THE IMPACT OF DEEPFAKES ON MISINFORMATION AND SOCIETY

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Engineering

By

Angus Chang

March 28, 2022

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.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

DEEPFAKES: A NEW TYPE OF CHALLENGE FOR OUR MEDIA

As the internet transitions into the main source of information for the general public, fake news and misinformation 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 fabricated and not real (Laishram et. al, 2021). This technology poses a serious threat to the legitimacy of content that the general public browses regularly.

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). One recent example of this comes in the midst of the conflict in Ukraine. A deepfake of the Ukrainian president Volodymyr Zelenzky telling citizens to "lay down arms and return to [their] families" was circulated on social media (Saxena, 2022, para. 2). Although Zelenzky never said these words, the fake was already viewed by hundreds of thousands of people and only debunked because he had the influence to speak out and rectify the misconceptions. Scholars have predicted a dark future where deepfakes are used at the expense of society. Konstantin Pantserev (2020) warns that these malicious actions could "threaten the psychological security of any state" (p. 38), and Hubert Etienne (2021) echoes that deepfakes have the potential to erode away our trust in online information.

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. Two STS frameworks will be used to illustrate these relationships. The first is Actor-Network Theory (ANT), where Michel Callon, Bruno Latour, and John Law were among the first writers to use this term (Cressman, 2009). This framework works as described - showcasing the complex web of actors that all have a stake in, or are otherwise affected by, some kind of technology. Second is the Social Construction of Technology (SCOT), first introduced by Wiebe E. Bijker (1984). SCOT focuses less on each individual actor and more on the larger groups in society who actively shape how a technology is developed, and how it evolves over time.

The goal of this research is to show a need for accessible and efficient deepfake detection strategies by first exploring how misinformation has already negatively affected society greatly. These are articles and posts that can be fact-checked to verify credibility, and yet millions of people still fall victim to false claims, and misinformation remains a massive threat to our trust in media and society (Gradoń et al., 2021). Deepfakes are currently extremely difficult to identify, and the methods that do exist are not widespread enough, making them inaccessible for the majority of the public. Once these AI-manipulated videos are added to the picture, the issue of misinformation will only grow worse. How should we as a society prepare for the oncoming erosion of credibility? More efficient and accessible deepfake detection methods must be developed. Being able to accurately discern between a real and faked video is a necessity, but this has to be coupled with the ability for the general public to use these tools. Otherwise, they will be just as susceptible to misinformation as right now.

THE EVER-GROWING THREAT OF MISINFORMATION

Using the internet to propagate, misinformation is becoming more widespread and more convincing than ever before. Even though fake news is seen by millions every day, Belluz and Lavis (2022) warn that "platforms, lawmakers and regulators aren't keeping up" (para. 4). Misinformation has made its way into our media while companies are not acting fast enough or efficiently enough to clean up these sources of false claims. In recent years, Facebook has been under fire as misinformation spreads across the platform. Among other conspiracies, Daniel E. Slotnik (2021) warns that those related to the COVID-19 pandemic pose an active health threat as users peddling anti-vaccine theories and unofficial home remedies are able to post unchecked. A whistleblower from Facebook, Frances Haugen (2021), testified that their current efforts in cleaning up fake news are likely to only remove "10 to 20 percent of content" (as cited in Slotnik, 2021, para. 7). Another large platform, Spotify, also faced scrutiny recently for refusing to take down a podcast that contained false claims regarding public health (Belluz & Lavis, 2022).

If our current institutions cannot adequately prevent fake news in its current form, what are the potential effects when deepfakes come into play and make the truth indiscernible? There exists a network of different groups and entities that play some part in the distribution, production, or consumption of misinformation. The Actor-Network Theory framework emphasizes how devastating the effects of deepfakes could be by showing the extent of who can be affected by fake news. Figure 1 on page 4 displays this network of common groups in society who can be considered 'actors' in relation to misinformation. Social media platforms and companies are both displayed as entities in Figure 1, even though they are not specifically human actors. Their importance in this network is that they assist in the *distribution* of misinformation

by providing a platform for malicious actors to spread or sell what they want. A good example of this distribution would be the two aforementioned cases involving Facebook and Spotify, where user-generated content about various topics was left unchecked and caused misconceptions about a variety of issues from health to politics. Even though these claims can be fact-checked and proven false, platforms still lag in their cleanup efforts. Deepfakes, which are already incredibly difficult to discern from real videos, would pose an even bigger challenge for these platforms to overcome. This is the advantage of using ANT to analyze these relationships - the framework takes into account both human and non-human 'actors' in the network, and the interactions mentioned above would not be considered under a traditional human-based analysis.

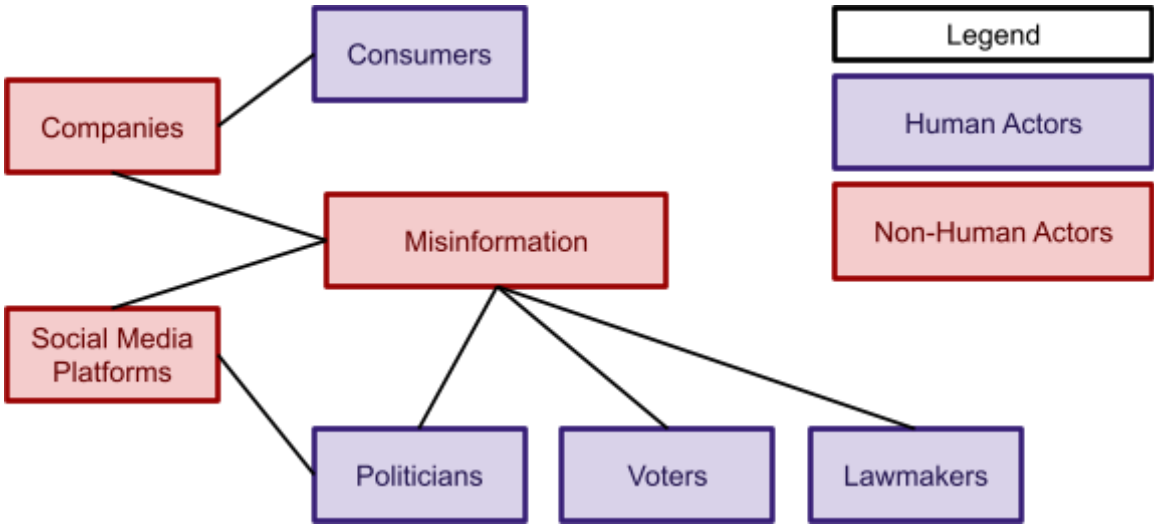

Figure 1: Misinformation ANT model. This figure shows the network of actors who play some part in the distribution, production, or consumption of misinformation. (Adapted by Chang (2021) from Cressman, 2009)

In the political sphere, there have been studies covering the extent of fake news during the 2016 US presidential election. Allcott & Gentzkow (2017) found that the average American citizen would have come across around one to three articles published by known distributors of fake news during the months surrounding the election. Additionally, Facebook estimated there to be over 60 million bots involved in posting political content around the 2016 elections (Lazer et

al., 2019). Misinformation can also come from the top, with political figures themselves distributing misleading facts directly through social media and bypassing traditional media institutions (Dan et al., 2021).

Echo chambers can also be caused by misinformation. Consider how people judge new information: they use “their existing set of beliefs, assessing whether an idea fits previously held notions” (Southwell & Thorson, 2015, p. 590). Essentially, people read what they want to read and will be quicker to reject something that goes against their current set of ideals. As these echo chambers expand, it becomes more and more difficult to debunk these false claims, since more people will be taking into account their peers’ beliefs (Southwell & Thorson, 2015). This will affect both voters and politicians alike.

Election tampering is no longer just a possibility, as these previous real-life examples show. If society is not built to handle plain text misinformation, then deepfakes will slip through the cracks even more easily, as they are made to be believable and undetectable. Given this exposition on the current state of misinformation in society, there is an evident vulnerability that deepfakes are made to exploit. The following sections will introduce the usage of deepfakes and the additional challenges they will create. Their relationships with societal groups, in political areas and beyond, will be covered in order to highlight likely negative outcomes and show a need for deepfake detection and regulation.

THE DETRIMENTAL EFFECTS OF DEEPFAKES

Deepfakes can be used in a variety of ways which are not necessarily all malicious. 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 instills a much scarier reality.

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 the wider issues that will affect larger groups, such as a region, nation, or society as a whole. These two sets of relationships between deepfakes and individuals, as well as deepfakes and larger groups, will be explored using the Social Construction of Technology (SCOT) framework, 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 also the threat of serious reputational sabotage (Chesney & Citron, 2018).

There is an inherent power gap between someone that can generate and use deepfakes and someone who receives 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 (Pinch & Bijker, 1984). This is illustrated in Figure 2 below.

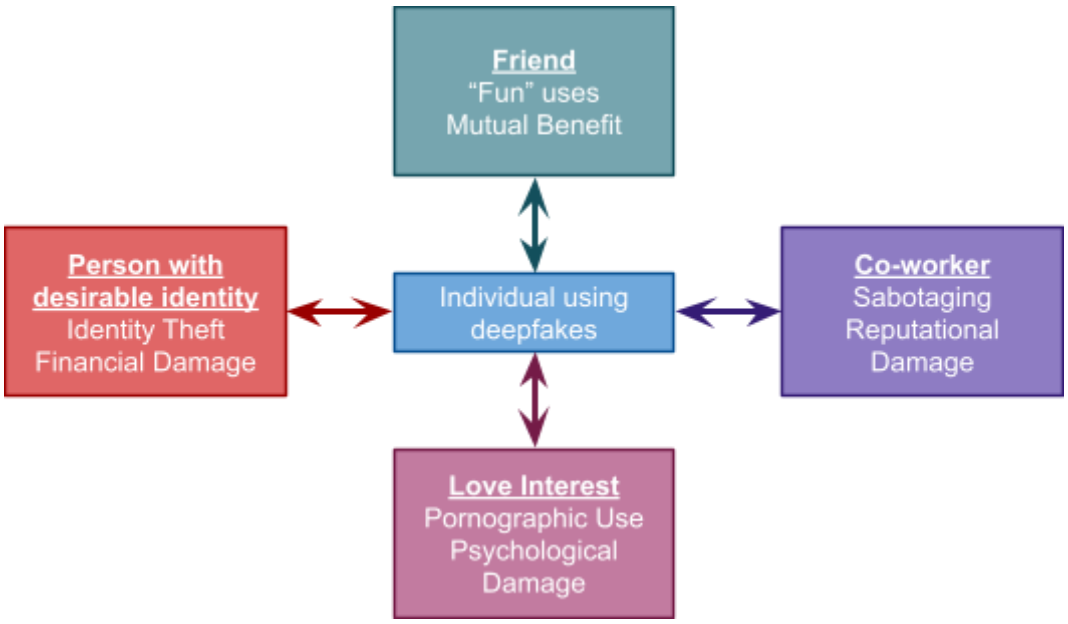

Figure 2: Deepfakes Technology and Social Relationships model. This figure shows how an individual uses deepfakes and the exchanges between them and other individuals. (Adapted by Chang (2021) from Carlson, 2009)

Individuals *can* use deepfakes in harmless ways. In this vein, the relationship between the individual and their friend can be beneficial as both experience the benefits of entertainment value without any serious harm. The same cannot be said for the other three potential uses. Starting from the right side in Figure 2, within 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, faked pornography 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. Finally, the left side shows 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. 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.

DETRIMENTS TO SOCIETY

Even without the widespread use of deepfakes, we have already seen fake news being spread on social media platforms 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 on this idea 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 3 below.

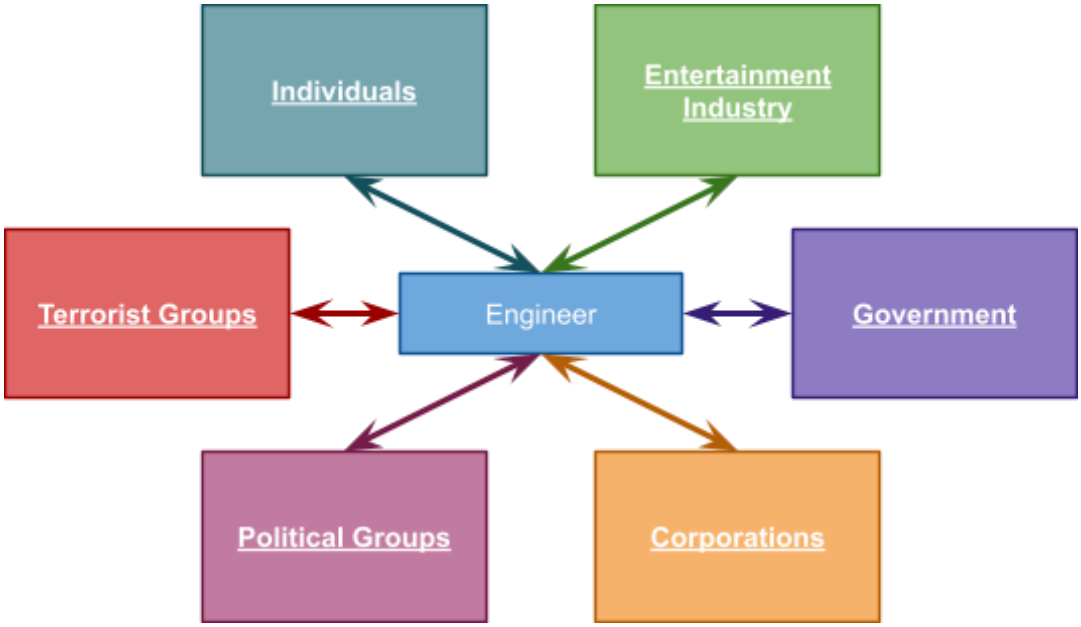

Figure 3: Deepfakes SCOT model. This figure shows some of the main groups that shape how deepfakes are being used and developed. (Adapted by Chang (2021) from Pinch & Bijker, 1984)

This model is used 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 will not 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). On the other hand, the rest of the groups shown in Figure 3 draw on malicious uses of deepfakes in order to benefit themselves at the expense of others. 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.

REGULATION AND DETECTION OF DEEPFAKES FOR THE FUTURE

If left unchecked without proper deepfake detection in place, the effects on society will be devastating. The easiest answer to this issue is building up preventative measures in the form of regulations against how deepfakes can be used, coupled with accurate deepfake detection methods to allow those rules to be enforced. It takes both a technical and societal effort in conjunction to eliminate the threat of deepfakes and misinformation.

There is already promising research surrounding the detection of deepfakes. 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. There are also temporal approaches that look at changes across frames. For example, Zhao et. al (2019) studied changes in facial expressions across frames, since the deepfake generation process

manipulates frames individually and does not accurately reflect gradual changes. Some have used biometric eyebrow matching as well (Nguyen & Derakhshani, 2020).

This is the direction that future technical research should be heading toward in order to properly brace society for deepfakes and lessen their impact once they inevitably become more widespread. If detection methods match the development speed of deepfake improvements, they will be much easier to keep under control. If deepfakes outpace their protective measures, we will witness the annihilation of trust in visual media.

REFERENCES

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

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. (2022). *Misinformation ANT model*. [Figure 1]. *STS Research Paper:* The impact of deepfakes on misinformation and society (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chang, A. (2022). *Deepfakes technology and social relationships model*. [Figure 2]. *STS Research Paper:* The impact of deepfakes on misinformation and society (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Chang, A. (2022). *Deepfakes SCOT model*. [Figure 3]. *STS Research Paper:* The impact of deepfakes on misinformation and society (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Cressman, D. (2009). *A brief overview of actor-network theory: Punctualization, heterogeneous engineering & translation*. ACT Lab/Centre for Policy Research on Science & Technology (CPROST).

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

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual mis- and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 641–664. https://doi.org/10.1177/10776990211035395

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

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

Gradoń, K. T., Hołyst, J. A., Moy, W. R., Sienkiewicz, J., & Suchecki, K. (2021). Countering misinformation: A multidisciplinary approach. *Big Data & Society*, *8*(1), 1–14. https://doi.org/10.1177/20539517211013848

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, A. S., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Nguyen, H. M., & Derakhshani, R. (Eds.) (2020). *2020 international conference of the biometrics special interest group (BIOSIG)*. IEEE. [Supplemental Material]. https://ieeexplore.ieee.org/document/9211068

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

Saxena, A. (2022, March 17). Despicable Zelensky deepfake ordering Ukrainians to 'lay down arms' taken offline. *Express Online*. https://bit.ly/36o0M63

Slotnik, D. E. (2021, October 5). Whistle-blower tells congress that facebook is not able to effectively police anti-vaccine misinformation. *The New York Times*. https://nyti.ms/3Lc8saB

Southwell, B. G., & Thorson, E. A. (2015). The prevalence, consequence, and remedy of misinformation in mass media systems. *Journal of Communication*, *65*(4), 589–595. https://doi.org/10.1111/jcom.12168

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

Wang, M., Guo, L., & Chen, W. Y. (2017). Blink detection using adaboost and contour circle for fatigue recognition. *Computers & Electrical Engineering*, *58*, 502–512. https://doi.org/10.1016/j.compeleceng.2016.09.008

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