Fooled Twice – People Cannot Detect Deepfakes But Think They Can

Abstract

Hyper-realistic manipulations of authentic audio and video content, aka deepfakes, present a new challenge for establishing veracity online. Research on the human impact of video deepfakes, addressing both *behaviors* in response to and cognitive *processing* of deepfakes, remains sparse. In a pre-registered behavioral experiment ($N = 210$), we show that (a) people largely fail to reliably detect deepfakes, and (b) neither raising awareness nor introducing financial incentives improves their detection accuracy. Taking a closer look into the underlying cognitive processes, we find that (c) people are biased towards mistaking deepfakes as authentic videos (rather than vice versa) and (d) overestimate their own detection abilities. Together these results suggest that people adopt a "seeing-is-believing" heuristic for deepfake detection while being overconfident in their (low) detection abilities. The combination renders people particularly susceptible to be influenced by inauthentic deepfake content.

New societal challenges arise from AI-manipulated media, particularly from deepfakes, the hyper-realistic imitation of authentic content (Chesney and Citron 2019). Consider recent high-profile online robbery cases in which scammers successfully used deepfake voice imitation to trick company's employees into wiring money to scammers — the cost of such defrauding schemes amounted to several hundred thousand Dollars (Damiani 2019). Indeed, a consortium of researchers, policymakers, and tech experts ranked the malicious use of deepfakes at the top of the list of AI-threats (Caldwell et al. 2020). Besides imitating voices, deepfakes are primarily used to imitate video content (Güera and Delp 2018). In many cases, the technology serves harmless entertainment purposes, such as Buzzfeed's popular deepfake video that put (curse) words in former president Barack Obama's mouth (Vaccari and Chadwick 2020). However, video deepfakes also have a dark side. For example, recent investigations revealed large-scale use of deepfakes to "undress" women (Hao 2020) and placing them porn videos (Cook 2019), hence turning deepfakes into a powerful weapon to attack people's reputation (Ayyub 2018).

Deepfakes allow manipulating content that was previously out of reach for forgery. Moreover, such tools have become ever more widely available, such as the FakeApp (https://fakeapp.com/) and Faceswap (https://github.com/MarekKowalski/FaceSwap), rendering deepfakes a tool accessible to the masses rather than only selected few experts. These developments raise new societal challenges and research questions (Köbis, Bonnefon, and Rahwan 2021). First, on a behavioral level, can people still reliably spot a deepfake? How to improve their detection accuracy? Does the strategy of appealing to the importance of discerning fake from authentic content, proven to be effective for fake news (Pennycook et al. 2021), also boost deepfake detection performance? Or do financial incentives increase accuracy?

Examining the cognitive processes involved in deepfake detection, do people underestimate or overestimate the occurrence of deepfakes? And, how accurately do they estimate their own abilities to identify deepfakes? In pursuit of first behavioral answers, we conducted a pre-registered, online experiment.

Methods

Subjects and experimental procedure

We recruited 233 participants ($M_{\text{age}} = 35.42$, $SD_{\text{age}} = 11.93$; female = 135) via the online participant platform Prolific. Each participant received 2.5 pounds as a participation fee for the study that, on average, took 18.59 minutes. The Ethics Review Board of *blinded for review* approved the study design. In total, 210 participants completed the study and were included in the analyses.

Videos

After being informed that the probability of each video being a deepfake was 50%, participants watched a video and indicated whether a given video was a deepfake or not. We randomly sampled 16 target videos from the MIT project DetectDeepfake project (www.detectfakes.media.mit.edu). It contains 3,000 of the most difficult videos for AI classifiers to identify as fake or real from Kaggle's DeepFake Detection Challenge (Dolhansky et al. 2020).The length of each video was constant (approx. 10 seconds). Participants could play the video as often as they desired. Each of the videos had an authentic and deepfake version. Each participant saw only one of the two versions during the experiment. We divided them into two sets, each containing eight authentic videos and eight deepfakes, presented in randomized order.

Treatments

Participants were randomly assigned to either the Control (CTRL: N = 74), the Awareness (AW: N = 64) or the Financial Incentive (FI: N = 72) treatment. In the Awareness treatment, participants read a piece by Chesney and Citron (2019) warning about the potentially harmful consequences of deepfakes. Theoretical work dealing with deepfakes suggests that such information about potential consequences of manipulated media raises awareness for the issue and, in turn, motivates people to detect deepfakes (Diakopoulos and Johnson 2019). To ensure that participants read the prompt, they had to answer a multiple-choice question about the text correctly to proceed to the detection task.

In the Financial Incentive treatment, participants received monetary rewards for accuracy as one of the 16 rounds was randomly chosen for payment. Participants received 3 pounds if the guess in the chosen round was correct. Inspiration for this intervention stems from behavioral science research, showing that financial incentives can increase motivation and accuracy (Schlag, Tremewan, and Van der Weele 2015). Hence, the complete design was a 3 (between-subjects: Control vs. Awareness vs. Financial Incentive Treatments) x 2 (within-subjects: Fake videos vs. Authentic videos) design. We pre-registered the hypothesis that accuracy levels in the Awareness treatment and the Financial Incentive treatment exceed accuracy levels in the Control Treatment (see https://aspredicted.org/blind.php?x=tk9qa7)

Confidence

To assess whether people accurately estimate their own detection abilities, we used two measures of confidence. First, after each video, participants rated their subjective confidence of guessing that video correctly on a scale from 50 (=as confident as flipping a coin) to 100 (=100% sure). This measure was not incentivized. Second, participants indicated how many videos they estimated to have guessed correctly after they completed all rounds. This guess was incentivized with 0.5 pounds for the correct answer within a range of ± 1 videos. Importantly, participants did not know that this incentivized measure would follow the detection task to avoid hedging. Namely, when informing participants about this question, they might indicate that they are correct in 50% of the rounds and then flip a coin each round.

We pre-registered the expectation that participants in our study show overconfidence, operationalized as subjective beliefs in one's detection abilities significantly exceeding actual abilities.

Exit Questions

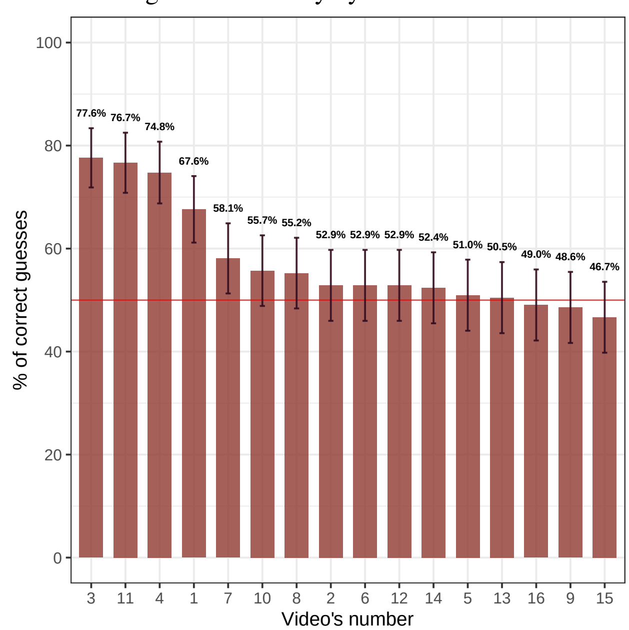
At the end of the study, participants indicated their level of motivation to classify the videos (1 = not motivated at all; 7 = very much motivated). This measure served as a manipulated check to test whether the two treatments (FI & AW) increased motivation. Finally, we assessed standard demographics.

Results

Detection Accuracy

Testing whether people can reliably spot a deepfake, the overall accuracy level of 57,6% exceeds chance levels according to a one-sample t-test ($t(209) = 9.539$, $p < .001$).[1] However, looking at the videos separately reveals that only for five out of the 16 videos, people guess at better than chance levels (see Figure 1).

Figure 1: Accuracy by video number



Notes: Average accuracy levels for each video in descending order. Error bars denote 95% confidence intervals. Reference line indicates the 50% accuracy rate of random guessing.

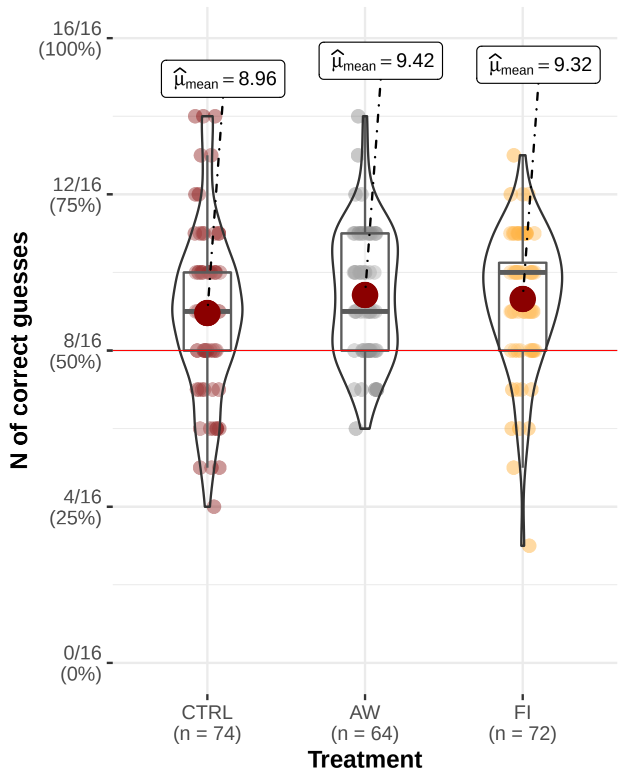
Interventions to Increase Accuracy

Contrary to our hypotheses, neither awareness and financial incentives increases detection accuracy (One-way ANOVA: $F(2, 207) = 0.276$, $p = 0.759$). Figure 2 shows that participants guess on average nine videos correct, independent of the treatment. Analysis of the self-rated motivation levels reveals high motivation levels across all three treatments as more than 75 percent of participants chose 6 or 7 on the seven-point scale anchored in 7 (= very much motivated). In fact, a one-way ANOVA reveals no significant differences in self-rated motivation levels ($F(2, 207) = 0.526$, $p = 0.592$). This finding suggests that the low detection

[1]The t-test uses the subject's average accuracy as independent data points.

rates stem from inability rather than from lack of motivation.

Figure 2: Number of correct guesses



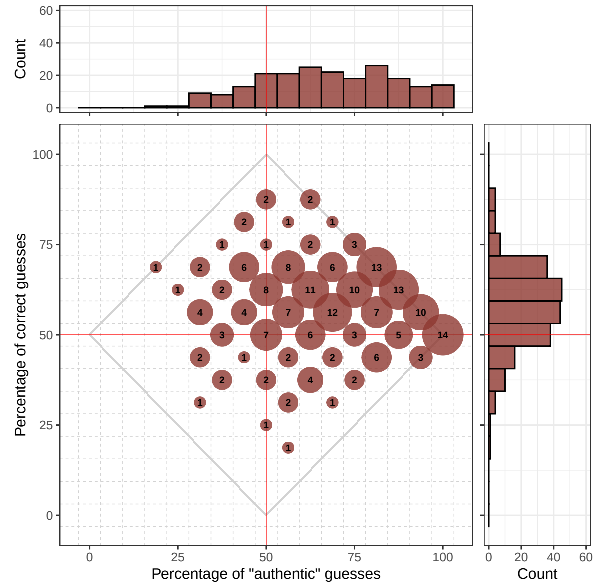
Notes: Violin plots showing the distribution of correct guesses (y axis) by treatments (x axis).CTRL = Control treatment, AW = Awareness treatment, FI = Financial incentives treatment. Black lines represent medians, red dots represent means and boxes indicate 95% confidence intervals. Plots created with the Ggstatsplot package (Patil 2018).

Detection Bias

Moving to the cognitive processes, analyses on guessing patterns reveal a bias towards guessing that the video is authentic. This bias is depicted by the contortion of bubbles towards the right side in the scatter-plot in Figure 3. Although knowing that only half of the videos are authentic, participants guess "authentic" 67,4% of the time which significantly exceeds equal guessing rates (one-sample $t(209) = 13.131$, $p < .001$).[2] The bias towards guessing "authentic" does not significantly differ across treatments (One-way ANOVA: $F(2, 207) = 0.618$, $p = 0.540$) Regression analyses further corroborate that the bias towards guessing "authentic" is robust to the number of views, demographic characteristics, and subjective motivation (see for regression results, Table 1 in the Appendix).

[2]The t-test uses the subject's proportion of "authentic" guesses as independent data points.

Figure 3: Fraction of correct guesses by fraction of "authentic" guesses.



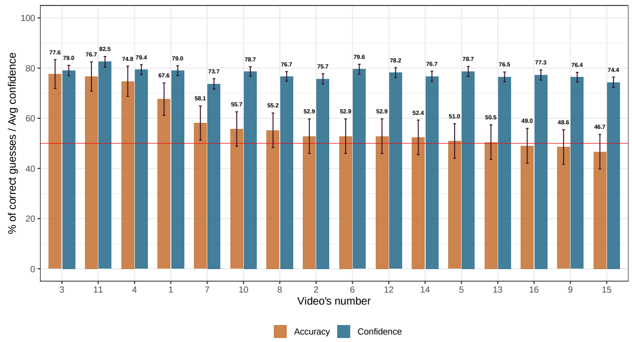
Notes: Distribution of participants according to "authentic" guesses (x axis) and fraction of correct guesses (y axis). Size of bubbles represent the number of participants (exact N within the bubble). The gray diamond delimit the maximum and minimum fraction of correct guesses conditional on the number of times the subject guessed that a video was "authentic". Histograms on each side represent marginal distributions.

Confidence

Comparing people's average accuracy levels to their subjective confidence levels per round reveals that they accurately estimate their detection ability for three (#3,11,4) of the 16 videos (see Figure 4). For the remaining videos, confidence (drastically) exceeds accuracy levels. Results using the incentivized measure of confidence corroborate this indication of overconfidence. Participants overestimate the number of deepfake videos they correctly identified ($t(209) = 2.621$, $p = 0.009$), a pattern that does not differ across treatments (one-way ANOVA: $F(2, 207) = 1.844$, $p = 0.161$). In fact, overconfidence and actual accuracy correlate negatively (Pearson's product-moment correlation: $r(208) = -0.475$, $p < 0.001$) This so-called Dunning-Kruger effect (Kruger and Dunning 1999), indicates that overconfidence is particularly pronounced among those who perform worse. Supporting our prediction, people appear overconfident in their detection abilities.

Exploratory analyses reveal that confidence and the bias towards authentic guesses are associated (product-moment correlation $r(208) = 0.427$, $p < 0.001$, see also Figure 3). The more people are biased towards guessing that a video is authentic, the more they are confident in their guesses (for more details, see Table 2 in the Appendix).

Figure 4: Accuracy and confidence by video number



Notes: Average accuracy levels—i.e., fraction of correct guesses—and average confidence—i.e., average belief about the probability to have guessed correctly the video—by video. Videos are ordered in descending order by level of accuracy. The error bars denote 95% confidence intervals. The red line indicates the 50% accuracy rate of random guessing.

Discussion

Recent developments in General Adversarial Networks (GANs) (Goodfellow et al. 2020) have revolutionized manipulative content generation, enabling the quick and realistic creation of synthetic content. While creating deepfakes is easier than ever, detecting them becomes increasingly difficult. Most research has approached the challenge to distinguish deepfakes and authentic content from a technical perspective by developing AI classifiers (Korshunov and Marcel 2018; Li et al. 2020; Afchar et al. 2018; Li and Lyu 2018; Yang, Li, and Lyu 2019). The current study is among the few that examine the human side of the equation. Results of our behavioral experiment reveal two main insights. First, detection of deepfakes is not a matter of low motivation but inability. People cannot reliably detect deepfakes, even though they are motivated to. Second, people are biased towards guessing that deepfakes are authentic. They are also overconfident in their detection abilities, suggesting that people apply an overly optimistic "seeing=believing" heuristic (Frenda et al. 2013). The combination of these findings puts people at particular risk of being influenced by deepfakes.

Motivated but inaccurate

Humans integrate (moving) visual information more effectively than other types of sensory data (Witten and Knudsen 2005). A recent meta-analysis underlines that visual input like pictures and videos is more convincing than text (Seo 2020). Especially online interactions heavily feature visual content. In fact, by now, videos constitute the vast majority of all consumer internet traffic (Aral 2020). The ability to discern fake from real videos thus marks an essential skill in an increasingly digital world. Extending previous studies on static images (Groh et al. 2019; Rössler et al. 2019) or AI-generated text (Köbis and Mossink 2021), our findings show a nuanced picture about people's ability to discern fake from authentic videos. While overall accuracy rates suggest that people can (still) guess better than chance, a closer look at the different videos indicates that for the vast majority they are not better than random guessing. Overall, we find lower accuracy rates than research on static images (Groh et al. 2019; Rössler et al. 2019).

Research in the adjacent field of misinformation has shown that the belief in fake news often stems from inattention (Pennycook et al. 2021; Bago, Rand, and Pennycook 2020). Similarly, attention and critical thinking have been proposed as a counter-strategy against deepfakes (Nolan and Kimball 2021; Diakopoulos and Johnson 2019). However, the current empirical evidence suggests that such appeals aiming to increase people's attention and awareness of the problem do not suffice to improve people's detection abilities. Nor do financial incentives for accuracy increase performance. A plausible explanation for the flat differences between the treatments lies in ceiling effects. Namely, also participants in the Control treatment were highly motivated to detect deepfakes. Although participants were highly motivated, their detection accuracy hardly exceeded random guessing.

Biased towards authenticity and overconfidence

Our results uncover two interrelated biases in human deepfake detection. First, people's guesses are skewed towards authenticity. Although participants were informed that half of the videos were authentic, they identified 67,4% of videos as authentic. In line with our expectation, confidence levels, independent whether elicited with or without incentives, systematically exceed people's abilities to detect deepfakes. Particularly low performers show overconfidence. This effect is often partially mechanic, in that worse performers have more room to be overconfident (Krajc and Ortmann 2008). Having said that, the evidence that people generally have inflated beliefs about whether they can spot a deepfake is remarkably robust across different measures of confidence.

Taken together, these two biases suggest that people adopt a "seeing = believing" heuristic (Frenda et al. 2013). Namely, people tend to take videos at face value unless they find clear-cut evidence of it being fake (Farid 2019). In doing so, they believe to be better able to spot such deepfakes than they actually are.

Conclusion

Technology is advancing at a pace that makes it hard for people, research and policy, to keep up. This experiment is a testament to this trend, showing that people can no longer reliably detect deepfakes. Some of the previously established strategies against misinformation and manipulation do not hold for the detection of deepfakes. As detecting deepfakes appears less a matter of motivation and attention, deepfakes warrant special attention for digital misinformation research and policy — especially when considering that people appear biased towards believing their eyes and mistake deepfakes as authentic, all while being overconfident in their detection abilities.

Appendix: Additional analyses

Regression Analyses - Detection Accuracy and Authenticity Bias

Estimating the robustness of the effects, Table 1 reports a series of linear probability models investigating the determinants of guessing behavior. Models (1) and (2) look at the determinants of guessing that a video is fake and Models (3) and (4) present regression models predicting the probability to make a correct guess. The regressions match model (2) and (4) drop the videos' fixed effects because of collinearity with the characteristics of the videos.

Table 1: Linear probability models of the likelihood to guess "fake" and to guess correctly.

	Dependent variable:			
	d(guessed fake)		*d*(correct guess)	
	Mod. (1)	Mod. (2)	Mod. (3)	Mod. (4)
d(AW tmt)	0.035	0.036	0.015	0.019
	(0.033)	(0.034)	(0.020)	(0.020)
d(FI tmt)	0.008	0.010	0.008	0.009
	(0.032)	(0.033)	(0.019)	(0.019)
d(video is fake)	0.153^{***}	0.151^{***}		
	(0.016)	(0.016)		
N. of views		-0.003		0.010
		(0.016)		(0.013)
d(verific. q.)		-0.076^{*}		0.002
		(0.040)		(0.039)
Motivation		-0.012		0.014^{*}
		(0.019)		(0.008)
Period N.		0.002		0.001
		(0.002)		(0.002)
Age		-0.002		-0.002^{***}
		(0.001)		(0.001)
d(Female)		0.004		-0.017
		(0.028)		(0.018)
d(Bachelor)		0.014		0.005
		(0.030)		(0.018)
d(Master or PhD)		-0.018		0.033
		(0.047)		(0.027)
Constant	0.236^{***}	0.546^{***}	0.569^{***}	0.639^{***}
	(0.024)	(0.127)	(0.013)	(0.067)
Videos' FE	No	Yes	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
Robust SE	Yes	Yes	Yes	Yes
Observations	3,360	3,360	3,360	3,360
Clusters	210	210	210	210
R^2	0.028^{***}	0.068^{***}	0.000	0.046^{***}

Notes: Robust standard errors clustered at individual level are in parentheses. d for dummy variables. The binary dependent variables "*d*(guessed fake)" and "*d*(correct guess)" are equal to 1 when the subject guesses that the video is fake and when the subject makes a correct guess, respectively. The variables "*d*(AW tmt)" and "*d*(FI tmt)" are equal to 1 in the awareness and the financial incentives treatments, respectively; "*d*(video is fake)" is equal to 1 when the video is a deepfake; "*d*(verific. q.)" is equal to 1 when the subject correctly answered the question about the content of the video; "N. of views" is the number of times the subject clicked to play the video; "Motivation" is the reported level of motivation to detect the videos on a scale 1-7. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

Regression Analyses - Confidence

Models (5) and (6) in Table 2 analyze the determinants of the reported confidence. Looking at Model (6), which includes the treatment dummies and some additional control variables, we see that treatment does not seem to have an effect on the stated confidence. As for the other variables, it seems that: (i) confidence is decreasing in the number of times subjects watch the video, which is not surprising; (ii) more educated subjects have lower confidence levels; and (iii) subjects with higher motivation to detect videos show higher confidence levels.

Table 2: Confidence in the guess made by the subject

	Dependent variable:	
	Confidence	
	Mod. (5)	Mod. (6)
d(wrong guess)	-0.543	-0.127
	(0.441)	(0.451)
d(guessed fake)	-4.076^{***}	-4.543^{***}
	(0.936)	(0.904)
d(wrong guess)$\times d$(guessed fake)	-4.582^{***}	-3.824^{***}
	(0.950)	(0.895)
d(AW tmt)		-1.135
		(1.745)
d(FI tmt)		-0.391
		(1.631)
N. of views		-4.030^{***}
		(0.798)
d(verific. q.)		-0.509
		(1.286)
Motivation		3.157^{***}
		(0.993)
Period N.		0.090^{**}
		(0.037)
Age		-0.036
		(0.061)
d(Female)		-1.297
		(1.365)
d(Bachelor)		-4.502^{***}
		(1.522)
d(Master or PhD)		-3.873^{*}
		(2.332)
Constant	79.786^{***}	72.775^{***}
	(0.815)	(6.492)
Videos' FE	No	Yes
Clustered SE	Yes	Yes
Robust SE	Yes	Yes
Observations	3,360	3,360
Clusters	210	210
R^2	0.044^{***}	0.153^{***}

Notes: Robust standard errors clustered at individual level are in parentheses. d for dummy variables. The dependent variable "Confidence" as the believed probability to have guessed correctly the video from 50% (random) to 100% (guessed for sure). The variables "d(AW tmt)" and "d(FI tmt)" are equal to 1 in the awareness and the financial incentives treatments, respectively; "d(guessed fake)" is equal to 1 when the subject guesses that the video is a fake; "d(wrong guess)" is equal to 1 when the subject's guess is wrong; "d(verific. q.)" is equal to 1 when the subject correctly answered the question about the content of the video; "N. of views" is the number of times the subject clicked to play the video; "Motivation" is the reported level of motivation to detect the videos on a scale 1-7. Significance is coded as follows: *p<0.1; **p<0.05; ***p<0.01.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7.

Aral, S. 2020. *The Hype Machine*. Currency.

Ayyub, R. 2018. I was the victim of a deepfake porn plot intended to silence me. Accessed: 2021-2-23.

Bago, B.; Rand, D. G.; and Pennycook, G. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*.

Caldwell, M.; Andrews, J. T.; Tanay, T.; and Griffin, L. D. 2020. AI-enabled future crime. *Crime Sci.* 9(1):1–13.

Chesney, R., and Citron, D. K. 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. Law Rev.* 107:1753–1820.

Cook, J. 2019. Here's what it's like to see yourself in a deepfake porn video. *Huffington Post*.

Damiani, J. 2019. A voice deepfake was used to scam a CEO out of $243,000. *Forbes Magazine*.

Diakopoulos, N., and Johnson, D. 2019. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society* 1–27.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge dataset.

Farid, H. 2019. *Fake Photos*. MIT Press.

Frenda, S. J.; Knowles, E. D.; Saletan, W.; and Loftus, E. F. 2013. False memories of fabricated political events. *J. Exp. Soc. Psychol.* 49(2):280–286.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Commun. ACM* 63(11):139–144.

Groh, M.; Epstein, Z.; Obradovich, N.; Cebrian, M.; and Rahwan, I. 2019. Human detection of machine manipulated media. *arXiv* 1–21.

Güera, D., and Delp, E. J. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE.

Hao, K. 2020. A deepfake bot is being used to "undress" underage girls. *MIT Technology Review*.

Köbis, N. C., and Mossink, L. D. 2021. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in Human Behavior* 114:106553.

Köbis, N. C.; Bonnefon, J.-F.; and Rahwan, I. 2021. Bad machines corrupt good morals. *Nature Human Behaviour* 1–31.

Korshunov, P., and Marcel, S. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Krajc, M., and Ortmann, A. 2008. Are the unskilled really that unaware? an alternative explanation. *Journal of Economic Psychology* 29(5):724–738.

Kruger, J., and Dunning, D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77(6):1121.

Li, Y., and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207–3216.

Nolan, S. A., and Kimball, M. 2021. Tom cruise, deepfakes, and the need for critical thinking. *Psychology Today*.

Patil, I. 2018. ggstatsplot: 'ggplot2' based plots with statistical details. *CRAN*.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 1–6.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.

Schlag, K. H.; Tremewan, J.; and Van der Weele, J. J. 2015. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics* 18(3):457–490.

Seo, K. 2020. Meta-Analysis on Visual Persuasion– Does Adding Images to Texts Influence Persuasion? *Athens J. Mass Media Commun.* 6:177–190.

Vaccari, C., and Chadwick, A. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Soc. Media Soc.* 6(1).

Witten, I. B., and Knudsen, E. I. 2005. Why seeing is believing: merging auditory and visual worlds. *Neuron* 48(3):489–496.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–8265.