The Detection of Political Deepfakes

Markus Appel & Fabian Prietzel

Julius-Maximilians-University Würzburg, Germany

Manuscript accepted to be published in the Journal of Computer-Mediated Communication

Corresponding author:

Prof. Dr. Markus Appel

Human-Computer-Media Institute

University of Würzburg, Germany.

E-mail: markus.appel@uni-wuerzburg.de

Phone: +49 931 31 88106

The stimuli, preregistrations, data, codes, and supplementary material can be found at

https://osf.io/fqdk4

Running Head: DEEPFAKES

Abstract

Deepfake technology, allowing manipulations of audiovisual content by means of artificial intelligence, is on the rise. This has sparked concerns about a weaponization of manipulated videos for malicious ends. A theory on deepfake detection is presented and three pre-registered studies examined the detection of deepfakes in the political realm (featuring UK's Prime Minister Boris Johnson, Studies 1-3, or former US president Barack Obama, Study 2). Based on two system models of information processing as well as recent theory and research on fake news, individual differences in analytic thinking and political interest were examined as predictors of correctly detecting deepfakes. Analytic thinking (Studies 1 and 2) and political interest (Study 1) were positively associated with identifying deepfakes and negatively associated with the perceived accuracy of a fake news piece about a leaked video (whether or not the deepfake video itself was presented, Study 3). Implications for research and practice are discussed.

Keywords: political deepfakes; misinformation; fake news; analytic thinking; artificial intelligence

Lay summary

With modern technology videos can be manipulated and show, for example, politicians that say things that they never said in real life. These manipulated videos are called deepfakes. In this manuscript, a theoretical model on the detection of deepfakes by ordinary citizens is introduced. The authors conducted three studies in which deepfakes with political content were presented. The deepfakes showed UK's Prime Minister Boris Johnson or Barack Obama. In the deepfake videos the two politicians said things they had never said in real life. The authors expected that people who regularly and automatically reflect on information they see (analytic thinking) are more likely to identify deepfakes correctly than people who tend to be less reflective, more intuitive. The authors further expected that interest in politics is positively related to detecting political deepfakes. Indeed, the higher participants' scores on analytic thinking (Studies 1-2) and political interest (Study 1), the better participants identified the deepfakes. Moreover, people with high analytic thinking and political interest were better at identifying a fake news article to be inaccurate (whether or not a warranting deepfake video was presented, Study 3). It is discussed how researchers, everyday people, and whole societies can deal with deepfakes.

The Detection of Political Deepfakes

Whereas convincing manipulations of video footage have been demonstrated by the movie industry for quite some time now, recent advances in artificial intelligence have made it easier than ever to create sophisticated and compelling fake videos—so called *deepfakes*—depicting individuals saying and doing things they never said or did. The earliest deepfake videos to gain notoriety were those featuring female celebrities, whose faces had been superimposed into sex videos without their consent (Harris, 2019; Spivak, 2019). The underlying technology has since migrated beyond the pornographic context, and its applications are expected to soon entail a disturbing array of malicious uses, e.g., in the realms of politics and international affairs, potentially eroding Western democracies (Chesney & Citron, 2019; Fletcher, 2018; Waisbord, 2018; Yadlin-Segal & Oppenheim, 2021). In 2019, for instance, the US intelligence community warned in its annual Worldwide Threat Assessment that "adversaries and strategic competitors probably will attempt to use deepfakes or similar machine-learning technologies to create convincing—but false—image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners" (Coats, 2019, p. 7). In line with this prediction, a deepfake surfaced on social media in March 2022 that showed Ukrainian president Volodymyr Zelenskyy telling his soldiers to surrender to Russian forces (Metz, 2022).

Theory and empirical research on the detection and the effects of deepfakes is scarce. We present a theoretical framework that is based on two-systems models of information processing and that connects to the literature on post-truth phenomena more generally, a topic that has received a substantial amount of attention in recent years (e.g., Ecker et al., 2022; Lazer et al., 2018; Lewandowsky et al., 2017). These lines of research suggests that individual differences predict responses to deepfakes, analytic thinking and general political interest in particular should reduce the

impact of deepfakes (e.g., Kahneman, 2011; Pennycook & Rand, 2020). Three studies are presented that examined deepfake detection in the political realm. Hypotheses and methods were preregistered.

The Advent of Deepfakes

In late May of 2019, a manipulated video of Nancy Pelosi that made the US House Speaker appear intoxicated and slurring her words spread across the internet. The video, which was viewed more than 2.5 million times on Facebook in a matter of days and shared by prominent political leaders (Mervosh, 2019), foreshadowed one type of disinformation that could disrupt political discourse and future elections: deliberately altered audiovisual content, amplified via social media. Deepfakes broke into the mainstream about a year before the Pelosi video, when a widely read technology blog published an article with the disturbing headline: "We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now" (Cole, 2018). An anonymous user posted several pornographic videos on the popular discussion board Reddit, purporting to feature famous actresses such as Gal Gadot (*Wonder Woman*) and Maisie Williams (*Game of Thrones*). In actual fact, the redditor had used artificial intelligence (AI) software to superimpose their faces onto the bodies of adult movie stars. Once Reddit got wind of this development, it banned deepfake porn and similar content from their platform. Despite this, the peculiar AI method of facial replacement quickly proliferated elsewhere on the internet, producing an ever-widening circle of actors, both technologically sophisticated and unsophisticated, capable of deploying it for a wide variety of purposes. By now, deepfake software is readily available as a free download (e.g., DeepFaceLab, DeepNude, FaceSwap, FakeApp, Zao), and hundreds of YouTube videos offer tutorials for the novice creator.

As for technology, deepfakes emerge from a specific type of *deep learning*—hence the term—in which sets of algorithms are pitted against one another in a generative adversarial network, or GAN (Goodfellow et al., 2014). Deep learning is a subset of AI in which sets of algorithms called

neural networks learn to infer rules and replicate patterns by analyzing large data sets. One algorithm, known as the generator, creates content modeled on source material (e.g., readily available pictures of a celebrity), while a second algorithm, the discriminator, seeks to detect flaws in the forgery. Executed in an iterative fashion, this method leads to rapid improvements addressing the flaws, allowing GANs to produce highly realistic yet fake video content. In essence, these AI-generated media fall into one of three categories: (1) face-swap, (2) lip-sync, and (3) real-time facial reenactment, also known as puppet-master, in which a target person is synthesized based on the input expressions (i.e., head positions, eye movements, facial expressions) from a source person sitting in front of a camera (Kim et al., 2018; Suwajanakorn et al., 2017; Thies et al., 2016). This method can be paired with the voice of an impersonating actor or a fully synthesized voice imitating the target person (Diakopoulos & Johnson, 2021). With proper post-processing, the resulting videos could be nearly undetectable—even with the aid of advanced forensic technology (Agarwal et al., 2019; Güera & Delp, 2018; Korshunov & Marcel, 2019).

Deepfakes are arriving at a perilous time. While the public's attention was exclusively in the hands of trusted media companies and professional journalists for much of the twentieth century, social networking sites and messengers such as Facebook, WhatsApp, or Telegram allow near instantaneous propagation of unverified content of almost any type. On these platforms, research found that people are especially prone to sharing negative and novel information, with false political stories being particularly effective in being spread (Vosoughi et al., 2018). With an unprecedented level of realism, speed, scale, and ability to personalize disinformation (Diakopoulos & Johnson, 2021), deepfakes could contribute to the broader problem of *fake news* on social media ("fabricated information that mimics news media content in form but not in organizational process or intent", Lazer et al., 2018, p. 1094; see also Egelhofer & Lecheler, 2019; Tandoc et al., 2018).

The Deepfake Detection Model

We assume that three sources of information can be used by recipients to detect deepfakes: context, audiovisual imperfections (i.e., technological glitches), and content. For the time being, the number of deepfakes that are circulating in the public is limited. And these videos are usually watched and shared because they are deepfakes. Thus, the deepfake is often accompanied by context information (paratext, Genette, 1987; Appel & Maleckar, 2012) signaling the manipulation. A popular deepfake video featuring former US President Barack Obama, for example, has often been embedded in journalistic pieces that deal with deepfakes. This context, however, may be missing more and more as technology progresses and a larger number of deepfakes are disseminated widely. Regarding deepfake indicators pertaining to the video itself, technical imperfections of the deepfake can indicate manipulation. Potential deepfake glitches are unnatural lip movements, asynchronicities of audio and lip movements, or traces of digital rendering. As a second indicator pertaining to the video itself, the content of the deepfake can evoke suspicion. A protagonist's actions and utterances in a deepfake can violate assumptions of what could reasonably be expected, leading to deepfake detection.

Importantly, we assume that deepfakes may remain undetected, even if relevant information (context, content, and/or technological glitches) is available for deepfake detection (see Figure 1). This assumption is based on classic reasoning models that distinguish between two cognitive systems (e.g., Bruner, 1986; Gawronski & Bodenhausen, 2006; Kahneman, 2011; Morewedge & Kahneman, 2010; Evans & Stanovich, 2013). System 1 is based on associative processes and intuitions. These are monitored by the slower, more effortful, and controlled System 2. Only the latter system is able to qualify incoming information as fabricated. Thus, the lower the activation of System 2, the more likely deepfake indicators should remain unnoticed. The distinction between both systems is illustrated by the following problem from the Cognitive Reflection Test (CRT; Frederick, 2005): "A bat and ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball

cost?" This problem elicits a fast, intuitive response (10 cents) that, upon reflection, cannot be the right answer (if the ball costed 10 cents, the bat would have to cost $1.10 and they would cost $1.20 in total). Coming up with the correct answer of 5 cents, however, requires most individuals to pause and reflect on the intuitively appealing response and eventually override it.

Individuals have a general tendency to suspend effortful cognitive processes and to rely on associative processes and intuitions (e.g., Pennycook et al., 2016, reported that 65% gave an incorrect answer to the 'bat and ball' problem). This general tendency is likely exacerbated when it comes to spotting falsified audiovisual content. Audiovisual content is met with credulity as videos are accepted almost axiomatically as accurate depictions of reality by journalists, politicians, jurisprudence, and everyday citizens (Farish, 2020; Fallis, 2021; Hancock & Bailenson, 2021). A study of Frenda et al. (2013), which illustrated the persuasive power of visual media, showed that doctored photos can lead people to believe having witnessed fabricated news events that never occurred. Given the fact that video is the superior medium in terms of persuasiveness ("seeing is believing"; e.g., Goldberg et al., 2019; Graber, 1990; Powell et al., 2015), the same is likely to apply to deepfake videos as well – unless System 2 is activated.

The use of more effortful and controlled System 2 processes (i.e., the propensity to engage in analytic thinking rather than relying on intuition) varies between individuals (e.g., Stanovich, 2012; Pennycook et al., 2016). In terms of CRT scores, 13% of web study participants gave correct answers to all three problems, 22% were correct twice, 25% had one answer correct, and 39% failed all three answers (Frederick, 2005; for the related concept of numeracy see Hutmacher et al., 2022). Recent research has linked individual differences in analytic thinking with skepticism about various epistemically suspect information, such as fake news (Pennycook & Rand, 2019; 2020), religious and paranormal beliefs (Pennycook et al., 2016; Shenhav et al., 2012), conspiracy theories (Swami et al., 2014), and pseudo-profound bullshit (Pennycook et al., 2015; Pennycook & Rand, 2020). Thus,

based on theory and results on fake news and related epistemically suspect information, we assume that the likelihood that deepfake indicators are processed, and deepfakes are identified as such, increases with participants' propensity to engage in analytic thinking.

In addition to a general proclivity to engage in analytic thinking, we assume that the thorough processing of information is a function of recipients' topic-specific motivation to do so. The motivational aspect of issue involvement is a key component to two-process models (e.g., Petty & Cacioppo, 1986) that overlap to a large degree with the two systems models outlined above. Whenever information is relevant to participants goals, beliefs, and interests, individuals tend to scrutinize a message more (e.g., Petty et al., 1983). In the field of political communication, individuals strongly interested in politics should therefore process political information more thoroughly than individuals less strongly interested in politics (e.g., Zaller, 1992). As a consequence, individuals who are more interested in politics should be better at identifying political deepfakes.

Current Evidence on Deepfake Detection

Empirical research on deepfakes is at its early stages and when our line of research was started in 2020, we were aware of only one study that focused on the processing of deepfakes (Vaccari & Chadwick, 2020). Since then, the number of studies has increased, but empirical evidence is still limited. We discuss the most relevant available studies in some detail, as they point out emerging questions and empirical paradigms. Much of the available research presented deepfakes of powerful and famous individuals, reflecting concerns that such videos could be particularly influential in distorting public opinion. Vaccari and Chadwick (2020) used a 2018 deepfake video created by Hollywood filmmaker Jordan Peele featuring former US President Barack Obama swearing (against Donald Trump) during a public service announcement. Overall, user responses to this video were mixed: Only 16% of the participants were misled by the deepfake while about twice as many were uncertain about the accuracy of its content (33.2%). The remaining 50.8%

were not deceived and correctly identified the video as a fake. The study was based on an experimental design; whereas two thirds of the participants saw the deceptive version (in a shorter or longer version), one third saw an educational version, in which it was revealed that it was Peele speaking, not Obama. Whereas the percentage of misled participants did not differ across conditions, the educational revelation reduced the percentage of those undecided (and increased the percentage of participants who were clearly not deceived). Being undecided about the epistemic status further mediated the effect of the treatment on trust in news on social media more generally: The video with the educational reveal increased trust in news on social media through reduced uncertainty.

Downstream effects of deepfake exposure in the political realm were examined more closely by Dobber et al. (2021). They presented a 5-second deepfake in which a well-known Dutch conservative politician ostensibly joked “But, as Christ would say: don’t crucify me for it” or an unmanipulated control clip in which no such assertion was made. Although not the focus of this investigation, answers to an open question suggested that only few participants identified the deepfake as such (even given the fact that the deepfake was imperfect, containing substantial technological glitches according to the authors). As predicted, the attitude towards the politician was more negative after watching the deepfake. Unexpectedly, being a Christian did not amplify the main effect. Regarding subgroups of Christians, the authors identified post-hoc that those who prayed very regularly (an indicator of religiousness) elaborated more on the content, but, it seems, not in a manner that was critical of the epistemic status of the deepfake: among Christians who were both very religious and past voters of the Christian party, the deepfake led to more negative attitudes towards the politician.

Deepfakes could be used to warrant the truth status of fake news – one of the arguably most relevant scenarios for the malicious use of deepfakes. Hwang et al. (2021) presented fake news about Marc Zuckerberg, who ostensibly stated that “the goal of Facebook was to control people” with or

without a warranting deepfake. In addition to the fake news article, the researchers showed one out of two educational videos (that informed about deepfakes or about disinformation more generally) or no video. Results suggest that the fake news article with deepfake was perceived to be more persuasive, more vivid, more credible, and led to higher sharing intentions than the fake news article without the deepfake. The educational videos (vs. control) reduced the dependent variable scores for the news article, irrespective of the presence of the deepfake.

Guided by the two-systems perspective outlined above, Köbis et al. (2021) followed a different experimental paradigm and presented 16 videos (8 deepfakes, 8 original) of unknown actors without political or ideological content. Participants had to indicate whether the video was a deepfake or not. They further received a warning about the potentially harmful consequences of deepfakes, a financial reward for a correct decision or no treatment (control). Participants performed a little above chance level and showed a bias toward guessing the video was authentic. The experimental treatment, however, did not affect detection accuracy.

In sum, the studies provide first evidence, that users tend to ascribe accuracy to deepfakes (Köbis et al., 2021), even if the deepfakes are imperfect (Dobber et al., 2021). They are further in line with the assumption that deepfakes shape recipients' thoughts and behavior (Dobber et al., 2021; Hwang et al., 2021) and that deepfakes may warrant information in a fake news article (Hwang et., 2021). Results on the effectiveness of treatments meant to foster critical scrutiny were mixed (Hwang et al., 2021; Köbis et al., 2021).

Study Overview and Predictions

Our theory and research are meant to contribute to the available literature by specifying deepfake indicators and introducing individual difference variables that predict the extent to which these indicators are used to detect a deepfake. In line with extant research (Dobber et al., 2021; Vaccari & Chadwick, 2020), we focused on the detection of deepfakes of well-known politicians

(UK's prime minister Boris Johnson, Studies 1-3, or former US president Barack Obama, Study 2). We assumed that analytic thinking predicts the detection of deepfake videos: Individuals who are more willing to engage effortful, deliberative, and reflective cognitive processes should be more likely to spot a fake video. Following our deepfake detection model, this increased detection should be due to a higher likelihood to identify fake content and to identify technological glitches (such as poor lip synch). We further expected that in the political realm, individuals who are more strongly interested in politics were more likely to identify a political deepfake, because higher interest in politics would motivate individuals to elaborate on and discern fake content. Finally, we assumed that fake news that discredit a politician and are warranted by a deepfake yield higher accuracy ratings than the same fake news that are not warranted by a deepfake (see also Hwang et al., 2021). This should lead to downstream effects on the politician's perceived leadership skills. The influence of the deepfake that accompanies the fake news article should decrease with analytic thinking propensity and interest in politics. We approached the topic empirically with an open-ended thought listing task (Study 1), closed-ended questions addressing the deepfake detection indicators (Study 2), and an experiment in which a deepfake video was embedded in a fake news story (Study 3). All three studies were conducted with German samples. The stimuli, preregistrations, data, codes, and supplementary material can be found at https://osf.io/fqdk4.

Study 1

Study 1 was based on a thought-listing task, an established method to collect cognitive responses to a media stimulus. A virtue of this method is that the instruction does not mention fakery. Like in relevant everyday contexts, participants are therefore not reminded or forewarned that an upcoming video could be fake. In this study, participants watched a deepfake video in which Boris Johnson gave a speech with delicate political content. After the video ended, participants were asked to list their thoughts, and these thoughts were content-analyzed with respect to the detection of the

deepfake (see Woelke & Pelzer, 2020). We predicted that the higher the participants' analytic thinking scores, the more likely participants were to suspect or explicitly state that the Boris Johnson deepfake-video was fake. We further predicted that the higher the participants' interest in politics, the more likely they were to notice the deepfake.

Method

Participants and Procedure

The number of participants was determined a priori using G*Power (Faul et al., 2009). The required sample size to detect a small to medium-sized point biserial correlation between analytic thinking and the identification of the deepfake video as fake amounted to 191 (two-tailed, given α = .05, power = .80, rho = .20). Accounting for potential careless responding, a total of 236 people from the German cloudworking service Clickworker were recruited to complete the online study (compensation €1.30 [$1.50]). Participants were removed from the final sample if they indicated to have responded carelessly by needing 180 seconds or less for the questionnaire (n = 17), answered the control question ("What is 11 + 8?") incorrectly (another 4 participants), reported that they used a search engine while participating (10 participants) or knew the video beforehand (6 participants).[1] The remaining sample consisted of 199 participants (48.7% female) between the ages of 18 and 69 years (M = 35.24, SD = 11.24). The percentage of participants excluded amounted to 15.6%. The literature on the reliability and validity of data obtained from crowdsourcing panels (such as Clickworker) emphasize the necessity of thorough data screening for low diligence participants (e.g., Chmielewski & Kucker, 2020; Kennedy et al., 2020). Low response times, which led to most exclusions in this study, are a particularly important criterion for excluding low-diligence responding (e.g., Curran, 2016). Our exclusion rate falls well within the margin observed in prior studies with similar participant pools (see Thomas et al., 2017; Chandler et al., 2019).

The survey started with demographic questions, including the political interest item. The deepfake video and the thought listing task followed. Next, analytic thinking and the control questions were assessed. Finally, participants were debriefed about the true nature of the video and the phenomenon of deepfakes, and they were provided with a link to the Future Advocacy initiative for further information.

Deepfake Stimulus

We aimed for a video that communicated a politically relevant message by a well-known politician. Given that the creation of a novel, high-quality deepfake video was beyond the scope of this project, we chose a pre-fabricated video. The vast majority of deepfakes that have been published are meant to be funny and have no political message. These videos did not meet our requirements. An even greater limitation was the fact that most online material falls short of sufficient production quality. Ultimately, we opted for a deepfake video from the research organization Future Advocacy (https://futureadvocacy.com/deepfakes) that was released in the run up to the 2019 UK general election (BBC News, 2019).[2] Not only did it meet the above criteria but it was also largely unknown at the time the study was conducted (just over 3,000 views on YouTube). The deepfake video shows British Prime Minister Boris Johnson sitting in his 10 Downing Street office, speaking directly to a camera and falsely endorsing his former opponent, ex-Labour Party leader Jeremy Corbyn, for prime minister (see Figure 2). We only used the first half of the original deepfake in this study (the second half contained gibberish).[3]

Measures

Thought Listing and Coding Procedure. After watching the video, participants were asked to write down in at least one or two sentences what went through their mind while watching. A free space to list the thoughts was provided. The participants' thoughts encompassed 15 words on average, with a range between 2 and 56 words. For coding the responses, we specified a thought unit

as the complete set of assertions made by each participant. Two independent coders coded the data after a training session led by the first author and the codebook was revised slightly after the session. Our main question was whether or not the participant identified the video as fake or staged. Three categories were provided, 1) no, 2) assertion that the video was fake, 3) assertion or question stating the possibility that the video was fake / doubting the authenticity of the video (Cohen's $\kappa = .84$). For the main analyses, categories 2) and 3) were collapsed. We further coded the reason underlying the authenticity doubts / fake assertion, providing categories for content (e.g., "he would have never said it"), technical issues (lip synch), knowledge of this specific deepfake, and combinations of these reasons (Cohen's $\kappa = .79$). We additionally coded evaluative statements regarding the politician (not further analyzed) and comments on comprehension. Twelve participants stated that they had problems to understand what Johnson was saying and among these, five participants mentioned difficulties associated with the English language / his pronunciation. We concluded that the video was intelligible for our German participants.

Analytic Thinking. Participants completed seven items from two versions of the Cognitive Reflection Test (CRT)—a set of questions with intuitively compelling but incorrect answers. The CRT has been shown to retain its predictive validity across time, and people continue to incorrectly respond to the same items after multiple exposures (Bialek & Pennycook, 2018; Meyer et al., 2018; Stagnaro et al., 2018). The seven items version of the scale included the original three-item CRT (Frederick, 2005) and the four-item non-numeric CRT by Thomson and Oppenheimer (2016). This seven item-scale is a common operationalization of analytic thinking (e.g., Pennycook & Rand, 2020) and was translated to German. Questions were presented in open-ended format and correct responses were summed up to create a score for each participant (minimum = 0, maximum = 7). Higher scores represent greater analytic thinking or a more analytic cognitive style (Cronbach's $\alpha = .67$, $M = 4.01$, $SD = 1.81$).

Political Interest. The participants' political interest was measured with the help of a single self-report item ("How interested are you in politics?"), as it has frequently been employed by the American National Election Studies (ANES) or the German Longitudinal Election Study (GLES; Prior, 2018). Response options ranged from *not at all* (1) to *extremely* (5) on a five-point scale (M = 3.48, SD = 0.85).

Results and Discussion

Out of the 199 participants, 82 (41.2%) stated or suspected that the video was fake. Among this group, 30 mentioned the video content as a reason for their judgment, 21 mentioned technical aspects, and 20 participants mentioned both. We expected that analytic thinking and political interest would increase the likelihood that the deepfake was identified as fake. Both continuous variables were entered in a logistic regression equation with deepfake identification (0 = no; 1 = yes) as the criterion. Both predictors taken together explained deepfake identification above chance, model fit, $\chi^2(2) = 10.88$, $p = .004$, Nagelkerke pseudo $R^2 = .072$. Analytic thinking was positively associated with the odds that the deepfake was discovered as such, $B = .21$, $SEB = .08$, Wald(1) = 6.13, $p = .013$, $OR = 1.23$. In a similar vein, political interest was positively associated with the odds of deepfake detection, $B = .40$, $SEB = .18$, Wald(1) = 5.05, $p = .025$, $OR = 1.49$ (see Figure 3). We further tested a model with an interaction between both predictors, which yielded no significant additional explanation, $B = -.02$, $SEB = .10$, Wald(1) = 0.04, $p = .845$, $OR = 0.98$.

In sum, using an open-ended question, around 40% of our participants spontaneously identified the deepfake video correctly to be fake. The results were in support of our deepfake detection model. As expected, the more individuals engage in analytic thinking the higher the likelihood of deepfake detection. This result is in line with recent research on fake news detection (Pennycook & Rand, 2019; 2020). Moreover, political interest increased the likelihood that the political deepfake was detected.

Study 2

In our second study, we examined deepfake detection using two stimulus videos and closed-ended questions. To gain deeper insight into the deepfake detection components specified in our model, we asked for the deepfake indicators of content and technological glitches. We assumed that analytic thinking and political interest predicted the detection of deepfakes. Moreover, individuals with higher analytic thinking scores should be more likely to identify fake content and to identify technological glitches (such as poor lip synch). Individuals with higher political interest scores should be more likely to identify deepfakes based on fake content. Finally, we were interested in a comparison between the two videos presented.

Method

Participants and Procedure

The required sample size to detect a small to medium-sized Pearson correlation between analytic thinking and our continuous deepfake identification measure amounted to 193 (two-tailed, given $\alpha = .05$, power = .80, rho = .20, G*Power, Faul et al., 2009). Accounting for potential careless responding, a total of 251 people from the same German crowdworking service as in Study 1 were recruited (compensation €1.30 [$1.50]). Participants were excluded if they answered the control question ("What is 11 + 8?") incorrectly (one participant), indicated to have responded carelessly by needing 240 seconds or less for the questionnaire (n = 9), or reported to have used a search engine while participating (n = 8**)**. The remaining sample consisted of 233 participants (38.2% female, 1.7% non-binary) between the ages of 18 and 69 years (M = 34.86, SD = 12.41). The study started with demographics, the political interest item, and the CRT. Next, both videos along with the dependent measures were presented. At the end of the survey, participants rated their political orientation and

answered the control questions. Participants were debriefed about the phenomenon of deepfakes and the fabrication of both videos specifically, including relevant links for further information.

Experimental Stimuli

We presented the Johnson deepfake of Study 1 and the Obama deepfake that was used in prior research (Vaccari & Chadwick, 2020). We presented the video without the educational reveal that follows the deepfake sequence. After each video, the dependent variables were assessed. Both videos were shown in random order (for 117 participants the Obama-video came first, for 116 participants the Johnson-video came first).

Measures

Analytic Thinking and Political Interest. The same measures as in Study 1 were used. The seven-item CRT had acceptable reliability (Cronbach's $\alpha = .69$, $M = 3.72$, $SD = 1.91$). Political interest averaged 3.54 ($SD = 0.81$) on the five-point scale.

Deepfake Detection. We introduced our dependent variables as follows: "Nowadays computer technology allows the manipulation of videos. For media users it is important to distinguish between manipulated and non-manipulated videos." The deepfake detection item ("How do you judge this video") went with a seven-point scale ranging from 1 (not technically manipulated) to 7 (technically manipulated). The scale midpoint was labelled *undecided*. We further asked how confident participants were in their latter judgment on a seven-point scale ranging from 1 (very unsure) to 7 (very sure). If participants scored 2 or higher on the deepfake detection item, they were asked to indicate what made them think the video was technically manipulated. Two assertions followed, one focused on technical issues ("Something was wrong technically with visuals sound [e.g. the lips moved unnaturally, were out of synch or the facial expressions were off"]), one focused on the deepfake's content ("The content, i.e., what the politician said, made me think that the video was manipulated"). Both items went with a seven-point scale ranging from 1 (do not agree) to 7

(fully agree). Participants who scored 1 on the deepfake detection item automatically received a score of 1 on the content and technical issues items. A final item was a forced choice between the Obama and the Johnson deepfake, asking "If you were required to decide between both videos, which of the videos was technically manipulated?".

Additional measures. Participants were further asked to indicate their political orientation on an eleven-point Likert-scale ranging from 1 = *left* to 11 = *right* (M = 5.09, SD = 1.83; "In political matters people talk of 'the left' and 'the right'. Where would you put yourself?"). We further asked for each video whether participants knew of or had seen the video before. An additional control items asked whether the participants had used a search engine while answering the questionnaire.

Results and Discussion

Twelve participants indicated that they had known of or seen the Obama video before whereas one participant indicated that they had known of or seen the Johnson video before. Thus, in our German sample, both videos were new to the great majority of the participants. We first inspected to what extent the deepfake videos were detected as such. On the scale ranging from not technically manipulated (score 1) to technically manipulated (7), the Johnson video received an average rating of 5.34 (SD = 1.70) which was significantly above the scale midpoint $t(232) = 12.06$, $p < .001$. The Obama video scored 4.91 (SD = 1.99) which was also above the scale midpoint $t(232) = 6.95$, $p < .001$. The detection ratings for both videos were slightly positively associated, $r(231) = .173$, $p = .008$). The Johnson video received a significantly higher average detection rating overall, $t(232) = -2.80$, $p = .005$, $d = .197$. The forced decision measure showed that there were more participants who thought that the Johnson rather than the Obama video was a deepfake (134 vs. 99, binomial test $p = .026$).

We predicted that the higher the participants' analytic thinking scores and the higher participants' interest in politics, the more likely they were to identify both deepfake videos correctly

as technologically manipulated. We excluded individuals who knew the respective video beforehand[4] and conducted a regression analysis with analytic thinking and political interest as predictors and the continuous deepfake detection measure as the criterion. For the Boris Johnson video, analytic thinking (B = .13, SEB = .06, p = .026) but not political interest (B = .13, SEB = .14, p = .338) predicted deepfake detection. For the Barack Obama video, neither analytic thinking (B = .02, SEB = .07, p = .835) nor political interest (B = -.06, SEB = .17, p = .710) predicted deepfake detection. When the scores for both videos were averaged, neither analytic thinking (B = .08, SEB = .05, p = .106) nor political interest (B = .08, SEB = .11, p = .486) were significant predictors. Interactions between both predictors were non-significant. Potential carry-over effects were examined by conducting additional analyses in which the order of presentation (Obama-video first or Johnson video first) served a moderating factor. As shown in Supplement S4, no interactions emerged.

In sum, our findings provided mixed support for the association found in Study 1. Whereas the role of analytic thinking for deepfake detection was supported (at least for the Johnson deepfake), political interest did not increase deepfake detection.

Our next analyses focused on the detection indicators, content and technological glitches. We assumed that participants' analytic thinking scores and interest in politics increased the likelihood that the deepfake was identified based on content. Again, we excluded individuals who knew the respective video beforehand and conducted a regression analysis with analytic thinking and political interest as predictors and the extent to which the video content made participants believe the video was fake as the criterion. For the Boris Johnson video, analytic thinking (B = .21, SEB = .06, p = .001) but not political interest (B = .03, SEB = .15, p = .845) predicted the content indicator DV. For the Barack Obama video, neither analytic thinking (B = .00, SEB = .07, p = .958) nor political interest (B = -.08, SEB = .17, p = .639) were significant predictors. When the scores for both videos were averaged, deepfake identification due to technology increased with analytic thinking scores (B

= .12, *SEB* = .05, *p* = .018), but not with political interest (*B* = .00, *SEB* = .11, *p* = .986). We further assumed that participants' analytic thinking scores increased the likelihood that the deepfake was identified based on technological glitches, like poor lip synch. However, both variables were unrelated for the responses averaged across both videos, *r*(231) = .027, *p* = .680 (see Supplement S3). Again, potential carry-over effects were examined, no interactions with the order in which the videos were presented emerged (Supplement S4).

In an exploratory analysis, we examined indirect effects between analytic thinking and the identification of deepfakes with the conspicuous content recognition measure and the technological glitches recognition measure as parallel mediators. Data for both videos were averaged. We used PROCESS version 3.5 (Hayes, 2018) for our mediation analyses, model 4 with default settings. A significant indirect effect, describing the path via the content identification emerged, effect estimate = .077, *SE* = .033, 95%CI[.015; .143]. Technological glitches identification did not serve as a significant path in the model, effect estimate = .005, *SE* = .012, 95%CI[-.019; .028]. Thus, we found that analytic thinking yielded an indirect effect on deepfake detection through implausible content identification whereas technological glitches detection was no alternative pathway.

Study 3

Deepfakes in the political realm could be particularly harmful in combination with fake news stories (Chesney & Citron, 2019). In Study 3, we randomly assigned participants to read a fake news story that was supported by a deepfake or to read the same fake news story without the deepfake (see also Hwang et al., 2021). We assumed that exposure to a deepfake video would increase the perceived accuracy of a related false news article, and that it would decrease the perceived leadership skills of a politician who is negatively portrayed in the deepfake (we used the incriminating deepfake of Boris Johnson as in Studies 1 and 2). However, given that a substantial part of the participants in Studies 1 and 2 correctly identified the deepfake video as such, a null effect or even a backfire effect were

plausible alternative predictions. Importantly, we assumed that analytic thinking and political interest would reduce the effect of the deepfake. Given that perceived accuracy of the fake news could predict perceived leadership skills, the effects outlined above constitute a moderated mediation assumption.

Method

Participants

The required sample size to detect a small to medium effect of $f^2 = .10$ was determined a priori with G*Power (Faul et al., 2009) to be at least 432, given $\alpha = .05$ and power $(1-\beta) = .90$. Accounting for potential careless responding, a total of 500 people from the German cloudworking service Clickworker.de were recruited to complete the online study (compensation €1.50 [$1.62]). Participants were removed from the final sample if they indicated to have responded carelessly by needing less than 240 seconds for the questionnaire ($N = 16$), using a search engine during the study ($N = 21$), or had seen the deepfake video before ($N = 11$). Seven participants further failed to correctly answer the control question ("What is 11 + 8?"), leading to an effective sample size of 445 participants (42.5% female) between the ages of 16 and 71 ($M = 37.72$ years, $SD = 12.30$).

Experimental Stimuli

The deepfake video of Boris Johnson was used. A plausible background story was created and presented in the format of a news article (see Online Supplements S5 and S6). According to the ostensible news, the video was leaked by a whistleblowing website shortly before the study was conducted. After a brief description of the video, the article went on to say that, while the exact circumstances were still unclear, the video might have been recorded at an earlier time when Johnson's political situation was less favorable (i.e., when his Brexit negotiations with the EU seemed to fail) and was never intended to be released. Johnson ostensibly decided against publication a few days after recording. The news text consisted of 265 words.

Measures

Analytic Thinking and Political Interest. The seven-item CRT had acceptable reliability (Cronbach's α = .71, *M* = 3.71, *SD* = 1.95), the average score for political interest was 3.59 (*SD* = 0.83).

Perceived Accuracy. The perceived accuracy of the presented news article was assessed using three bipolar items. One item followed the wording of Pennycook et al. (2018), while two more items were added for the sake of this study. Responses were given on a seven-point Likert-scale (1 to 7) with the extremes: *inaccurate – accurate*, *untrue – true*, and *did not happened as described – happened as described* (α = .93, *M* = 3.05, *SD* = 1.51).

Perceived Leadership Ability. Participants rated their perception of Boris Johnson's leadership ability ("How would you rate Boris Johnson's leadership?") on a seven-point scale (1 to 7) consisting of five bipolar trait items (*weak-minded – decisive*, *not capable of strong leadership – capable of strong leadership*, *incompetent – competent*, *unreliable – reliable*, and *ineffective – effective*). Attributes on the positive side of the scale originated from prior research on relevant personality traits of political leaders in terms of overall candidate evaluations and individual vote choice (Clarke et al., 2004; Funk, 1999; Mondak, 1995). The scores of the five items were averaged. The reliability of the scale was good (α = .88, *M* = 3.64, *SD* = 1.31).

Procedure

After sociodemographic questions, the interest in politics measure, and the seven-item CRT were administered, participants were randomly assigned to the control condition (news article only) or the deepfake condition in which the deepfake video was presented prior to the news article. Participants had to stay on the survey page for at least the duration of the video. Once watched, the video could not be replayed. Participants of both groups were subsequently instructed to rate the perceived leadership of Boris Johnson and the perceived accuracy of the news article. At the end of

the survey, participants rated their political orientation, answered the control questions and were debriefed that the news story was entirely fabricated and that the video was a so-called deepfake. All participants were provided with a link to the Future Advocacy initiative for further information. Zero-order correlations of the main variables and descriptives per condition are shown in Table S7 of the Online Supplement.

Results and Discussion

Deepfake Main Effects

Two between-subjects analyses of variance (ANOVAs) were conducted to compare the effect of the deepfake on perceived accuracy and perceived leadership, respectively. Participants in the deepfake condition (M = 3.13; SD = 1.54) and in the control condition (M = 2.98; SD = 1.49) did not differ in the perceived accuracy of the fake news article, $F(1, 443) = 1.12$, $p = .290$, $\eta_p^2 < .01$. Likewise, the perceived leadership ability of Boris Johnson was unaffected by the Deepfake treatment (M = 3.68; SD = 1.33, control condition M = 3.60; SD = 1.28), $F(1, 443) = 0.38$, $p = .537$, $\eta_p^2 < .01$.

Analytic Thinking, Political Interest, and the Interplay with Deepfakes

To examine the influence of analytic thinking and political interest, separate hierarchical (two-step) regression analyses were conducted. In the first step, the experimental condition as a categorical predictor (dummy coded with 0 = no deepfake; 1 = deepfake) and either the analytic thinking score or political interest (continuous, z-standardized; Aiken & West, 1991) were included. The second step entailed the respective interaction term.

In our first hierarchical regression analysis with perceived accuracy as the criterion, we first entered the factor deepfake and analytic thinking as predictors: Analytic thinking turned out to be a significant predictor of perceived accuracy, B = -.14, SEB = .07, p = .045. The higher participants' analytic thinking the lower the perceived accuracy of the fake news article. There was, however, no

interaction between the deepfake and analytic thinking on the perceived accuracy of the news article, B = -.18, SEB = .14, p = .200, $\Delta R^2 < .01$ (Online Supplement S8). Associations between analytic thinking and accuracy amounted to r (222) = -.153, p = .022 in the deepfake condition, and to r (219) = -.038, p = .576 in the control condition. Given the absence of a significant interaction, we do not appraise these slopes as supporting the hypothesis that analytic thinking reduces the effect of the fake news article on accuracy ratings specifically deepfake condition.

Regarding the perceived leadership ability of Boris Johnson as the criterion, the deepfake factor and analytic thinking were entered first, but no significant effect of analytic thinking was found, B = .07, SEB = .06, p = .285. We expected analytic thinking to reduce the negative effect of the deepfake video on the perceived leadership of Boris Johnson, but the interaction between both predictors did not yield a significant effect, B = .03, SEB = .13, p = .824, $\Delta R^2 < .01$. Thus, no support for our assumptions regarding perceived leadership ability was found.

We repeated the analyses with political interest as an individual difference factor (Online Supplement S9). In the first step of the hierarchical regression, political interest predicted perceived accuracy overall, B = -.40, SEB = .07, $p < .001$. There was, however, no interaction between the deepfake and political interest on perceived accuracy of the news article; B = .10, SEB = .14, p = .456, $\Delta R^2 < .01$. Political interest reduced the perceived accuracy of the fake news article, irrespective of the presence of the deepfake. Regarding the perceived leadership ability of Boris Johnson, no significant effect of political interest was found, B = -.05, SEB = .06, p = .447. There was also no interaction between the deepfake manipulation and political interest, B = -.09, SEB = .13, p = .464, $\Delta R^2 < .01$.

In sum, analytic thinking and political interest were associated with lower accuracy ratings in response to a false news article, irrespective of exposure to the deepfake video. This result supports previous studies that found analytic thinking to predict a more critical stance towards fake news and

other post-truth phenomena (e.g., Pennycook & Rand, 2019; 2020). Perceived leadership abilities of Boris Johnson were independent of the deepfake manipulation and our individual difference measures.

General Discussion

The advent of deepfake technology has marked a turning point in the ability to distort reality (Chesney & Citron, 2019). Whereas video footage was rightfully considered solid evidence in the past, we are now faced with new generations of software that apply deep learning algorithms to manipulate audiovisual content. Deepfakes could be a distressing factor in many fields, most notably in the political realm in which audio and video footage have been pivotal to the fate of politicians and governments (e.g., *the 2019 Austrian Ibiza Affair,* Oltermann, 2019; see Fallis, 2021). The emergence of the 2022 Zelenskyy deepfake has demonstrated the current relevance of the deepfake topic. Understanding and explaining user responses to deepfakes is a key challenge to the social sciences and a first step at preparing citizens to better deal with this post-truth phenomenon (Lewandowsky et al., 2017). In the deepfake detection model presented here, we assume that deepfakes can be identified with the help of three clusters of indicators, context, content, and technological glitches. Importantly, we further assume that the use of these indicators is associated with the activation of systematic, effortful processing (System 2, cf. Kahneman, 2011; Gawronski & Bodenhausen, 2006).

Our first two studies show that, in the absence of relevant context such as prior knowledge about the deepfake or its source, content and technological glitches are used to detect deepfakes. Across both studies, individuals who regularly engage in analytic thinking had a higher likelihood of successfully detecting the deepfake. This result extends prior research that associated analytic thinking with higher scrutiny towards other post-truth phenomena, such as fake news (Pennycook & Rand, 2019; 2020) and conspiracy theories (Swami et al., 2014). Study 2 showed that individuals

with higher analytic thinking scores were more likely to take note of suspicious content of the deepfake shown, whereas analytic thinking was unrelated to recognizing technological imperfections. In the future, the development of new generations of deep learning software will likely decrease the prevalence of technological imperfections, which will put even more weight on content and context as deepfake indicators. Although our results on political interest were a bit more mixed, we believe that in addition to general individual differences, interests (and potentially knowledge) concerning a deepfake's topic can increase the likelihood of deepfake detection.

Study 3 showed that the perceived accuracy of a fake news article was negatively associated with analytic thinking and political interest. These relationships were found irrespective of the presence of a deepfake that warranted the assertions made in the fake news piece. Moreover, including the deepfake did not increase acceptance or influence the perceived leadership abilities of the politician. The lack of an effect of including a deepfake to fake news contrasts with our assumptions and the results of a study by Hwang et al. (2021). Future research is encouraged to identify under which conditions deepfakes embedded in a fake news story increase the perceived accuracy of the story, as compared to fake news without a deepfake.

Our theory and research could be a starting point for developing countermeasures against the acceptance and spread of deepfakes:

1. As indicated across three studies, participants who are inclined to think analytically rather than trusting intuition have a higher likelihood to use deepfake indicators effectively and to correctly identify deepfakes (Studies 1 and 2) or to classify a fake news article to be inaccurate (Study 3). Study 2 showed that analytic thinking was particularly linked to identifying the deepfake through suspicious content (see Pennycook & Rand, 2020, for a similar conclusion about the identification of fake news). Political interest was associated with identifying a deepfake (Study 1) and to ascribing lower accuracy to a fake news article (Study 3). Media education and training as well as measures to

increase interest in politics could encourage citizens to reflect habitually on audiovisual content against the background of deepfake technology, rather than accepting, liking, and sharing videos intuitively.

2. We believe that citizens require context information on the general phenomenon of deepfakes as well as information on specific deepfakes. To this end, citizens should be informed of the possibility of increasingly circulating deepfakes, due to developments in AI. This should be accompanied with information on the likely motivations underlying deepfake creation and spreading, and the sources that tend to disseminate fake media content (cf. Lazer et al., 2018; Lewandowsky & van der Linden, 2021). As soon as specific deepfakes are identified, this knowledge should be made available to citizens (e.g., via mass media and on fact-checker websites such as snopes.com) and to social media companies. This recommendation is based on our deepfake detection model, but we acknowledge that context information had not be an empirical focus in our set of studies. Please note that one available empirical study demonstrated that forewarnings are able to combat the influence of deepfakes (Hwang et al. 2021) whereas another one showed no effect (Köbis et al., 2021). Recent research on the related topics of fake news and conspiracy theories demonstrated the effectiveness of inoculation or prebunking interventions that take place prior to exposure to the misinformation (e.g., Lewandowsky & van der Linden, 2021). As a caveat, we acknowledge that warning citizens about deepfakes could unintendedly increase distrust regarding real content (e.g., Carey et al., 2020, on such effects when correcting false assertions; but see Pennycook et al., 2020). In the introduction of this manuscript, the case of the March 2022 Zelenskyy deepfake was mentioned. Prior to the actual dissemination of the deepfake, government agencies had warned Ukrainians about the upcoming deepfake (Metz, 2022). At this point, little can be said about the role this prebunking information played in the swift and widespread identification and condemnation of this deepfake. More research on providing effective prebunking of deepfakes is certainly warranted.

3. For the time being, many circulating deepfakes with political content are imperfect and allow citizens to identify deepfakes by spotting technological glitches. Thus, it would be helpful to encourage citizens to pay close attention to likely imperfections. That said, software that enable the creation of deepfakes by applying deep learning algorithms are getting better and more efficient. In the next years, non-professionals could be able to create deepfakes that are technically flawless. Thus, technological glitches may become a less reliable deepfake detection indicator in the future.

This is one of the first sets of studies to examine the psychology underlying responses to deepfakes in the political realm. Our work, however, is not without its limitations. First, our work had to rest on stimuli that were available in 2020 when this series of studies started. Due to a scarcity of well-made political deepfakes that convey a politically relevant message, our results are largely based on one video, the Boris Johnson deepfake. Although we do not expect that the main findings would have differed for different politicians or topics, future research is encouraged to profit from the growing number of deepfakes that are available to researchers (see, for example, Sankaranarayanan et al., 2021). With technological advances, testing a larger selection of videos (featuring different politicians and covering different topics) will be possible in future studies. On a related note, we acknowledge that the deepfakes chosen (most notably the Obama deepfake) may be perceived as satirical, possibly funny, and entertaining, particularly in the case that the deepfake is identified as such.

Second, our results are based on self-report deepfake detection measures. There is an inherent flaw in mentioning manipulation in a question, given that providing this cue likely allocates attention to the possibility of deception, leading to inflated detection ratings among many respondents. We tried to minimize this potential method bias by omitting explicit questions about manipulation in Studies 1 and 3. To reduce this method bias in Study 2, the distinction between manipulated and original videos was introduced as our study goal. Future research could make use of unobtrusive

measures to identify respondents' identification of deepfakes in the lab. Ideally, these measures could be used to measure detection in the moment detection indicators become salient.

Conclusion

Political deepfakes are a challenge to democracies. Citizens can and do use deepfake indicators based on context, content, and technological imperfections. Analytic thinking and political interest are associated with the identification of deepfakes.

Data Availability Statement

The data that support the findings of this paper are openly available in https://osf.io/fqdk4/.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 38–45). IEEE. https://perma.cc/2bcs-m8j3

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Appel, M., & Malečkar, B. (2012). The influence of paratext on narrative persuasion: Fact, fiction, or fake? *Human Communication Research, 38*(4), 459-484. https://doi.org/10.1111/j.1468-2958.2012.01432.x

BBC News (2019). *The fake video where Johnson and Corbyn endorse each other*. https://perma.cc/zw3t-4uhs

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Bruner, J. (1986). *Actual minds, possible worlds.* Harvard University Press.

Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B., & Zeitzoff, T. (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science Advances*, *6*(5), eaaw7449. doi: 10.1126/sciadv.aaw7449

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods, 51*(5), 2022-2038.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753–1820. https://doi.org/10.2139/ssrn.3213954

Clarke, H. D., Sanders, D., Stewart, M. C., & Whiteley, P. (2004). *Political choice in Britain*. Oxford University Press.

Coats, D. R. (2019). Worldwide threat assessment of the US intelligence community. *Senate Select Committee on Intelligence*. https://perma.cc/y673-3ep4

Cole, S. (2018). *We are truly fucked: Everyone is making ai-generated fake porn now*. VICE. https://perma.cc/mwn2-z655

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

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

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

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29. https://doi.org/10.1038/s44159-021-00006-y

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97-116. https://doi.org/10.1080/23808985.2019.1602782

Evans, G., Heath, A., & Lalljee, M. (1996). Measuring left-right and libertarian-authoritarian values in the British electorate. *British Journal of Sociology*, *47*(1), 93–112. https://doi.org/10.2307/591118

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology, 34*(4), 623-643. https://doi.org/10.1007/s13347-020-00419-2

Farish, K. (2020). Do deepfakes pose a golden opportunity? Considering whether English law should adopt California's publicity right in the age of the deepfake. *Journal of Intellectual Property Law & Practice*, *15*(1), 40–48. https://doi.org/10.1093/jiplp/jpz139

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, *70*(4), 455–471. https://doi.org/10.1353/tj.2018.0097

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. *Journal of Experimental Social Psychology*, *49*(2), 280–286. https://doi.org/10.1016/j.jesp.2012.10.013

Funk, C. L. (1999). Bringing the candidate into models of candidate evaluation. *The Journal of Politics*, *61*(3), 700–720. https://doi.org/10.2307/2647824

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692-731. https://doi.org/10.1037/0033-2909.132.5.692

Genette, G. (1987). *Seuils.* Edition du Seuil. (English language version: Genette, G. [1997] Paratexts. Thresholds of interpretation. Cambridge University Press).

GitHub (2020). *DeepFaceLab*. https://github.com/iperov/DeepFaceLab

Goldberg, M. H., van der Linden, S., Ballew, M. T., Rosenthal, S. A., Gustafson, A., & Leiserowitz, A. (2019). The experience of consensus: Video as an effective medium to communicate scientific agreement on climate change. *Science Communication*, *41*(5), 659–673. https://doi.org/10.1177/1075547019874361

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2672–2680). MIT Press.

Graber, D. A. (1990). Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication, 40*(3), 134–156. https://doi.org/10.1111/j.1460-2466.1990.tb02275.x

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE. https://doi.org/10.1109/avss.2018.8639163

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149-152. https://doi.org/10.1089/cyber.2021.29208.jth

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review*, *17*(1), 99–128. https://perma.cc/5xs6-gejq

Hayes, D. (2010). Trait voting in US senate elections. *American Politics Research*, *38*(6), 1102–1129. https://doi.org/10.1177/1532673x10371298

Hutmacher, F., Reichardt, R., & Appel, M. (2022). The role of motivated science reception and numeracy in the context of the COVID-19 pandemic. *Public Understanding of Science*, *31*(1), 19-34. https://doi.org/10.1177/09636625211047974

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188-193. https://doi.org/10.1089/cyber.2020.0174

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics (TOG)*, *37*, 1–14. https://doi.org/10.1145/3197517.3201283

Köbis, N., Doležalová, B., & Soraperra, I. (2021). Fooled twice–people cannot detect deepfakes but think they can. *iScience, 24* (11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Korshunov, P., & Marcel, S. (2019, June). Vulnerability assessment and detection of deepfake videos. In *The 12th IAPR International Conference on Biometrics (ICB)* (pp. 1–6). IEEE. https://doi.org/10.1109/icb45273.2019.8987375

Lazer, D. M., 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, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348-384. https://doi.org/10.1080/10463283.2021.1876983

Mervosh, S. (2019, May 24). *Distorted videos of Nancy Pelosi spread on Facebook and Twitter, helped by Trump*. The New York Times. https://perma.cc/a26h-4pf3

Metz, R. (2022, March 16). Facebook and YouTube say they removed Zelensky deepfake. *CNN*. https://edition.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html

Meyer, A., Zhou, E., & Shane, F. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making*, *13*(3), 246–259.

Mondak, J. J. (1995). Competence, integrity, and the electoral success of congressional incumbents. *The Journal of Politics*, *57*(4), 1043–1069. https://doi.org/10.2307/2960401

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences, 14*(10), 435–440. https://doi.org/10.1016/j.tics.2010.07.004

Oltermann, P. (2019, May 20). Austria's 'Ibiza scandal': what happened and why does it matter? *The Guardian.* Retrieved from https://www.theguardian.com/world/2019/may/20/austria-ibiza-scandal-sting-operation-what-happened-why-does-it-matter

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944-4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563. https://journal.sjdm.org/15/15923a/jdm15923a.pdf

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*, 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2016). Atheists and agnostics are more reflective than religious believers: Four empirical studies and a meta-analysis. *PloS One*, *11*, 1–18. https://doi.org/10.1371/journal.pone.0153039

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 123–205). New York: Academic Press.

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, *10*(2), 135-146. https://doi.org/10.1086/208954

Powell, T. E., Boomgaarden, H. G., De Swert, K., & de Vreese, C. H. (2015). A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication*, *65*(6), 997-1017. https://doi.org/10.1111/jcom.12184

Prior, M. (2018). *Hooked: How politics captures people's interest*. Cambridge University Press.

Sankaranarayanan, A., Groh, M., Picard, R., & Lippman, A. (2021, August). *The presidential deepfakes dataset.* Paper presented at the workshop 'AIofAI: 1st workshop on adverse impacts and collateral effects of artificial intelligence technologies'. http://ceur-ws.org/Vol-2942/paper3.pdf

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391

Spivak, R. (2019). "Deepfakes": The newest way to commit one of the oldest crimes. *The Georgetown Law Technology Review*, *3*(2), 339–400. https://perma.cc/y32v-y4x9

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, *13*(3), 260–267. https://doi.org/10.2139/ssrn.3115809

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, *36*, 1–13. https://doi.org/10.1145/3072959.3073640

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition*, *133*(3), 572–585. https://doi.org/10.1016/j.cognition.2014.08.006

Tandoc Jr., E. C., Lim, Z. W., & Ling, R. (2018). Defining 'fake news': A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2378–2395. https://doi.org/10.1109/cvpr.2016.262

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184-197. https://doi.org/10.1016/j.chb.2017.08.038

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the Cognitive Reflection Test. *Judgment and Decision Making, 11*(1), 99–113. https://doi.org/10.1037/t49856-000

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1), 1–13. https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380)*,* 1146–1151. https://doi.org/10.1126/science.aap9559

Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism Studies, 19*(13), 1866–1878. https://doi.org/10.1080/1461670X.2018.1492881

Woelke, J., & Pelzer, E. (2020). Cognitive assessment: Think-aloud and thought-listing technique. *The International Encyclopedia of Media Psychology*, 1-6.

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence, 27*(1), 36-51. https://doi.org/10.1177/1354856520923963

Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

Endnotes

[1] Due to an oversight, the latter two exclusion criteria were not pre-registered. We excluded participants who knew of the video before the study, because our focus was on deepfake detection without prior knowledge of the status of the video. We further excluded participants who indicated that they did not follow instructions and used external sources (such as google) when answering the questionnaire, because this behavior likely overrides the associations of interest. We re-ran all analyses with the pre-registered exclusion criteria, that is, a sample of 215 participants (16 participants more than reported here). The results remain virtually unchanged. These results are reported in detail in Supplement S1.

[2] Future Advocacy is a non-partisan consultancy and artificial intelligence think tank specializing at the intersection of technology and global affairs. The deepfake video used in this study has been produced in collaboration with the UK-based artist Bill Posters with the aim to apply pressure on British lawmakers to address the potential dangers of deepfakes online.

[3] In this video Johnson appears to say: “Hi folks, I am here with a very special message. Since that momentous day in 2016, division has coursed through our country as we argue with fantastic passion, vim and vigour [about Brexit]. My friends, I wish to rise above this divide, and endorse my worthy opponent, the Right Honourable Jeremy Corbyn, to be prime minister of our United Kingdom. Only he, not I, can make Britain great again.” We altered a short section (in brackets) and changed the video to simulate video buffering because in this part of the video, the mouth region did not match the voice track (poor lip-sync).

[4] Due to an oversight, we did not pre-register that individuals who knew the video beforehand (12 participants in the case of the Obama video, one participant in the case of the Johnson video) were excluded in the following analyses. These participants were excluded, because our focus was on deepfake detection without prior knowledge of the status of the video. Results did not change when this criterion was suspended (see Supplement S2).

Figure 1

Graphical Representation of the Deepfake Detection Model

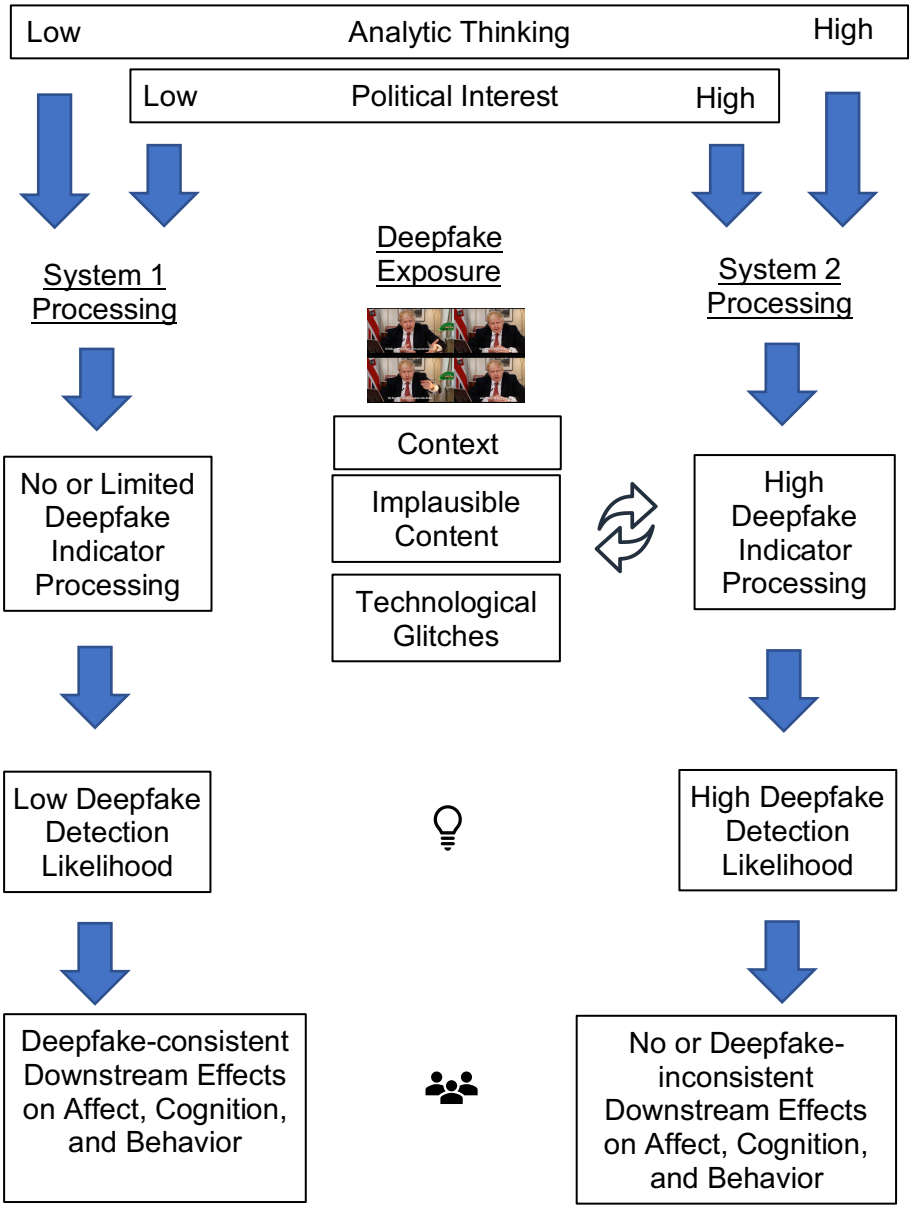

Figure 2

Deepfake – Screenshots

Figure 3

Plot of Conditional Deepfake Detection Probabilities (Study 1)