Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media

Soubhik Barari, University of Chicago
Christopher Lucas, Washington University, St. Louis
Kevin Munger, European University Institute

There is widespread concern that political "deepfakes"—fabricated videos synthesized by deep learning—pose an epistemic threat to democracy as a uniquely credible form of misinformation. To test this hypothesis, we created novel deepfakes in collaboration with industry partners and a professional actor. We then experimentally assess whether deepfakes are distinctly deceptive and find that deepfakes are approximately as credible as misinformation communicated through text or audio. However, in a follow-up discernment task, subjects often confuse authentic videos for deepfakes if the video depicts an elite in their political party in a scandal. Moreover, informational interventions and accuracy primes only sometimes (and somewhat) attenuate deepfakes' effects. In sum, our results show that while deepfakes may not be uniquely deceptive, they may still erode trust in media and increase partisan polarization.

Deepfakes pose an especially grave threat to the public's trust in the information it consumes . . . if the public can no longer trust recorded events or images, it will have a corrosive impact on our democracy.
—Senators Marco Rubio and Mark Warner, in letters to social media companies (Rubio and Warner 2019).

Societal concerns about misinformation have recently centered on novel deep learning technologies capable of synthesizing realistic videos of politicians making statements that they never said, colloquially termed deepfakes. Unlike previous video manipulation tools, contemporary deepfake tools are open source, and thereby unlicensed, unregulated, and able to be harnessed by hobbyists (rather than visual effect specialists) with relatively basic computational skills and resources. Figure 1 graphically summarizes the two major technologies to produce deepfakes, which, by many counts, are responsible for the production of the vast majority of political deepfakes at the time of writing (Ajder et al. 2019; Davis 2020; Lewis 2018).[1]

Because deepfakes let ordinary users produce media that falsely depicts someone saying and doing that which they never said nor did, it is commonly suggested that deepfakes uniquely threaten the electorate's trust in the information it consumes. This concern is not without cause; since the advent of open-source deepfake technologies, political elites around the world have been targeted in deepfake video scandals. For example, the Russia–Ukraine war of 2022 saw an escalation in the usage of deepfakes for wartime propaganda: Deepfakes of both Ukrainian President Volodymyr Zelenskyy and Russian President Vladimir Putin circulated on mainstream social media sites before being identified and banned (Wakefield 2022). It is unknown whether these

Soubhik Barari (barari-soubhik@norc.org) is a research methodologist at NORC at the University of Chicago, Chicago, IL 60603. Christopher Lucas (christopher.lucas@wustl.edu) is an associate professor at the Department of Political Science and Division of Computational and Data Sciences, Washington University in St. Louis, St. Louis, MO 63130. Kevin Munger (kevin.munger@eui.eu) is an assistant professor at the Department of Political and Social Sciences, European University Institute, Florence, Italy 50014.

All aspects of the research protocol were approved by the institutional review boards of Harvard University, Washington University in St. Louis, and Pennsylvania State University. All data and code are publicly available at https://github.com/soubhikbarari/Political-Deepfakes-Fx. Replication files are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the *JOP* replication analyst. An online appendix with supplementary material is available at https://doi.org/10.1086/732990. Christopher Lucas gratefully acknowledges financial support through the National Science Foundation (award no. 2120087 through the Methodology, Measurement, and Statistics Program).

[1] We summarize the most up-to-date empirical knowledge about the current circulation, intended purpose, authorship, and population distribution of political deepfakes in app. A.

Published online March 13, 2025.

The Journal of Politics, volume 87, number 2, April 2025. © 2025 Southern Political Science Association. All rights reserved. Published by The University of Chicago Press for the Southern Political Science Association. https://doi.org/10.1086/732990

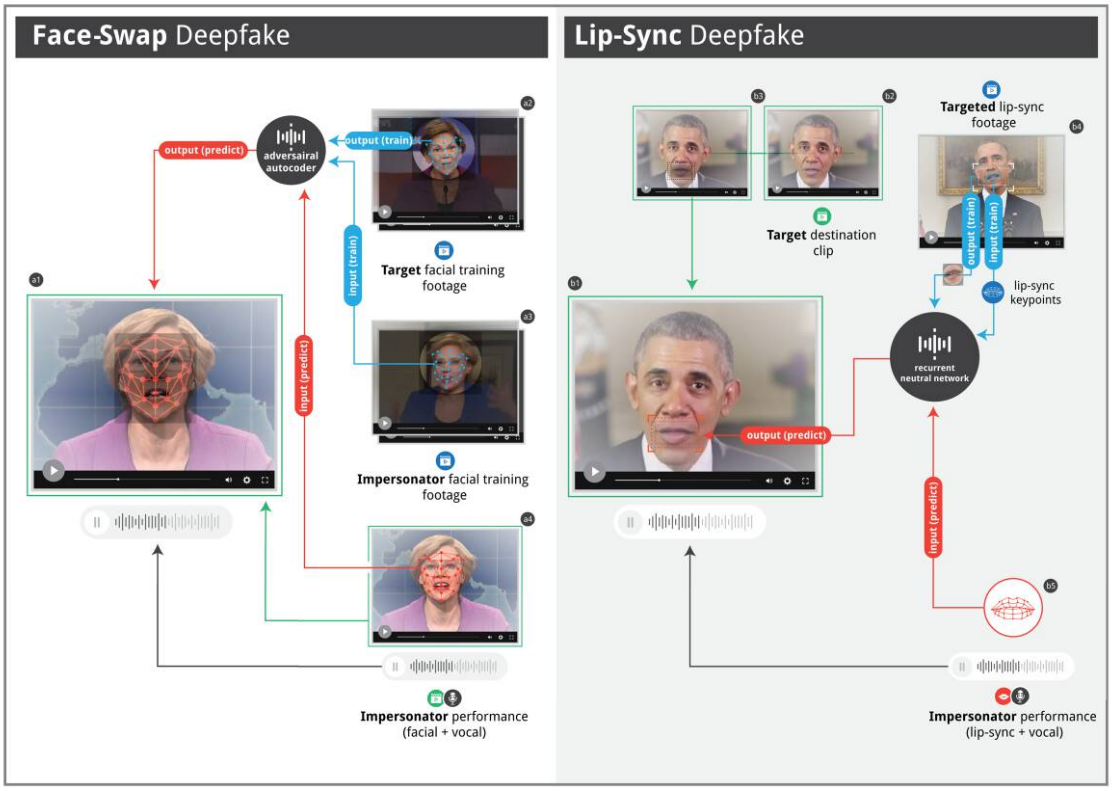


Figure 1. How deepfake videos are generated. Shown are two major methods of producing deepfakes. The left illustrates the production of a face-swap deepfake, which requires a full clip featuring the impersonator's performance including the audio and the background context for the clip where the facial features are swapped via a trained deep learning model called an autoencoder. The right illustrates a lip-sync deepfake, which requires a destination clip of the target and a vocal impersonator's performance, including their audio and lip-sync key points; these key points are transferred into a matching synthetic lip-sync video of the target via a deep convolutional neural network model trained on the target.

deepfakes continued to circulate on channels like Signal, WhatsApp, or Telegram.

Because of this threat, lawmakers (Brown 2019; Galston 2020; Gazis and Becket 2019), news outlets (Frum 2020; Hwang and Watts 2020; Parkin 2019), and civil society groups (Ajder et al. 2019; Bateman 2020; Davis 2020; Lewis 2018) have all emphasized the potential harm that deepfakes may cause to democracy, and legislation exists in more than a dozen states to regulate the production and dissemination of deepfake videos (Prochaska, Grass, and West 2020).

This article evaluates whether or not these concerns are warranted by answering a series of fundamental research questions. First, are deepfake videos of salient public officials more credible (i.e., not appearing fake or doctored) than equivalent information faked in existing media modalities such as textual headlines or audio recordings? We denote this question as research question 1 (RQ1) throughout the text. Second, are deepfakes more credible to certain subgroups (RQ2)? Third, are deepfake videos as credible as authentic videos of political elites (RQ3)?

Although the scope of possible deepfakes, political or nonpolitical, is vast, our experiment chiefly employs deepfake scandal videos of political elites given their prominence in contemporary debates and the disproportionate number in the discernible population of deepfakes relative to other forms of misinformation (see app. sec. A). Scandals—or "public revelation(s) of previously concealed misconduct" (Dziuda and Howell 2021)—have demonstrable effects on a variety of important outcomes: mass public opinion (Berinsky et al. 2011; Darr et al. 2019), national media outlets' agendas (Galvis, Snyder Jr., and Song 2016; Puglisi and Snyder Jr. 2011), election outcomes (Basinger 2013; Hamel and Miller 2019), the afflicted individuals' career trajectories, the legislative behavior of copartisans (Dewan and Myatt 2007; Dziuda and Howell 2021), and others. If the answers to the research questions we pose are "yes," ensuing scandals from circulated deepfake videos may shape the behaviors and activities of political elites, in addition to misinforming the public and eroding institutional trust.

However, we also note that scandals do not always have significant consequences for politicians (Zaller 1998), perhaps because of the proliferation of choice in media (Bennett and Iyengar 2008) or partisan attachment and resistance to counterattitudinal information (Bartels 2002a). Even if true, our results are nonetheless interesting: Our research questions are chiefly about media credibility, not attitudes regarding public officials. There is no reason to suspect that the credibility of deepfake scandals (relative to text stories) differs much from that of deepfake policy statements, relative to a textual equivalent.

On the question of credibility effects, after running a large, carefully controlled online survey experiment, we find little evidence to suggest that deepfakes are uniquely credible or affectively manipulative compared to the same misinformation communicated through text or audio. However, in a follow-up discernment task, we find that subjects confused authentic videos of political elites for deepfakes if the elites were in-partisan politicians depicted in a scandal. Throughout the experiment, we staged interventions—broad informational messages, specific debriefs, and an accuracy prime—that only somewhat attenuated deepfakes' effects. Above all else, broader literacy in politics and digital technology increased discernment between deepfakes and authentic videos of political elites.

To be clear, results based on temporally constrained experiments like ours cannot guarantee that deepfakes will not eventually change the broader informational environment, nor can we perfectly anticipate how the technology will evolve. For example, prior to the widespread adoption of deep learning, it was common to manipulate video with conventional video editing software. These videos, now termed "cheapfakes," can be understood as a part of a continuum spanning from cheapfakes to deepfakes. Increasingly, popular social media platforms like TikTok, Snapchat, and Instagram incorporate video manipulation techniques that exist somewhere on this continuum (including face-swap, lip-sync varieties, and many others). Within this broader environment, manipulated videos made their way into political discourse well before widespread access to deepfake technology. For example, in the 2016 election, a video posted to YouTube was edited to create unfounded rumors that Hillary Clinton had Parkinson's, and in 2019, a video was deceptively edited to make Nancy Pelosi appear unwell. Around the same time, a video of Jim Acosta was sped up to appear as if CNN reporter Jim Acosta struck a White House staffer (Chesney, Citron, and Jurecic 2019).

Already, many of the most-viewed faces on social media platforms have been digitally altered with nearly the same realism as the deepfake videos in the current experiment. The long-term effects of this shift and others made possible by digital manipulation technology are difficult to discern, but we endorse broader theory-building in service of hypotheses that are potentially orthogonal to the effects of earlier media technologies, that is, to "reconcile the categories of normal political communication research with [newly] important aspects of lived political experience" (Bennett and Iyengar 2008, 714). We thus present our research as a direct intervention into an immediately policy-relevant debate, one in which popular attention has not yet been met with sufficient empirical evidence. We hope that these results help drive future theorization about other possible effects that seemingly potent video manipulation technologies may have.

MEDIA EFFECTS OR MEDIUM EFFECTS?

McLuhan (1964) famously quipped that "the medium is the message," proposing that the form and method of communication is at least as important as its message in how it affects both the receiver and society more broadly. This insight was significantly refined and empirically tested in Iyengar's and Kinder's (1987) pioneering analysis of the role of television in American politics. As audiovisual political communication has evolved, scholars have identified certain novel attributes of the medium that produce previously unobserved effects. For example, Mutz (2016) finds that the combination of close-up camera shots, large television sets in the household, and uncivil political talk in political news programs induces anxiety in viewers and amplifies partisan responses to its content. Television campaign ads have been demonstrated to successfully persuade with emotional appeals through affective language, visual frames, and musical cues (Brader 2006). Similarly, Damann, Knox, and Lucas (2023) demonstrate that audio and video reporting of political statements elicits emotional responses that are not present in equivalent textual summaries. Other research shows political "infotainment" (e.g., satire, late night talk shows, comedy) is the main source of political news for a large swath of Americans (Mitchell et al. 2016) and engages audiences by cultivating both positive and negative emotional attachments to political figures and concepts (Baym and Holbert 2020; Boukes et al. 2015). Moreover, comedic impersonations that depict caricatured negative traits of politicians prime viewers of those traits and can also influence viewers' electoral support (Esralew and Young 2012).

Finally, beyond political science, a broad literature documents how audiovisual information is the prima facie medium for persuasion in a variety of contexts: recall of emotionally charged or traumatic events (Christianson and Loftus 1987; Kassin and Garfield 1991), courtroom testimony (Kassin and Garfield 1991), and persuasion in election campaigns (Grabe and Bucy 2009).

Despite this large body of research, we did not find justification for strong expectations on RQ1 ("Are deepfake videos of salient public officials more credible than equivalent information faked in existing media modalities?"). While the aforementioned work investigates the effect of different mediums of communication, it is not obvious how this research relates to novel technology for generating synthesized video (i.e.,

Table 1. Subgroups Hypothesized to Perceive Deepfakes As Credible

Subgroup	Mechanism(s) of Credibility
Intervenable	
Older adults (≥65 y)	Inability to evaluate accuracy of digital information (Barbera 2018; Guess et al. 2019; Osmundsen et al. 2021)
Partisans (with out-partisan target)	• Directional motivated reasoning about out-partisans (Enders and Smallpage 2019; Leeper and Slothuus 2014) • Accuracy motivated reasoning about out-partisans (Druckman and McGrath 2019; Tappin, Pennycook, and Rand 2021)
Sexists (with female target)	• Consistency with prior hostile beliefs about women (Cassese and Holman 2019; Glick and Fiske 1996; Schaffner et al. 2018) • Consistency with prior benevolent beliefs about women (Cassese and Holman 2019; Glick and Fiske 1996; Schaffner et al. 2018)
Low cognitive reflection	Overreliance on intuition over analytical thinking in making judgments (Pennycook and Rand 2019)
Low political knowledge	• Inability to evaluate plausibility of political events • Inability to recognize real facial features of target (Brenton et al. 2005; Lupia 2016)
Low digital literacy	• Inability to evaluate accuracy of digital information • Limited/no recognition of deepfake technology (Guess et al. 2020; Munger et al. 2020)
Nonintervenable	
Low accuracy salience	Limited/no attention to factual accuracy of media (Pennycook et al. 2020)
Uninformed about deepfakes	Limited/no recognition of deepfake technology

Note. This list is neither exhaustive nor mutually exclusive but rather enumerates substantively important subgroups in American politics. We clarify possible mechanisms for each groups' susceptibility, but proving these and not alternative mechanisms is beyond the scope of this article.

deepfakes). At the time of fielding, consistent with the related literature and contemporary popular press, we hypothesized that deepfake videos are more deceptive than other formats and therefore would be perceived as more credible than equivalent information in text or audio formats.[2]

Susceptible subgroups

A robust literature has identified a number of "at-risk" subgroups with heightened susceptibility to misinformation in the political context of the United States. We summarize the most-studied groups in table 1 and hypothesized these groups would also be susceptible to deepfakes (RQ2).

The first category—older adults—draws on the observation that "users over 65 shared nearly seven times as many articles from fake news domains as the youngest age group" during the 2016 US Presidential election (Guess, Nagler, and Tucker 2019, 1). Similarly, Barbera (2018) finds that people over 65 shared roughly 4.5 as many fake news stories on Twitter as people 18 to 24. Matching Twitter users to voter files, Osmundsen et al. (2021) find that the oldest age group was 13 times more likely to share fake news than the youngest. If the primary mechanism of this susceptibility is inability to evaluate digital information, we expect this will be exacerbated when exposed to more complex information in the form of video.

Next, research identifies that motivated reasoning, or the selective acceptance of information based on consistency with prior beliefs, powerfully shapes how individuals respond to information. We identified mechanisms for how two types of substantively important prior dispositions (although many more exist) may predict deception by deepfake: partisan group identity and sexist attitudes. A large literature documents how partisan identity—either by way of strong directional motivations to reject new evidence or differing prior beliefs about the credibility of new evidence—directs voters' attitudes about events, issues, and candidates (Druckman and McGrath 2019; Enders and Smallpage 2019; Leeper and Slothuus 2014). Moreover, voters' evaluations of candidates or events can be driven by prior negative stereotypes (Cassese and Holman 2019; Teele, Kalla, and Rosenbluth 2018). Women are a

[2] In our study, to prevent survey fatigue and reduce priming across outcomes, we elicit direct credibility evaluations of media upon exposure rather than asking if the depicted events truly occurred, which may be evaluated on their perceived plausibility independent of the information presented. See the "External Validity" section for further discussion.

particularly salient group in the post-Trump era: A recent survey finds that, next to partisanship, ambivalent sexist views[3] most strongly predicted support for Donald Trump in the 2016 US Presidential election (Schaffner, MacWilliams, and Nteta 2018). For both groups, the affective and evidentiary appeal of videos may interact with the need to maintain consistent beliefs and heighten the credibility of deepfakes.

Another set of subgroups may be especially susceptible to deepfakes due to constraints on cognitive resources or knowledge. Performance in cognitive reflection tasks measures reliance on "gut" intuition, which may preclude careful examination of video evidence (Pennycook and Rand 2019). Similarly, those with little political knowledge may have little prior exposure to the targeted political figure, rendering them unable to discern "uncanny" deepfake artifacts that resemble but do not perfectly replicate their intended facial features (Brenton et al. 2005). Finally, the last two categories describe traits that we can intervene on via direct information provision—or raising the salience of deepfakes conceptually or by example—and accuracy priming—or raising the salience or normative value of engaging with accurate news—each of which we expect to reduce deepfakes' credibility (Pennycook et al. 2020, 2021).

Consistent with our expectations for RQ1, we preregistered the prediction that all subgroups in table 1 would be differentially susceptible to deepfake misinformation over text and audio misinformation.

Discerning authentic from fake

Lastly, on RQ3—as with RQ1—if popular claims about deepfakes are correct, they should be nearly indistinguishable from authentic video clips in a shared context (e.g., a news feed about politics). Thus, we expected that deepfakes should be perceived as equally credible as authentic video clips in the same context.

EXPERIMENTAL DESIGN

To test our hypotheses, we employed two experiments embedded in a survey fielded to a nationally representative sample of 5,724 respondents on the Lucid[4] survey research platform. The first experiment (incidental exposure) presents respondents with a news feed of apparently authentic video clips, audio clips, and text headlines about candidates in the 2020 Democratic presidential primary, in which a deepfake video of one of the candidates may or may not be embedded. The second experiment (detection task) asks the same respondents to scroll through a feed of eight news videos—randomized to contain either no deepfakes (dubbed the no-fake feed), two deepfakes (low-fake), or six deepfakes (high-fake)—and discern deepfakes from the authentic video clips. Table 2 describes our overall design, and appendix figure B6 provides a graphical illustration of the survey flow.

Our design is motivated by a number of considerations. First, the two experiments capture different quantities of interest by way of comparing different types of randomized media exposure. The incidental exposure experiment measures the perceived credibility of a single, carefully masked deepfake video relative to the equivalent scandal depicted via other formats or similar reference stimuli about the candidate in question (RQ1, RQ2). In the incidental exposure experiment, we also compare affect toward the politicians in each clip as an auxiliary outcome. In contrast, the detection task captures the credibility of deepfakes relative to authentic videos (RQ3) measured by overall discernment accuracy and errors due to false positives.

Second, the experiments both inherently and by their ordering allow us to test credibility perceptions across differing levels of information provision. The first experiment simulates exposure to a deepfake "in the wild" with, at most, the following verbal description about deepfakes for those randomized to receive information: "During the 2016 Presidential campaign, many people learned about the risk of fake or zero-credibility news: fabricated news stories posted on websites that imitated traditional news websites. While this is still a problem, there is now also the issue of digitally manipulated videos (sometimes called 'deepfakes'). Tech experts are warning everyone not to automatically believe everything they read or watch online." All participants in the detection task, on the other hand, are explicitly told about deepfakes, and some are even provided visual examples of deepfakes if randomly assigned to be debriefed about their incidental exposure before the task.

Third, and arguably most important for external validity, our two experiments allow us to test credibility perceptions across multiple deepfakes that differ in their targets, quality, and technology. In the first experiment, as we will describe in the next section, we hired a professional firm to produce several novel deepfakes of a single politician depicted in several realistic scandals via the face-swap method depicted on the left side of figure 1. In the second experiment, we used a representative

[3] Ambivalent sexism describes a bundle of both outright hostile (e.g., "women are physically inferior to men") and deceptively benevolent views about women (e.g., "women are objects of desire") (Glick and Fiske 1996).

[4] At the time of fielding, Ternovski and Orr (2022) noted systematic trends in inattentive survey respondents on Lucid. We describe the battery of attention checks we employ to maintain a high-quality sample in app. F; subjects who failed the simple attention checks at the beginning of the survey were not allowed to complete the survey. All findings are consistent across samples divided by performance in mid-survey attention checks or duration spent evaluating stimuli, though slightly smaller in magnitude for less attentive respondents.

Table 2. Overview of Experiments Embedded in Survey

	Exposure(s)	Pre-exposure Interventions	Respondent Outcomes
(1) Incidental exposure	1. Pre-exposure authentic coverage of 2020 Democratic primary candidates 2. Randomized exposure to text, audio, video, skit clip of Elizabeth Warren scandal, attack ad, or control (no stimuli) 3. Postexposure authentic coverage of 2020 Democratic primary candidates	• Information about deepfakes	• Belief that candidate clippings are not fake/doctored (credibility) • Favorability of candidates (affect)
(2) Detection task	Randomized task environment: • No-fake feed: 8 authentic clips of political elites • Low-fake feed: 6 authentic clips, 2 deepfakes of political elites • High-fake feed: 2 authentic clips, 6 deepfakes of political elites	• Debrief of deepfakes exposed to in (1) before task • Accuracy prime	• Deepfake detection accuracy • Deepfake false-positive rate • Deepfake false-negative rate

set of pre-existing deepfakes of many different elites made by experts and amateurs alike via lip-sync and face-swap. To draw our conclusions from a realistic, externally valid set of deepfakes, we use existing knowledge of the population of deepfakes “in the wild” (see app. sec. A) to guide the creation and selection of stimuli in the exposure and detection experiments, respectively.

To adjust for observable demographic skews in our respondent pool, all analyses are replicated using poststratification weights estimated from the US Census in appendix G. Details of this poststratification and other characteristics of the sample are given in appendix F.

Incidental exposure experiment

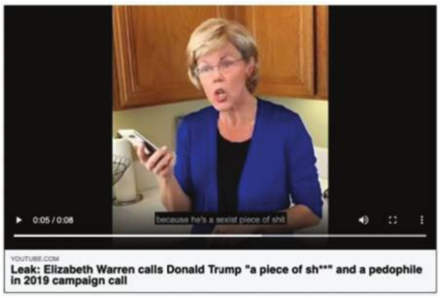
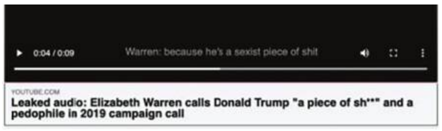
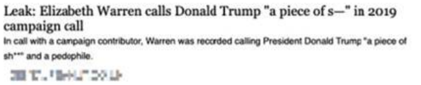
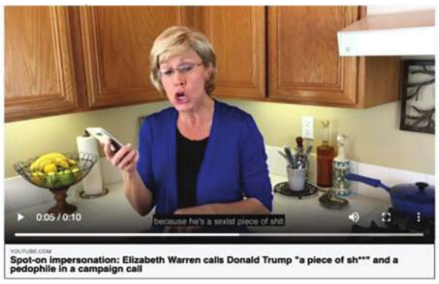
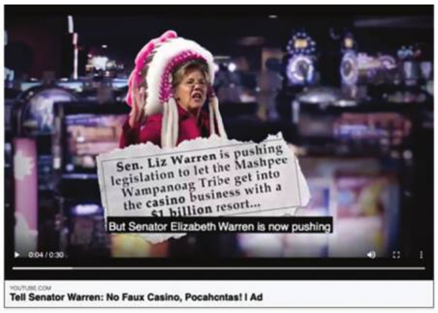
In the first experiment, we implement a 2 × 6 factorial design, pairing a randomized informational message about deepfakes with randomization into one of six conditions—a deepfake video (presented as a leaked mobile phone recording) or, alternatively, audio, text, or skit of a political scandal involving a 2020 Democratic primary candidate Elizabeth Warren, a campaign attack ad against Warren, or a control condition of no clip at all—after which we measure several outcomes. In the incidental exposure experiment, we selected Elizabeth Warren because she was both a salient politician during the primary election and (at the time of fielding) had not been the target of any visible deepfake online. Thus, credibility perceptions would not be contaminated from prior exposure as would be the case if we recycled an existing deepfake.

To create a natural environment for media consumption, we surround the experimentally manipulated media exposure with four media clips, two before and two after.[5] These reports are all real coverage of different Democratic primary candidates, presented either in audio, textual, or video form. The order and content of these media are fixed and primarily serve to mask the main manipulation, replicating the visual style of Facebook posts. The six conditions of our manipulation (video, audio, text, skit, ad, control) and their exact differences from each other are shown in table 3, where video is the group assigned to the deepfake.

Participants in the video, audio, and skit conditions are randomly exposed to one of five different scandal events to reduce the possibility that our results are being driven by a

[5] App. sec. K displays these surrounding clips, which were included to better represent a real-world scenario in which a subject is scrolling through a news feed and to permit a naturalistic presentation of deepfake videos. These surrounding media were fixed for all conditions and contained a text story of Klobuchar, a video of Biden, and similar media.

Table 3. Experimental Conditions in Incidental Exposure Experiment

	Condition	Description of Variation	Example Clip
Scandal Clips (Script Head Constant)	**video** (n = 872)	Face-swap performed on video in skit condition; title and video edited to resemble leaked video footage.	 YOUTUBE.COM Leak: Elizabeth Warren calls Donald Trump "a piece of sh**" and a pedophile in 2019 campaign call
	audio (n = 954)	Visuals stripped from video condition; title edited to resemble leaked hot mic.	 YOUTUBE.COM Leaked audio: Elizabeth Warren calls Donald Trump "a piece of sh**" and a pedophile in 2019 campaign call
	text (n = 950)	Visuals and sound stripped from video condition; title describes scandal as a leak; subtitle describes event captured on video.	 Leak: Elizabeth Warren calls Donald Trump "a piece of s—" in 2019 campaign call In call with a campaign contributor, Warren was recorded calling President Donald Trump "a piece of sh**" and a pedophile. [redacted]
	skit (n = 956)	Filmed impersonator portraying a campaign scandal event; used to create video and audio conditions.	 YOUTUBE.COM Spot-on impersonation: Elizabeth Warren calls Donald Trump "a piece of sh**" and a pedophile in a campaign call
Reference Stimuli	**ad** (n = 935)	Campaign attack advertisement describing real scandal event.	 Sen. Liz Warren is pushing legislation to let the Mashpee Wampanoag Tribe get into the casino business with a $1 billion resort... But Senator Elizabeth Warren is now pushing 0:04 / 0:30 YOUTUBE.COM Tell Senator Warren: No Faux Casino, Pocahontas! I Ad
	control (n = 916)	No stimulus presented.	N/A

single scandal. Each scandal is entirely fictitious, written to maximize realism and capture a range of plausible candidate scandals according to our best assessments, and each respective video was created in collaboration with a professional actor and a tech industry partner, both typical of the kinds that produce current political deepfake videos.[6]

[6] We discuss this collaboration further in app. C. While an academic–industry partnership may be a unique source for a deepfake video, app. sec. A demonstrates that the type of deepfake we create—a face-swap deepfake—is in fact the most common in circulation. Moreover, to the extent that our deepfake videos differ from the population of deepfakes as a result of this collaboration, it is likely more compelling than the average deepfake. Despite this, we still did not find that this deepfake was more deceptive than either audio or text versions of the same content (see "Results"). We therefore argue that any bias that results from this collaboration is likely conservative (i.e., relative to deepfakes that are not produced by industry partners, we are arguably less likely to observe null results).

Specifically, the audio condition consists of the audio recording of the actor making a scandalous statement. Participants in the skit condition are exposed to the original videos used in the creation of the deepfake video prior to the modifications made by the neural network algorithm. That is, this condition displays the unaltered video of the paid actress hired to impersonate Elizabeth Warren, which is clearly framed as a skit: The title of the corresponding deepfake in the video condition is shown, but "Leak" is replaced with "Spot On Impersonation." Finally, the video condition employs a deepfake constructed from the footage used in the skit condition. Details on the production of these stimuli are provided in appendix C, and each of the five scripts are provided in table C5. We do not register any hypotheses about heterogeneous effects across these particular scandals within condition but conduct exploratory analyses that show small differences across conditions (app. J).

Finally, in the ad condition, subjects are exposed to a real negative campaign ad titled, "Tell Senator Warren: No Faux Casino, Pocahontas!", which criticizes Senator Warren's supposedly illicit support for federally funding a local casino owned by an Indian tribe despite her previous opposition to such legislation and her disputed claims of Cherokee heritage. Although the ad frames Warren as politically insincere, similar to script (e), and primes the viewer of her Cherokee heritage controversy, similar to script (c), it stylistically and informationally differs in many other ways and thus is not an exact ad counterfactual of our deepfake. Nevertheless, the ad serves as a benchmark comparison for a deepfake's affective effect, since it is an actual campaign stimulus used in the primary election to activate negative emotions toward Warren.

Following the feed, respondents are asked to evaluate the credibility of each textual, audio, or video clip in the feed (the extent to which they believe the clip is "fake or doctored" on a five-point scale) in between other distraction evaluations (funny, offensive, informative). Consequently, respondents are also asked to evaluate how warmly or coldly they feel toward each of the Democratic candidates on a continuous 100-point feeling thermometer.

Our main counterfactuals of the deepfake video condition are the text and audio conditions. Importantly, we do not make a comparison of credibility ("is this fake or doctored?") of the skit and ad stimuli with the three scandal clippings because of concerns about differential item functioning: It is possible that respondents say the ad or skit is "fake or doctored" because they correctly perceive the skit as a staged depiction or the ad as an edited video rather than because they incorrectly perceive it as depicting Warren participating in a fabricated event. However, we can still usefully compare affective responses toward Warren between the scandal clippings and these reference stimuli.

Detection task experiment

After completing the battery of questions in which we measure our primary outcomes of interest and ask another attention check question, the subjects begin the second experimental task that measures their ability to discriminate between authentic and deepfake videos.

Before this task, half of the subjects (in addition to all the subjects not taking part in this task) are debriefed about whether they were exposed to a deepfake in the first experiment. The other half are debriefed after this final task. This randomization allows us to test the effect of the debrief, which, unlike the verbal information randomly provided in the first stage, provides visual examples of deepfakes. Additionally, half of all respondents are provided an accuracy prime—an intervention designed to increase the salience of information accuracy (Pennycook and Rand 2019).

Subjects were randomly assigned to one of three environmental conditions: The percentage of deepfakes in their video feed was either 75% (high-fake), 25% (low-fake), or 0% (no-fake). Appendix D displays screenshots and descriptions of each of these videos. Misclassifications (reductions in accuracy) in the detection task can be decomposed into false negatives (misclassifications of deepfakes as authentic), and false positives, (misclassification of authentic clips as deepfakes). We measure both, in addition to overall accuracy, to gauge respondents' discernment abilities and the source of their errors.

In the task itself, we employ videos created by Agarwal et al. (2019) and a mix of other publicly available deepfake videos of both lip-sync and face-swap varieties. To the extent that respondents have previously viewed these videos, we should expect detection performance to be biased upward, although no respondent explicitly indicated as such in open feedback. For the pool of authentic videos, we primarily selected, where possible, real-world video scandals of the elites used in the deepfake pool. Unlike in the incidental exposure experiment, in both the deepfake and non-deepfake pools, we have clips of Republicans (Donald Trump) and Democrats (Barack Obama, Joe Biden, Elizabeth Warren), creating both Democratic and Republican out-partisans in the detection task.

Ethical considerations

Creating deepfakes raises important ethical concerns, which we aimed to address at every stage of our research design. First, given the risk of deepfakes disrupting elections, understanding their effects is of the utmost importance: This research has the potential to improve the resilience of democratic politics to this technological threat by better informing policy and consumer behavior. Second, we created deepfakes of a candidate who was not currently running for office to ensure that our experiment could not plausibly influence the outcome of an election. Third, we designed "active debriefs" that required subjects to affirm in writing whether they were exposed to false media. Fourth, deepfakes are increasingly part of the standard media environment, so our study only exposes subjects to things they should be prepared to encounter online. Finally, to ensure that our study does not contribute to the existing supply of online misinformation, we made it impossible for respondents to download our videos and have searched extensively for our stimuli online after our experiment. We can find no evidence that we have contributed to the supply of misinformation with our stimuli. We discuss these points in more detail in appendix E.

RESULTS

Figures 2 through 5 summarize our main results, which robustly reject our hypotheses for RQ1 and RQ2 but produce a nuanced answer to RQ3. Figure 2 compares baseline and relative subgroup credibility evaluations and affect toward Warren from respondents in all the Warren clip conditions. Figure 3 compares performance in the detection task across environments and subgroups, while figure 4 and figure 5 break down performance differences by our preregistered subgroup traits and by clips, respectively. We organize our results into three main findings, each of which we discuss in detail in

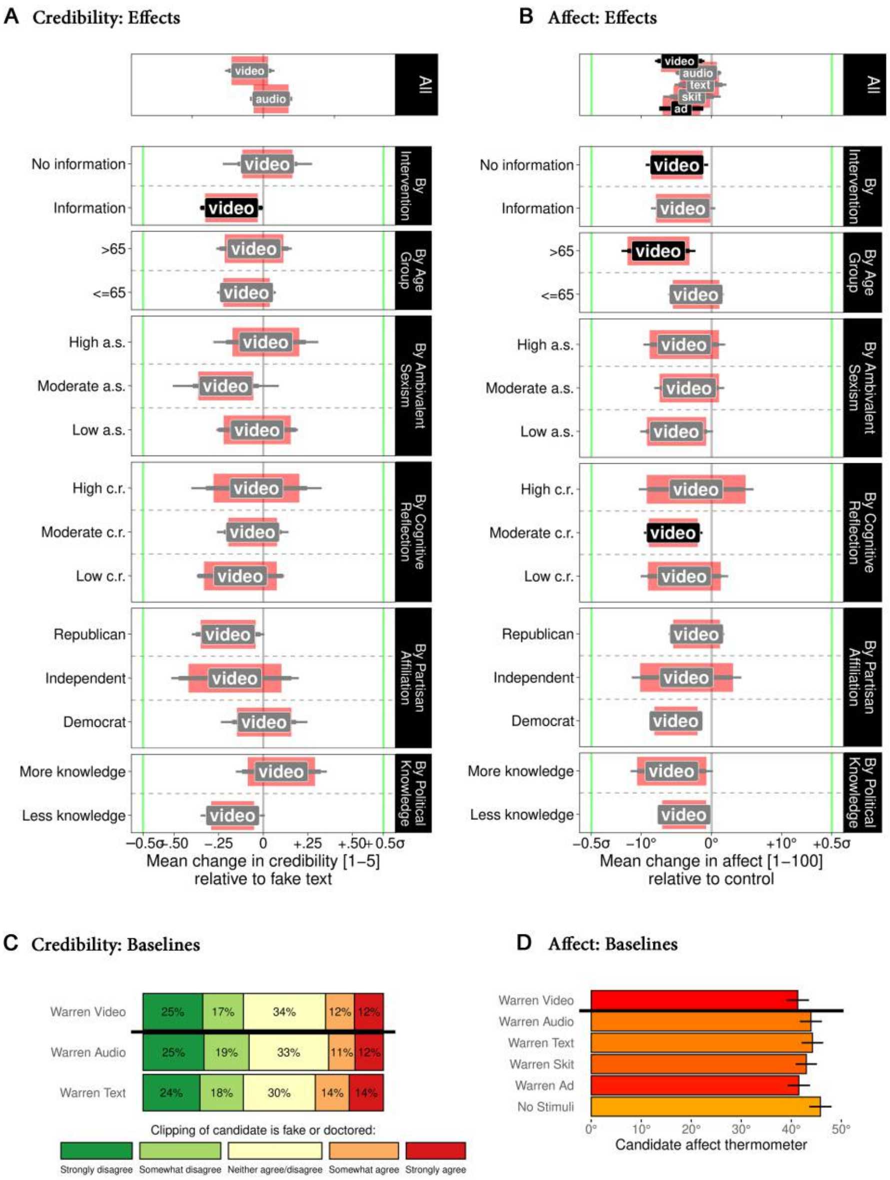


Figure 2. Relative to other stimuli, effects of incidental exposure to a deepfake video are minimal overall and across subgroups. Categories for ambivalent sexism are constructed as equal-sized percentiles from sample values. Thicker lines denote 95% CIs, thinner lines denote 95% CIs adjusted for multiple subgroup comparisons (Benjamini and Hochberg 1995), and red blocks indicate 95% CIs from two one-sided equivalence *t*-tests with equivalence bounds (Wellek 2010). For brevity, the text condition is not shown in the subgroup results for affect; however, in no subgroup condition did text produce a significant effect relative to control.

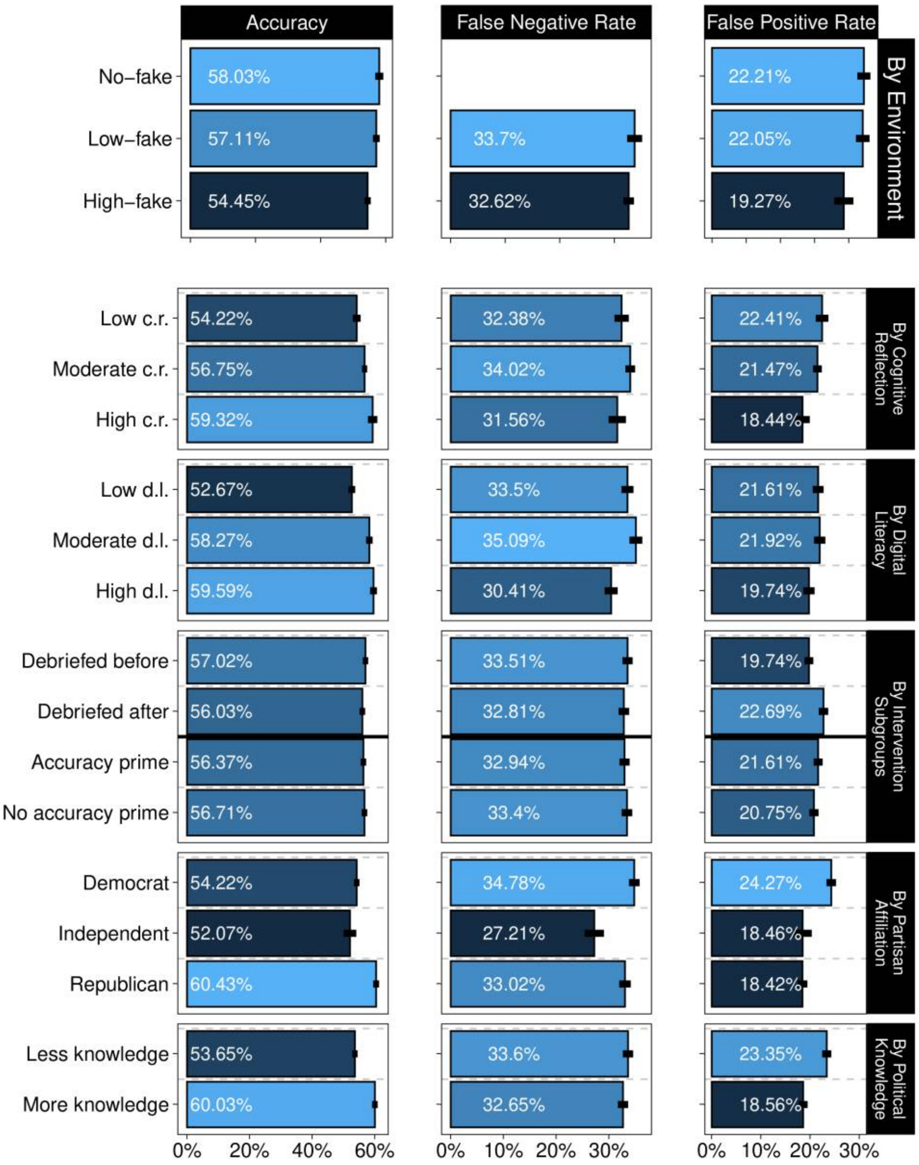


Figure 3. Performance comparisons in deepfake detection task by subgroup. Shown are three different measures for $n = 5{,}497$ (99%) of respondents who provide a response to at least one video in the detection experiment task. Coefficient estimates are given in appendix G and are robust to the choice of missingness threshold. Accuracy is the percent of all videos in the task correctly classified as either fake or real. False-negative rate is the percent of deepfakes in the task incorrectly classified as authentic (as such, this quantity is degenerate in the no-fake condition). False-positive rate is the percent of authentic videos in the task incorrectly classified as deepfakes.

relation to our original hypotheses, and conclude with a brief discussion of external validity.

For all results involving multiple group-wise comparisons or estimating multiple substantive coefficients, we adjust *p*-values according to the Benjamini-Hochberg "step-up" procedure, which bounds each group of tests' false discovery rate at $\alpha = .05$ without as strict of a correction as the Bonferroni procedure, which assumes no dependence between hypotheses (Benjamini and Hochberg 1995). Additionally, we conduct equivalence tests to test whether estimated effects, statistically null or not, are substantively null in magnitude (Wellek 2010). For consistency, we deem an effect "substantively null" if it fails to explain half of a standard deviation or more of the outcome, that is, falls within the equivalence bounds of $\pm .5\sigma$. We now summarize our findings.

Deepfake scandal videos are no more credible or affectively appealing than comparable fake media

In the incidental exposure experiment, just under half of subjects (42%) found our deepfake videos of Warren at least somewhat credible (top left of fig. 2). However, the videos were, on average, less credible than the faked audio (44%) and

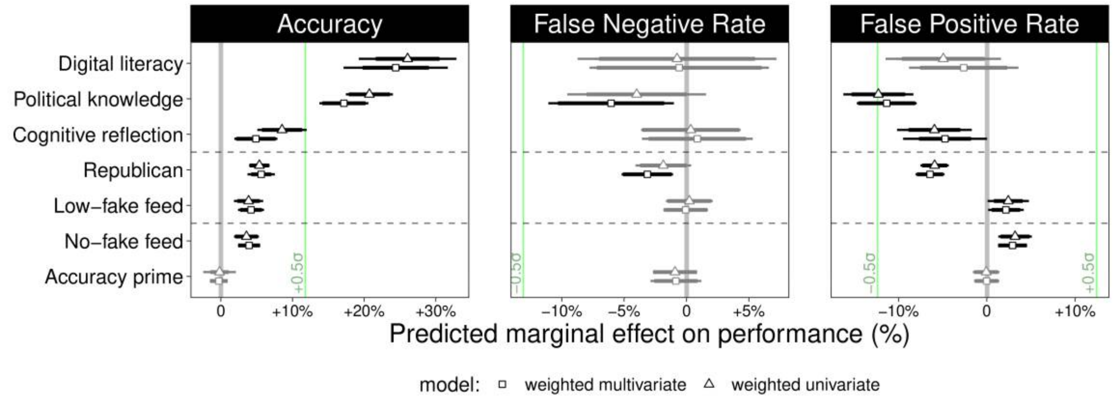


Figure 4. Predictors of detection task performance. Predictors are grouped by dashed grey lines into respondent traits (all rescaled to the [0, 1] range), detection environment (relative to high-fake), and intervention assignment. Predictors include all group indicators from table 1, excluding age, which has no significant effects on performance (see app. tables G24, G25, G25). The multivariate model estimates the effects of all predictors jointly and additionally controls for age group, education, and internet usage. Both models are weighted via a poststratification model (see app. sec. F). Appendix figure I11 shows that dropping nonrespondents in the task does not change the substantive interpretation of detection experiment results.

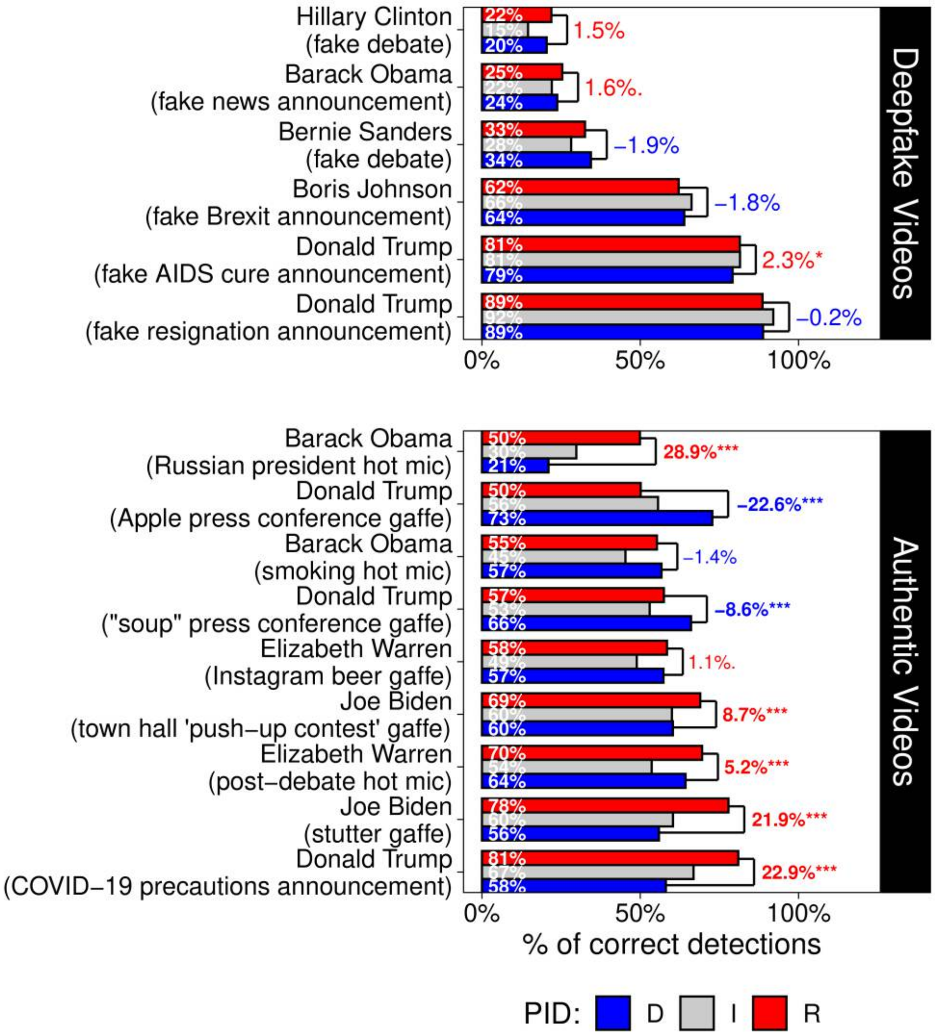


Figure 5. Detection performance comparisons across partisanship and clip authenticity. The *p*-values for differences in correct detection proportions between Democrats and Republicans (*** indicates $p' < .001$) derived from two-proportions *z*-tests, where p' is the transformed larger *p*-value after adjusting for multiple comparisons (Benjamini and Hochberg 1995).

comparable in credibility to the fake text (42%). Both the fake audio and video clippings not only fail to reject a traditional null hypothesis of no effect relative to the fake text headline but also reject the null hypothesis of a minimal change of $\pm .5\sigma$ ($\approx$.68) in credibility confidence, let alone a full point step between confidence categories. Appendix tables G7 and G8 show that these differences are robust to a variety of model-based adjustments. Our best answer to RQ1 is, thus, "no."

Even if deepfakes are not more credible than comparable fake media, can they still move affect toward the target elite? Relative to no exposure, videos do slightly decrease Elizabeth Warren's favorability as measured by the 0 to 100 feeling thermometer, though this still fails to clear our equivalence bounds for a null effect. However, there are demonstrably null effects of the deepfake video on affect when compared to text and audio, as seen in the top-right cell in figure 2. Deepfake videos are also at least as affectively triggering as negative attack advertisements, a decades-old technology, of the same target. Appendix table G9 produces this same null effect with model-based controls.

Investigating whether the previous null results mask any credibility or affect heterogeneities for subgroups specified in table 1 (panels 2–7 in fig. 2), we find few. The answer we give to RQ2 is then also "no." This is not to say that these subgroups are not moved by a scandal of Elizabeth Warren in general. To take the most notable examples, sexist attitudes and out-party identification predict increases in the credibility (substantively large in the latter case) of the scandal stimulus (app. fig. J14, tables G18–G19, tables G21–G22) but not disproportionately so for the deepfake relative to the headline or audio clipping.

Digital literacy and political knowledge improve discernment more than information

Baseline performance accuracy (fig. 3) in the detection task (52%–60% across all groups) and error rates of less than 50% suggest that their discernment capabilities are better than random. Though, notably, the false-negative rate for our clips is consistently larger than the false-positive rate despite the average distribution across conditions of one-third deepfakes and two-thirds authentic clips. A little more than one-third of all deepfakes in our feed are undetected, while a little under one-third of authentic clips are falsely flagged across all subgroups.

Examining whether subgroup traits in table 1 predict performance, we find that neither of our interventions improves discernment accuracy during the detection task (see estimated marginal effects on accuracy in fig. 4). While information and accuracy salience fail, figure 4 shows that respondent traits—specifically digital literacy, political knowledge, and, to a lesser extent, cognitive reflection—predict the most substantively meaningful improvements. Republicans also appear to marginally outperform Democrats and Independents but scored little less than a full clip higher in correct classifications than the rest.

Discernment of authentic videos varies significantly by partisanship more than deepfakes

Remarkably, although partisanship overall predicts small effects on performance relative to other traits, an examination of individual clips (fig. 5) reveals some massive performance gaps between Democrats and Republicans but only for real videos. Fifty percent of Republicans believed that real leaked footage of Obama caught insinuating a postelection deal with the Russian president was authentic compared to 21% of Democrats, a highly significant differential according to a simple Chi-squared test ($\chi 2 = 338.3$, $p < .01$). Performance is flipped for the clip of Donald Trump's public misnaming of Apple CEO Tim Cook, which was correctly identified by 73% of Democrats but only 50% of Republicans ($\chi 2 = 78.5$, $p < .01$). Most striking is that for an authentic clip from a presidential address of Trump urging Americans to take cautions around the COVID-19 pandemic, the finding holds in the opposite direction: Although a positive portrayal,[7] at least for Democrats who by and large hold similarly cautionary attitudes toward COVID-19 (Clinton et al. 2021), only 58% of Democratic viewers flagged it as authentic, whereas 81% of Republicans believed it to be real ($\chi 2 = 167.89$, $p < .01$). Controlling for both clip and respondent characteristics, appendix figure J23 shows that Republican identity only predicts a boost in performance when asked to corroborate real scandal video clippings of Obama.

Thus, individual clips' performance suggests that partisans fare much worse in correctly identifying real clips, but not deepfakes, portraying their own party's elites in a scandal. In contrast, digital literacy, political knowledge, and cognitive reflection bolster correct detections roughly evenly for all clips (app. fig. J22).

Taken together with the previous finding, this provides a nuanced answer to RQ3. Baseline discernment accuracy is not particularly high for any subgroup; however, performance varies significantly by subgroup. Literacy (both political and technological) reduces false skepticism, while partisanship increases skepticism about real scandal videos of in-party elites.

DISCUSSION

To summarize, we have demonstrated that deepfakes, even when designed specifically to depict a prominent politician in a

[7] Positive portrayal, here, means depiction of valence traits or characteristics that, all else equal, voters should unanimously prefer more of rather than less of (Bartels 2002b).

scandal, are not uniquely credible or emotionally manipulative. They are no more effective than the same misinformation presented as text or audio or the same target attacked via a campaign ad or mocked in a satirical skit. Our experiments reveal that several characteristics are essential components of how citizens process both authentic and fake video media. In particular, at least two types of prior beliefs (partisanship, sexism) can enhance the credibility of fake media, while general knowledge about politics, literacy in digital technology, and propensity for cognitive reflection can bolster discernment.

Theoretical implications

Our results for RQ1 and RQ2 concord with a growing body of research on video media effects (Dobber et al. 2021; Vaccari and Chadwick 2020; Wittenberg et al. 2021), which, taken together, cast doubt on the fear that manipulated videos themselves will directly deceive the public of false events on a mass scale. The emergence of "misinformation" as a phenomenon of public interest has led to an understandable emphasis on credibility and deception as outcomes in the broader study of political media. However, for motivated respondents, these outcomes are in flux even when exposed to both authentic media and analogously falsified nonvideo media. That is not to say the effects of video media in particular are not worthy of further scholarship: Video media varies on many theoretically relevant dimensions beyond facticity, including presentation of gender, dynamics in vocal tone, and patterns of facial expressions known to influence perceptions of its subject (Boussalis et al. 2021; Knox and Lucas 2021). Given the "primacy of visual communication for human cognition" (Hancock and Bailenson 2021, 150), the downstream impact of deepfakes could be deeper and more complex than our design can infer.

Our results for RQ3 in particular reinforce a broader scholarly view on public opinion: When evaluating information, voters are more perceptive of the congeniality of information (e.g., whether a copartisan is negatively portrayed) than its other attributes (e.g., authenticity). In fact, the detection task results suggest that this motivated reasoning occurs more often with authentic videos than with deepfakes. Without further assumptions or subjective assessments, we cannot pinpoint exactly which other attributes that widely differ across our real and fake stimuli (e.g., plausibility of event, magnitude of scandal, policy area, issue salience) explain this difference. However, we rule out attributes such as source cue (fig. J20) and the type of scandal (fig. J16 and J18).

That said, we find strong evidence that certain subject attributes significantly affect deepfake detection capacity, independently of partisan-motivated reasoning. In keeping with a now-robust literature on the correlates of the persuasiveness of "fake news" and other contemporary media, we find substantively large heterogeneities in deepfake detection. The largest is in subject digital literacy, further advancing the case that this construct is a key moderator of digital media effects (Guess and Munger 2023; Luca et al. 2022; Munger et al. 2021). This result agrees with the scope conditions of digital literacy proposed by Sirlin et al. (2021), who find that it is useful for understanding accuracy discernment but not for sharing behaviors. We find smaller but still significant heterogeneities by subject cognitive reflection, in agreement with a large, related literature (Mosleh et al. 2021; Pennycook and Rand 2019; Stecula and Pickup 2021).

In contrast, we do not find that accuracy priming subjects influences their overall performance. Increases in the true detection of deepfakes are outweighed by increases in false positives. Our finding thus disagrees with the conclusion from a related literature (Pennycook and Rand 2022; Pennycook et al. 2021), although the scope conditions of our experiment do not perfectly overlap with previous studies. Future research should probe the limits of these accuracy primes.

It is tempting to conclude from our topline results that scandals do not matter. More accurately, our findings imply that the exact details and the medium through which they are initially communicated may not matter, at least on first reaction and in an experimental setting. In this view, the latest deepfake technology needn't be harnessed to implicate one's political adversaries in a scandal: A far less sophisticated attack ad or a satirical skit priming the same character traits may be equally effective. Given recent evidence that Americans may be more responsive to the policy preferences and constituency activities of their representatives (Costa 2021), future studies might evaluate the effectiveness of deepfake scandals that highlight policy incongruences between candidates and their audience.

In light of our findings, policymakers should devote more time and resources to bolster the credibility of real news videos and curb the development and spread of deepfake videos that cause psychological or social damages for their targets. Recent counts of deepfakes on the Internet find that most are nonconsensual pornographic clips of women (app. A), suggesting that perhaps the greater, more novel harm of deepfakes is the harassment of its targets, not the deception of its viewers.

At the same time, we follow Ternovski, Kalla, and Aronow (2022) in cautioning against the indiscriminate deployment of interventions warning the public about deepfakes. Our findings suggest that targeted informational interventions cause a small reduction in the credibility of deepfakes but at a cost to the credibility of nonmanipulated videos, in concurrence with Vaccari and Chadwick (2020). The trade-off between these "false negatives" and "false positives" has implications for the health of democratic information environments and thus should not be made lightly. The design of optimal misinformation interventions on these and other dimensions remains an open problem (Saltz et al. 2021).

External validity

External validity is a central concern for all experimental research, especially tightly controlled media effects experiments like the ones we conduct here. We therefore address four external validity considerations about our results. First, it is possible that deepfakes of other less salient elites may produce larger effects relative to text or audio than the ones seen here. However, thus far, deepfakes of this kind (at least accessible to the public) have been exceedingly rare, possibly because of technical limitations: As we describe in appendix C, deepfakes require a large training set of high-definition facial images, which may be unavailable for a city councilor or a low-profile Congressman. We believe our effects are representative of the kind likely to be seen in the present population of deepfakes (app. A), though research on "downballot deepfakes" would be valuable. Furthermore, the population of future deepfakes may well be different from the population of present deepfakes. This "temporal validity" aspect of external validity is a fundamental constraint on the scope of social scientific knowledge (Munger 2019, 2023).

Although we created and selected, to the best of our ability, a diverse and representative set of publicly accessible deepfakes, we cannot control for all idiosyncratic features of each clip. Future scholars may wish to decompose our multidimensional treatments into their constituent causal attributes, but that requires careful identification assumptions that the present design cannot afford (Egami et al. 2018). We also cannot demonstrate that either our deepfakes or authentic clips are exactly representative of these features in the news environment. There is a fundamental trade-off between experimental control and external validity on every possible dimension, and our study insists on high levels of the former. Relatedly, according to the Brutger et al. (2020) framework of experimental abstraction, our design choices along the dimensions of situational hypotheticality and contextual detail are unlikely to have substantially influenced our results. One thing we can consider is how our results might differ if elites in our detection task were shown in proportion to how often they were actually involved in scandals. For example, according to journalistic (Leonhardt and Thompson 2017) and scholarly (Bode et al. 2020) accounts of President Trump's behavior, it is possible that news consumers during this period would encounter many more authentic scandal videos of Trump than of other elites. Given the unique nature of President Trump's relationship with the media and traditional standards for evidentiary claims, we have reason to expect that the effect of these videos might differ from those of other Republican elites.

Similarly, we cannot test for heterogeneous effects according to the gender of the targeted politician. However, it is possible that deepfake effects on male targets are smaller than those for female targets because of sexism on the part of voters. Indeed, when we regress affect toward Warren on a measure of ambivalent sexism, ambivalent sexism is more predictive than the effect of the treatment condition (e.g., text, audio, deepfake).[8] Given the current trajectory of female candidate emergence (Bernhard and de Benedictis-Kessner 2021), the prevalence and potency of gender-based attacks received during their campaigns (Cassese and Holman 2018), and the fact that women are in general more likely to be the targets of online harassment, it is important to understand the potential effects of deepfakes for female candidates in particular. However, future studies may better disentangle the degree to which the effects we do and do not observe are due to implicitly conditioning on a female target as opposed to being generally true of deepfake effects.

Relatedly, Republicans and Democrats disproportionately encounter favorable media coverage of their party's elites to begin with, which suggests respondents' detection of deepfakes may look different in the wild. If all Democrats' and Republicans' false-positive rates were regraded by dropping noncongenial clips in their detection task, Democrats improve false positives from 24.3% to 13.7%, while Republicans improve from 18% to 17.8%. Ideological segregation and selective exposure in media consumption—to the extent that it exists—may thus attenuate rates of false skepticism about authentic media.

In this study, we elicited credibility perceptions of clippings ("is this clip real?"), which may be distinct from belief in the occurrence of the depicted event ("did X happen?"). In theory, someone could flag a video as a deepfake yet believe that the event still occurred. However, manipulation checks on two clips in our detection task suggest that respondents who believe the video is fake generally believe the event did not occur and vice versa (fig. J24 in app. J). Exploring the theoretical and empirical distinctions between these outcomes is a research agenda of its own.

Finally, we recognize that deepfake technology will continue to improve beyond the scope of this experiment. Although we have faithfully replicated the deepfake production process using the best available technology at the time of fielding, readers may

[8] Table G23 reports this regression, where ambivalent sexism is a measure created from a short question battery, shown in app. sec. K.1.

live in a world where open-source deepfake technology is capable of generating photorealistic deepfakes completely indistinguishable from authentic videos. In this case, reactions to deepfakes may more closely resemble the responses to real videos we have seen here, where cognitive effort and literacy still improve discernment, while partisanship continues to drive false beliefs depending on what is shown. Thus, while we encourage technological solutions to contain the spread of manipulated videos as well investments in both crowd workers and algorithms to detect deepfakes to begin with (Groh et al. 2022), there will never be a substitute for an informed, digitally literate, and reflective public for the practice of democracy.

ACKNOWLEDGMENTS

For excellent research assistance, we thank Jordan Duffin Wong. We thank the Wiedenbaum Center at Washington University in St. Louis for generously funding this experiment. For helpful comments, we thank the Political Data Science Lab and the Junior Faculty Reading Group at Washington University in St. Louis; the Imai Research Group; the Enos Research Design Happy Hour; the American Politics Research Workshop at Harvard University; the Harvard Experiments Working Group; and Jacob Brown, Andy Guess, Connor Huff, Yphtach Lelkes, Jacob Montgomery, and Steven Webster for helpful comments. We thank Hany Farid for sharing video clips used in this project. We are especially grateful to Sid Gandhi, Rashi Ranka, and the entire Deepfakeblue team for their collaboration on the production of videos used in this project. We thank Gary King for access to Brandwatch Twitter data.

REFERENCES

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. “Protecting World Leaders Against Deep Fakes.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 38–45.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. “The State of Deepfakes: Landscape, Threats, and Impact.” http://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

Barbera, Pablo. 2018. “Explaining the Spread of Misinformation on Social Media: Evidence From the 2016 US Presidential Election.” In *Symposium: Fake News and the Politics of Misinformation*. Washington, DC: APSA (American Political Science Association).

Bartels, Larry. 2002a. “Beyond the Running Tally: Partisan Bias in Political Perceptions.” *Political Behavior* 24 (2): 117–50.

Bartels, Larry. 2002b. “The Impact of Candidate Traits in American Presidential Elections.” In Anthony King, ed., *Leaders’ Personalities and the Outcomes of Democratic Elections*. Oxford University Press, 44–69.

Basinger, Scott. 2013. “Scandals and Congressional Elections in the Post-Watergate Era.” *Political Research Quarterly* 66 (2): 385–98.

Bateman, Jon. 2020. “Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.” Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2020/07/deepfakes-and-synthetic-media-in-the-financial-system-assessing-threat-scenarios?lang=en.

Baym, Geoffrey, and Lance Holbert. 2020. “Beyond Infotainment.” In Elizabeth Suhay, Bernard Grofman, and Alexander H. Trechsel, eds., *The Oxford Handbook of Electoral Persuasion*. Oxford University Press, 455.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Bennett, Lance, and Shanto Iyengar. 2008. “A New Era of Minimal Effects? The Changing Foundations of Political Communication.” *Journal of Communication* 58 (4): 707–731.

Berinsky, Adam, Vincent Hutchings, Tali Mendelberg, Lee Shaker, and Nicholas Valentino. 2011. “Sex and Race: Are Black Candidates More Likely to Be Disadvantaged by Sex Scandals?” *Political Behavior* 33 (2): 179–202.

Bernhard, Rachel, and Justin de Benedictis-Kessner. 2021. “Men and Women Candidates Are Similarly Persistent After Losing elections.” *Proceedings of the National Academy of Sciences* 118(26).

Bode, Leticia, Ceren Budak, Jonathan M. Ladd, Frank Newport, Josh Pasek, Lisa O. Singh, Stuart N. Soroka, and Michael W. Traugott. 2020. *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*. Brookings Institution Press.

Boukes, Mark, Hajo Boomgaarden, Marjolein Moorman, and Claes De Vreese. 2015. “At Odds: Laughing and Thinking? The Appreciation, Processing, and Persuasiveness of Political Satire.” *Journal of Communication* 65 (5): 721–44.

Boussalis, Constantine, Travis Coan, Mirya Holman, and Stefan Müller. 2021. “Gender, Candidate Emotional Expression, and Voter Reactions during Televised Debates.” *American Political Science Review* 115 (4): 1242–57.

Brader, Ted. 2006. *Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work*. Chicago: University of Chicago Press.

Brenton, Harry, Marco Gillies, Daniel Ballin, and David Chatting. 2005. “The Uncanny Valley: Does It Exist?” In *Proceedings of the Conference of Human Computer Interaction*.

Brown, Nina. 2019. “Congress Wants to Solve Deepfakes by 2020. That Should Worry Us.” *Slate Magazine*. https://slate.com/technology/2019/07/congress-deepfake-regulation-230-2020.html.

Brutger, Ryan, Joshua Kertzer, Jonathan Renshon, Dustin Tingley, and Chagai Weiss. 2020. “Abstraction and Detail in Experimental Design.” *American Journal of Political Science* 67 (4): 979–95.

Cassese, Erin, and Mirya Holman. 2018. “Party and Gender Stereotypes in Campaign Attacks.” *Political Behavior* 40 (3): 785–807.

Cassese, Erin, and Mirya Holman. 2019. “Playing the Woman Card: Ambivalent Sexism in the 2016 US Presidential Race.” *Political Psychology* 40 (1): 55–74.

Chesney, Robert, Danielle Citron, and Quinta Jurecic. 2019. “About That Pelosi Video: What to Do About ‘Cheapfakes’ in 2020.” *Lawfare*. https://www.lawfaremedia.org/article/about-pelosi-video-what-do-about-cheapfakes-2020.

Christianson, Sven-åke, and Elizabeth Loftus. 1987. “Memory for Traumatic Events.” *Applied Cognitive Psychology* 1 (4): 225–39.

Clinton, J., J. Cohen, J. Lapinski, and M. Trussler. 2021. “Partisan Pandemic: How Partisanship and Public Health Concerns Affect Individuals’ Social Mobility During COVID-19.” *Science Advances* 7 (2).

Costa, Mia. 2021. “Ideology, not Affect: What Americans Want from Political Representation.” *American Journal of Political Science* 65 (2): 342–58.

Damann, Taylor, Dean Knox, and Christopher Lucas. 2023. “A Framework for Studying Causal Effects of Speech Style: Application to US Presidential Campaigns.” Working paper, August 17. https://dcknox.github.io/files/DamannKnoxLucas_CausalEffectsSpeechStyle.pdf.

Darr, Joshua, Nathan Kalmoe, Kathleen Searles, Mingxiao Sui, Raymond Pingree, Brian Watson, Kirill Bryanov, and Martina Santia. 2019. “Collision with Collusion: Partisan Reaction to the Trump-Russia Scandal.” *Perspectives on Politics* 17 (3): 772–87.

Davis, Raina. 2020. “Technology Factsheet: Deepfakes.” https://www.belfercenter.org/publication/technology-factsheet-deepfakes.

Dewan, Torun, and David Myatt. 2007. “Scandal, Protection, and Recovery in the Cabinet.” *American Political Science Review* 101 (1): 63–77.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. “Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?” *The International Journal of Press/Politics* 26 (1): 69–91.

Druckman, James, and Mary McGrath. 2019. “The Evidence for Motivated Reasoning in Climate Change Preference Formation.” *Nature Climate Change* 9 (2): 111–19.

Dziuda, Wioletta, and William Howell. 2021. “Political Scandal: A Theory.” *American Journal of Political Science* 65 (1): 197–209.

Egami, Naoki, Christian Fong, Justin Grimmer, Margaret Roberts, and Brandon Stewart. 2018. “How to Make Causal Inferences Using Texts.” Preprint, *arXiv*. https://arxiv.org/abs/1802.02163.

Enders, Adam M., and Steven M. Smallpage. 2019. “Informational Cues, Partisan Motivated Reasoning, and the Manipulation of Conspiracy Beliefs.” *Political Communication* 36 (1): 83–102.

Esralew, Sarah, and Dannagal Goldthwaite Young. 2012. “The Influence of Parodies on Mental Models: Exploring the Tina Fey–Sarah Palin Phenomenon.” *Communication Quarterly* 60 (3): 338–52.

Frum, David. 2020. “The Very Real Threat of Trump’s Deepfake.” *The Atlantic.* https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/.

Galston, William A. 2020. “Is Seeing Still believing? The Deepfake Challenge to Truth in Politics.” *Brookings.* https://www.brookings.edu/articles/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

Galvis, Ángela Fonseca, James M. Snyder Jr., and BK Song. 2016. “Newspaper Market Structure and Behavior: Partisan Coverage of Political Scandals in the United States from 1870 to 1910.” *The Journal of Politics* 78 (2): 368–81.

Gazis, Olivia, and Stefan Becket. 2019. “Senators Pressure Social Media Giants to Crack Down on “Deepfakes.” *CBS News.* https://www.cbsnews.com/news/deepfakes-mark-warner-marco-rubio-pressure-social-media-giants-to-crack-down/.

Glick, Peter, and Susan Fiske. 1996. “The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism.” *Journal of Personality and Social Psychology* 70 (3): 491.

Grabe, Maria Elizabeth, and Erik Page Bucy. 2009. *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press.

Groh, Matthew, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. “Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds.” *Proceedings of the National Academy of Sciences* 119 (1).

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. “Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook.” *Science Advances* 5 (1).

Guess, Andrew M., and Kevin Munger. 2023. “Digital Literacy and Online Political Behavior.” *Political Science Research and Methods* 11 (1): 110–28.

Guess, Andrew, Michael Lerner, Benjamin Lyons, Jacob Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. “A Digital Media Literacy Intervention Increases Discernment Between Mainstream and False News in the United States and India.” *Proceedings of the National Academy of Sciences* 117 (27): 15536–45.

Hamel, Brian, and Michael Miller. 2019. “How Voters Punish and Donors Protect Legislators Embroiled in Scandal.” *Political Research Quarterly* 72 (1): 117–31.

Hancock, Jeffrey, and Jeremy Bailenson. 2021. “The Social Impact of Deepfakes.” *Cyberpsychology, Behavior, and Social Networking* 24 (3): 149–52.

Hwang, Tim, and Clint Watts. 2020. “Opinion: Deepfakes Are Coming for American Democracy. Here’s How We Can Prepare.” *The Washington Post.* https://www.washingtonpost.com/opinions/2020/09/10/deepfakes-are-coming-american-democracy-heres-how-we-can-prepare/.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Kassin, Saul M., and David A. Garfield. 1991. “Blood and Guts: General and Trial-Specific Effects of Videotaped Crime Scenes on Mock Jurors.” *Journal of Applied Social Psychology* 21 (18): 1459–72.

Knox, Dean, and Christopher Lucas. 2021. “A Dynamic Model of Speech for the Social Sciences.” *American Political Science Review* 115 (2): 649–66.

Leeper, Thomas J., and Rune Slothuus. 2014. “Political Parties, Motivated Reasoning, and Public Opinion Formation.” *Political Psychology* 35:129–56.

Leonhardt, David, and Stuart Thompson. 2017. “Trump’s Lies.” *New York Times.* https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html.

Lewis, Rebecca. 2018. “Alternative Influence: Broadcasting the Reactionary Right on YouTube.” https://apo.org.au/sites/default/files/resource-files/2018-09/apo-nid193281.pdf.

Luca, Mario, Kevin Munger, Jonathan Nagler, and Joshua Tucker. 2022. “You Won’t Believe Our Results! But They Might: Heterogeneity in Beliefs About the Accuracy of Online Media.” *Journal of Experimental Political Science* 9 (2): 267–77.

Lupia, Arthur. 2016. *Uninformed: Why People Know So Little About Politics and What We Can Do About It*. Oxford, UK: Oxford University Press.

McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. McGraw-Hill.

Mitchell, Amy, Elisa Shearer, Jeffrey Gottfried, and Michael Barthel. 2016. “Where Americans Are Getting News About the 2016 Presidential Election.” *Pew Research Center.* https://www.pewresearch.org/wp-content/uploads/sites/8/2016/02/PJ_2016.02.04_election-news_FINAL.pdf.

Mosleh, Mohsen, Gordon Pennycook, Antonio Arechar, and David Rand. 2021. “Cognitive Reflection Correlates with Behavior on Twitter.” *Nature Communications* 12 (1): 1–10.

Munger, Kevin. 2019. “The Limited Value of Non-Replicable Field Experiments in Contexts with Low Temporal Validity.” *Social Media+Society* 5 (3).

Munger, Kevin. 2023. “Temporal Validity as Meta-science.” *Research & Politics* 10 (3). https://doi.org/10.1177/20531680231187271.

Munger, Kevin, Ishita Gopal, Jonathan Nagler, and Joshua Tucker. 2021. “Accessibility and Generalizability: Are Social Media Effects Moderated by Age or Digital Literacy?” *Research & Politics* 8 (2).

Munger, Kevin, Mario Luca, Jonathan Nagler, and Joshua Tucker. 2020. “The (Null) Effects of Clickbait Headlines on Polarization, Trust, and Learning.” *Public Opinion Quarterly* 84 (1): 49–73.

Mutz, Diana C. 2016. *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton, NJ: Princeton University Press.

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. “Partisan Polarization Is the Primary Psychological Motivation Behind Fake News Sharing on Twitter.” *American Political Science Review* 115 (3): 999–1015.

Parkin, Simon. 2019. “The Rise of the Deepfake and the Threat to Democracy.” *The Guardian.* https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

Pennycook, Gordon, and David G. Rand. 2019. “Lazy, not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning.” *Cognition* 188:39–50.

Pennycook, Gordon, and David G. Rand. 2022. “Accuracy Prompts Are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation.” *Nature Communications* 13 (1): 1–12.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, Dean Eckles, and David G. Rand. 2021. “Shifting Attention to Accuracy Can Reduce Misinformation Online.” *Nature* 592 (7855): 590–95.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. 2020. "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science* 31 (7): 770–80.

Prochaska, Stephen, Michael Grass, and Jevin West. 2020. "Deepfakes in the 2020 Election and Beyond: Lessons from the 2020 Workshop Series." *Center for an Informed Republic*. https://jevinwest.org/papers/Prochaska2020CIP_Deepfake_Report.pdf

Puglisi, Riccardo, and James M. Snyder Jr. 2011. "Newspaper Coverage of Political Scandals." *The Journal of Politics* 73 (3): 931–50.

Rubio, Marco, and Mark Warner. 2019. "Warner, Rubio Express Concern Over Growing Threat Posed by Deepfakes." https://www.warner.senate.gov/public/index.cfm/2019/10/warner-rubio-express-concern-over-growing-threat-posed-by-deepfakes (accessed September 19, 2021).

Saltz, Emily, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. "Misinformation Interventions Are Common, Divisive, and Poorly Understood." *HKS Misinformation Review*. https://misinforeview.hks.harvard.edu/article/misinformation-interventions-are-common-divisive-and-poorly-understood/.

Schaffner, Brian F., Matthew MacWilliams, and Tatishe Nteta. 2018. "Understanding White Polarization in the 2016 Vote for President: The Sobering Role of Racism and Sexism." *Political Science Quarterly* 133 (1): 9–34.

Sirlin, Nathaniel, Ziv Epstein, Antonio A. Arechar, and David G. Rand. 2021. "Digital Literacy Is Associated with More Discerning Accuracy Judgments but not Sharing Intentions." *HKS Misinformation Review*. https://misinforeview.hks.harvard.edu/article/digital-literacy-is-associated-with-more-discerning-accuracy-judgments-but-not-sharing-intentions/.

Stecula, Dominik A., and Mark Pickup. 2021. "Social Media, Cognitive Reflection, and Conspiracy Beliefs." *Frontiers in Political Science* 3.

Tappin, Ben, Gordon Pennycook, and David Rand. 2021. "Rethinking the Link Between Cognitive Sophistication and Politically Motivated Reasoning." *Journal of Experimental Psychology: General* 150 (6): 1095–114.

Teele, Dawn, Joshua Kalla, and Frances McCall Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112 (3): 525–41.

Ternovski, John, Joshua Kalla, and P. Aronow. 2022. "The Negative Consequences of Informing Voters About Deepfakes: Evidence From Two Survey Experiments." *Journal of Online Trust and Safety* 1 (2).

Ternovski, John, and Lilla Orr. 2022. "A Note on Increases in Inattentive Online Survey Takers Since 2020." *Journal of Quantitative Description: Digital Media* 2. https://doi.org/10.51685/jqd.2022.002.

Vaccari, Cristian, and Andrew Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media+ Society* 6 (1).

Wakefield, Jane. 2022. "Deepfake Presidents Used in Russia-Ukraine War." *BBC News*. https://www.bbc.com/news/technology-60780142.

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press.

Wittenberg, Chloe, Ben M. Tappin, Adam J. Berinsky, and David Rand. 2021. "The (Minimal) Persuasive Advantage of Political Video Over Text." *Proceedings of the National Academy of Sciences* 118 (47). https://doi.org/10.1073/pnas.2114388118.

Zaller, John. 1998. "Monica Lewinsky's Contribution to Political Science." *PS: Political Science & Politics* 31 (2): 182–89.