A REVIEW OF CURRENT DEEPFAKE DETECTION AND MITIGATION METHODS

THE IMPACT OF DEEPFAKES ON MISINFORMATION AND SOCIETY

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Engineering

By
Angus Chang

November 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Catherine Baritaud, Department of Engineering and Society

Harry Powell, Department of Electrical and Computer Engineering

As the internet transitions into the main source of information for the general public, fake news and misinformation also become easier to spread as a result. The newest threats on this front are deepfake videos. These are the product of artificial intelligence (AI), and can result in "realistic looking and sounding video or audio files of individuals doing or saying things they did not necessarily do or say" (de Ruiter, 2021, p. 2). Their name derives from the phrase "deep learning" which is a type of AI, combined with the word "fake" because its products are all fabricated and not real (Laishram et. al, 2021). This technology poses a serious threat to the legitimacy of content we browse. The negative possibilities posed by these fakes include "identity theft and exploitation, defamation, and manipulation of legal evidence" (Katarya & Lal, 2021, p. 486). Deepfakes could also be used in fabricating the actions of political figures and influencing voter behavior (Diakopoulos & Johnson, 2020). These malicious actions could "threaten the psychological security of any state" (Pantserev, 2020, p. 38) as well as "annihilate any trust in online information" (Etienne, 2021, p. 1).

Technology to detect deepfakes and distinguish real imagery from doctored products must stay ahead of the curve to prevent the onset of the aforementioned scenarios. In order to shine a light on the urgency of the situation, the STS topic will focus on how deepfakes negatively affect society, such as the impedance of judicial proceedings or politically motivated mass-misinformation campaigns, among other threats. The goal is to show a need for accessible and efficient deepfake detection strategies, and the tightly coupled technical paper will assist in this aspect. By compiling the latest developments in deepfake detection from a variety of sources, this state-of-the-art report will show how prepared or underprepared the industry is, as well as provide a source for future research to build off of. This work will be completed across

two semesters, from fall 2021 through spring 2022, and the anticipated schedule of deliverables is displayed in Figure 1.

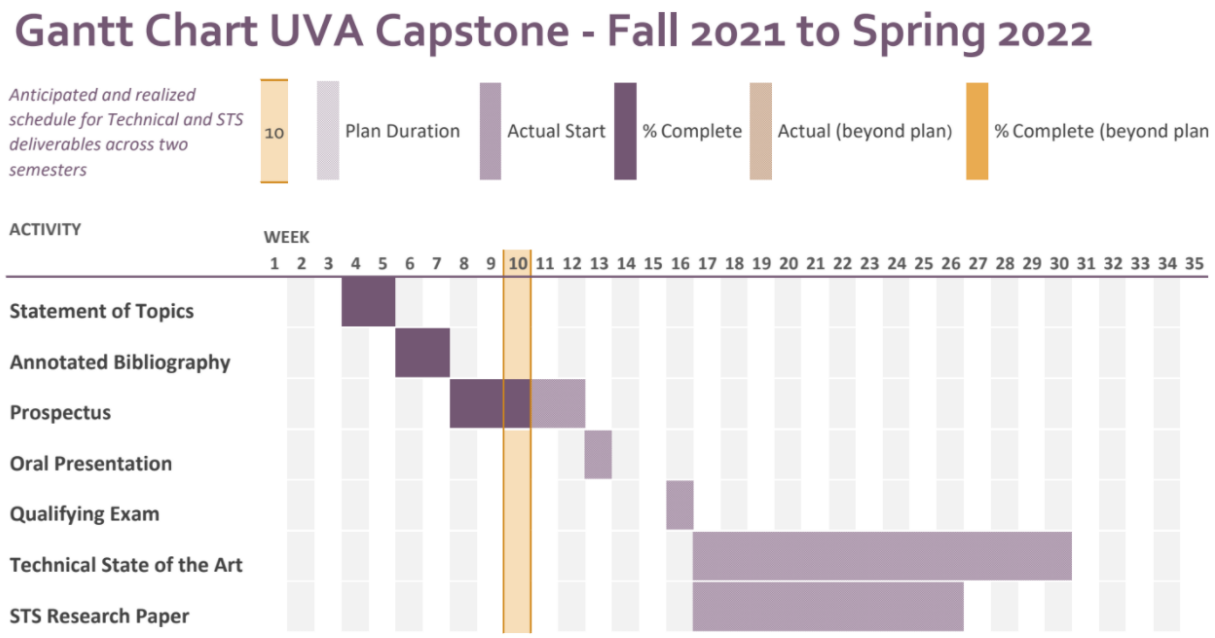

Figure 1: Gantt Chart UVA Capstone: shows the anticipated and realized schedule for technical and STS deliverables. (Chang, 2021)

STATE OF THE ART IN DEEPFAKE DETECTION

In anticipation of the dangers that deepfakes will pose to our society, many different routes for detecting them and/or lessening their negative impacts have been explored. Researchers have found success in analyzing the residual noise of a deepfake, which differs from that of a normal video as a result of AI manipulation (El Rai et. al, 2020). Wang et. al (2016) analyzed inconsistencies of eye blinking in deepfakes and distinguished fakes with 96.6% accuracy (as cited in Katarya & Lal, 2021, p. 488). Zhao et. al (2019) looked at changes in facial expressions across frames, since the deepfake generation process manipulates frames individually and doesn't accurately reflect gradual changes. With a less technical aspect in mind,

Ahmed et. al (2021) found success with simple human detection of deepfakes after exposing them to many examples of said fakes.

Without a doubt, there are a host of options available right now; however, there is no definitive path going forward yet. Each piece of research is isolated and not much has been put into practice yet. At the same time, deepfake technology is becoming more accurate; they will continue to evolve (Katarya & Lal, 2021, p. 486). Accessibility is improving too, as a number of apps already provide the opportunity for the general public to create low-level deepfakes (Fowler, 2021). If deepfake detection technology is to stay ahead of the curve, effective solutions need to be agreed upon and adopted sooner rather than later. This state-of-the-art report will lay out these methods to create one source for ease of future reference. Unique approaches across different categories that are detailed below will be examined, and then their respective methodologies and findings will be presented.

METHOD CLASSIFICATIONS

Attempts at dealing with deepfakes generally fall under two broad categories: detection and awareness. Both will be analyzed in the state-of-the-art report. Detection is just as it sounds - any process that can be used to reliably identify whether or not a video is fake and AI-generated. The more technical approaches usually use comparisons between real and faked videos in order to learn the inconsistencies. Because of this, a large collection of example videos, both real and fake, are needed to train the models (Deshmukh & Wankhade, 2020). Within detection itself, there are three main subcategories that describe what each method analyzes or trains as shown in Figure 2 on page 4: visual, temporal, and human-based.

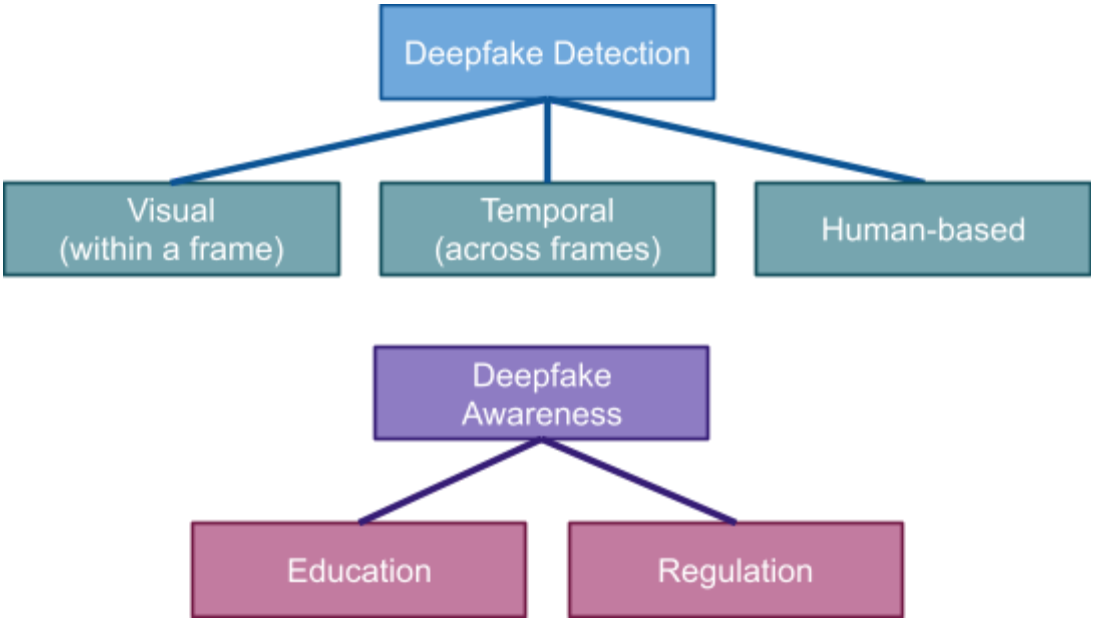

Figure 2: Method classifications: shows the two broad categories to fight deepfakes - detection and awareness - as well as their most relevant subcategories. (Chang, 2021)

Visual and temporal detection are the more algorithm-focused methods. A visual method analyzes frames independently to look for signs of manipulation, while temporal methods look for inconsistencies in how objects shift across frames. The third category is specifically for human-based methods, such as what El Rai et. al (2020) explored, with the training of the human eye to detect deepfakes. These are worth examining because they are much easier to implement, since the main component - humans - are already available. There is no need for any additional hardware or software.

An additional category of deepfake awareness also has relevance in terms of dealing with this threat. It would be difficult to research deepfake detection techniques if not enough people even knew about deepfakes in the first place, hence the importance of awareness. A survey conducted by Chi et. al (2020) on students studying cybersecurity showed that 52% of them had not even heard of deepfake technology before. There is also the legal route of regulation regarding the production and distribution of deepfakes. Farid (2021) argues that “the burden here has to be on the companies and our government and our regulators” (as cited in Fowler, 2021,

para. 33). Due to the non-technical nature of this category, the state-of-the-art report will focus on the detection section and its three subsets only.

THE DETRIMENTAL EFFECTS OF DEEPFAKES

There exists a wide range of usability with deepfakes. The most prominent examples involve faking the speech or behavior of real people; however, it is one thing to see “de-aged actors with million-dollar digital faces" in the entertainment industry (Bode et. al, 2021, p. 849). The darker side of misinformation caused by deepfakes is a whole different story.

A beneficial side to deepfakes does exist. In the field of education, Chesney and Citron (2018) mention them being used to “manufacture videos of historical figures speaking directly to students” (p. 1769), and in the entertainment industry deepfake technology has already been applied to “use images of actors who have died to make new films or improve scenes of low quality” (Pantserev, 2020, p. 51). Unfortunately, these pale in comparison to the long list of malicious uses. There are two broad categorizations of such harms: the more specifically targeted issues that affect individuals, and then the wider issues that will affect larger groups, such as a region, nation, or society as a whole. These will be explored using the framework of the Social Construction of Technology (SCOT), first introduced by Pinch & Bijker (1984), to illustrate the need for robust deepfake detection and regulation.

DETRIMENTS TO INDIVIDUALS

Against individuals, deepfake technology can be used for “stealing people’s identities to extract financial or some other benefit” (Chesney & Citron, 2018, p. 1772). There are already documented cases of deepfakes “being weaponized, particularly against women, to create humiliating, nonconsensual fake pornography” (Fowler, 2021, para. 12). Aside from the direct

psychological damage that these examples can inflict on victims, there is the reputational sabotage that comes with it (Chesney & Citron, 2018).

There is an inherent power gap between someone that can generate and use deepfakes, versus someone who is on the receiving end of the effects. The Technology and Social Relationships model is a form of the SCOT framework which focuses on how the user of a technology interacts with other parties as a result. The user and those that have a relationship with them can benefit or lose. This is illustrated in Figure 3.

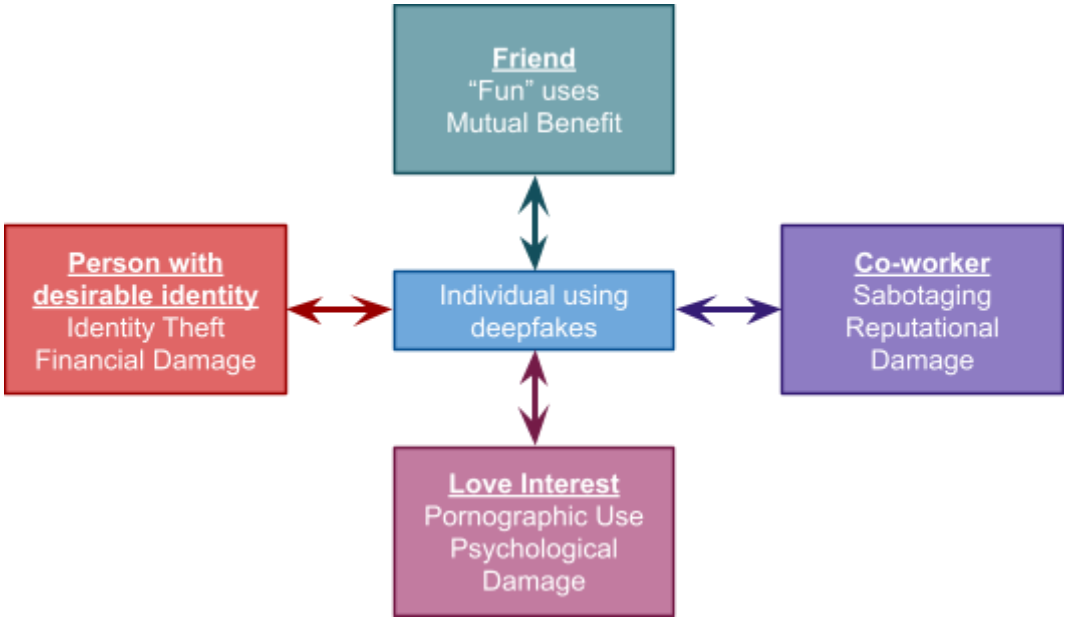

Figure 3: Deepfakes Technology and Social Relationships model: shows how an individual uses deepfakes and the exchanges between them and other individuals. (Adapted by Chang (2021) from Carlson, 2009)

Yes, there are harmless "fun" uses that an individual could do with deepfakes. In this vein, the relationship between the individual and their friend can be beneficial as both get some entertainment value out of it and nobody is seriously harmed. The same cannot be said for the other three potential uses. Starting from the right side in Figure 3, the co-worker section: as mentioned above, individuals could use deepfakes to show someone performing unacceptable actions which could damage their reputation. On the bottom, there is the case of fake

pornography which can have serious psychological effects on the victim of the fake if they were ever to find out that their likeness was being used in explicit ways. With the left side comes identity theft, where the individual committing the act stands to have financial gains from doing so. If the victim's face can be transposed seamlessly onto the user, they could successfully pose as said victim.

The importance of this model comes from asking the question of "who wins, and who loses?" Most technologies, when studied under this framework, will show potential benefit to all parties. In this case, deepfakes are incredibly one-sided in terms of benefits. The individual using this technology stands to gain any and all benefits, while the victims who do not hold the technology will not only gain nothing, but in fact take a significant loss to their assets, health, or well-being. For this reason, it is important to highlight how deepfakes concentrate power in the hands of the user, at the expense of anyone else. The STS paper will research these connections more in-depth and tie them together using the aforementioned Technology and Social Relationships model.

DETRIMENTS TO SOCIETY

Even without the widespread use of deepfakes we have already seen fake news being spread on social media such as Facebook. This is often politically motivated, as "...political actors can use illegitimate means such as disinformation to further their goals" (Dobber et. al, 2020, p. 71). Pantserev (2020) warns that the "distribution of fake news represents a real and very serious threat to the psychological security of any country" (p. 39). Faking the actions of political figures to influence voter behavior could have devastating impacts on democratic elections (Diakopoulos & Johnson, 2020). Not only does this undermine the integrity of

elections, but it also destroys the trust of the general public. As Chesney and Citron (2018) lay out:

> Deep fakes will erode trust in a wide range of both public and private institutions and such trust will become harder to maintain. The list of public institutions for which this will matter runs the gamut, including elected officials, appointed officials, judges, juries, legislators, staffers, and agencies. (p. 1779)

Vaccari and Chadwick (2020) build off this by explaining that when people no longer have a concrete source of true information, public discourse becomes meaningless as "citizens struggle to reconcile the human tendency to believe visual content with the need to maintain vigilance against manipulative deepfakes" (p. 9). Any sort of terrorist group could use this to their advantage in order to sow discord and "disturb relations between countries and thereby undermine international stability" (Pantserev, 2020, p. 52).

Similar to the Technology and Social Relationships model used in the "Detriments to Individuals" section above, it is helpful to look at the broader group connections using the SCOT model, detailed in Figure 4 below.

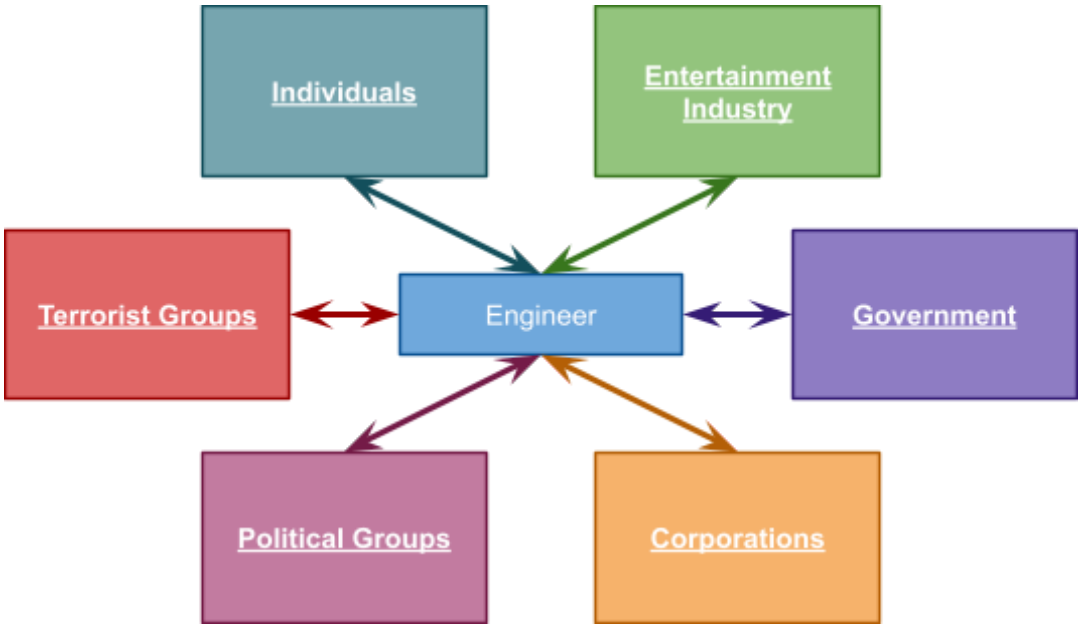

Figure 4: Deepfakes SCOT model: shows some of the main groups that shape how deepfakes are being used and developed. (Adapted by Chang (2021) from Pinch & Bijker, 1984)

The big takeaway from this model is to show how many different groups could have interests in using deepfake technology, and that each of their unique needs will shape how this technology develops. Out of the six examples, only the entertainment industry might have primarily beneficial uses for deepfakes as they won't necessarily come at the expense of anyone else. As mentioned previously, the technology has been used to provide scenes of now-deceased actors (Pantserev, 2020). For the other groups though, they all stand to benefit themselves from overall malicious use of deepfakes. The government and political groups can make use of these fakes in similar ways, to defame their opponents and create propaganda out of misinformation since the average citizen is likely to fall for these (Dobber et. al, 2020). Corporations can play a similar role against their competitors. Terrorist groups, both domestic and foreign, can spread chaos by degrading trust in informational institutions (Chesney & Citron, 2018). In return, the engineer provides improved accuracy of these deepfakes, theoretically to the point where they are no longer detectable by current methods.

Because so many groups may want this technology, it is likely to improve rapidly in terms of effectiveness, while its counterpart of deepfake detection will be left in the dust lacking proper attention. With all of the aforementioned negative impacts in mind, including lack of public trust, rigged elections, and threats to the security of a state, the need for deepfake detection methods to stay ahead of the curve is clear. The STS research paper will highlight how individual users of deepfakes stand to benefit at the expense of others, and how multiple societal groups require rapid development of the technology. If left unchecked without proper deepfake detection in place, the effects on society will be devastating.

REFERENCES

Ahmed, M. F. B., Miah, M. S. U., Bhowmik, A., & Sulaiman, J. B. (Eds.) (2021). *2021 international congress of advanced technology and engineering (ICOTEN)*. IEEE. [Supplemental Material]. https://doi.org/10.1109/ICOTEN52080.2021.9493549

Bode, L., Lees, D., & Golding, D. (2021). The digital face and deepfakes on screen. *Convergence: The International Journal of Research into New Media Technologies*, *27*(4), 849–854. https://doi.org/10.1177/13548565211034044

Chang, A. (2021). *Gantt Chart UVA Capstone*. [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chang, A. (2021). *Method classifications*. [Figure 2]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chang, A. (2021). *Deepfakes technology and social relationships model*. [Figure 3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chang, A. (2021). *Deepfakes SCOT model*. [Figure 4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy. *California Law Review*, *107*(6), 1753–1819. https://doi.org/10.15779/Z38RV0D15J

de Ruiter, A. Philosophy & Technology. (2021). *The distinct wrong of deepfakes*. https://doi.org/10.1007/s13347-021-00459-2

Deshmukh, A., & Wankhade, S. B. (2021). Deepfake detection approaches using deep learning: A systematic review. In V. E. Balas, V. B. Semwal, A. Khandare, & M. Patil (Eds.), *Intelligent Computing and Networking* (pp. 293–302). Springer. https://doi.org/10.1007/978-981-15-7421-4_27

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

El Rai, M. C., Al Ahmad, H., Gouda, O., Jamal, D., Talib, M. A., & Nasir, Q. (Eds.) (2020) *Third international conference on signal processing and information security (ICSPIS)*. IEEE. [Supplemental Material]. https://doi.org/10.1109/ICSPIS51252.2020.9340138

Etienne, H. AI and Ethics. (2021). *The future of online trust (and why deepfake is advancing it)*. https://doi.org/10.1007/s43681-021-00072-1

Fowler, G. A. (2021, March 28). Easy deepfake tech is fun - and unsettling. *The Washington Post*. https://bit.ly/3m7D4jF

Katarya, R., & Lal, A. (Eds.) (2020). *Fourth international conference on IoT in social, mobile, analytics and cloud (I-SMAC)*. IEEE. [Supplemental Material]. https://doi.org/10.1109/I-SMAC49090.2020.9243588

Laishram, L., Rahman, M., & Jung, S. K. (Eds.) (2021). *27th international workshop on frontiers of computer vision*. Springer. [Supplemental Material]. https://doi.org/10.1007/978-3-030-81638-4_11

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Eds.), *Cyber defence in the age of AI, smart societies and augmented humanity* (pp. 37–55). Springer. https://doi.org/10.1007/978-3-030-35746-7

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Vaccari, C., & Chadwick, A. Social Media + Society. (2020). *Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news*. https://doi.org/10.1177/2056305120903408

Zhao, Y., Ge, W., Li, W., Wang, R., Zhao, L., & Ming, J. (Eds.) (2019). *21st international conference on information and communications security*. Springer. [Supplemental Material]. https://doi.org/10.1007/978-3-030-41579-2_37