that can then be adapted to match the voice of new speakers using only a few minutes or seconds of data from that person (Arik et al., 2018). Interactive versions of voice editing allow text transcripts to be edited and the voice of a narrator seamlessly blended into the audio track to match the text edits (Jin et al., 2017). Voice synthesis capabilities, including target speaker cloning, are now offered as commodity services by several providers.

Although this article focuses primarily on video and audio deepfakes, images and texts can also be synthesized using some of the same deep learning techniques. For instance, convincing portrait images of people who do not exist can be generated (Karras et al., 2018), with control over facial emotions, hair, gender, age, and skin color (Choi et al., 2018). In the domain of text, techniques have been developed for synthesizing convincing online comments, such as Yelp reviews (Yao et al., 2017), as well as news articles that mimic a given news outlet's style (Zellers et al., 2019).

Election 2020 scenario development methodology

Using an anticipatory governance approach, we developed eight hypothetical scenarios that serve to create strategic foresight into the potential of deepfakes to negatively affect the US presidential elections (primaries and general) of 2020. Anticipatory governance calls for examination of the social implications of emerging technologies, while they are still in the early stages of development when such examination can influence future development (Guston, 2014). Anticipatory ethics is a form of anticipatory governance that focuses specifically on the ethical implications of emerging technologies (Brey, 2017; Johnson, 2011). Scenarios are one of the several methods typically used to develop the foresight needed for anticipatory ethics studies (Brey, 2017). In this case, the scenarios we developed help anticipate the space of possible action for various actors to use deepfakes to affect elections, operating as an instrument to expose ethical issues and harms. We use the scenarios as focal points for examining cross-cutting ethical issues and for understanding how intervention strategies might address those issues.

Because the scenarios were to be used as material for our analysis, a systematic and disciplined approach to their development was essential. The overarching goal was to ensure that the scenarios were plausible and covered a broad range of uses. To arrive at credible renditions, the approach was systematic and iterative in researching, writing (or re-writing), and evaluating scenarios (Selin, 2006). We began by identifying the state of the art in the technology (as described in the preceding section), so that the scenarios were grounded in current and near-future technical capabilities. We then developed a typology of conceptual dimensions to ensure that the scenarios covered a range of sociotechnical configurations. The typology included different *actors* (i.e. a candidate or campaign staff, external entities such as political action committees (PACs), foreign governments, or regular citizens), *motivations* for those actors (e.g. support a candidate, hurt a competitor, thwart the process, express a political opinion), *modalities* of media (e.g. audio, video, image), *phases* of a campaign (e.g. early or late in the election cycle), *channels* for distribution (e.g. social media, podcasts, chat apps), and *mechanisms* for influencing voters (e.g. discrediting a candidate's reputation by association, exaggerating a candidate's views, suggesting a candidate's corruption or hypocrisy, inciting a campaign's base, intimidating voters, and undermining or attacking the election process).

We also incorporated a participatory approach, harnessing the insights and feedback of non-expert crowd workers to help expand the range and diversity of perspectives represented beyond those of typical specialist stakeholders and thereby enhancing pluralistic reflection on ethical issues (Brey, 2017; Nikolova, 2013; Sarewitz, 2011). To do this, we asked workers on Amazon Mechanical Turk (AMT) to creatively write scenarios about the use of media synthesis in the 2020 elections. Workers were instructed to "write a ~300 word fictional scenario to explore the future of how a new media technology could be used to impact the outcome of the 2020 U.S. presidential election." The technology of either video face-swapping or audio voice synthesis was first described to workers and examples were shown. Then, we suggested workers consider various different actors and goals for those actors as well as setting the election context for the scenario. In total, we collected 20 written scenarios, 10 focused on video and 10 on audio. Workers were paid US$4 for each scenario with up to US$3 bonus paid for exceptional scenarios that were "specific, believable, and plausible." We then read these and noted plausibly conveyed uses of the technology that were then added to our typology as additional elements to take into consideration.

Our scenario writing process then systematically combined different dimensions of the typology to explore plausible configurations. For instance, one scenario could explore a motivation related to undermining the election process using video and produced by a foreign nation, while another might outline how a candidate could be hurt using audio produced by an opposing campaign. In the late spring and early summer of 2019, we produced 13 scenarios this way, covering many of the various motives, media modalities, and phases from our typology.

We then evaluated the scenarios based on the criteria of *consistency* (i.e. no internal contradictions), *completeness* (i.e. no gaps in the narrative), and *plausibility* (i.e. reasonable to conclude the scenario could happen given technical and social constraints; Urueña, 2019). We focus on *plausible* scenarios, rather than the merely *possible* or the more stringent *probable*, as it balances opening up the future while not being so broad as to become unfocused or create a need to rank based on likely risk (Ramírez and Selin, 2014). We additionally assessed how each of the 13 scenarios touched other aspects of the typology to ensure we had adequate representation of the concepts. We dropped four of the 13 that were not compelling according to these criteria, or which could be condensed to more efficiently convey ideas. The next step in evaluation was again participatory and utilized crowd workers to assess the plausibility of the scenarios on a scale from 1 "not plausible" to 5 "very plausible," as well as explain their evaluation of what specifically made each scenario plausible or not. This is important as the notion of plausibility is not intrinsic, but is rather subjective and its evaluation therefore benefits from diverse input. Much like in the previous crowd task, here we also explained what the technology could do and provided examples. Ten independent ratings for each scenario were gathered from AMT workers (paid US$.75 per rating). If a scenario received an average rating below 4.0, we revisited it and edited aspects that workers indicated as implausible. In one case, four of the raters gave a score of 1 or 2 out of 5 and suggested that the crux of the scenario was not believable, and so we decided to drop that scenario. This left us with eight scenarios.

The eight scenarios are provided in Appendix A along with short reflections on the ethical implications of each. The reflections consider previous ethical analyses (Bendel, 2019) as well as ethics frameworks sometimes used in evaluating emerging technologies (Brey, 2017; Wright, 2010). Readers are encouraged to read the scenarios and reflections before going on to the next sections of the article because the scenarios served as the basis for our subsequent analysis.

What harms could result from the use of deepfakes in elections? An ethical analysis

Chesney and Citron (2019) identify two categories of harms generally resulting from deepfakes: harms to individuals and organizations and harms to society. Our analysis refines these two categories further as we examine the effects of deepfakes in the domain of elections. We identify three categories of harm: (1) harms to viewers/listeners (i.e. those who watch or listen to deepfakes); (2) harms to subjects (i.e. the targeted subjects of deepfakes as well as those who are auxiliary to the intended target); and (3) harms to social institutions (i.e. the domain in which a deepfake operates). In the election context, the primary viewers/listeners are voters; the primary subjects are candidates and campaigns; and the social institution is the democratic election system. Table 1 lists the specific harms to voters, candidates and campaigns, and elections together with the scenarios used to illustrate each.

Amplification plays an important role in the harms of deepfakes, but amplification is not a harm in itself. Amplification expands other harms. In the case of elections, broad distribution compounds harm to voters, candidates and campaigns, and the integrity of

Table 1. This matrix summarizes the scenarios referenced in the text to illustrate each ethical harm (the reader is encouraged to read the full scenarios in Appendix A).

Harms to viewers/listeners	
Deception	• *Scenario 1*. A deepfake video is used to exaggerate the military record of a candidate • *Scenario 2*. A deepfake video of a candidate falsely shows the candidate making disparaging and hateful comments about the opponent's base • *Scenario 3*. A video of deepfaked testimonials is used to show individuals falsely claiming they had an affair with the candidate • *Scenario 4*. A targeted email campaign deceives recipients about their correct polling location and uses a deepfake video of a respected public figure to lend credibility to the message • *Scenario 5*. A deepfake audio clip of a candidate falsely represents the candidate saying disparaging things about voters • *Scenario 6*. A deepfake image is used to threaten to deceive an individual voter's friends and family about their participation in pornography • *Scenario 7*. A deepfake audio clip of an exchange between campaign staff and a debate moderator falsely suggest a candidate cheated in a public debate • *Scenario 8*. A deepfake parody video of a candidate is taken out of context on social media and some viewers no longer know it is a parody
Intimidation	• *Scenario 6*. Individuals are threatened via a spoofed text message that a deepfake pornographic image of them will be shared publicly and with friends and family if they vote on election day
Harms to subjects	
Reputation	• *Scenario 2*. A deepfake video of a candidate shows them making disparaging and hateful comments about the opponent's base, negatively affecting their reputation amongst that group • *Scenario 3*. A video is released showing deepfaked testimonials of individuals claiming they had an affair with the candidate, tarnishing the candidate's reputation amongst those who are morally opposed • *Scenario 5*. A deepfake audio clip of a candidate saying disparaging things about a politically important group of voters hurts the candidate's reputation amongst that group • *Scenario 6*. A deepfake image is used to threaten to deceive an individual voter's friends and family about their participation in pornography, thus tarnishing the voter's reputation • *Scenario 7*. A deepfake audio clip of an exchange between campaign staff and a debate moderator suggest a candidate cheated in a public debate
Misattribution	• *Scenario 8*. A deepfake parody video exaggerates the views of a candidate and when taken out of context creates misattributed negative credit for those exaggerated views
Harms to social institutions	
Undermining trust	• *Scenario 6*. Individuals threatened with a deepfake pornographic image are deterred from voting, but it is unclear how many people were affected and therefore how it may have affected the outcome • *Scenario 7*. A fabricated exchange between campaign staff and a debate moderator suggests a candidate cheated in a public debate, calling into question the legitimacy of the political process.

elections. For this reason, we discuss amplification within each category of harm rather than as a separate category of harm. Nevertheless, amplification (and platforms' facilitation of it) is a key reason that deepfakes have drawn public attention and one of the reasons they have the potential to erode the integrity of elections.

Harms to viewers/listeners

Deception

Deception is the most dominant harm that cuts across the scenarios.[1] Many of the deepfakes depicted in our eight scenarios are used to induce voters to believe something that is not true about a candidate. For example, in scenario 2, the deepfake falsely documents a candidate saying disparaging and hateful things, and in scenario 3, a video depicts testimonials falsely attesting to seeing a candidate engage in unseemly behavior. However, deception need not be only about the candidate. Scenario 4, for instance, involves deception about polling locations conveyed by a trusted leader, and in scenario 6, a compromising deepfake targeting an individual voter involves the threat of deception about that voter's behavior to friends, family, and the public.

Importantly, in most of the scenarios, the deception is unambiguously intentional. This is because producing a deepfake is an intentional act. The role of intent in deception is important here because it grounds the distinction between misinformation (no intention to deceive) and disinformation (intention to deceive) (Wardle and Derakhshan, 2017). The deepfakes used in scenarios 2 through 7 involve disinformation, since they were intentionally produced to deceive voters or voters' family and friends (as in scenario 6). In scenario 1, the authors of the synthesized depiction of a battle scene might have believed that they were accurately recreating what happened, so the synthesized video cannot simply be considered disinformation. Moreover, insofar as the video is an accurate recreation, it may not qualify as misinformation. The problem with the video in scenario 1 is that viewers were not informed that some of the footage was synthesized so they might presume the video consisted entirely of real footage. Viewers may have a different stance toward depictions once they are recognized as interpreted representations. Intent also plays a role when it comes to deepfake parodies, as illustrated in scenario 8. Parodies are not intended to deceive; those who create them expect viewers to share the joke (Tandoc et al., 2018). In scenario 8, however, deception occurs because a deepfake parody is moved from its original context to another context, that is, social media platforms, where there is no indication that the video is a parody. Consequently, some viewers take what they see to be real. In this situation, whether or not the deepfake could be considered misinformation or disinformation would depend on the intentions of the person who moved the deepfake into a new context without labeling. If the sharer did not intend deception, but a viewer was nonetheless deceived this would be misinformation, whereas if the sharer intended viewers to be deceived, this would be an instance of disinformation.

Deception is a profound harm to individuals because it impedes their ability to make informed decisions in their own best interests. Intentionally distributing false information about a candidate in an election manipulates voters into serving the interests of the deceiver. False information about candidates may distort a voter's reasoning about how to vote. Even if a person would have chosen the candidate regardless of the deception, the voter's autonomy has been impeded in the sense that the voter has been disabled from making a decision on the basis of accurate information. Although this is true for any false information distributed during an election, deepfakes are unusual insofar as they increase the ostensible authenticity and credence of their misrepresentations.

A single act of deception may be harmful to a single individual, but when deepfakes are widely distributed, harm is amplified, that is, many more voters are deceived. Obviously, the difference between one-person to one-person deception and amplified deception is the number of people harmed. Both the breadth and speed of amplification facilitated by digital and social channels make it more difficult to counteract the harm.

Intimidation

To intimidate is to make a person fearful (of injury or harm); intimidation involves a threat that generates fear. The use of deepfakes to intimidate people has already emerged as an ethical issue related to deepfake pornography (Cook, 2019). Indeed, as mentioned earlier, deepfakes initially gained notoriety for their use in creating pornography. Previous research has found that they are overwhelmingly gendered and sexualized in their current usage, and that much of the real harm already done by the technology has been directed toward women (Ajder et al., 2019; Paris and Donovan, 2019).

Scenario 6 illustrates how in the election context deepfakes can be used to threaten targeted voters into *not* voting. The threat is to release a defamatory pornographic deepfake of the voter if they vote. Here, the intent is to instill fear of the possibility that family, friends, and others will see and believe the deepfake. Of course, even if those who were to see the distributed deepfake would not believe it, the voter would still likely experience a loss of dignity and a feeling of humiliation or shame. The more broadly the deepfake spread, the more humiliation. Intimidation is harmful in itself and in the election context, if the intimidation works, it deprives voters of a fundamental right, the right to vote.

Amplification is limited in scenario 6 both because the distribution medium is text messaging and because the deepfakes are personalized. However, the technical capability to personalize deepfakes may become easier in the future, and this could lead to a different kind of amplification. That is, in addition to amplification by means of a single deepfake being spread broadly through public or social media (i.e. one-to-many amplification), the ease of personalized deepfakes could lead to a profusion of personalized deepfakes spread through more narrow non-public (e.g. chat) channels (i.e. one-to-few amplification repeated many times).

Harms to subjects

In the election context, the primary targets (and subjects) of deepfakes are campaigns and candidates. However, a deepfake video may also include auxiliary subjects who are not the targets, but nonetheless serve a role and are depicted in the synthesized media. For instance, the soldiers depicted alongside the candidate in scenario 1 are auxiliary subjects that are included to make the reenactment more realistic. Similarly in scenario 7 to suggest that a candidate cheated, the debate moderator (a journalist) and a staff member on a candidate's campaign are auxiliary subjects and made to appear to be perpetrators or at least complicit in the cheating. Auxiliary subjects may be harmed in the same way that primary subjects are. The harms here are harms to the subject's reputation and misattribution.

Reputational harm

In scenario 6, deepfakes are used to intimidate *voters* by threatening harm to their reputations. In four of the other scenarios, reputational harm is also at issue but the harm is to *candidates*. In scenarios 2 and 5, videos depict a candidate saying something that suggests the candidate is not who she or he purports to be; in scenario 3, people testify that a candidate behaved in a way that makes the candidate seem hypocritical and an unsavory person; and in scenario 7, a fabricated incident suggests that a candidate cheated in a public debate. In these scenarios, the subject of the deepfake is not threatened with release of an unattractive video, the video or audio is simply released. The harm is not a threat, the harm is done through the act of publication.

Legally, deepfakes of this kind could fall under the concept of defamation (the act of damaging a person's good reputation); however, the courts have been reluctant to interfere with campaign speech. Courts in the United States have been divided on the constitutionality of laws that prohibit false campaign speech (Hasen, 2013; Marshall, 2004). This has rendered the law with regard to campaign speech extremely complex with the result that false speech is generally not disallowed. Yet, reputation is, in an important sense, the commodity of campaigns; campaigns are interested in shaping beliefs about who a candidate is, what the candidate has and has not done, what kind of character the candidate has, and what the candidate is likely to do in the future. Competitors are interested in countering their rival's positive reputation. The reluctance of US law to ban false campaign speech has more to do with the dangers and challenges of trying to *regulate* false claims than denying the harmfulness of false speech. Since our concern here is with the ethical rather than legal implications of deepfakes, the harmfulness of false speech as constituted in deepfakes is incontrovertible.

The amplified consequences of reputational harm result both from the large number of people reached and the speed with which the damage is done, making it difficult to counteract. Deepfakes that are released in the final weeks of an election could be nearly impossible to effectively counter, while deepfakes released early in an election might still be difficult to counteract but at least there would be time to publicize their falsity.

Misattribution

Producing a deepfake involves using a subject's image and/or voice either by manipulating available renderings or by synthesizing entirely new ones. Because of this, deepfakes seem to violate ownership rights of the subjects. That is, independent of the reputational harm that might be done to an individual using their image or likeness in a deepfake, there seems to be an additional harm in using and distributing the individual's image or likeness without permission. Because the law in this area is complex and varies across jurisdictions (Rothman, 2018), it does not provide an incisive approach to this harm. Public figures have publicity rights in their images, but this right is based on a commercial interest, that is, the public figure is protected to be able to reap economic benefits from their image (Berkman, 1976). Although candidates for public office are public figures and therefore have publicity rights in their image, it is unclear how a legal suit claiming harm from the use of a candidate's image in a deepfake would fare in an election context where the harm is not commercial or even pecuniary (Spivak, 2019), and, as already mentioned, courts are reluctant to regulate false campaign speech.

What, then, is the additional harm—beyond reputational harm—to using another person's image or likeness (without their permission) in a deepfake? The harm can, perhaps, best be understood as "persona plagiarism"—an inversion of plagiarism focused on the source rather than the content of a message. Plagiarism is the act of "using or closely imitating the language and thoughts of another author without authorization and the representation of that author's work as one's own."[2] Legal scholars have attempted to define it simply as "non-consensual fraudulent copying" (Posner, 2007). In plagiarism, A presents B's words and takes credit for them, claiming the words to be her own. Plagiarism is misattribution; B has created a text and A misattributes the text to herself. In the act of producing a deepfake, A manipulates B's image and/or voice to create something new and then misattributes it to B. In short, in plagiarism, A takes B's words and misattributes them to A; in deepfakes, A takes B's persona (likeness and voice), inserts A's own expression according to A's goals, and misattributes that expression to B. This can also be thought of in terms of credit. In plagiarism, the original author of words is not given credit for what he or she created; in deepfakes, credit is not given to the creator of the deepfake and is falsely given to the subject of the deepfake. In plagiarism, credit has a positive valence and the real author is denied that. In persona plagiarism, credit with a positive *or negative* valence is misattributed to the subject without their consent. The typical punishment for plagiarism in society involves social sanctions against the plagiarist once their copying is detected (Posner, 2007); however, in the case of deepfakes, the lack of attribution to the creator of the deepfake frustrates a comparable normative sanctioning against that creator.

Misattribution is also at issue with parodies, though in a complicated way. The deepfake parody creator takes a candidate's likeness and voice and inserts his or her own expression into it by attributing exaggerated, unattractive or embarrassing characteristics or behavior to the candidate. Although there is a lack of consent, there is no deceptive attribution of credit, and therefore, there is no persona plagiarism. In other words, as long as viewers understand they are viewing parody there has been no misattribution. However, when the deepfake is moved to another context in which viewers do not recognize it as a parody and believe the attributed characteristics and behavior to be real, the misattribution becomes a form of persona plagiarism. Of course, identifying who committed the plagiarism is problematic because it was not the intent of the parody creator to deceive, and those who moved it to another context may or may not have understood the implications of what they were doing.

To understand the significance of amplification of misattribution, the parallel with plagiarism is helpful again. When a person plagiarizes another's work in a single document only seen by one or a few people (e.g. a student turning in a plagiarized paper to a teacher), the harm is already done. But when a person plagiarizes a document that is read by thousands (e.g. in a widely read newspaper), the harm is greater because the number of people led to misattribute credit is greater. It is the same with deepfakes. A deepfake depicting a candidate doing something he or she did not do shared in a non-public channel with a few individuals is not as harmful as a deepfake spread widely across social media thereby creating much more misattributed negative credit.

Harms to social institutions

Undermining trust in elections

In their broad discussion of deepfakes, Chesney and Citron (2019) list nine different harms to society that could result from deepfakes, including election manipulation. They also mention distortion of democratic discourse and eroding trust in institutions, both of which are highly relevant to democratic elections. Insofar as deepfakes are a form of disinformation, the National Democratic Institute's (NDI) statement on Disinformation and Electoral Integrity unambiguously articulates the threat: "Deliberately blurred lines between truth and fiction amplify voter confusion and devalue fact-based political debate . . . Manipulation of voter and civic information dampens participation and degrades trust in election management bodies."[3]

As already discussed above, the law with regard to campaign speech is not very helpful in addressing the threat of deepfakes because election law is shaped by a compelling concern for the protection of first amendment rights. This is especially so when it comes to deepfake parodies for they are generally seen as a form of free expression. The law favors false campaign speech over violations of free speech for fear that regulating campaign speech would become political (Marshall, 2004; Rowbottom, 2012). Despite the law, however, the harm that deepfakes cause in undermining trust in electoral outcomes stands. Deepfakes along with other kinds of false speech distort campaign results and threaten public trust in those results.

Scenarios 6 and 7 illustrate specific ways in which deepfakes can be used to manipulate voters and dampen participation which in turn generates mistrust in the integrity of electoral processes. In scenario 6, deepfakes are used to deter voters from voting by means of a threat to release deepfaked pornographic images. Even if it was unclear how many people had been deterred, once voters become aware of the tactic, trust in the integrity of election results may be eroded (Daniels, 2009). In scenario 7, a deepfake is used to instill mistrust by falsely suggesting that a candidate cheated in a public debate, thereby calling into question the legitimacy of the political process.

Although these two scenarios illustrate specific ways in which deepfakes can be used to thwart electoral processes, it is also important to note that all of the harms exposed in the other scenarios can also lead voters to mistrust the processes by which officials are elected. For instance, deception undermines individual decision autonomy and when discovered may undermine the trust a voter has in her fellow voters to make well-informed decisions. Similarly, reputational harm and misattribution distort how voters perceive and understand candidates and, even if an individual viewer is aware of the manipulations, he or she may believe that others are not, which could further degrade their trust in the ability of others to make well-informed voting decisions. With new uncertainties injected into the question of whether voters think their peers are well-informed, trust in democratic processes is undermined.

How could the harmful effects of deepfakes be limited? Addressing the ethical challenges

In this section, we consider four intervention strategies that could be used to address the harms identified above: education and media literacy, subject defense, verification, and publicity modulation. Rather than frame these as recommendations, we present them as broad approaches and consider their limitations as well as their benefits. These approaches are discussed with the election context specifically in mind though they are applicable to other contexts in which deepfakes are or will be used. Although these four are not the only possible forms of intervention (e.g. see Ovadya (2019) for specific tool-related suggestions), they are potentially important strategies for mitigating the harmful effects of deepfakes. More research will need to be done to determine the extent to which they are viable and useful for political campaigns and in other domains.

Our analysis of these four intervention strategies is framed in terms of forward-looking responsibility. This involves identifying how responsibilities (duties, obligations) might be assigned to various actors. Although those who create and distribute deepfakes with the intention to deceive, intimidate, damage reputations, or generate distrust in elections have to be held accountable, assignments of blame are often undermined because of lack of clarity about who was supposed to do what. In keeping with the anticipatory stance, our focus is on how responsibility could be ascribed in the future. Forward-looking responsibility is concerned with delegating responsibility to actors in particular roles. This involves locating the actors that are in a position to contribute to preventing harms in the future (Fahlquist, 2017). Once duties are assigned and expectations set, blame and liability are a matter of identifying those who failed to fulfill their responsibilities.

To define appropriate and responsible use of an emerging technology such as synthetic media, either new norms must be created or established norms must be extended to accommodate new technical capabilities (Johnson, 2009). In this case, existing norms around image manipulation as practiced by professional publishers or the wider public might be extended (Paris and Donovan, 2019; Solaroli, 2015). But whether developing new norms or extending existing ones, it is helpful to clarify the delegation of specific responsibilities to actors in distinct roles. This delegation often happens informally through social discourse that expresses and solidifies expectations as well as formally through law or legal precedents and other professional or industry self-regulatory codes. For each of the intervention strategies, we consider how a range of actors, including social media platforms, journalists and news media organizations, technologists, political campaigns, policy makers, and individual citizens could be ascribed responsibility to mitigate the harms of deepfakes.

Education and media literacy

Education and media literacy training have a role to play among the strategies necessary to address deepfakes (Witness & First Draft, 2018). The idea here is that individuals could be encouraged to develop an awareness of the capabilities of the technology so as to be able to spot characteristic flaws in deepfakes, or more generally acquire skills in how to verify, fact check, and do research to assess online sources of information. Such increased literacy would allow individuals to take more responsibility for their consumption and sharing of information (synthesized or not). Even if it does little to address the sources and motivations for the production of deepfakes, by equipping viewers with critical assessment abilities this strategy is compelling because it directly mitigates the potential for deception and furthermore diminishes reputational harm by reducing amplification via social channels.

The challenge of educating the public and promoting media literacy is daunting. However, from the perspective of forward-looking responsibility, there are many actors that could facilitate this including news organizations, media platforms, political campaigns, policy makers, technologists, and other education-oriented organizations. Journalists, for instance, can contribute by raising awareness of the capabilities of deepfake technology and the ethical harms they inflict. Platforms can develop media literacy modules that are interjected into users' experience at opportune moments. Campaigns can contribute by talking about the issues of deception and synthetic media. Policy makers can create or fund programs which help educate the public. And technologists can help educate the public as well as other specialized stakeholders on what to look for to determine if a piece of media is synthesized.

Of course, there may be limits to the effectiveness of this strategy because deepfake technologies are becoming increasingly more realistic. For instance, when it comes to out of context parodies, it may be difficult to train individuals to uniformly identify what might be considered comically absurd. Moreover, an unintended consequence of increased public education about deepfakes is the possibility of what Chesney and Citron (2019) have termed the "liar's dividend." This is the idea that as individuals learn to be more critical and skeptical of media, they may begin to doubt real video and audio evidence. This in turn makes it easier for some to deny and cast doubt on the occurrence of real events (Golingai, 2019). In short, while education serves to mitigate deception directly, it could also generally depress trust in social institutions such as democratic elections. Hence, educational interventions will need to be paired with other interventions discussed below.

Subject defense

The harms to the subjects of deepfakes suggest a suite of interventions enabling subject defense, either *reactively* or *proactively*. These measures apply to primary as well as auxiliary subjects.

In the election context, campaigns could adopt a strategy of monitoring platforms where deepfakes circulate, so that any manipulated representations of the candidate can be quickly debunked. In addition to detection, campaigns could also create crisis response teams that have plans in place for how to react quickly should a deepfake of the candidate be used to attack the candidate's reputation. The plans could include legal response strategies for harms specifically relating to defamation, false light, or right of publicity (Chesney and Citron, 2019; Spivak, 2019).

Of course, each of these reactive approaches has limitations. Not all media platforms can be monitored because deepfakes can circulate in private personal messaging systems, groups, and chat apps that have even been encrypted. The effectiveness of legal responses will be limited because correctly identifying the legally accountable source of the deepfake, and/or reaching perpetrators outside the jurisdiction of US legal processes is challenging (Chesney and Citron, 2019). And, of course, if any of the identified deepfakes have been created as parody, they may be protected as free speech. The problem with reactive defense interventions in general is that they do not stop the deepfake from being seen; they either counter the disinformation with accurate information or hinder its spread. Many of the harms are still done as in the case of intimidation and undermining of trust; others are only lessened as with deception and misattribution because all viewers cannot be reached. Reactive strategies also cannot operate at a speed commensurate with the spread of media on many platforms.

In terms of proactive defense, options are more limited. Since deepfakes use audiovisual material as training data, campaigns could try to restrict the audiovisual material of a candidate in a way that would prevent the creation of deepfakes. However, this would work against freedom of expression, would limit a campaign's capacity to reach voters and, in any case, the technology is moving too quickly toward realistic synthesis with only small amounts of training data. Another proactive strategy would be for candidates to employ immutable life logs—verifiable recordings documenting their every utterance—which could be used to authoritatively debunk any potential deepfake (Chesney and Citron, 2019). An even more futuristic scenario would entail candidates using facial or vocal digital implants that imprint a verifiable encrypted signal in any audiovisual recording of the individual. Audiovisual recordings that did not contain the watermark signal could then be identified as fakes. The advantage of these proactive measures is that they address harms both to subjects and to viewers. But at the same time, proposals such as immutable life logs or implants create new harm to subjects by requiring them to relinquish some degree of privacy, that is, with respect to bodily intrusions from implants or as a result of tracking their behavior with life logs.

Although subject defense primarily assigns responsibility to subjects, other actors are in a position to help. Policy makers can craft legal protections that buttress legal actions by campaigns while being careful not to chill free expression (Hasen, 2019). Policy makers might also create and fund programs that support technologists in developing technologies to support proactive defenses.

Although campaigns may be motivated to monitor for deepfakes of a candidate, their ability to do so will be constrained by limited access to data from the major platforms. Platforms are in a better position to enable effective and timely monitoring of candidate personas, which provides a strong argument for ascribing responsibility to them. Their responsibility might take the form of having individuals enroll with the platform (e.g. through the disclosure of a photo, video, or voice sample) to be alerted when synthesized media of the individual is detected. The role of the enrolled individual in this arrangement would then be to monitor, evaluate, and react to alerts from the platform. Alternatively, an intermediary monitoring entity might bring together these alerts across different platforms as well as monitor the broader Internet for deepfakes of the enrolled individual.

Verification

A potentially powerful antidote to deepfakes is the use of verification techniques and technologies that help reveal how the audio and/or video was constructed. These provide evidence about whether a piece of media was synthesized (entirely or in part), and include automated algorithms as well as semi-automated forensics procedures for detection and provenance determination. Verification techniques increase the likelihood that fakes will be identified as such and, therefore, minimize harm to viewers. Verification can also provide support to other intervention strategies. For instance, it can counteract erosion of trust and the liar's dividend by supporting established and rigorous verification procedures, such as those practiced by professional news media and fact checking organizations, augment the efficacy of subject defense strategies by helping identify misattribution, and work in the service of legal interventions by tracing provenance and helping attribute media to an accountable entity.

Despite recent advances in automated deepfake detection and verification algorithms, for example, reaching as high as 92% to 98% accuracy when applied to synthesized faces of leading politicians (Agarwal et al., 2019), these strategies are nonetheless limited in scope. For instance, they can have degraded performance with reduced encoding quality and exhibit statistical uncertainties that confound clear conclusions (Johnston and Elyan, 2019; Rössler et al., 2019). According to Farid (2019), "No technique can prove authenticity . . . authentication is simply a failure to find evidence of tampering."

Yet another problem with automated verification is that it can introduce new harms having to do with curtailment of free speech. Automated techniques for detecting and verifying media at scale introduce probabilistic uncertainties (Ananny, 2019), which means some false positives, that is, some media will be detected as fake when it is not. Since automated techniques are unable to detect the intent of a deepfake's author, parody and satire may be weeded out along with deceptive content. Hence, although verification techniques can help reduce some of the harms of deepfakes, they will need to be paired with expert human moderators capable of assessing intent and appropriateness (Paris and Donovan, 2019). In general, verification interventions will benefit from humans in-the-loop to interpret forensics evidence, reach sound conclusions, and flexibly cope with an ever evolving range of deceptive practices and deepfake technologies.

Publishers of media will need to take some responsibility for media verification, be the publisher a social media platform, a phone chat app, or a news organization. This is not a matter of blameworthiness but rather because publishers are in the best position to stop the negative effects of deepfakes. Any harmful effects of deepfakes can be largely subverted if they are simply not published. Nevertheless, publishers need not bear this responsibility for verification on their own. Their effectiveness can be supported by campaigns monitoring media for deepfakes of their candidates, news media investing in debunking and fact checking initiatives, platforms limiting the spread of deepfakes, and individuals becoming media literate. Technologists, especially those who are building, and therefore most familiar with, media synthesis techniques, can take responsibility for developing new and better automated detection algorithms and semi-automated verification tools that are understandable and easy to use by these various stakeholders (Partnership on AI, 2020). Policy makers can contribute by funding research and development on such forensics techniques.

Publicity modulation

With the advent of the Internet and social media, everyone is a publisher with the potential to add publicity (i.e. to amplify public attention) to a piece of media via any number of digital channels. As noted previously, publicity is a key driver of both subject and viewer ethical harms. A potential intervention strategy is, therefore, to modulate the publicity that deepfakes receive. An extreme version of this would be to *ban* all deepfakes. A softer approach would be to *throttle* the degree of publicity a deepfake can receive by strategically moderating a deepfake's amplification so as to diminish its potential harm (Donovan and boyd, 2019). Publicity can also be affected through *counterspeech*, including disclosure notices or labels (e.g. indicating nature and veracity) and separate debunks.

Whether the intervention is banning, throttling, or counterspeech, the question arises as to whether paid publicity, such as advertising using deepfakes, should be treated differently. This in turn points to a broader question: Should all deepfakes have their publicity modulated, or only those that meet particular criteria, such as those that are unlabeled or which have been debunked? Answering these questions to arrive at a publicity policy for deepfakes involves two non-trivial steps: defining what types of deepfakes should be subject to moderated publicity and then identifying which pieces of media fit that definition. If automated detection algorithms are not able to accurately identify media according to the definitions, this intervention will require significant human labor and resources.

The three options, banning, throttling, and counterspeech, each address ethical harms somewhat differently. Banning would be problematic because it would threaten free speech. It would, among other things, stifle satire, parody, and entertainment. In comparison, counterspeech would address potential harm to viewers by facilitating the correction of deception, though it has the problem that it still allows individuals to view the original deepfake, thus, risks having influence and causing reputational damage and misattribution, albeit with a smaller number of individuals. Counterspeech could be used at the point of distribution of deepfakes through labeling, or it could involve debunks that are distributed afterwards, though the efficacy of these different approaches vary in their ability to prevent deception. Counterspeech privileges individual autonomy for the viewer in choosing to heed it (or not), but can impinge on the autonomy of creators and distributors in their free speech when it is compelled (Volokh, 2018). Finally, throttling the amplification of a deepfake would address harms to subjects (i.e. reducing breadth of reputational damage), as well as to viewers (i.e. diminishing opportunities for deception). Throttling would minimally affect free speech, since creators are still entitled to make and post deepfakes. Each intervention would have to also be considered in light of the temporal dynamics of publicity. For instance, counterspeech may be appropriate in situations where time is less critical, such as in the earlier phases of an election, when there is adequate time for debunks to circulate and expose individuals to corrective information. On the contrary, throttling may be more attuned to situations of high time pressure right before an election because it directly addresses harms to subjects and viewers and does not rely on the publicity of counterspeech subject to complex algorithmically mediated distribution environments (Napoli, 2019).

As with the other interventions discussed, several types of stakeholders can be assigned responsibility for modulating the publicity of deepfakes. During election periods, social media platforms might configure distribution algorithms to throttle the spread of synthetic media intended to deceive. News media might limit publicity of deepfakes and only publish those that are being clearly debunked. Individuals could curtail their amplification of deepfakes and campaigns could pledge not to amplify them (or create them for that matter).[4] Policy makers, such as those in California, have already begun to regulate the specifics of publishing and labeling deceptive audiovisual media of candidates during election periods. The forward-looking responsibility for modulating publicity could be shared, but would be most effective if the publishers that produce the largest amount of publicity had the largest responsibility, for example, the social media platforms and search engines that drive the majority of online attention.

Conclusion

In this article, we applied an anticipatory ethics approach to the study of deepfake technology. To do so, we systematically developed eight hypothetical scenarios describing how deepfakes could be used in the 2020 US presidential elections and used those scenarios as material from which to identify ethical harms and think through how negative impacts could be minimized. We first identified how deepfakes can result in harms to viewers, subjects, and electoral integrity including deception, intimidation, reputational harm, misattribution, and undermining of trust in electoral processes. Although, as mentioned earlier, by far the most prominent use of deepfakes today has been to create pornography and use it to target women, our analysis makes clear that deepfakes also have the potential to do harm in elections. Our conceptualization of ethical harms provides a contribution to the nascent literature on deepfakes, including the identification of a novel misattribution harm we term "persona plagiarism." We hope this framework will spur future work addressing the social consequences of synthetic media, both in the domain of elections and beyond.

Furthermore, we elaborated four general intervention strategies that could be used to address the harms resulting from deepfakes in elections, including education and media literacy, subject defense, verification, and publicity modulation. Our analysis of the limitations and benefits of these approaches in how they affect the ethical issues can serve to inform future pragmatic work by a variety of stakeholders. Importantly, when it comes to forward-looking responsibility, a variety of actors will need to take on new or modified responsibilities. The analysis makes clear that future work should seek to explore intervention strategies further by advancing technology, fleshing out policy and legal frameworks, and establishing an empirical basis for comparing intervention efficacy.

Finally, our scenario development methodology—grounding scenarios in state-of-the-art technology, emphasizing the criterion of plausibility, and iterating using participatory inputs—provides an example and a model for future research on the ethics of emerging technologies. We hope this methodology will inspire more research which is able to systematically examine the future trajectories of new media technologies.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Norwegian Research Council's SAMANSVAR Program in the context of the VisMedia project at the University of Bergen.

ORCID iD

Nicholas Diakopoulos https://orcid.org/0000-0001-5005-6123

Notes

1. We are using deception to refer to an act or omission that wrongfully causes a false belief in another (Klass, 2018).
2. https://www.dictionary.com/browse/plagiarism?s=t
3. This statement of the National Democratic Institute is found at: https://www.ndi.org/sites/default/files/Disinformation%20and%20Electoral%20Integrity_NDI_External_Updated%20May%202019%20%281%29.pdf
4. One such pledge was drafted by the Association of State Democratic Committees: https://www.ndn.org/blog/2019/06/asdc-resolution-protecting-our-elections-foreign-manipulation

References

Agarwal S, Farid H, Gu Y, et al. (2019) Protecting world leaders against deep fakes. In: *Workshop on media forensics at CVPR*, Long Beach, CA, 23 April.

Ajder H, Patrini G, Cavalli F, et al. (2019) *The State of Deepfakes: Landscape, Threats, and Impact*. Amsterdam: Deeptrace.

Ananny M (2019) *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance*. New York: Knight First Amendment Institute.

Arik S, Chen J, Peng K, et al. (2018) Neural voice cloning with a few samples. In: *Proceedings of NeurIPS*. arXiv preprint arXiv:1802.06006

Bendel O (2019) The synthetization of human voices. *AI & Society* 34: 83–89.

Berkman H (1976) The right of publicity—protection for public figures and celebrities. *Brooklyn Law Review* 42(3): 527–557.

Bregler C, Covell M and Slaney M (1997) Video rewrite: driving visual speech with audio. In: *Proceedings of SIGGRAPH '97*, Los Angeles, CA, 3–8 August.

Brey P (2017) Ethics of emerging technology. In: Hansson SO (ed.) *Methods for the Ethics of Technology: Methods and Approaches*. London; New York: Rowman & Littlefield International, pp. 175–192.

Chan C, Ginosar S, Zhou T, et al. (2018) Everybody dance now. In: *IEEE international conference on computer vision (ICCV)*. Available at: https://arxiv.org/pdf/1808.07371

Chesney R and Citron DK (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review* 107: 1753.

Choi Y, Choi M, Kim M, et al. (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: *Proceedings of the computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, 18–23 June. Available at: https://ieeexplore.ieee.org/document/8579014

Cook J (2019) Here's what it's like to see yourself in a deepfake porn video. *Huffington Post*. Available at: https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0faee4b0a3941861fced

Daniels GR (2009) Voter deception. *Indiana Law Review* 43: 343–388.

Diakopoulos N (2019) *Automating the News: How Algorithms Are Rewriting the Media*. Cambridge, MA: Harvard University Press.

Donovan J and boyd d (2019) Stop the presses? Moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist*. Epub ahead of print 29 September. DOI: 10.1177/0002764219878229.

Fahlquist JN (2017) Responsibility analysis. In: Hansson SO (ed.) *Methods for the Ethics of Technology: Methods and Approaches*. London: Rowman & Littlefield International, pp. 129–142.

Farid H (2019) *Fake Photos*. Cambridge, MA: MIT Press, pp. 129–142.

Fried O, Tewari A, Zollhofer M, et al. (2019) Text-based editing of talking-head video. *ACM Transactions on Graphics* 38(4): 1–14.

Golingai P (2019) Is it Azmin or a deepfake? *The Star Online*. Available at: https://www.thestar.com.my/opinion/columnists/one-mans-meat/2019/06/15/is-it-azmin-or-a-deepfake

Guston DH (2014) Understanding "anticipatory governance." *Social Studies of Science* 44(2): 218–242.

Hasen RL (2013) A constitutional right to lie in campaigns and elections. *Montana Law Review* 74: 53.

Hasen RL (2019) *Deep Fakes, Bots, and Siloed Justices: American Election Law in a Post-Truth World*. St. Louis, MO: St. Louis University Law Review.

Jin Z, Mysore GJ, Diverdi S, et al. (2017) Voco: text-based insertion and replacement in audio narration. *ACM Transactions on Graphics* 36(5): 96.

Johnson DG (2009) *Computer Ethics*. 4th ed. Upper Saddle River, NJ: Prentice Hall.

Johnson DG (2011) Software agents, anticipatory ethics, and accountability. In: *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight*. Dordrecht: Springer, pp. 61–76.

Johnston P and Elyan E (2019) A review of digital video tampering: from simple editing to full synthesis. *Digital Investigation* 29: 67–81.

Karras T, Aila T, Laine S, et al. (2018) Progressive growing of GANs for improved quality, stability, and variation. *Proceedings of the international conference on learning representations (ICLR)*. Available at: https://arxiv.org/abs/1710.10196

Kim H, Garrido P, Tewari A, et al. (2018) Deep video portraits. *ACM Transactions on Graphics* 37(4): 163–114.

Klass G (2018) The law of deception: a research agenda. *University of Colorado Law Review* 89: 707–740.

Lee T (2019) I created my own deepfake—it took two weeks and cost $552. *Ars Technica*. Available at: https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/

Marshall WP (2004) False campaign speech and the first Amendment. *University of Pennsylvania Law Review* 153: 285–323.

Napoli PM (2019) *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. New York: Columbia University Press.

Nikolova B (2013) The rise and promise of participatory foresight. *European Journal of Futures Research* 2(1): 1–9.

Nirkin Y, Keller Y and Hassner T (2019) FSGAN: subject agnostic face swapping and reenactment. In: *Proceedings of international conference on computer vision (ICCV)*. Available at: https://scholar.google.com/citations?user=PAf4OCkAAAAJ&hl=es#d=gs_md_cita-d&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Des%26user%3DPAf4OCkAAAAJ%26citation_for_view%3DPAf4OCkAAAAJ%3A9yKSN-GCB0IC%26tzom%3D-330

Ovadya A (2019) Making deepfake tools doesn't have to be irresponsible. Here's how. *MIT Technology Review*. Available at: https://www.technologyreview.com/2019/12/12/131605/ethical-deepfake-tools-a-manifesto/

Paris B and Donovan J (2019) *Deepfakes and Cheap Fakes*. New York: Data & Society.

Partnership on AI (2020) A report on the deepfake detection challenge. Available at: https://www.partnershiponai.org/wp-content/uploads/2020/03/671004_Format-Report-for-PDF_031120-1.pdf

Posner RA (2007) *The Little Book of Plagiarism*. New York: Pantheon Books.

Ramírez R and Selin C (2014) Plausibility and probability in scenario planning. *Foresight* 16(1): 54–74.

Rössler A, Cozzolino D, Verdoliva L, et al. (2019) FaceForensics++: learning to detect manipulated facial images. Available at: arXiv.org.

Rothman JE (2018) *The Right of Publicity: Privacy Reimagined for a Public World*. Cambridge, MA: Harvard University Press.

Rowbottom J (2012) Lies, manipulation and elections —controlling false campaign statements. *Oxford Journal of Legal Studies* 32(3): 508–535.

Sarewitz D (2011) Anticipatory governance of emerging technologies. In: Marchant GE, Allenby B and Herkert JR (eds) *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight, the International Library of Ethics, Law and Technology*. Dordrecht: Springer, pp. 95–105.

Selin C (2006) Trust and the illusive force of scenarios. *Futures* 38(1): 1–14.

Solaroli M (2015) Toward a new visual culture of the news. *Digital Journalism* 3(4): 513–532.

Spivak R (2019) "Deepfakes": the newest way to commit one of the oldest crimes. *Georgetown Law Technology Review* 3: 339–400.

Suwajanakorn S, Seitz SM and Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics* 36(4): 95–13.

Tandoc EC Jr, Zheng WL and Ling R (2018) Defining 'Fake News': a typology of scholarly definitions. *Digital Journalism* 6(2): 137–153.

Thies J, Zollhöfer M, Stamminger M, et al. (2016) Face2face: real-time face capture and reenactment of RGB videos. In: *Proceeding of 2016 IEEE conference on computer vision and pattern recognition (CVPR)*. Available at: https://ieeexplore.ieee.org/document/7780631

Urueña S (2019) Understanding "plausibility": a relational approach to the anticipatory heuristics of future scenarios. *Futures* 111: 15–25.

Vaccari C and Chadwick A (2020) Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media & Society*. Epub ahead of print 19 February. DOI: 10.1177/2056305120903408.

Van Den Oord A, Dieleman S, Zen H, et al. (2016) WaveNet: a generative model for raw audio. Available at: *arXiv.org*.

Volokh E (2018) The law of compelled speech. *Texas Law Review* 97(2): 355–392.

Wardle C and Derakhshan H (2017) *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Brussels: Council of Europe.

Witness & First Draft (2018) Mal-uses of AI-generated synthetic media and deepfakes: pragmatic solutions discovery convening. Available at: http://witness.mediafire.com/file/q5juw7dc3a2w8p7/Deepfakes_Final.pdf/file

Wright D (2010) A framework for the ethical impact assessment of information technology. *Ethics and Information Technology* 13(3): 199–226.

Yao Y, Viswanath B, Cryan J, et al. (2017) Automated crowdturfing attacks and defenses in online review systems. In: *Proceedings of conference on computer and communications security (CCS)*. Available at: https://dl.acm.org/doi/10.1145/3133956.3133990

Zellers R, Holtzman A, Rashkin H, et al. (2019) Defending against neural fake news. *Advances in Neural Information Processing Systems* 32: 9054–9065.

Author biographies

Nicholas Diakopoulos is an assistant professor in Communication Studies and Computer Science (by courtesy) at Northwestern University where he is director of the Computational Journalism Lab (CJL). He is the author of the book *Automating the News: How Algorithms are Rewriting the Media*, published by Harvard University Press in 2019.

Deborah Johnson is Anne Shirley Carter Olsson professor Emeritus in the Science, Technology and Society Program in the School of Engineering of the University of Virginia. Her latest book, *Engineering Ethics*, *Contemporary and Enduring Debates*, was just published by Yale University Press, 2020.

Appendix A: Scenarios and Ethical Reflections

In this appendix section we present the final eight scenarios that emerged from our scenario writing process, as well as offering a short ethical reflection on each.

Scenario 1

A small veterans' organization would like to see a particular candidate win the democratic primary because although he is not the only candidate with military experience, he is the only one with significant combat experience and the only one making veterans issues a central component of his campaign. The group becomes a Political Action Committee (PAC) and raises funds to make a promotional video for the candidate. The video consists of a combination of video clips with voice over that valorize the candidate's bravery. One of the clips is a synthesized depiction of an incident in Iraq when the candidate heroically saved the lives of several members of their platoon. The video is posted on YouTube without any indication that one portion is synthesized. Thousands view the video; hundreds make comments. From the comments, it is apparent that some viewers believe the video is real footage taken by a reporter present at the event. Other comments include complaints from soldiers in the candidate's platoon who claim that the depiction is an exaggeration of what happened.

This scenario illustrates a situation in which a deepfake video is used by a PAC that is motivated to promote a candidate by valorizing their military record. It is used early in the campaign so there is time for reaction. One of the most salient ethical questions posed by this scenario is whether (and ultimately how) the synthesized nature of a video should be disclosed. The scenario is complicated on this matter because only one component of the video is synthesized. Another, perhaps more subtle issue has to do with the extent to which exaggerations of a candidate's record are okay, particularly when the candidate has not consented. Exaggeration and inflation of a candidate's qualifications has always been an issue in elections but deepfakes expand the possibilities for doing this.

Scenario 2

The race is neck and neck with only 3 days before the 2020 general election between Candidate A and Candidate B. The winner may largely come down to who turns out to vote. Candidate A's campaign advisors develop a strategy to get out the vote amongst their base: disaffected white voters. Candidate A's campaign staff synthesize a deepfake video of Candidate B in a supposed close-door meeting with a few select members of the Black Congressional Caucus, and post it on Twitter and YouTube. In the cellphone-quality video Candidate B is heard saying disparaging and hateful things about white men in the US. In a matter of hours, Candidate B as well as other people falsely depicted in the video, publicly proclaim that the video is a fake. CNN and MSNBC quickly spread the word that the video is likely a fake. Nevertheless it spreads virally throughout social media, propelled and further amplified by troll and bot accounts. The video enrages Candidate A's base, many of whom are unaware of the debunks or simply do not care, and spurs them to the polls for record turn-out amongst their party.

This scenario illustrates a situation in which a deepfake is used by campaign staff to hurt an opposing candidate by attributing to the opposing candidate extreme views that will incite the campaign's base. Platforms play different roles here, as distributors and amplifiers not only of the lie, but also of its debunks. This scenario depicts the generic concern about the use of deepfakes, that is, that they will be used by one candidate (and/ or their supporters) to distort and misrepresent a competitor candidate. In the past candidates have mischaracterized their competitors, but deepfakes provide ostensible authenticity and credence to such mischaracterizations. The scenario also points to the challenge of countering the effects of a deepfake deception when it is deployed late in a campaign. If the same deepfake had been posted four months before the election, there would be more time to effectively counter the lie and the revelation of its intentional distortion might have the effect of hurting the perpetrator.

Scenario 3

The general election has come down to Candidate A and Candidate B. While Candidate B has enjoyed strong evangelical support, Candidate A has had difficulty garnering support from evangelicals. He was raised as a Catholic, is now a practicing Episcopalian, and in his campaign, has emphasized his Christian faith. In particular, he has repeatedly mentioned (in support of gay marriage) that his marriage (to a man) has made him a better Christian. Six weeks before the election, the internal polls of Candidate B's party begin to show that a small but not insignificant portion of evangelicals are moving towards Candidate A. And then, ten days before the election, a deepfake video appears featuring several testimonials by men who say that they had sex with Candidate A while he has been married. A Political Action Committee (PAC) made the video and leaked it to a few evangelical groups via posts to their closed members-only Facebook groups using sock-puppet (i.e. fake) accounts. "Wow, look what I found — how can this guy say he's christian?" reads one of the posts. Several religious leaders from Evangelical and other churches denounce Candidate A to their congregations and followers.

As with others, this scenario illustrates how a deepfake might be used to harm a candidate by misrepresenting the candidate. However, in this case the responsible party is unknown, the strategy is to show the candidate to be a hypocrite, and instead of showing the candidate saying or doing something, the deepfake has others making statements as if they were real testimonials about a behavior for which there would be no other witnesses. The distribution of the deepfake is limited to a small number of people (members of a closed group). Hence, the scenario illustrates that deepfakes can have a significant impact even when the scale of distribution is relatively small. Indeed, with this mechanism of distribution, counteracting the effects is further challenged by their lack of public observability.

Scenario 4

Two weeks before the 2020 election, a targeted demographic group of voters throughout Iowa, Michigan, and Wisconsin receive emails spoofed to seemingly come from their local party. The email addresses each recipient by name and tells them the location of their voting place. It also includes a personalized embedded video in which a respected public figure from their demographic group addresses recipients by name and tells them that there have been attempts to undermine the election with false information about polling locations in their state. The respected figure encourages voters to use the information provided in the email when they go to vote. However, the video has been synthesized and the location information is bogus. On election day some voters are confused when they arrive to vote but are turned away because they're at the incorrect location for their voter registration. The source of the deepfake is unknown until after the election when it is traced back to a radical group that would like to see the opposing candidate re-elected.

This scenario is unlike most of the others in that the deepfake is not used to attack or support a particular candidate but to undermine the democratic process by interfering with individuals' autonomy in exercising their right to vote. It therefore indicates how deepfakes can be used to subvert the integrity of elections. Like the other scenarios, use of the deepfake involves deception, but not deception about candidates, rather deception about what an influential and recognized figure is claiming. The deepfake exploits the figure's countenance to bolster trust in deceptive information and it raises complex questions about consent, property rights, and *publicity rights around the use of an individual's facsimile even if the individual is a public figure.*

Scenario 5

Things are heating up and there's a lot of competition from a wide group of contenders for the Democratic Party nomination. Candidate A is campaigning heavily in Iowa and Nevada, two of the earliest primaries to be held, and races they know they need to win. One of Candidate A's competitors in the primary resolves to leak synthesized audio of Candidate A saying disparaging things about Iowan farmers, calling them "hillbillies" and "rednecks" to a small group of mexican-american voters the candidate met with in Nevada at a campaign stop. The audio clip gets uploaded onto Tumblr by an anonymous blogger who claims that he wants the world to know about "the real Candidate A" but is worried about retaliation against himself and his family. The clip is picked up and played on several podcasts with followings on the left and right. Candidate A's campaign is able to produce a complete recording of the purported event in Nevada which proves the candidate never said those things. Major media outlets air debunkings of the clip, but it continues to circulate online and on some smaller podcasts.

In this scenario, an audiofake is anonymously produced and distributed to hurt a candidate by suggesting that the candidate is not who they purport to be, in the sense that they have said demeaning things about groups that they publicly support. The case might be thought of as a simple case of using an audiofake to misrepresent a candidate; however, the scenario involves anonymity and raises a question about when anonymity is legitimate and how anonymous information should be treated. The case might also be seen in a somewhat positive light in that a real audio recording is used to counter the audiofake. However, it is difficult to say that the impact of the audiofake can be effectively counteracted by the real recording. How will listeners know which is real?

Scenario 6

It's a tight race. Turn out in some key swing districts could make all the difference. Two days before the election, a foreign power unleashes a campaign to suppress certain voter demographics in battleground states. They send spoofed text messages with fake synthesized images to targeted individuals, primarily women. The message says that if they vote on Tuesday, the attached image depicting the targeted individual participating in a pornographic video will be released publicly online and sent to their friends and family. These images are synthesized using a database of pornographic scenes with a face swapped using photos of the targeted person scraped from their Facebook page or Instagram account. The still images are convincing enough to intimidate and coerce some targeted individuals into not voting. A few individuals contact the police or their phone service provider, but the spoofed messages continue until election day. Since no one knows how many voters may have been affected, the incident undermines public perception of the legitimacy of the election.

In this scenario, we see a foreign actor interfering in an election campaign in a powerful way. The deepfake is not aimed at hurting or supporting a candidate, rather it is superficially aimed at voter suppression – superficial because the number of voters *impacted would probably be rather small. Yet once the public becomes aware of this activity, the broader effect would be to cast doubt on the legitimacy of the election by suggesting voter suppression while making it difficult to understand its extent. In addition to eroding trust in election outcomes, this scenario shows how manipulated visual images can threaten individuals, and do so in private communication channels away from public observation. Like all the other scenarios, this one involves deception (because a synthesized video falsely represents behavior), but this one also illustrates how deepfakes can be used to coerce and intimidate.*

Scenario 7

It's the morning after the first presidential debate between Candidate A and Candidate B. Candidate A is basking in the media attention as the clear winner of the exchange when CNN suddenly airs an audio clip of the moderator of the debate which purports to be from an exchange the moderator had with one of Candidate A's campaign officials before the debate. The clip was synthesized by the Candidate B campaign to discredit Candidate A's performance. In the audio the moderator is heard asking about whether "your candidate has any questions about the questions I sent yesterday?" Pundits interpret it as evidence of trying to fix the debate by sharing questions with Candidate A's staff beforehand. The moderator, a respected journalist, firmly denies the account, but not before the #RiggedDebates hashtag starts trending. Candidate B amplifies the idea that the debates are rigged to their Twitter following and refuses to participate in subsequent debates which are then cancelled.

This scenario illustrates how faked audio can be used simultaneously to hurt a competitor and undermine the integrity of (a component of) the election process, in this case a public debate. In this case, the attack was initiated by campaign staff suggesting corruption in the organization of the debate and participation in that corruption on the part of an opposing candidate. Historically, accusations of corruption have been used in campaigns and, to be credible, accusers have had to produce some sort of evidence to support their accusation. Deepfakes of the kind described in this scenario enable accusers to fabricate evidence that looks credible and and is widely distributed and amplified through social media. This expands the power of a false accusation both by making it seemingly real and by the quick and wide distribution that it has.

Scenario 8

As part of its coverage of the primaries, a satirical news show develops a website where viewers can interactively create synthesized videos of the various candidates. The users can act like a puppet-master with their web camera, making gestures and mouth movements that are mapped onto a candidate in a synthesized video. If the user records their voice, it is also dubbed and lip synced onto the candidate. Several users create humorous parody videos of candidates, exaggerating characteristic facial expressions and gestures while adding passable voice impersonations. Some of the caricatures reveal genuine concerns and reflect critical commentaries of the candidates, while others more vacuously mock or ridicule. Users download the videos and re-upload them to social media platforms such as YouTube, Facebook, and Reddit where they begin circulating. While most viewers (e.g.~90%) can easily recognize the videos as parodies, a few are convincing enough that some people can't tell. In particular, one of the more visually believable videos is an exaggerated portrayal of a female candidate's approach to health policy that goes viral.

Unlike the others, this scenario focuses on the use of deepfakes by politically motivated individuals exercising their free speech rights by creating parodies, which they intend as a form of political commentary and critique. Yet while it is individuals creating the parodies, it is an intermediary (a satirical news show) that enables those individuals by making the technology available and easy to use, and it is another intermediary (a social media platform) that amplifies the audience for the videos. The contribution of the intermediaries complicates attributions of responsibility for any negative effects of the videos. Another ethical question is whether the intermediaries have violated the publicity rights of the candidates by distributing the videos without their consent. The uploading of the parodies onto social media also highlights the importance of context and the misunderstanding that can result when synthesized video moves from one context to another. While a video might have easily been recognized as parody on the site of the satirical news show, when taken out of that context it may become more difficult to recognize. Some people are duped by the parodies, believing them to be genuine reflections of candidate behavior. If most people can tell it's a parody, but a few, 10% in this case, are (unintentionally) misled — does that change the ethical calculus?