Fooled twice – People cannot detect deepfakes but think they can

Article *in* iScience · October 2021

DOI: 10.1016/j.isci.2021.103364

CITATIONS
41

READS
800

3 authors, including:

Nils Kobis
University of Duisburg-Essen
83 PUBLICATIONS **1,654** CITATIONS
SEE PROFILE

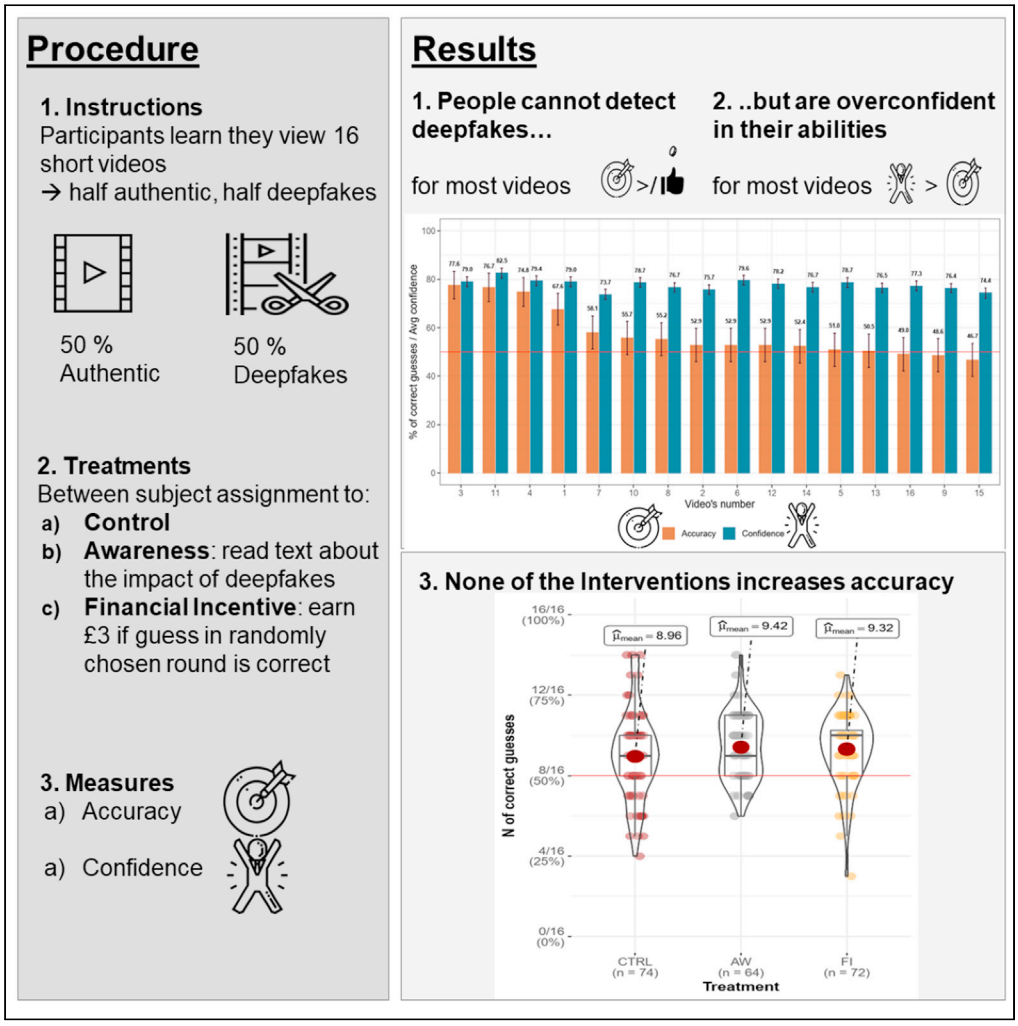
Ivan Soraperra
Max Planck Institute for Human Development
49 PUBLICATIONS **332** CITATIONS
SEE PROFILE

iScience

CellPress
OPEN ACCESS

Article

Fooled twice: People cannot detect deepfakes but think they can



Nils C. Köbis, Barbora Doležalová, Ivan Soraperra

koebis@mpib-berlin.mpg.de

Highlights

People cannot reliably detect deepfakes

Raising awareness and financial incentives do not improve people's detection accuracy

People tend to mistake deepfakes as authentic videos (rather than vice versa)

People overestimate their own detection deepfake abilities

Köbis et al., iScience 24, 103364
November 19, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.103364

iScience

CellPress
OPEN ACCESS

Article

Fooled twice: People cannot detect deepfakes but think they can

Nils C. Köbis,[1,3,*] Barbora Doležalová,[2] and Ivan Soraperra[2]

SUMMARY

Hyper-realistic manipulations of audio-visual content, i.e., deepfakes, present new challenges for establishing the veracity of online content. Research on the human impact of deepfakes remains sparse. In a pre-registered behavioral experiment (*N* = 210), we show that (1) people cannot reliably detect deepfakes and (2) neither raising awareness nor introducing financial incentives improves their detection accuracy. Zeroing in on the underlying cognitive processes, we find that (3) people are biased toward mistaking deepfakes as authentic videos (rather than vice versa) and (4) they 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 deepfake content.

INTRODUCTION

New societal challenges arise from artificial intelligence (AI)-manipulated media, particularly from deepfakes, the hyper-realistic imitation of authentic audio-visual content (Chesney and Citron, 2019). Although 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), deepfakes also have a dark side. For example, recent investigations revealed large-scale use of deepfakes to "undress" women (Hao 2020) and place them in porn videos (Cook 2019), hence turning deepfakes into a threat to people's reputation (Ayyub 2018). Relatedly, consider high-profile online theft cases in which scammers successfully used deepfake voice imitation to trick the company's employees into wiring money to the scammers, the cost of such defrauding schemes amounting to several hundred thousand dollars (Damiani 2019). Indeed, a consortium of researchers, policymakers, and tech experts ranked the malicious use of deepfakes as the number one emerging AI threat (Caldwell et al., 2020).

By changing audio-visual content, deepfakes allow new types of manipulations (Hancock and Bailenson, 2021). Moreover, tools to generate deepfakes such as FaceApp, (>500 million downloads) and FaceSwap have become widely available. Such tools increasingly render deepfakes accessible to the masses rather than only a selected few experts. These developments raise new societal challenges and research questions. First, can people still reliably differentiate authentic from deepfake videos? What are effective ways to increase people's detection accuracy? Does an appeal to the importance of detecting deepfakes improve detection accuracy, or do financial incentives help? Examining the underlying cognitive processes, do people underestimate or overestimate the occurrence of deepfakes? Moreover, how accurately do they estimate their own abilities to identify deepfakes?

In pursuit of empirical answers, we conducted an online experiment. In the experiment, participants watched a series of deepfake and authentic videos and guessed which ones were deepfakes. We introduced two experimental manipulations to boost detection accuracy. Namely, we sought to increase people's motivation to detect deepfakes by providing a short prompt emphasizing the danger of deepfakes (Awareness Treatment) or by financially incentivizing accuracy (Financial Incentive Treatment). We compared both interventions to a control treatment without motivational interventions (Control). We further assessed participants' confidence in their guesses. In line with the recently proposed machine behavior approach advocating the use of real algorithms instead of hypothetical stimuli (Rahwan et al., 2019), we used a data-set of deepfake videos that contain comparable deepfake and their corresponding original short clips (Dolhansky et al., 2020).

[1]Center for Humans and Machines, Max Planck Institute for Human Development, 14195 Berlin, Germany

[2]Amsterdam School of Economics, University of Amsterdam, 1001 NJ Amsterdam, The Netherlands

[3]Lead contact

*Correspondence: koebis@mpib-berlin.mpg.de

https://doi.org/10.1016/j.isci.2021.103364

The rising challenge of detecting deepfakes

Fake information, gossip, and smear campaigns have existed throughout much of human history, and even methods to alter facial images have been around for more than 20 years (Blanz et al., 2004; Bregler et al., 1997). However, recent developments in Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) have revolutionized manipulative content generation. By harnessing the potential of pitting neural networks against each other, GANs enable the realistic creation of synthetic audio-visual content: voice and moving images can now be instantaneously manipulated. For example, the Face2Face system allows taking an input video of a face and transferring the mouth shape and expressions onto synthesized target faces in real time (Thies et al., 2016). As a consequence, videos have ceased to be the gold standard for establishing veracity (online).

Creating deepfakes also no longer requires expert knowledge (Alexander et al., 2009) or even large training datasets (Nirkin et al., 2019). Applications like Face swapping GAN are democratizing the creation of deepfakes (Nirkin et al., 2019). Identifying who is behind the production of a particular deepfake becomes increasingly demanding, effectively undermining accountability and often resulting in impunity (King et al., 2020; Köbis et al., 2021). At the same time, deepfakes can reach an unprecedented audience (Diakopoulos and Johnson, 2019). Since false information often spreads faster and permeates social media networks deeper than accurate information (Vosoughi et al., 2018), correcting falsehoods like deepfakes online presents a complex challenge (Korshunov and Marcel, 2018). Therefore, worry emerges about using deepfakes as a new deceptive tool (Köbis et al., 2021), for example, in mass voter misinformation campaigns before elections (Dobber et al., 2021).

But, can people still spot a deepfake? Research addressing this question is small but growing (Hancock and Bailenson, 2021). Empirical research using *static images* suggests that people can (still) spot AI-manipulated pictures. In one study Groh et al. (2021) built a platform containing authentic and manipulated pictures. Over 15,000 respondents guessed which image is fake and which one is real. Over several rounds, they received feedback about their accuracy and could try again with a new set of pictures. The results reveal that people's accuracy increases over time, stabilizing at a rate of 88%. Similarly, studies using synthetic images generated by a novel face-swapping technology indicate that people's detection accuracy equally exceeds chance levels (Rössler et al., 2019), albeit the accuracy levels were overall lower than the results obtained by Groh et al. (2021), with participants scoring 68.7% for high-quality images and merely 58.7% for low-quality images.

Moving from static images to *deepfake videos*, research on detection accuracy is sparse. Although not assessing whether people could differentiate between deepfakes and authentic content, a related study by Vaccari and Chadwick (2020) provides some insights into whether viewers believe the content of deepfakes. Using edits of the aforementioned Buzzfeed deepfake video of Barack Obama, the experiment tested whether the video tricked participants into believing that Obama, indeed, cursed. Compared with a version containing a disclaimer about the video's fake nature, two versions lacking such disclaimers did not successfully deceive participants. The study also revealed that the disclaimer-less versions of the deepfake video increased participants' uncertainty. In the absence of a control group without a deepfake, it remains somewhat unclear whether the deepfake video successfully misled viewers into believing the fake content and whether they were aware that it was a deepfake.

Two other studies use deepfake videos of famous individuals and examine the perceived similarity between the remake and the original (Iacobucci et al., 2021) and self-rated persuasiveness of the content (Hwang et al., 2021). Using famous deepfakes has the disadvantage that people might have seen the deepfake before or be otherwise motivated to believe or discredit the video (e.g., sympathy toward the protagonist or belief in the message). Avoiding such possible confounds requires using videos that people have not seen before. Here, we draw on a unique set of short, unknown deepfake videos and their respective original (for more details, see Video Section) to tackle the fundamental question of whether people can (still) spot a deepfake video.

With established theoretical frameworks on deepfake detection lacking, a promising approach comes from the related field of fake news identification. Here, signal detection theory (SDT) provides a unifying theoretical framework for the various approaches studying fake news detection (Batailler et al., 2021). The SDT framework has a long history of perceptual studies on the factors influencing people's ability to distinguish

signals from noise (Green and Swets, 1966), particularly in ambiguous situations (Wickens, 2002). It thus also helps to understand when and how people distinguish between fake and real news (Batailler et al., 2021). Although the stimuli are novel for deepfakes, SDT provides a conceptual underpinning to study people's abilities to discern authentic videos (=signals) from deepfake videos (=noise).

SDT defines two main indices. First, the sensitivity *d'* captures people's ability to discriminate between deepfakes and authentic videos, i.e., their detection *accuracy*. Second, the *bias c* captures how conservative people are when deciding to report that a video is fake. Based on these two main indices of SDT, next, we motivate which interventions could help increase detection accuracy and outline possible cognitive biases involved in deepfake detection.

Interventions to improve detection accuracy

The pursuit of strategies to increase deepfake video detection has so far been a predominantly technical one. Research in computer science has developed machine classifiers for deepfake videos (Korshunov and Marcel, 2018; Li et al., 2020; Afchar et al., 2018; Li and Lyu, 2018; Yang et al., 2019). Although highly relevant, such efforts often neglect the human side of the equation. Namely, although such tools are helpful, they have not yet reached the broad public. Hence, people are often left to their own devices—their eyes and ears—to detect deepfakes (Nygren et al., 2021). Are there ways to boost people's detection accuracy?

A recent literature review on the adjacent field of fake news detection identifies four main accounts that explain fake information detection (Batailler et al., 2021). First, partisan bias describes how political or otherwise ideological content leads people to believe congruent information while disregarding incongruent information. Second, prior exposure proposes that having previously seen fake content increases people's tendency to believe it. Third, cognitive reflection accounts postulate that inattentiveness and lack of deliberation largely drive the belief in fake news. And, fourth, motivated reflection suggests that people employ their cognitive resources for information discernment selectively according to their goals, motivated to obtain the desired outcome in the process (Batailler et al., 2021).

Here, we focus on the latter two accounts that put cognitive reflection center stage because they have been proposed as counter-strategies against deepfake deception too. Namely, conceptual work suggests that educational interventions, such as improving media literacy, help people to navigate digital news and detect deepfakes (García Lozano et al., 2020; Diakopoulos and Johnson, 2019). Informing people about the potential consequences of manipulated media content helps raise awareness for the issue. This awareness, in turn, motivates people to invest cognitive resources to detect deepfakes, so the argument reads. Indeed, empirical studies on fake news detection indicate that encouraging people to pay attention to accuracy and deliberate about their decisions increases their ability to identify veracity and reduces their sharing intentions of fake news (Pennycook et al., 2021; Bago et al., 2020). In the context of deepfakes, first findings suggest that people's self-reported willingness to share deepfakes decreases after a media literacy intervention (Hwang et al., 2021). To test whether awareness can also improve detection accuracy, in one of our treatments (i.e., the Awareness treatment), participants read a short awareness prompt informing them about the harmful consequences of deepfakes before engaging in the detection task.

Another strategy consists of providing financial incentives for accuracy. Although experiments testing people's abilities to detect manipulated media have not used financial incentives for accuracy, extensive evidence from behavioral science shows that incentivization can increase people's motivation and accuracy (Schlag et al., 2015; Gächter and Renner, 2010; Köbis and Mossink, 2021). Here, we follow that tradition and test whether financial incentives could increase deepfake detection accuracy in another treatment (i.e., Financial Incentive treatment).

We compare these two treatments with a control treatment in which no intervention to increase accuracy exists. We pre-registered (see As.Predicted, https://aspredicted.org/am5rn.pdf) the hypotheses that, compared with the Control treatment, participants in the Awareness treatment and the Financial Incentive treatment would perform significantly better at detecting deepfakes.

Detection bias

Extensive research in judgment and decision making has revealed that people often use heuristics (mental shortcuts), also when establishing veracity online (Pennycook and Rand, 2021). Yet, studies specifically

looking at the cognitive processes underlying deepfake detection are still lacking. It is conceivable that people make systematic mistakes when classifying deepfake and authentic videos. When it comes to the distribution of these mistakes, previous work suggests two possible directions. First, once people learn about the existence of deepfakes and thus are aware of the possibility of fabricated content, they might become oversensitive to it. The notion of a "liar's dividend" (Chesney and Citron, 2019) describes the idea that when people become more critical and skeptical of media, they may even doubt authentic content. Implications of such a bias would be that establishing veracity online becomes a demanding challenge. For example, even authentic announcements by politicians might be quickly dismissed as (deep) fakes. Hence, one hypothesis reads that people might become overly sensitive to potentially manipulated media and overestimate the actual frequency of deepfakes.

Second, the counter-argument reads that people primarily view authentic video content online. Fake videos still represent the rare exception to the rule. In particular, for visual input (rather than written or auditory input), the heuristic of "seeing is believing" remains dominant (Frenda et al., 2013). In general, people trust audiovisual content more than verbal content as it resembles the real world more closely (Sundar 2008). Thus, people might consider video content authentic until they find clear-cut evidence of tampering (Farid 2019). Consequently, this argument suggests that people rather err on the conservative side and underestimate the frequency of deepfakes. This bias would imply that people become susceptible to being duped by deepfakes as they consider fake content real, rendering deepfake a potent tool for misinformation campaigns.

By using an equal proportion of authentic and deepfake videos in our sample (and informing participants about it), the current study tests both competing biases—liar's dividend versus seeing is believing.

(Over)confidence in detection accuracy

Another possible cognitive bias occurs if people overestimate their abilities to identify deepfake content. Overconfidence constitutes one of the most widespread and costly biases in human decision-making (Kahneman 2011; Kruger and Dunning, 1999). Indeed, events as diverse as global stock market crashes, entrepreneurial failures, and even the nuclear accident in Chernobyl have all been attributed to overconfidence (Moore and Healy, 2008). When it comes to detecting synthetic content, overconfidence might make people particularly susceptible to being manipulated. If people believe they can spot a deepfake, but in fact, cannot, they might consume manipulated content without being aware of it. Recent experiments have examined whether people overestimate their abilities to discern AI-generated from human-written text (Köbis and Mossink, 2021). Results reveal that participants have exaggerated beliefs in their detection abilities when such beliefs are elicited in an unincentivized way. However, overconfidence disappeared when participants stood to gain financially from accurately gauging their abilities.

Cognitive processing of textual input, however, differs from cognitive processing of audio-visual input. For one, a recent meta-analysis underlines that information encoded in visual form, e.g., via pictures and videos, is more persuasive than information encoded in text (Seo 2020). Similarly, media framing effects for visual content exceed those for text content (Powell et al., 2015). One reason lies in more effective cognitive integration of visual input compared with other types of sensory data (Witten and Knudsen, 2005). For example, visual messages produce higher retention rates than verbal messages (Graber 1990). Such cognitive integration likely increases people's confidence in discerning fake from authentic visual content.

Thus, it is conceivable that overconfidence in one's detection abilities might be particularly pronounced for deepfake video input. We thus pre-registered the expectation that participants in our study show overconfidence, operationalized as perceived detection abilities significantly exceeding actual detection abilities.

RESULTS

Participants, design, and procedures

To measure people's ability to detect deepfakes, we ran an incentivized online experiment. In line with our pre-registered sample size, 210 participants (M_{age} = 35.42, SD_{age} = 11.93; female = 135, 83.8% = UK citizens) were recruited via the online participant platform Prolific.co and completed the study. This sample allowed detection of small to medium effect sizes (see more details Table S1). Each participant received 2.5 pounds as a participation fee, plus potential bonus payments. The study took, on average, 18.59 min

to complete. The Ethics Review Board of the University of Amsterdam, Center for Experimental Economics and political Decision Making (CREED) approved the study design. Materials, data, and R code to reproduce the results are available on the Open Science Framework (OSF) (https://osf.io/mz974/). For more details on the methodology and the participants, see STAR Methods section.

As a first step, participants read the instructions informing them that the probability of each video being a deepfake was 50%. Then they had to answer several comprehension check questions correctly before they could continue. Next, participants watched 16 videos and indicated for each one whether it was a deepfake or not. After each video, we asked a question about the content of the video to assess participants' attentiveness. Overall, participants answered 94.2% of these verification questions correctly, suggesting conscientious completion.

We sampled videos from the MIT project DetectDeepfake, which contains 3,000 of the most difficult videos for AI classifiers to classify from Kaggle's DeepFake Detection Challenge. We chose videos without political or ideological content to rule out motivational biases in detection accuracy. In contrast to previous work on deepfake videos (Vaccari and Chadwick, 2020; Hwang et al., 2021; Iacobucci et al., 2021), we intentionally sampled videos with unknown actors to avoid that participants knew which videos were fake.

We randomly sampled 16 target videos from this pool of videos, keeping the length of the video constant across all videos ($\approx$10 s). Participants could play the video as often as they desired. Each of these videos had an authentic version and a deepfake version. Each participant saw only one of the two versions during the experiment as we divided them into two sets, each containing eight authentic videos and eight deepfakes. In each set, we randomized the order of the videos (see Figure 1 for an illustration).

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 written by Chesney and Citron (2019) that warns about the potentially harmful consequences of deepfakes (see OSF for the text, https://osf.io/ye83h/). By raising awareness about deepfakes' dangerous consequences, the prompt aimed to increase participants' motivation for deepfake detection. To ensure that participants read the prompt, they had to correctly answer a multiple-choice question about its content before proceeding to the detection task.

In the Financial Incentive treatment, participants received monetary rewards for accuracy. Namely, at the end of the experiment, one of the 16 rounds was randomly chosen for payment. Participants received a bonus of 3 pounds if the guess in the chosen round was correct. 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.

To assess whether people accurately estimate their own detection abilities, we used two measures of confidence. First, after each video, participants rated their subjective probability of guessing correctly on a scale from 50 (= the same as flipping a coin) to 100 (= 100% probability of having guessed correctly). This measure allowed us to assess confidence on a video-by-video basis. Second, participants indicated how many videos they estimated to have guessed correctly after they completed all rounds. This guess was incentivized. The participant obtained 0.5 pounds if their guessed number was within a range of ± 1 from the actual number of videos correctly detected. Of importance, participants did not know that this incentivized measure would follow the detection task to avoid hedging. Namely, had we informed participants about this question, they might have indicated to be correct in half of the rounds and then flipped a coin each round. We included such an incentivized measure of confidence to make participants carefully think about the answer and encourage them to provide their best, unbiased estimate of their performance.

At the end of the study, participants indicated their level of motivation to classify the videos on a seven-point Likert scale (1 = not motivated at all; 7 = very much motivated). This measure served as a manipulation check to test whether the two treatments (FI & AW) increased motivation compared with the control treatment. We also assessed demographics of age, gender, education, nationality, and current employment status and if the participants' work/study focuses on visual media production. Finally, using an open text field, we invited participants to provide any feedback about the study. None of the replies in the open text field indicated any technical difficulties in watching the videos. Afterward, the participants were thanked and redirected to Prolific.

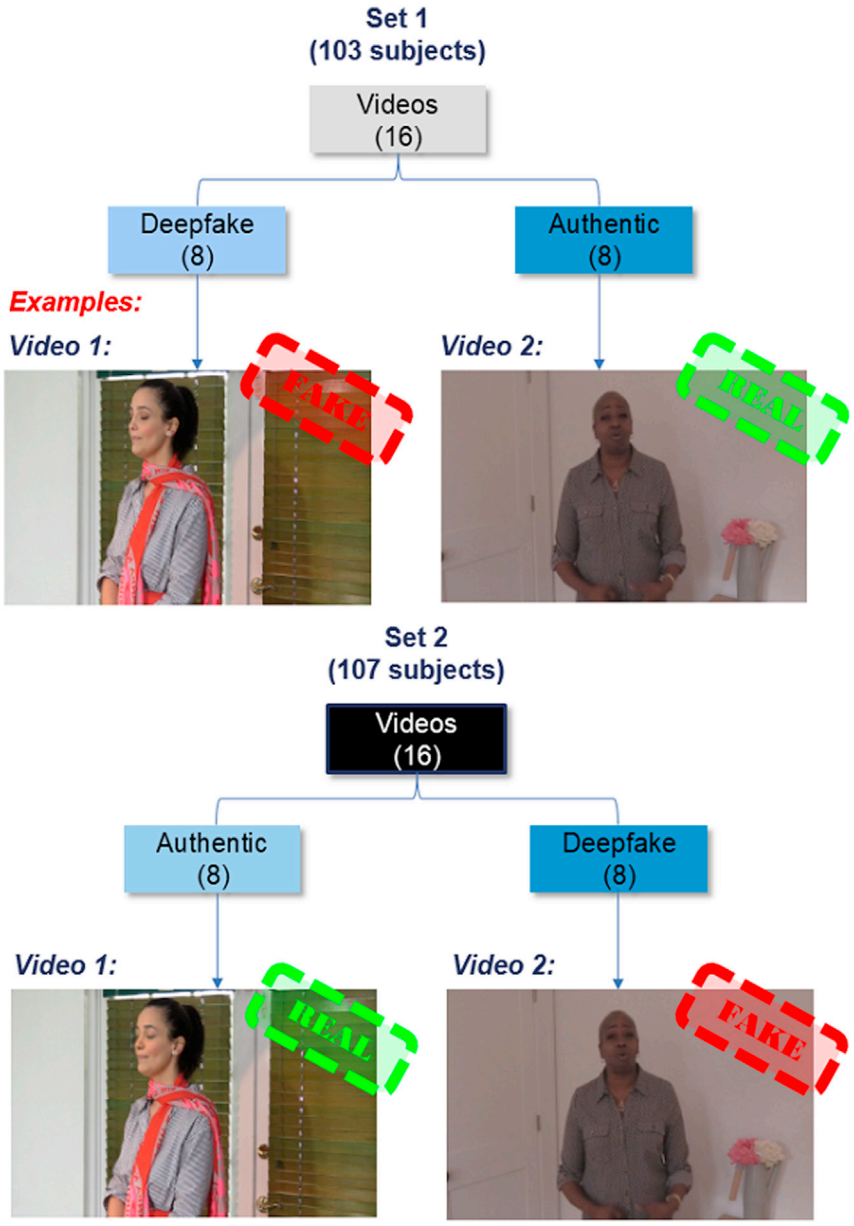


Figure 1. Illustration of video treatments

Design of the experiment. Videos were separated into two sets, each consisting of eight deepfakes and eight authentic videos. Deepfake and authentic versions of the same videos were presented at the same slot in different sets.

Detection accuracy

As a first test of whether people can reliably spot a deepfake, we compared participants' average accuracy as independent data points to random guessing. The overall accuracy level of 57.6% (SD = 11.6) exceeds chance levels according to a one-sample t test ($t(209) = 9.539$, $p < .001$, Cohen's $d = 0.658$). In signal detection theory terms, we observe a positive level of sensitivity, with an average d' of 0.484. A one-sample t test testing the null hypothesis of non-discriminability (H0: $d' = 0$) confirms that participants are sensitive to the differences between fake and authentic videos ($t(209) = 9.874$, $p < .001$).

However, the average sensitivity also shows that participants are overall not very good at discriminating. The average d' implies a probability of 59.6% to guess correctly for an unbiased participant (for further results using the SDT framework, see Figure S1). Also, looking at the videos separately reveals that only

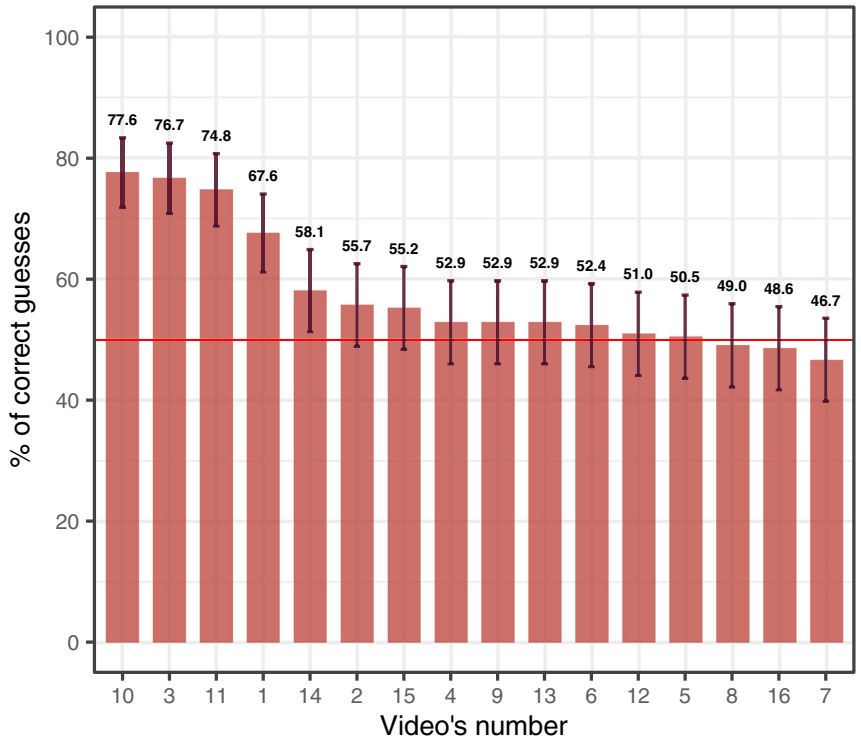


Figure 2. Accuracy by video number

Average accuracy levels for videos in descending order. The error bars denote 95% confidence intervals. The red line indicates the 50% accuracy rate of random guessing.

for 5 of the 16 videos, participants' guesses are significantly more accurate than flipping a coin (see Figure 2).

Looking at detection accuracy dis-aggregated by video type provides deeper insights into how participants guessed. Namely, for most authentic videos (13 of 16), accuracy levels exceed chance levels. However, when it comes to deepfakes, participants achieve higher-than-chance accuracy only for four videos. For the vast majority of deepfake videos, participants are correct at lower-than-chance levels (see Figure 3).

Note that, in our setup, the accuracy for fake and authentic videos is mechanically correlated. Consider, for example, a participant who guesses "authentic" in 100% of the rounds. This person would have a 100% accuracy rate for authentic videos but a 0% accuracy rate for deepfake videos. As we will discuss in the Subsection detection bias the main reason why accuracy in detecting fake videos is low stems from participants' tendency to guess that a video is authentic.

Interventions to increase accuracy

Contrary to our hypotheses, neither raising awareness nor financial incentives increase detection accuracy (one-way ANOVA: $F(2,207) = 0.276$, $p = 0.759$). Figure 4 shows that, on average, participants guess correctly 9 of the 16 videos, independent of the treatment. Also a comparison of the d' values across treatments reveals no differences ($F(2,207) = 0.190$, $p = 0.827$, see also Figure S2). Analysis of the self-reported motivation to detect deepfakes reveals high motivation levels across all treatments as more than 75% 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 the self-reported motivation levels across treatments ($F(2,207) = 0.526$, $p = 0.592$, see also Figure S3). This finding suggests that the low detection rates stem from inability rather than from lack of motivation. In sum, this pattern of results shows that for the large majority of videos, people cannot reliably discern deepfakes from authentic videos, and the interventions intended to raise motivation do not significantly increase accuracy levels.

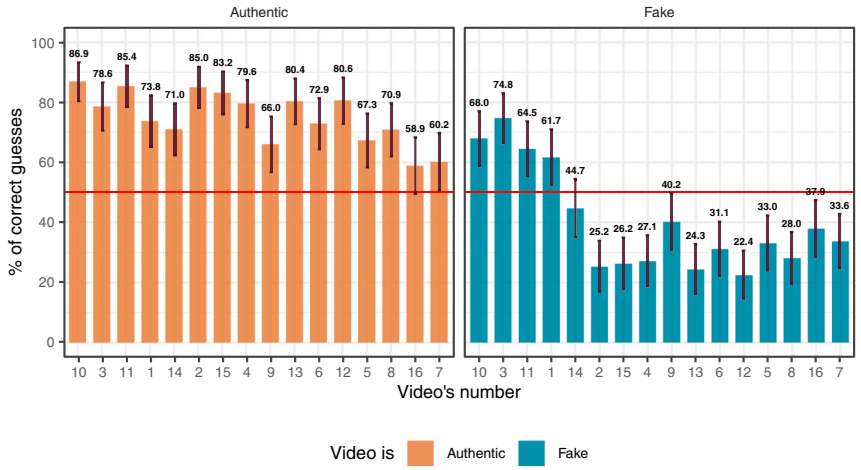


Figure 3. Accuracy dis-aggregated by video's number and video's type

Average accuracy levels for each video, dis-aggregated for authentic (left) and fake (right) videos. The error bars denote 95% confidence intervals. The red line indicates the 50% accuracy rate of random guessing.

Detection bias

Moving to the cognitive processes underlying deepfake detection, analyses on guessing patterns reveal a strong bias toward guessing that a video is authentic. Although we informed participants that only half of the videos were authentic, they guessed authentic 67.4% of the time, which significantly exceeds equal guessing (one-sample $t(209) = 13.131$, $p < 0.001$). This bias is depicted by the high density of observations to the right side of the vertical red line in the scatter-plot in Figure 5. This figure makes apparent that many participants strongly believe that videos are authentic, with 14 participants even guessing that all videos are authentic. In SDT terms, we observe an average bias c of 0.518 ($SD = 0.578$, Cohen's $d = 0.897$), which confirms that our participants are very conservative when reporting that a video is a deepfake, i.e., people have a tendency toward guessing authentic. Accordingly, a one-sample t test strongly rejects the null hypothesis of no bias ($t(209) = 13.000$, $p < .001$).

Furthermore, the likelihood of guessing authentic does not significantly differ across treatments (one-way ANOVA: $F(2,207) = 0.618$, $p = 0.540$, see Figure 4B), suggesting comparable biases in guessing across treatments. Regression analyses further corroborate that the bias toward guessing authentic is robust to the number of views, demographic characteristics, subjective motivation, and video characteristics (see for full regression in Tables S2 and S4).

(Over)confidence

To estimate whether participants accurately perceive their abilities to detect deepfakes, we first compared their actual accuracy rate to the reported subjective probability of having guessed correctly for each video. As illustrated in Figure 6, participants accurately estimate their detection ability for only 3 of the 16 videos (#3,11,4). For the remaining videos, confidence largely exceeds accuracy levels. For regression analyses predicting the subjective confidence level for each video, see Tables S3 and S5.

Results using the incentivized measure of confidence corroborate this indication of overconfidence. Participants significantly 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 negatively correlate (Pearson's product-moment correlation: $r(208) = -0.475$, $p < 0.001$, see Figure 7, see also Figure S4). This typical pattern, often referred to as the Dunning-Kruger effect (Kruger and Dunning, 1999), indicates that overconfidence is particularly pronounced among those who perform worse.

Taken together, analyses on the cognitive processes behind deepfake detection reveal that people do not only have a bias toward guessing authentic but also show overconfidence in their detection abilities.

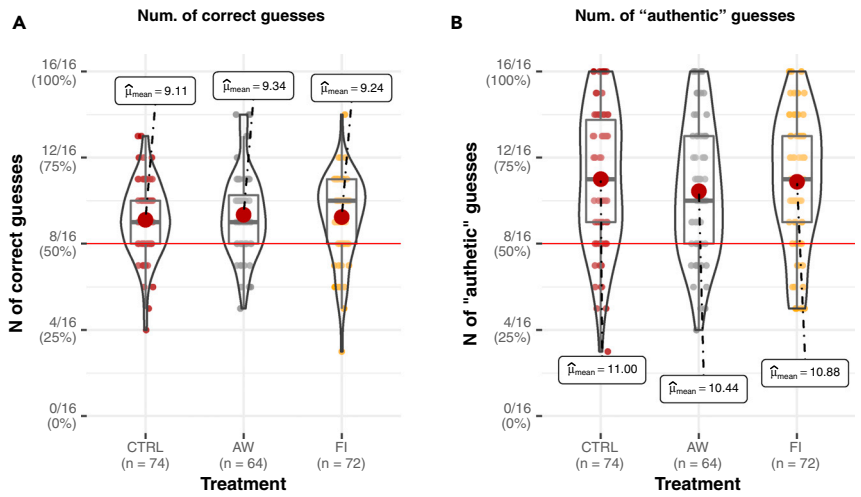


Figure 4. Guessing behavior—Correct and "authentic" guesses

(A) Violin plots of the distribution of the number of correct guesses a participant made (y axis) by the treatment (x axis). (B) Violin plots of the distribution of frequency of "authentic" guesses (y axis) by the treatment (x axis). Dark gray lines represent medians; the dark red dots represent means. Boxes indicate the interquartile range; each dot shows a raw data point. Plots created with the Ggstatsplot package (Patil, 2018).

DISCUSSION

Creating deepfakes is easier than ever, yet detecting them becomes increasingly difficult. While most research has approached the challenge of distinguishing deepfakes and authentic content from a technical perspective, this study examines the human side of the equation. Our experiment reveals two main insights. First, detection of deepfakes is not a matter of lacking motivation but inability. Second, we find a systematic bias toward guessing that videos are authentic and an indication that people are overconfident in their detection abilities. Together, these findings suggest that people apply an overly optimistic seeing-is-believing heuristic, which might put them at a particular risk of being influenced by deepfakes. We discuss these findings below, emphasizing the need for more research on the interplay of behavior by humans and machines (Rahwan et al., 2019; Köbis et al., 2021).

Motivated but inaccurate

Humans integrate (moving) visual information more effectively than other sensory data (Witten and Knudsen, 2005; Seo 2020). Especially online interactions heavily feature audio-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 AI-manipulated static images (Groh et al., 2021; Rössler et al., 2019) or AI-generated text (Köbis and Mossink, 2021), our findings show a nuanced picture about people's ability to discern deepfake from authentic videos. Across all videos, people can (still) guess marginally better than chance. However, a closer look at the different videos indicates that they are not better than random guessing for most videos. Overall, we find lower detection accuracy than research on static images (Groh et al., 2021; 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 et al., 2020). Similarly, attention and critical thinking have been proposed as counter-strategies against deepfakes (Nolan and Kimball, 2021; Diakopoulos and Johnson, 2019). However, the current empirical evidence suggests that such appeals aiming to increase people's 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. Therefore, little room for the interventions to increase their motivation in the two treatments exists. Such high levels of motivation might stem from the sample characteristics. Prolific participants are typically conscientious in completing surveys (Peer et al., 2017). Future work drawing on samples with a

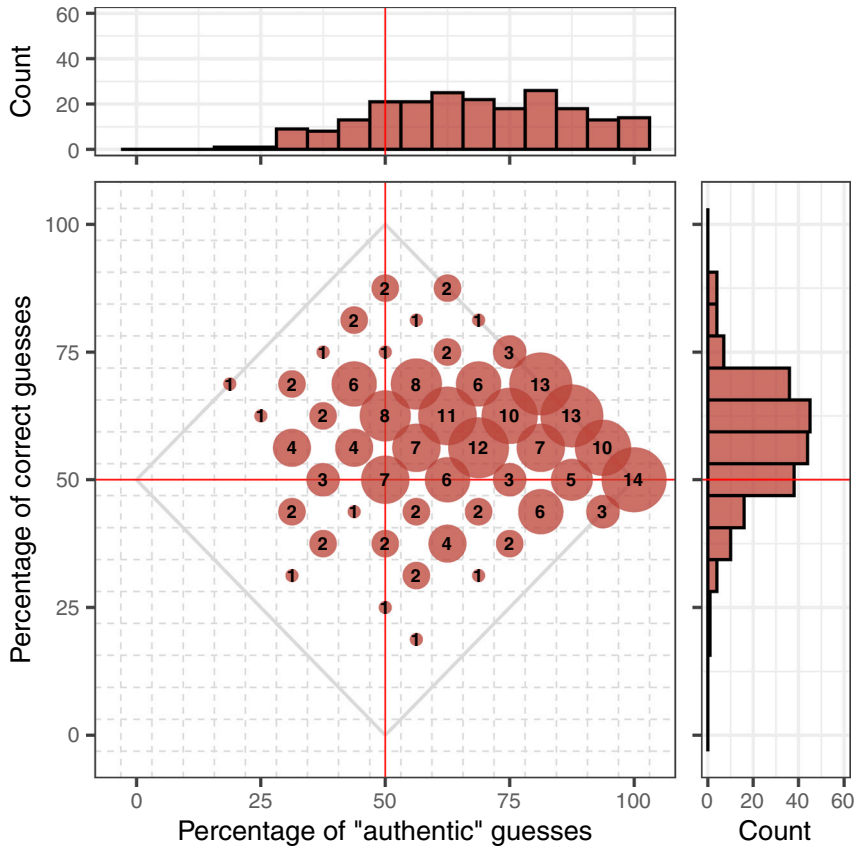


Figure 5. Fraction of correct guesses by fraction of "authentic" guesses

Distribution of participants according to the percentage of authentic guesses (x axis) and the percentage of correct guesses (y axis) they made. The size of the bubble is proportional to the number of participants (exact *N* reported inside the bubble). The gray diamond delimits the maximum and minimum fractions of correct guesses conditional on the number of times the participant guessed that the video was authentic. Histograms on each side represent marginal distributions.

more heterogeneous distribution of motivation might help to alleviate this ceiling effect, for example, by using deepfakes with politicized content among different political groups (Dobber et al., 2021).

Although participants were highly motivated, their guesses were, for the most part, as good as flipping a coin. In line with findings from the adjacent literature on deception detection (Hancock and Bailenson, 2021; Verschuere et al., 2018), our results suggest that people do not reliably detect manipulated video content. As detecting deepfakes appears less a matter of motivation and attention, our results suggest that deepfakes warrant special attention in the research and policy field of digital misinformation. In line with the widely voiced concern about deepfakes becoming a new AI threat (Caldwell et al., 2020; Chesney and Citron, 2019; Köbis et al., 2021), our findings suggest some of the previously established strategies against manipulation do not hold for detection of deepfakes. A need exists for more research on human-centered strategies against deepfakes. For example, how fast can people learn to detect deepfakes when they receive feedback? Can awareness about common deepfake artifacts improve their detection performance? How do people perform when having an AI classifier as a decision aid?

Biased toward authenticity and overconfidence

We further uncover two related biases in human deepfake detection. First, people's guesses are skewed toward authenticity. Although participants were informed that half of the videos were authentic, they guessed that 67.4% of videos were authentic. It is possible that people still used the "ecological base rate" in the task. That is, outside of the current experiment, the vast majority of videos are authentic.

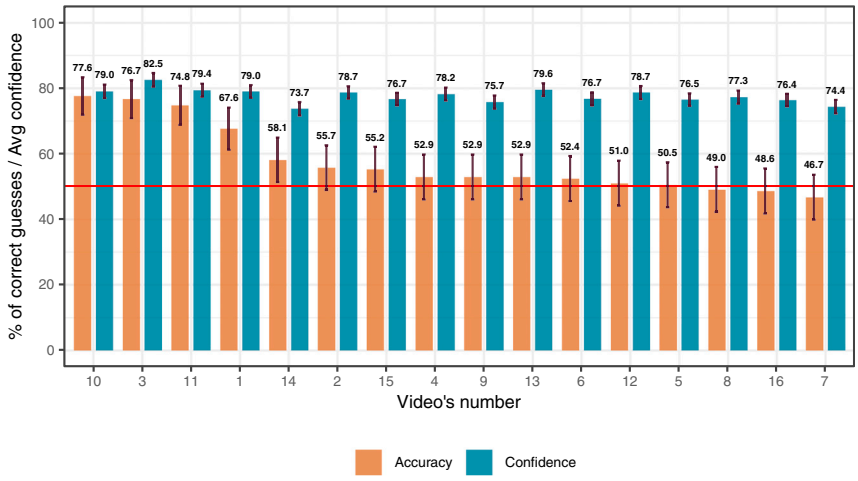


Figure 6. Accuracy and confidence by video number

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.

Therefore, people might have difficulty adjusting their guessing pattern to a new base rate of 50% deepfakes. Having said that, we intentionally included a comprehension question to ensure that participants are aware of the proportion of the deepfake videos. Hence, our findings indicate that the tendency to guess authentic likely does not stem from a deliberate process.

Second, people systematically overestimate their abilities to detect deepfakes. In line with our expectation, people's detection performance does not live up to their confidence, independent whether elicited with or without incentives. Indicative of the Dunning-Kruger effect (Kruger and Dunning, 1999), we even find a negative correlation between performance and overconfidence. Hence, particularly low performers show overconfidence. This effect is often partially mechanical, in that worse performers have more room to be overconfident (Krajc and Ortmann, 2008). The evidence that people generally have inflated beliefs about their deepfake detection abilities, however, is remarkably robust.

Taken together, these two biases suggest that people adopt a seeing-is-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 have exaggerated beliefs about their detection abilities. One fruitful avenue for future research is to examine the motivational roots of both biases. Do both biases stem from self-enhancement tendencies (Taylor and Brown, 1988) (i.e., to make them feel better about themselves)? Inspired by work on the cultural differences of overconfidence (Muthukrishna et al., 2018), are participants from Non-Western societies similarly overconfident in their deepfake detection abilities? Finally, how much are both biases influenced by political, ideological video content?

Conclusions

Philosophers have argued that the advent of deepfakes presents a new epistemic threat (Fallis 2020). Namely, if we can no longer believe what we see in a video, deepfakes will undermine our knowledge acquisition through media consumption. Our findings are a testament to this trend, showing that people can no longer reliably detect deepfakes. Some of the previously established strategies against misinformation do not hold for the detection of deepfakes. Spotting a deepfake appears less a matter of motivation and attention but rather inability. Hence, deepfakes warrant special attention for digital misinformation research and policy, especially in light of people's bias toward believing their eyes and mistake deepfakes as authentic, all while being overconfident in their detection abilities.

Limitations of the study

As outlined above, the study is not without limitations. For one, the lack of differences between treatments might stem from ceiling effects, as also participants in the Control treatment were highly motivated to

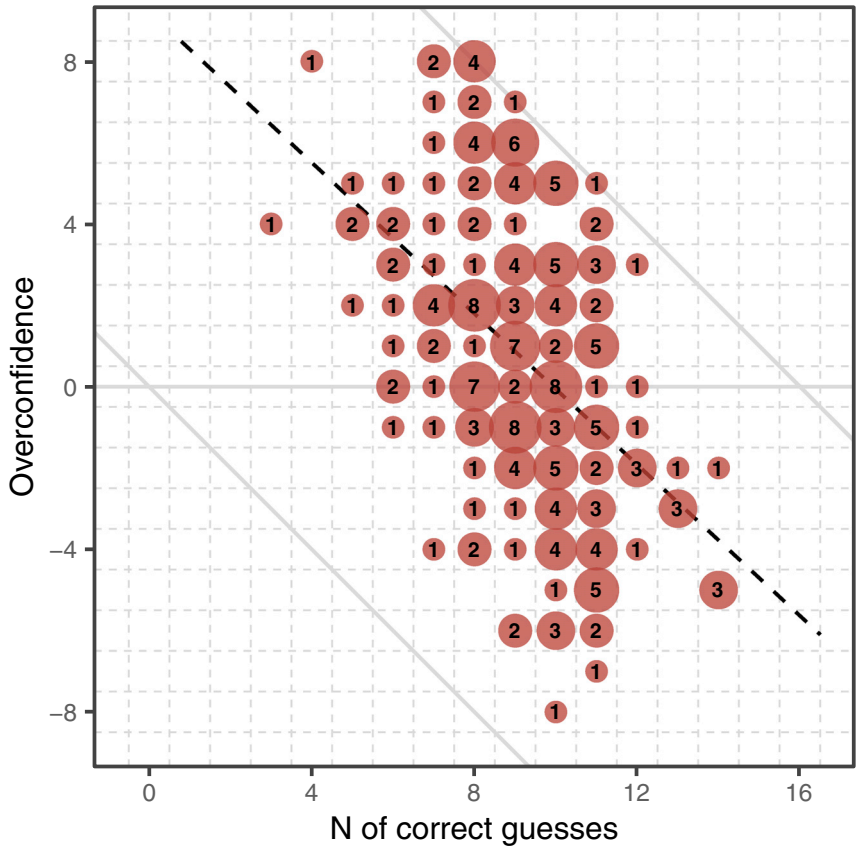


Figure 7. Overconfidence by number of correct guesses

The figure displays the distribution of participants according to the number of correct guesses they made (x axis) and the overconfidence, i.e., believed minus actual number of correct guesses (y axis). The size of the bubble is proportional to the number of participants. The number of participants is reported inside the bubble. The downward sloping gray lines delimit the maximum and minimum levels of overconfidence that can be achieved conditional on the number of correct guesses. The dashed black line shows the linear regression.

detect deepfakes. Moreover, even though many deepfake videos "in the real world" have political, ideological, or counter-intuitive content, our study does not feature such content because our research focuses on people's cognitive detection abilities void of emotional context.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability:
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103364.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from the Research Priority Area Behavioural Economics of the University of Amsterdam (proposal nr. 202006110906) and financial support by the Max Planck Institute for Human Development. We are also grateful for useful comments by attendees of the Moral AI lab (Max Planck Institute for Human Development & Toulouse School of Economics), the CREED lunch seminar (University of Amsterdam), Information Credibility Workshop (at ICWSM 2021), as well as by Margarita Leib and Christopher Starke.

AUTHOR CONTRIBUTIONS

Conceptualization, B.D., I.S., and N.C.K.; methodology, B.D., I.S., and N.C.K.; investigation, B.D.; data curation, I.S.; visualization, I.S.; formal analysis, I.S.; writing – original draft, N.C.K.; writing – review & editing, B.D., I.S., and N.C.K.; funding acquisition, B.D., I.S., and N.C.K.; supervision, I.S. and N.C.K.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: April 19, 2021
Revised: July 28, 2021
Accepted: October 25, 2021
Published: November 19, 2021

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Buolamwini and Gebru, 2018; Makowski, 2018.

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), pp. 1–7.

Alexander, O., Rogers, M., Lambeth, W., Chiang, M., and Debevec, P. (2009). Creating a photoreal digital actor: the digital emily project. In CVMP 2009 - 6th European Conference on Visual Media Production, pp. 176–187.

Aral, S. (2020). The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt. Currency.

Ayyub, R. (2018). I was the victim of a deepfake porn plot intended to silence me. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

Bago, B., Rand, D.G., and Pennycook, G. (2020). Fake news, fast and slow: deliberation reduces belief in false (but not true) news headlines. J. Exp. Psychol. Gen. *149*, 1608.

Batailler, C., Brannon, S.M., Teas, P.E., and Gawronski, B. (2021). A signal detection approach to understanding the identification of fake news. Perspect. Psychol. Sci. *22*, 1–21.

Blanz, V., Scherbaum, K., Vetter, T., and Seidel, H.-P. (2004). Exchanging faces in images. Comput. Graph. Forum *23*, 669–676.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: driving visual speech with audio. In Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 353–360.

Buolamwini, J., and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (PMLR), pp. 77–91.

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

Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. *107*, 1753–1820.

Cook, J. (2019). Heres what its like to see yourself in a deepfake porn video. https://www.huffingtonpost.co.uk/entry/deepfake-porn-video_uk_5d106e03e4b0aa375f4f1ea7.

Damiani, J. (2019). A Voice Deepfake Was Used to Scam a CEO Out of $243,000. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

Diakopoulos, N., and Johnson, D. (2019). Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections (New Media & Society), pp. 1–27.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., and de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? Int. J. Press/Politics *26*, 69–91.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C.C. (2020). The Deepfake Detection Challenge Dataset. arXiv. https://arxiv.org/abs/2006.07397.

Fallis, D. (2020). The Epistemic Threat of Deepfakes (Philosophy & Technology), pp. 1–21.

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. Social Psychol. *49*, 280–286.

Gächter, S., and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. Exp. Econ. *13*, 364–377.

García Lozano, M., Brynielsson, J., Franke, U., Rosell, M., Tjörnhammar, E., Varga, S., and Vlassov, V. (2020). Veracity assessment of online data. Decis. Support Syst. *129*, 113132.

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*, 139–144.

Graber, D.A. (1990). Seeing is remembering: how visuals contribute to learning from television news. J. Commun. *40*, 134–156.

Green, D.M., and Swets, J.A. (1966). Signal Detection Theory and Psychophysics, *Vol. 1* (Wiley).

Groh, M., Epstein, Z., Obradovich, N., Cebrian, M., and Rahwan, I. (2021). Human detection of machine-manipulated media. Commun. ACM *64*, 40–47.

Hancock, J.T., and Bailenson, J.N. (2021). The social impact of deepfakes. Cyberpsychol. Behav. Social Network. *24*, 149–152.

Hao, K. (2020). A Deepfake Bot Is Being Used to "Undress" Underage Girls. https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/.

Hwang, Y., Ryu, J.Y., and Jeong, S.-H. (2021). Effects of disinformation using deepfake: the protective effect of media literacy education. Cyberpsychol. Behav. Social Netw. *24*, 188–193.

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., and Pagliaro, S. (2021). Deepfakes unmasked: the effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. Cyberpsychol. Behav. Social Netw. *24*, 194–202.

Kahneman, D. (2011). Thinking, Fast and Slow (Macmillan).

King, T.C., Aggarwal, N., Taddeo, M., and Floridi, L. (2020). Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. Sci. Eng. Ethics *26*, 89–120.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad machines corrupt good morals. Nat. Hum. Behav. *5*, 679–685.

Köbis, N., and Mossink, L.D. (2021). Artificial intelligence versus maya angelou: experimental evidence that people cannot differentiate ai-generated from human-written poetry. Comput. Hum. Behav. *114*, 106553.

Korshunov, P., and Marcel, S. (2018). Deepfakes: a new threat to face recognition? Assessment and detection. https://arxiv.org/abs/1812.08685.

Krajc, M., and Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. J. Econ. Psychol. *29*, 724–738.

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing ones own incompetence lead to inflated self-assessments. J. Personal. Social Psychol. *77*, 1121.

Li, Y., and Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. https://arxiv.org/abs/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, pp. 3207–3216.

Makowski, D. (2018). The psycho package: an efficient and publishing-oriented workflow for psychological science. J. Open Source Softw. *3*, 470.

Moore, D.A., and Healy, P.J. (2008). The trouble with overconfidence. Psychol. Rev. *115*, 502–517.

Muthukrishna, M., Henrich, J., Toyokawa, W., Hamamura, T., Kameda, T., and Heine, S.J. (2018). Overconfidence is universal? elicitation of genuine overconfidence (ego) procedure reveals systematic differences across domain, task knowledge, and incentives in four populations. PLoS ONE *13*, e0202288.

Nirkin, Y., Keller, Y., and Hassner, T. (2019). FSGAN: subject agnostic face swapping and reenactment. In Proc. IEEE/CVF International Conference on Computer Vision, pp. 7184–7193.

Nolan, S.A., and Kimball, M. (2021). Tom cruise, deepfakes, and the need for critical thinking. https://www.psychologytoday.com/intl/blog/misinformation-desk/202103/tom-cruise-deepfakes-and-the-need-critical-thinking.

Nygren, T., Guath, M., Axelsson, C.-A.W., and Frau-Meigs, D. (2021). Combatting visual fake news with a professional fact-checking tool in education in France, Romania, Spain and Sweden. Information *12*, 201.

Patil, I. (2018). ggstatsplot: ggplot2 based plots with statistical details. CRAN. https://CRAN.R-project.org/package=ggstatsplot.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: alternative platforms for crowdsourcing behavioral research. J. Exp. Social Psychol. *70*, 153–163.

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 *592*, 590–595.

Pennycook, G., and Rand, D.G. (2021). The psychology of fake news. Trends Cogn. Sci. *25*, 388–402.

Powell, T.E., Boomgaarden, H.G., De Swert, K., and de Vreese, C.H. (2015). A clearer picture: the contribution of visuals and text to framing effects. J. Commun. *65*, 997–1017.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., et al. (2019). Machine behaviour. Nature *568*, 477–486.

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, pp. 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. Exp. Econ. *18*, 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.

Sundar, S.S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility (MacArthur Foundation Digital Media and Learning Initiative).

Taylor, S.E., and Brown, J.D. (1988). Illusion and well-being: a social psychological perspective on mental health. Psychol. Bull. *103*, 193.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2Face: real-time face capture and reenactment of RGB videos. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395.

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 + Soc. *6*.

Verschuere, B., Köbis, N.C., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2018). Taxing the brain to uncover lying? meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. J. Appl. Res. Mem. Cogn. *7*, 462–469.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. Science *359*, 1146–1151.

Wickens, T.D. (2002). Elementary Signal Detection Theory (Oxford University Press).

Witten, I.B., and Knudsen, E.I. (2005). Why seeing is believing: merging auditory and visual worlds. Neuron *48*, 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), pp. 8261–8265.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data of the main experiment		https://osf.io/hx2vt/
Data for the video coding		https://osf.io/ngk8j/
Read.me file explaining the data		https://osf.io/5t8c9/
R code to reproduce the statistical analyses and figures		https://osf.io/rew2q/
Software and algorithms		
Qualtrics Version 09/2020	Qualtrics	www.Qualtrics.com
R version 4.0.4	R Project	https://www.r-project.org/
PyCharm 2021.1.3	PyCharm	https://www.jetbrains.com/pycharm/
NotePad++ 8.1.1	NotePad	https://notepad-plus-plus.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nils Köbis, (koebis@mpib-berlin.mpg.de; n.c.kobis@gmail.com).

Materials availability

All instruction texts of the materials used, the videos used as well as the full programming .qsf file of the experiment software, executable on Qualtrics, are available via the project's page on the Open Science Framework https://osf.io/mz974/.

Data and code availability:

- data of the main experiment: https://osf.io/hx2vt/
- data of the video coding: https://osf.io/ngk8j/
- read.me file explaining the data: https://osf.io/5t8c9/
- R code to reproduce the statistical analyses and figures: https://osf.io/rew2q/
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

For the experiment reported in this manuscript, we recruited human participants from the online platform Prolific.co that is specialized in enabling high-quality online behavioral research. As specified in our pre-registration (see As.Predicted), we aimed to recruit 210 English-speaking participants. While 233 participants originally started the study, we stopped recruitment when 210 completes were achieved. The average age of the final sample was M_{age} = 35.42 (SD_{age} = 11.93) of which 135 participants self-identified as female. This sample allowed to detect small to medium effect sizes (see more details Table S1). Of the 210 participants 83.8% were from the United Kingdom, 6.7% from the United States of America, 2.9% from Australia, 2.4% from Ireland, 1.9% from New Zealand, 1.4% from Canada, and one participant from France and South Africa each (=0.5%). The average duration of the study was 18.59 minutes, and each participant received 2.5 pounds as a participation fee, plus potential bonuses based on their behavior in the tasks. The Ethics Review Board of the University of Amsterdam, Center for Experimental Economics and political Decision Making (CREED) approved the study design. We obtained informed consent from all participants prior to participation.

METHOD DETAILS

We used Qualtrics to program the study. To facilitate reproducibility, the .qsf file to execute the experimental procedure is available on OSF. The experiment consisted of six steps.

1. Participants read an information sheet outlining that:
   - study would last less than 20 minutes and pays 2.5 pounds,
   - participants' anonymity was ensured,
   - no known physical or psychological risks are associated with the study,
   - the sole purpose of data collection was academic research
   - participants could end the study at any point,
   - participants could contact one of the authors if any questions arise,
   - the study entails control questions to detect inattentive responding.
   - participation in the study required participants to be able to play media content.
   - to participate in the study, they needed to actively consent and on the next page provide their Prolific ID.
2. Participants read the instructions to the study, informing them that:
   - they would watch 16 videos,
   - guess which video is authentic and which one is a deepfake,
   - answers some questions about each video, and
   - that there was a 50% chance that the video is a deepfake.
3. We randomly assigned participants to one of three treatments, namely:
   (a) in the Awareness Treatment they read a short text by Chesney and Citron (2019) that highlights the importance of detecting deepfakes
   (b) in the Financial Incentive Treatment participants were informed that they could earn a bonus of 3 pounds. They were informed that one of the videos would be randomly selected at the end of the experiment and, if they guessed correctly, they would obtain the bonus.
   (c) in the Control Treatment no further information was provided.
4. After engaging in one practice trial, participants started the guessing task consisting of 16 rounds. In each round, they:
   - saw a video of approximately 10 seconds that was either the authentic or deepfake version. The videos were randomly sampled from the MIT project DetectDeepfake, which contains 3,000 of the most difficult videos for AI classifiers to classify from Kaggle's DeepFake Detection Challenge,
   - answered whether the video is a deepfake (Y/N),
   - indicated their subjective confidence about the guess (continuous slider measure ranging from, 50 = As confident as flipping a coin - 100 = 100% sure),
   - answered a content-related question about the video with one correct and one incorrect answer option. This question was included to ensure that participants watched the videos attentively. The overall accuracy rate for these verification questions was 94.2%, indicating high attentiveness of the participants. We include them as control variables in the regression tables reported in the supplemental information section.
5. After completing all 16 rounds, participants:
   - indicated their subjective level of motivation to recognize the video,
   - estimated their performance in the guessing task by indicating how many videos they thought they guessed correctly. Instruction text and an example accompanying the question informed participants that if their answer was within +/- 1 video, they would receive a bonus of 0.5 pounds.

This measure served as an incentivized elicitation of confidence in the guessing task. After estimating their own performance, participants were informed about their actual performance.

6. Participants answered several exit questions, assessing their:
    - age (in years),
    - gender (m/f/other/ rather not say),
    - highest obtained level of education, (ranging from "Less than high school degree" to "Professional degree (JD, MD)",
    - nationality (drop down list for all countries recognized by the UN),
    - occupation status (student, employed, unemployed, other),
    - experience with visual media (works / studies related field, Y/N),
    - comments about the study (open text field).

QUANTIFICATION AND STATISTICAL ANALYSIS

We conducted all statistical analyses in R. All statistical details and sample sizes are provided. The exact statistical tests and variables used are described in the text and the legends of the tables and figures.

ADDITIONAL RESOURCES

The study was pre-registered at AsPredicted (see https://aspredicted.org/am5rn.pdf).