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

Soubhik Barari† Christopher Lucas‡ Kevin Munger§

The Journal of Politics 87, no. 2 (2025): 510-526

Abstract

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.

Keywords: Media effects; deepfakes; scandals; experimental methods

*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. All data and code is publicly available here. Replication files are available in the JOP Data Archive on Dataverse.The empirical analysis has been successfully replicated by the JOP replication analyst. 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. Supplementary material for this article is available in the appendix in the online edition.

†Research Methodologist, NORC at the University of Chicago, 55 E Monroe St, Chicago IL, 60603; URL: soubhikbarari.com, Email: barari-soubhik@norc.org

‡Associate Professor, Department of Political Science and Division of Computational and Data Sciences; Washington University in St. Louis, St. Louis, Missouri, 63130; URL: christopherlucas.org, Email: christopher.lucas@wustl.edu

§Assistant Professor, Department of Political and Social Sciences, European University Institute, Florence, Italy 50014; URL: kevinmunger.com, Email: kevin.munger@eui.eu

> *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 previously available 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 summarises the two major technologies for the production of deepfakes, which, by many counts, are responsible for the production of the vast majority of political deepfakes at the time of writing (Lewis, 2018; Davis, 2020; Ajder et al., 2019).[1]

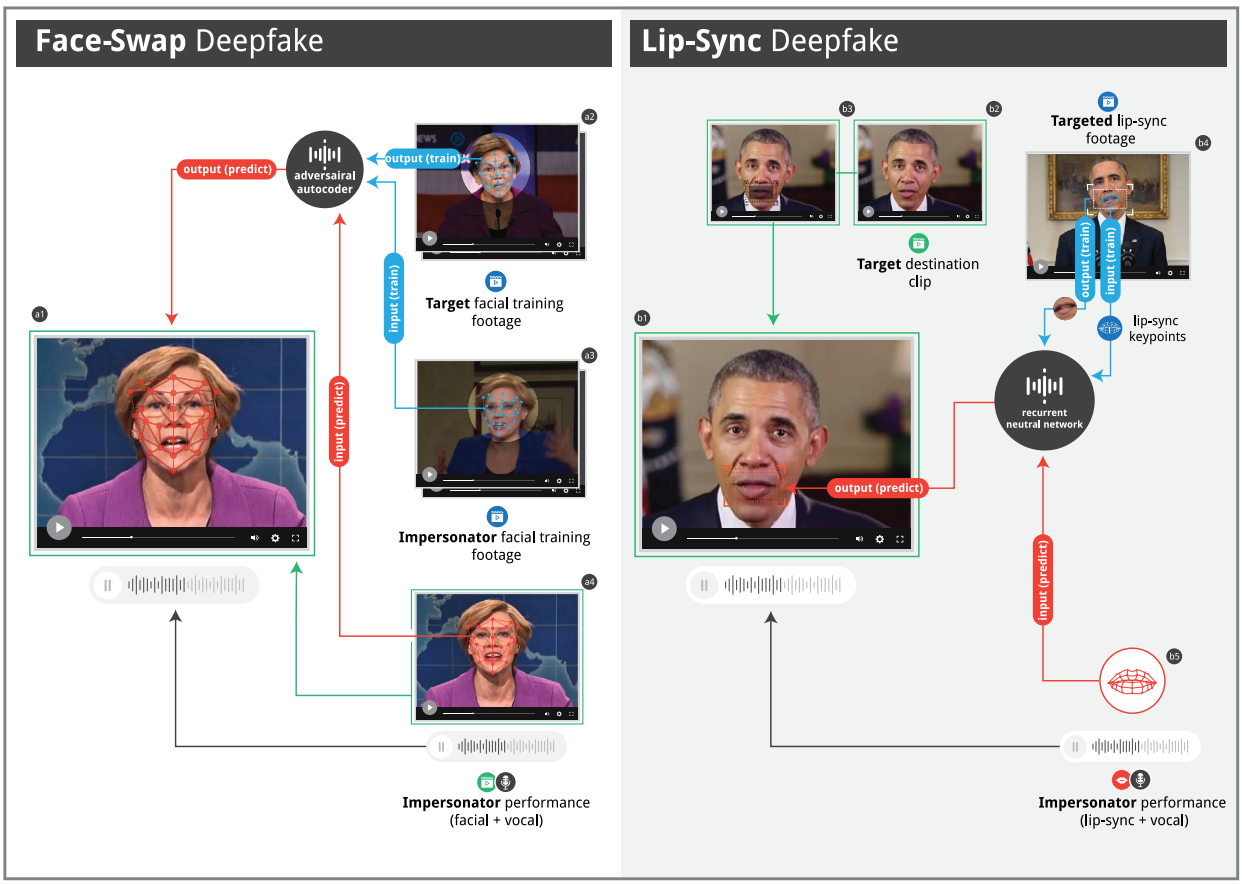
Because deepfakes let ordinary users produce media that falsely depicts someone saying and doing that which they never said nor did, it is commonly suggested that deepfakes uniquely threaten the electorate's trust in the information it consumes. This concern is not without cause; since the advent of 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 deepfakes continued to circulate on less-moderated

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

channels like Signal, WhatsApp or Telegram.

Because of this threat, lawmakers (Gazis and Becket, 2019; Brown, 2019; Lum, 2019; Galston, 2020), news outlets (Harwell, 2019; Parkin, 2019; Frum, 2020; Hwang and Watts, 2020; Schick, 2020) and civil society groups (Lewis, 2018; Davis, 2020; Ajder et al., 2019; Bateman, 2020) 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).

Figure 1: **How Deepfake Videos are Generated**



Notes: 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 (**black**) and the background context for the clip (**green**) where the facial features are swapped (**red**) via a trained (**blue**) deep learning model called an autoencoder. The right illustrates a *lip-sync deepfake* which requires a destination clip of the target (**green**) and a vocal impersonator's performance including their audio (**black**) and lip-sync keypoints (**red**); these keypoints are transferred into a matching synthetic lip-sync video of the target via a deep convolutional neural network model trained on the target (**blue**).

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 non-political, 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 Appendix Section 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 (Puglisi and Snyder Jr, 2011; Galvis, Snyder Jr and Song, 2016), election outcomes (Basinger, 2013; Hamel and Miller, 2019), the afflicted individuals' career trajectories, the legislative behavior of co-partisans (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 significantly 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 due to the proliferation of choice in media (Bennett and Iyengar, 2008) or partisan attachment and resistance to counter-attitudinal information (Bartels, 2002*a*). 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 Parksinson'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). We thus present our research as a direct intervention into a 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.

1 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 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) demonstrates 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 effectively 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), persuasion in election campaigns (Grabe and Bucy, 2009), and encouraging belief in climate change (Goldberg et al., 2019).

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., 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]

1.1 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 summarise 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 7 times as many articles from fake news domains as the youngest age group" during the 2016 US Presidential election (Guess, Nagler and Tucker, 2019). 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. (2020) find that

[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 Section 4.2 for further discussion.

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 priors about the credibility of new evidence – directs voters' attitudes about events, issues, and candidates (Druckman and McGrath, 2019; Leeper and Slothuus, 2014; Enders and Smallpage, 2019). Moreover, voters' evaluations of candidates or events can be driven by prior negative stereotypes towards groups including women (Teele, Kalla and Rosenbluth, 2017; Cassese and Holman, 2019). Women are a 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 U.S. 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; Pennycook et al., 2019). Similarly, those with little political knowledge may have little prior exposure to the targetted 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

[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).

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, 2019, 2021).

Consistent with our expectations for **RQ1**, we pre-registered the prediction that all subgroups in Table 1 would be differentially susceptible to deepfake misinformation over text and audio misinformation.

1.2 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.

2 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

[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 Appendix 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 1: **Subgroups Hypothesized to Perceive Deepfakes As Credible**

Subgroup	Mechanism(s) of Credibility
Intervenable	
Older adults ($\geq$ 65 y.o.)	Inability to evaluate accuracy of digital information (Guess, Nagler and Tucker, 2019; Barbera, 2018; Osmundsen et al., 2020)
Partisans (with out-partisan target)	• Directional motivated reasoning about out-partisans (Leeper and Slothuus, 2014; Enders and Smallpage, 2019) • Accuracy motivated reasoning about out-partisans (Druckman and McGrath, 2019; Tappin, Pennycook and Rand, 2020)
Sexists (with female target)	• Consistency with prior hostile beliefs about women (Glick and Fiske, 1996; Schaffner, MacWilliams and Nteta, 2018; Cassese and Holman, 2019) • Consistency with prior benevolent beliefs about women (Glick and Fiske, 1996; Schaffner, MacWilliams and Nteta, 2018; Cassese and Holman, 2019)
Low cognitive reflection	Overreliance on intuition over analytical thinking in making judgments (Pennycook and Rand, 2019; Pennycook et al., 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; Tucker et al., 2018)
Low digital literacy	• Inability to evaluate accuracy of digital information • Limited/no recognition of deepfake technology (Guess et al., 2020; Munger et al., 2020)
Non-Intervenable	
Low accuracy salience	Limited/no attention to factual accuracy of media (Pennycook et al., 2020, 2019)
Uninformed about deepfakes	Limited/no recognition of deepfake technology

Notes: 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 paper.

and Appendix Figure B6 provides a graphical illustration of the survey flow.

Our design is motivated by a number of considerations. Firstly, 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 set of pre-existing deepfakes of many different elites made by experts and amateurs alike made 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 Appendix Section 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 post-stratification weights estimated from the U.S. Census in Appendix G. Details of this post-stratification and other characteristics of the sample are given in Appendix F.

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. Post-exposure 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: eight authentic clips of political elites • Low-fake feed: six authentic clips, two deepfakes of political elites • High-fake feed: two authentic clips, six 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

2.1 Incidental exposure experiment

In the first experiment, we implement a 2 x 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

[5]Appendix Section K displays these surrounding clips, which were included in order 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

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 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]

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

Klobuchar, a video of Biden, and similar media.

[6]We discuss this collaboration further in Appendix C. While an academic-industry partnership may be a unique source for a deepfake video, Appendix Section 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 these 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 Section 3). 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).

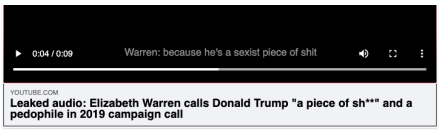
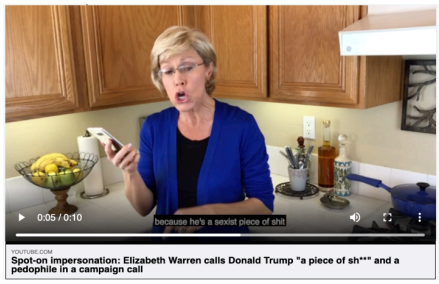
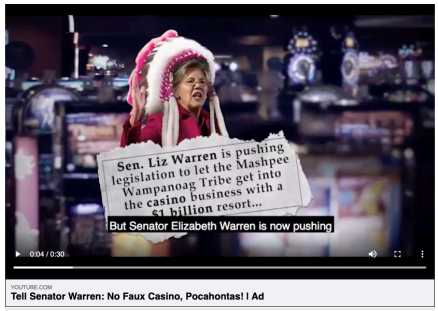
conduct exploratory analyses which show small differences across conditions (Appendix J).

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 towards 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 5 point scale) in between other distraction evaluations (funny, offensive, informative). Consequently, respondents are also asked to evaluate how warmly or coldly they feel towards 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, due to 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 towards Warren between the scandal clippings and these reference stimuli.

Table 3: **Experimental Conditions in Incidental Exposure Experiment**

	Condition	Description of Variation	Example Clip
Scandal Clips (Script Held Constant)	`video` ($n = 872$)	Face-swap performed on video in `skit` condition; title and video edited to resemble leaked video footage.	
	`audio` ($n = 954$)	Visuals stripped from `video` condition; title edited to resemble leaked hot mic.	
	`text` ($n = 950$)	Visuals and sound stripped from `video` condition; title describes scandal as a leak; subtitle describes event captured on video.	
	`skit` ($n = 956$)	Filmed impersonator portraying a campaign scandal event; used to create `video` and `audio` conditions.	
Reference Stimuli	`ad` ($n = 935$)	Campaign attack advertisement describing real scandal event.	
	`control` ($n = 916$)	No stimulus presented.	N/A

2.2 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 of the subjects not taking part in this task) are debriefed about whether or not 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 for 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 (or reductions in accuracy) in the detection task can be decomposed into false negatives, or misclassifications of deepfakes as authentic, and false positives, or misclassification of authentic clips as deepfakes. We measure both, in addition to overall accuracy, to gauge our 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 upwards, 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.

2.3 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.

3 Results

Figures 2–5 summarise 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 towards 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 pre-registered subgroup traits and by clips respectively. We organize our results into three main findings, each of which we discuss in detail in relation to our original hypotheses, and conclude with a brief discussion of external validity.

For all of results involving multiple group-wise comparisons or estimating multiple substantive coefficients, we adjust p-values according to the Benjamini-Hochberg "step-up" pro-

cedure which bounds each group of tests' false discovery rate at $\alpha = 0.05$ 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, i.e. falls within the equivalence bounds of $\pm 0.5\sigma$. We now summarize our findings.

1. 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 Figure 2). However, the videos were, on average, less credible than the faked audio (44%) and 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 $\pm 0.5\sigma$ (≈ 0.68) 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 towards the target elite? Relative to no exposure, videos do slightly decreaase Elizabeth Warren's favorability as measured by the 0-100 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 Figure 2), we find few. The

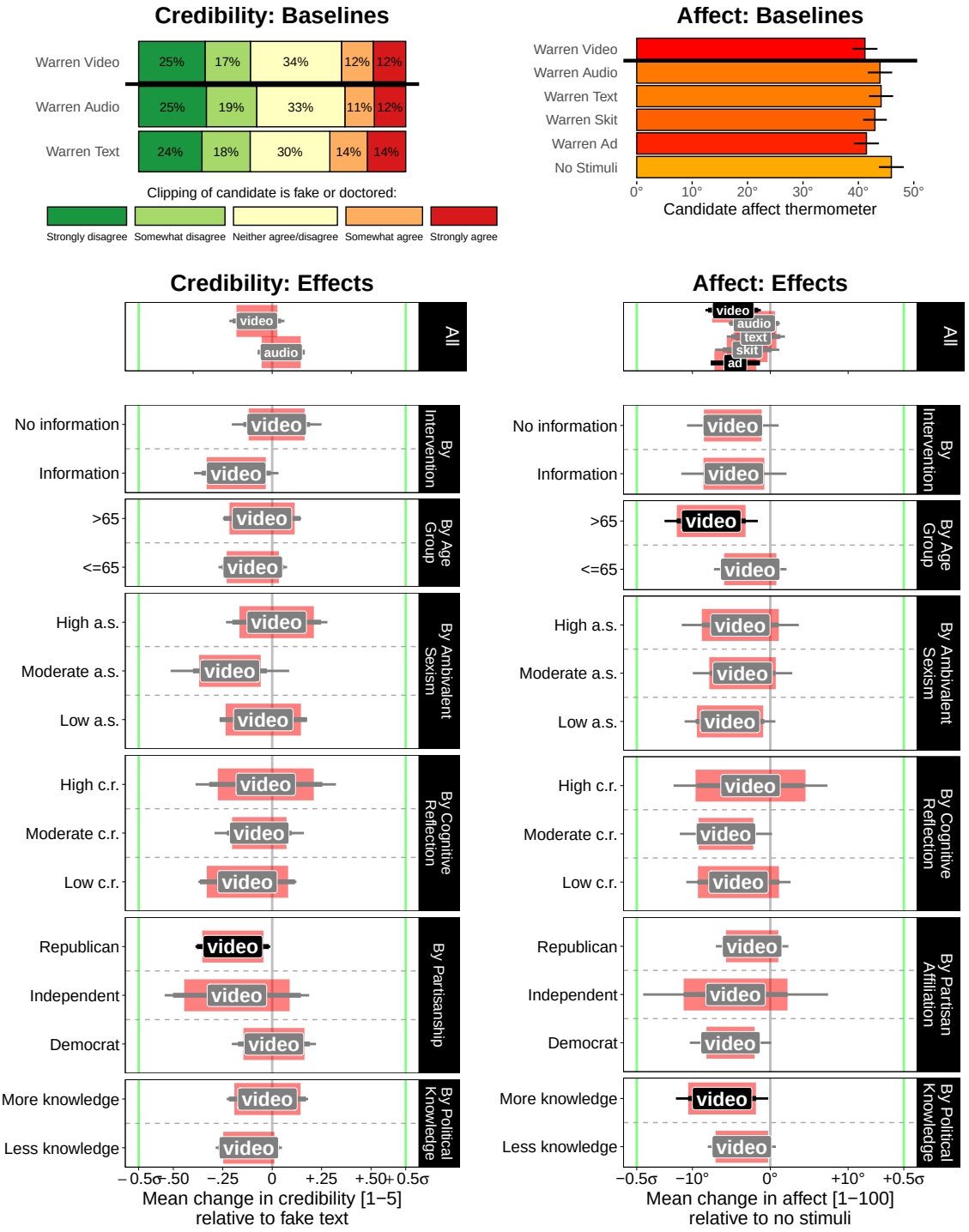
answer we give to **RQ2** is then also "no". This is not to say that these subgroups are not moved by a scandal of Elizabeth Warren, *in general*. To take the most notable examples, sexist attitudes and out-party identification predict increases in the credibility (substantively large in the latter case) of the scandal stimulus (Appendix Figure J14, Tables G18–G19, Tables G21–G22), but not disproportionately so for the deepfake relative to the headline or audio clipping.

2. Digital literacy and political knowledge improve discernment more than information. Baseline performance accuracy (Figure 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 1/3 deepfakes, 2/3 authentic clips. A little more than $1/3^{rd}$ of all deepfakes in our feed are undetected while a little under $1/3^{rd}$ 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 Figure 4). While information and accuracy salience fail, Figure 4 shows that respondent traits – specifically digital literacy[7], political knowledge and, to a lesser extent, cognitive reflection – predict the most substantively meaningful improvements. Republicans also appear to marginally outperform Democrats and Independents, though scoring little less than a full clip higher in correct classifications than the rest.

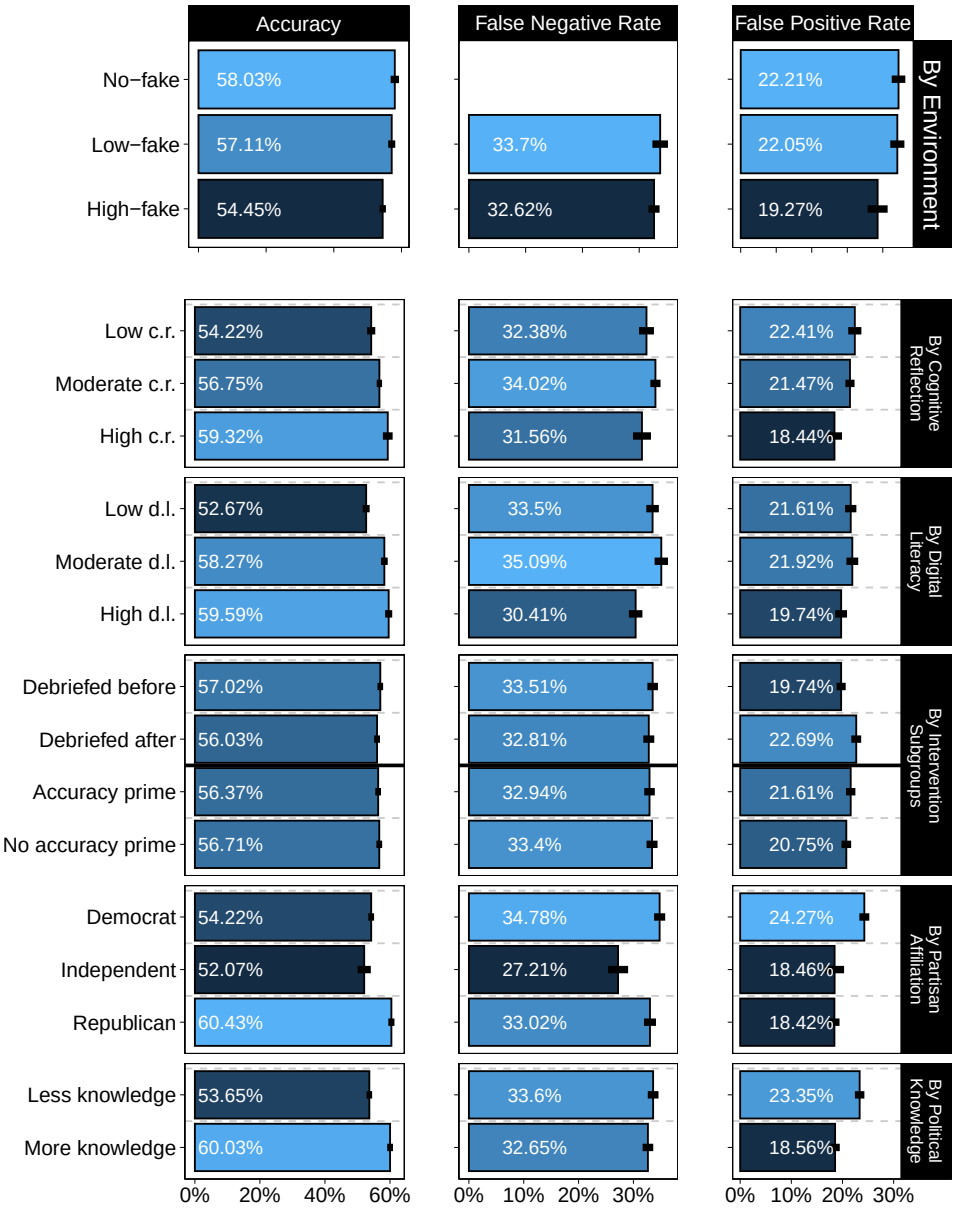
[7]Note that digital literacy predicts significant gains in accuracy, but no significant reductions in false negatives or false positives – one reason is that digital literacy predicts fewer "I don't know" responses which improves a respondent's accuracy in the detection task, but does not improve their false negative rate or false positive rate. Another is that respondents in the no-fake condition can only have a zero false negative rate; large accuracy gains in this condition would only improve the false positive rate.

Figure 2: **Relative to Other Stimuli, Effects of Incidental Exposure to a Deepfake Video are Minimal Overall and across Subgroups**



Notes: Categories for ambivalent sexism are constructed as equal-sized percentiles from sample values. Thicker lines denote 95% confidence intervals (CIs), thinner lines denote 95% CIs adjusted for multiple subgroup comparisons (Benjamini and Hochberg, 1995), **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`.

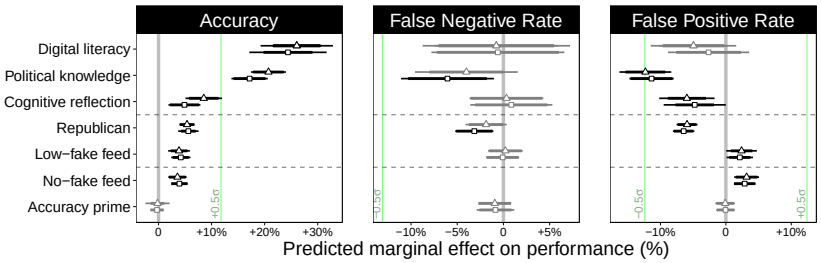
Figure 3: **Performance Comparisons in Deepfake Detection Task by Subgroup**



Notes: 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 missing-ness threshold. Accuracy is the % of all videos in the task correctly classified as either fake or real. False negative rate is the % 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 % of authentic videos in the task incorrectly classified as deepfakes.

3. Discernment of authentic videos varies significantly by partisanship more than deepfakes. Remarkably, although partisanship overall predicts small effects on per-

Figure 4: **Predictors of Detection Task Performance**

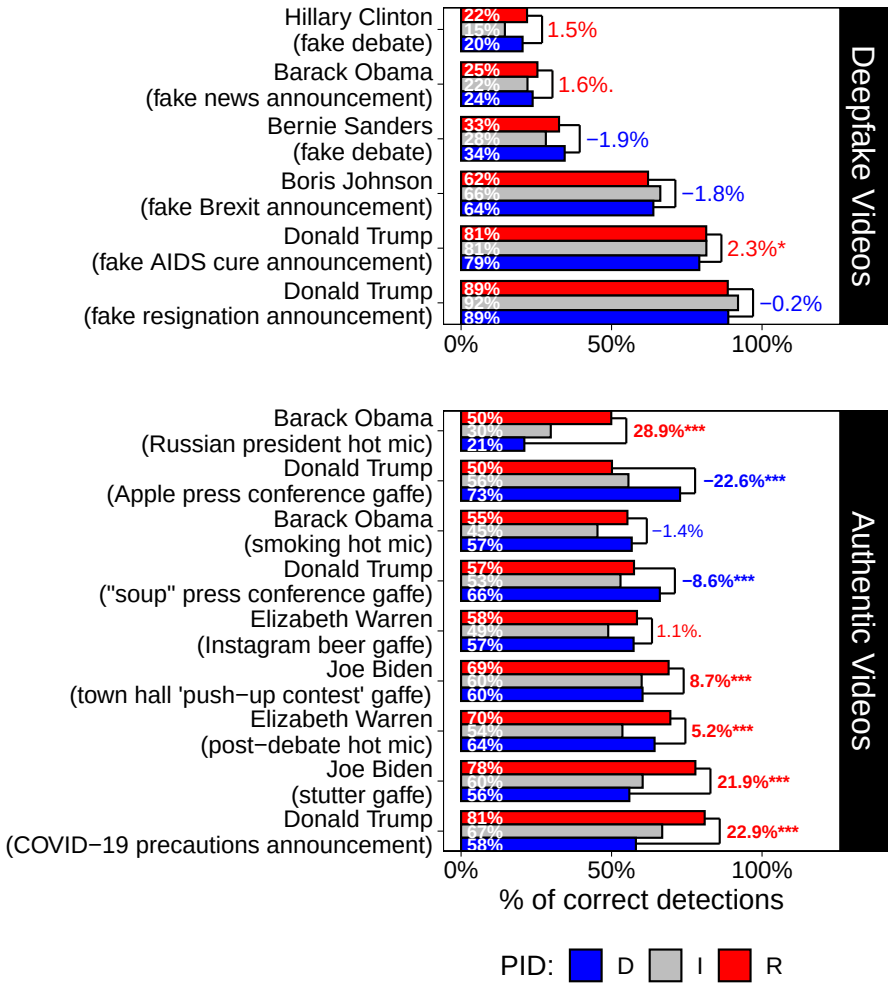


Notes: Predictors are grouped by dashed grey lines into respondent traits (all re-scaled to the [0,1] range), detection environment (relative to high-fake), and intervention assignment. Predictors include all group indicators from Table 1 excluding age which has no significant effects on performance (see Appendix 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 post-stratification model (see Appendix Section F). Appendix Figure I11 shows that dropping non-respondents in the task does not change the substantive interpretation of detection experiment results.

formance relative to other traits, an examination of individual clips (Figure 5) reveals some massive performance gaps between Democrats and Republicans, but *only* for real videos. 50% of Republicans believed that real leaked footage of Obama caught insinuating a post-election deal with the Russian president was authentic compared to 21% of Democrats, a highly significant differential according to a simple Chi-squared test ($\chi^2 = 338.3, p < 0.01$). 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 ($\chi^2 = 78.5, p < 0.01$). 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[8], at least for Democrats who by and large hold similarly cautionary attitudes towards COVID-19 (Clinton et al., 2020), only 58% of Democratic viewers flagged it as authentic whereas fully 81% of Republicans believed it to be real ($\chi^2 = 167.89, p < 0.01$). 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'

[8]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, 2002*b*).

Figure 5: **Detection Performance Comparisons Across Partisanship and Clip Authenticity**



Notes: p-values for differences in correct detection proportions between Democrats and Republicans (*** indicates $p' < 0.001$) derived from two-proportions z-tests where p' is the transformed larger p-value after adjusting for multiple comparisons via (Benjamini and Hochberg, 1995).

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 (Appendix Figure 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.

4 Discussion

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

4.1 Theoretical Implications

Our results for **RQ1** and **RQ2** concord with a growing body of research on video media effects (Vaccari and Chadwick, 2020; Dobber et al., 2020; Wittenberg et al., 2020), 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 non-video 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), 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 co-partisan 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 (Figure J20) and the type of scandal (Figures 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, 2022; Munger et al., 2021; Luca 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 (Pennycook and Rand, 2019; Stecula and Pickup, 2021; Mosleh et al., 2021).

In contrast, however, we do not find that priming subjects for accuracy has an effect on 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[9] 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, policy-makers should devote more time and resources to bolster the credibility of real news videos and curb the development and spread of deepfake videos that perpetrate psychological or social damages against their targets. Recent counts of deepfakes on the Internet find that most are non-consensual pornographic clips of women (Appendix 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 non-manipulated 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).

[9]As Figure J15 shows, the scandal scenario that re-affirms a past controversy (Elizabeth Warren misleading the public about her Cherokee heritage) appeared most credible to viewers, however this difference is statistically insignificant between all but the least credible situation (an instance of in-party incivility).

4.2 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 for 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 (Appendix 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).

Firstly, 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 require 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 tradeoff 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 suggests, 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; Quealy, 2021) 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, due to 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).[10] 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 re-graded by dropping non-congenial 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

[10]Table G23 reports this regression, where Ambivalent Sexism is a measure created from a short question battery, shown in Appendix Section K.1.

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 (Figure J24 in Appendix 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 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 still continues to drive false beliefs depending on what is shown. Thus, while we encourage technological solutions to constrain the spread of manipulated video as well investments in both crowdworkers 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.

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*. pp. 38–45.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli and Laurence Cullen. 2019. "The State of Deepfakes: Landscape, Threats, and Impact." *Policy Brief* .
URL: `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. APSA*.

Bartels, Larry. 2002*a*. "Beyond the running tally: Partisan bias in political perceptions." *Political behavior* 24(2):117–150.

Bartels, Larry. 2002*b*. "The Impact of Candidate Traits in American Presidential Elections." *Leaders' Personalities and the Outcomes of Democratic Elections* pp. 44–69.

Basinger, Scott. 2013. "Scandals and congressional elections in the post-Watergate era." *Political Research Quarterly* 66(2):385–398.

Bateman, Jon. 2020. "Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.".

Baym, Geoffrey and Lance Holbert. 2020. "Beyond Infotainment." *The Oxford Handbook of Electoral Persuasion* p. 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):721–744.

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):1242–1257.

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

Brutger, Ryan, Joshua Kertzer, Jonathan Renshon, Dustin Tingley and Chagai Weiss. 2020. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* .

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* .

Christianson, Sven-åke and Elizabeth Loftus. 1987. "Memory for Traumatic Events." *Applied Cognitive Psychology* 1(4):225–239.

Clinton, J., J. Cohen, J. Lapinski and M. Trussler. 2020. "Partisan Pandemic: How Partisanship and Public Health Concerns Affect Individuals' Social Mobility During COVID-19." *Science Advances* .

Costa, Mia. 2021. "Ideology, Not Affect: What Americans Want From Political Representation." *American Journal of Political Science* 65(2):342–358.

Damann, Taylor, Dean Knox and Christopher Lucas. 2023. "A Framework for Studying Causal Effects of Speech Style: Application to US Presidential Campaigns.".

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):772–787.

Davis, Raina. 2020. "Technology Factsheet: Deepfakes." *Policy Brief* .
URL: `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. 2020. "Do (Microtargeted) Deepfakes Have Real Éffects on Political Attitudes?" *The International Journal of Press/Politics* p. 1940161220944364.

Druckman, James and Mary McGrath. 2019. "The Evidence for Motivated Reasoning in Climate Change Preference Formation." *Nature Climate Change* 9(2):111–119.

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." *arXiv preprint arXiv: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):338–352.

Frum, David. 2020. "The Very Real Threat of Trump's Deepfake." *The Atlantic* .
URL: *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* .

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):368–381.

Gazis, Olivia and Stefan Becket. 2019. "Senators Pressure Social Media Giants to Crack Down on "Deepfakes"." *CBS News* .
URL: `https://www.cbsnews.com/news/deepfakes-mark-warner-marco-rubio-pressure-social-med`

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

Goldberg, Matthew, Sander van der Linden, Matthew Ballew, Seth Rosenthal, Abel Gustafson and Anthony Leiserowitz. 2019. "The Experience of Consensus: Video as an Effective Medium to Communicate Scientific Agreement on Climate Change." *Science Communication* 41(5):659–673.

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: Preva-

lence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5(1).

Guess, Andrew and Kevin Munger. 2022. "Digital literacy and online political behavior." *Political Science Research and Methods* pp. 1–19.

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):15536–15545.

Hamel, Brian and Michael Miller. 2019. "How voters punish and donors protect legislators embroiled in scandal." *Political Research Quarterly* 72(1):117–131.

Hancock, Jeffrey and Jeremy Bailenson. 2021. "The Social Impact of Deepfakes." *Cyberpsychology, Behavior, and Social Networking* 24(3):149–152.

Harwell, Drew. 2019. "Top AI Researchers Race to Detect 'Deepfake' Videos: 'We are outgunned'." *The Washington Post* .

Hwang, Tim and Clint Watts. 2020. "Opinion: Deepfakes Are Coming for American Democracy. Here's How We Can Prepare." *The Washington Post* .

Iyengar, Shanto and Donald R Kinder. 1987. *News That Matters: Television and American Opinion*. 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):1459–1472.

Knox, Dean and Christopher Lucas. 2021. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* 115(2):649–666.

Leeper, Thomas J and Rune Slothuus. 2014. "Political Parties, Motivated Reasoning, and Public Opinion Formation." *Political Psychology* 35:129–156.

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

Lewis, Rebecca. 2018. "Alternative Influence: Broadcasting the Reactionary Right on YouTube." *Data & Society* 18.

Luca, Mario, Kevin Munger, Jonathan Nagler and Joshua Tucker. 2021. "You Won't Believe Our Results! But They Might: Heterogeneity in Beliefs About the Accuracy of Online Media." *Journal of Experimental Political Science* pp. 1–11.

Lum, Zi-Ann. 2019. "Obama Tells Canadian Crowd He's Worried About 'Deepfake' Videos." *HuffPost Canada* .

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

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, 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* .

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

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2020. "Partisan Polarization Is the Primary Psychological Motivation Behind "Fake News" Sharing on Twitter.".

Parkin, Simon. 2019. "The Rise of the Deepfake and the Threat to Democracy." *The Guardian* .

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, 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):770–780.

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):590–595.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, Dean Eckles and David Rand. 2019. "Understanding and Reducing the Spread of Misinformation Online." *Working Paper* .

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* .

Puglisi, Riccardo and James M Snyder Jr. 2011. "Newspaper Coverage of Political Scandals." *The Journal of Politics* 73(3):931–950.

Quealy, Kevin. 2021. "Trump's Lies." *New York Times* .
URL: `https://www.nytimes.com/interactive/2021/01/19/upshot/trump-complete-insult-list.html`

Rubio, Marco and Mark Warner. 2019. "Warner, Rubio Express Concern Over Growing Threat Posed by Deepfakes.". [Online; accessed 19-September-2021].

Saltz, Emily, Soubhik Barari, Claire Leibowicz and Claire Wardle. 2021. "Misinformation Interventions are Common, Divisive, and Poorly Understood." *HKS Misinformation Review* .

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.

Schick, Nina. 2020. "Deepfakes Are Jumping From Porn to Politics. It's Time to Fight Back." *Wired United Kingdom* .

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* .

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. 2020. "Rethinking the Link Between Cognitive Sophistication and Politically Motivated Reasoning." *Journal of Experimental Psychology: General* .

Teele, Dawn, Joshua Kalla and Frances McCall Rosenbluth. 2017. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political*

Science Review .

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.

Tucker, Joshua, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *Hewlett Foundation* .

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* .

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

Wittenberg, Chloe, Jonathan Zong, David Rand et al. 2020. "The (Minimal) Persuasive Advantage of Political Video Over Text." *Working Paper* .

Zaller, John. 1998. "Monica Lewinsky's Contribution to Political Science." *PS: Political Science & Politics* 31(2):182–189.

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

Online Appendix

Contents

A Background on Deepfakes

To provide a broader context about the current state of deepfakes (e.g., their significance, circulation, intended purpose, characteristics in the wild, etc.), this section reviews a series of important empirical claims about political deepfake videos that we cite throughout the main text and use to motivate our experimental design. These facts are corroborated by the latest research at the time of writing and verified by our own original research from a variety of data sources (i.e. social media attention, search results, fact checks). We emphasize that the validity of these claims only holds at the time of writing and may require further investigation as deepfake technologies progress.

Deepfake videos and production technologies are rapidly growing in circulation. According to the artificial intelligence think tank Sentinel.ai, the number of deepfakes in circulation has doubled every six months between 2018 and 2020 (Tammekänd, Thomas and Peterson, 2020). The number of software repositories on GitHub to produce such videos has grown more than eight-fold during this period. In late 2019 – the time period when a deepfake similar to our stimulus could have surfaced of a Democratic primary candidates – an estimated 14,678 deepfake videos were in circulation on the public Internet. At the time of writing, there are estimated to be roughly 145,000 deepfake videos in circulation on the public Internet; for comparison, at this time, there are roughly 70,000 official Fox News video clips on YouTube and 154,000 official CNN video clips on YouTube.

Face-swap deepfakes are more widely produced and circulated than are lip-sync deepfakes. While conducting Google searches at the time of fielding the experiment in 2020 and again in 2021, and in 2022 we consistently found that software for producing face-swap deepfakes (e.g `faceswap`, `DeepFaceLab`) surface before do software packages for lip-sync deepfakes (e.g. `ObamaNet`, `wav2lip`). The authors find the same is true for the types of deepfake videos surfaced by YouTube search results. Similarly, demand for face-swap deepfakes seems to be higher than for lip-sync deepfakes: Google Trends shows that searches for "deepfake face-swap" (and associated synonyms and individual software) far exceed searches for "deepfake lip-sync" (and associated synonyms and individual software). These findings are consistent with (Tammekänd, Thomas and Peterson, 2020)'s comprehensive counts of face-swap and lip-sync produced deepfake videos across platforms in 2020. According to many other reports (Lewis, 2018; Davis, 2020; Ajder et al., 2019), face-swap softwares were the first deepfake technologies to receive popular press `Face2Face` in 2016; see also Suwajanakorn, Seitz and Kemelmacher-Shlizerman (2017) in 2017, `FakeApp` in 2018, `Faceswap` and `DeepFaceLab` in 2019.

Political deepfakes remain a minority of all circulated deepfakes, target a

handful of political elites, and exhibit power law dynamics. Of the thousands of deepfake videos in circulation, researchers estimate that 93% (Tammekänd, Thomas and Peterson, 2020) to 96% (Ajder et al., 2019) are non-consensual *pornography* videos of women, largely celebrities. The remaining 7% of non-pornographic deepfakes in circulation are mostly characterized (63%) as *comedy/entertainment*, rather than explicit disinformation (Westerlund, 2019). These videos mostly feature male (roughly 61% on YouTube) rather than female targets.

This is not to say that misinformative deepfake scandal videos that are the focus of our study do not actually exist: as we note in the main text and as Tammekänd, Thomas and Peterson (2020) enumerate, several deepfakes depicting state actors in scandals have circulated virally and account for the vast majority of views and shares on social media sites like Facebook and Twitter. The popularity and targets of political deepfakes, thus, exhibit *power law* distributions consistent with many other forms of Internet media (Adamic and Huberman, 2000).

By our best count, most videos claimed to be deepfakes on Twitter appear to target a small number of political elites. We collected all tweets in the United States mentioning the keyword `deepfake` between 2016 and 2021 and containing a link to a video[11], and we extracted the 100 most mentioned entities in these tweets. We find that Donald Trump is the single most mentioned political elite in this context, but – consistent with Tammekänd, Thomas and Peterson (2020) and Ajder et al. (2019) – find that the most mentioned individuals *overall* are celebrities (e.g. Tom Cruise, Leonardo DiCaprio, Jim Carrey, Mark Zuckerberg, Anthony Bourdain) The political elites as detected by a proprietary named-entity recognition algorithm are shown below in Table A4.

Table A4: **Top Political Actors Mentioned in the Context of Deepfakes in Tweets (2016-2021)**

Elite	Mentions
Donald Trump	30,125
Joe Biden	14,838
Queen Elizabeth II	11,796
Vladimir Putin	6,740
Barack Obama	5,702
Richard Nixon	3,460
Boris Johnson	2,624

Notes: Names identified from the list of the 100 most mentioned entities in tweets during this period according to Brandwatch's proprietary named-entity recognition software. This entity recognition and count was conducted before the circulation of deepfakes depicting President Volodymyr Zelenskyy during the Russian-Ukraine War.

This count should only be taken as a crude list of the elite targets of deepfakes. The

[11]Data provided by Brandwatch. Thanks to Gary King for access.

underlying tweets are not restricted to verified political deepfakes; many of these mentions may simply be confabulations of authentic videos with deepfakes or mere references of deepfake technology. Moreover, this crude count does not capture the circulation of undetected deepfakes on social media or elsewhere on the Internet. Overcoming these challenges remains the subject of current misinformation research.

Less technologically intensive video misinformation (e.g. "cheapfakes") surpass deepfakes in circulation. Using a dataset of publicly available fact checked claims from the Google Fact Check Markup Tool, we find that the vast majority of fact checks do not involve deepfake videos.[12] The few claims that do involve video (less than 1%) involve "cheapfakes" or videos edited (e.g. sped up, slowed down, selectively cut) to portray an event that did not occur (Nancy Pelosi slurring her speech, CNN reporter Jim Acosta acting aggressively towards staff, Obama claiming he is not a U.S. citizen). Tammekänd, Thomas and Peterson (2020) similarly theorize that the supply of cheapfakes likely outnumbers deepfakes on the Internet.

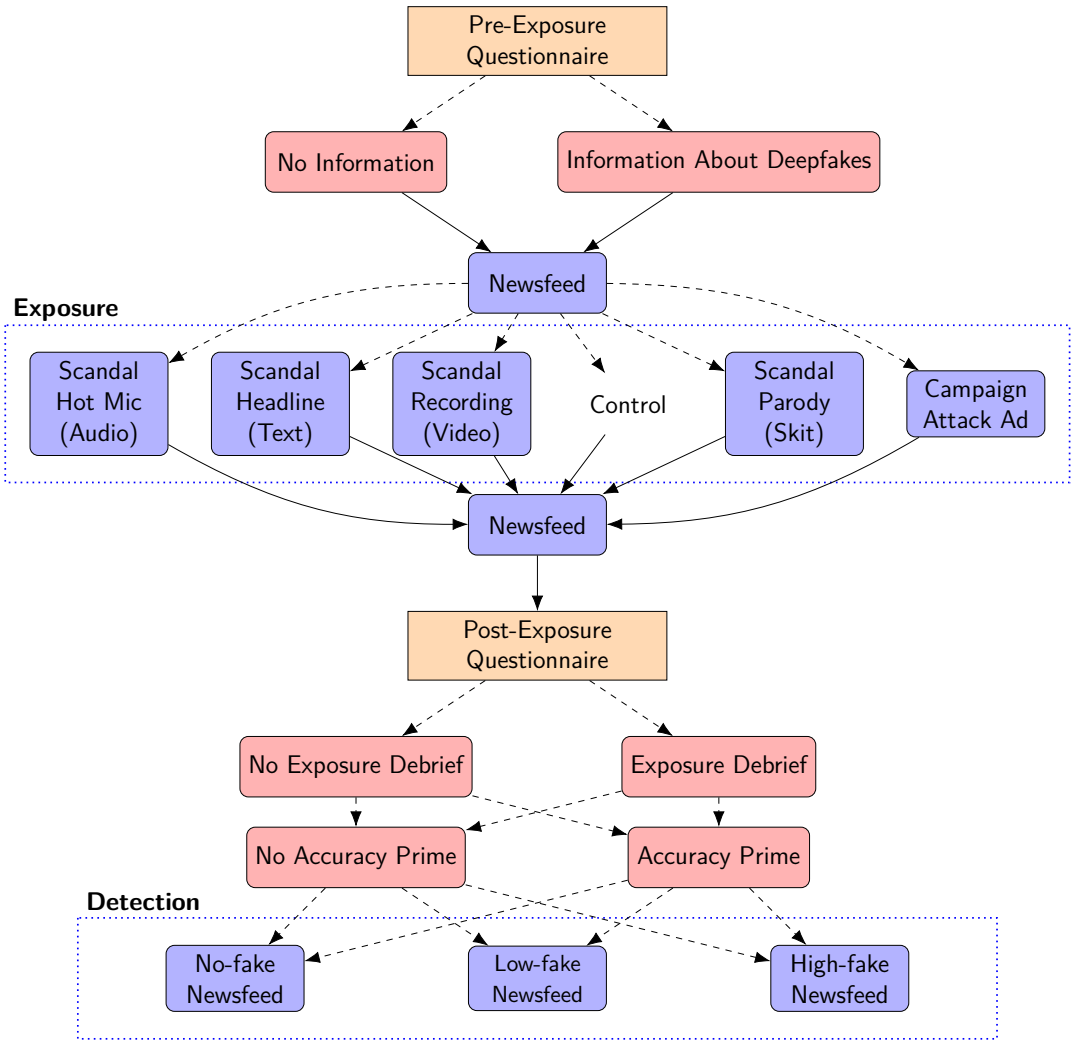
The true distribution of producer type for political deepfakes remains unknown. As Westerlund (2019) theorize, there are at least four kinds of deepfake producers with different intents: (1) *private hobbyists* and (2) *legitimate media outlets* producing deepfakes primarily for educational, recreational, or entertainment purposes; (3) *political interest groups* such as foreign governments and activists producing deepfakes to target opposition elites, and (4) *private bad actors* targeting members of the public and elites for the purposes of fraud, disinformation, or defamation. Westerlund (2019) note that although individual hobbyists are difficult to track down, online Reddit communities of deepfake producers have upwards of 100,000 members.

To the best of our knowledge, most explicitly labelled political deepfakes, as found on YouTube, are created by private hobbyists without a clear intent to falsely depict an actual event. After collecting all unique and explicitly identified deepfake videos from top 1,000 YouTube search results in 2022 and fact-checked claims from independent organizations, we link roughly 90% of them to a named individual, production company, or unincorporated media organization. The remainder are stylized deepfakes produced by media corporations (e.g. BuzzFeed or Bloomberg). However, given that these procedure has not systematically replicated across other major video-hosting portals and that this procedure cannot, of course, detect undetectable deepfakes masquerading as authentic videos, we cannot conclusively attribute most political deepfakes to private hobbyists.

[12]This dataset consists of all internally demarcated fact checks located on the sites of verified fact check organizations (e.g. PolitiFact, FactCheck) between March 31, 2016 and June 4, 2019. Fact checks from these organizations are typically conducted on "prominent claims" made by electoral candidates and political elites, that are given notable coverage in mainstream media.

B Experimental Setup

Figure B6: **Diagram of Experimental Flow**



Notes: In the **audio**, **text**, **skit**, and **video** exposure cells, respondents are further randomized to one of the 5 clippings in Table C5. Red cells denote interventions to minimize credibility in the Exposure experiment and improve discernment in the Detection experiment. Subjects who do not receive an exposure debrief prior to the detection task receive it immediately after in overall debrief.

C Stimuli in Exposure Experiment

C.1 Production details

We discuss the ethical reasoning behind our research design in more detail in Appendix E, but we first highlight here our selection of Elizabeth Warren for both ethical, practical, and substantive grounds. At the time of fielding, Senator Elizabeth Warren was a salient politician, making our experiment more ecologically valid than one with a low-profile or hypothetical politician – nearly all political deepfakes target high-profile politicians as we show in Section A. At the same time, no prominent, detectable deepfake of Warren was in current circulation, which avoids any bias in credibility perceptions if most respondents are already exposed and debriefed of a pre-existing deepfake. Finally, although most political deepfakes in contemporary circulation are of men, women are more likely to be the targets of non-political deepfakes (Ajder et al., 2019; Abram, 2020) and political harassment more broadly (Krook and Sanín, 2020). Using Warren allows us to specifically test whether motivated reasoning against women among subjects changes the effect of a deepfake.

In order to produce a high-quality deepfake, albeit demonstrative of current technological capabilities, we consulted *Buzzfeed* CEO Jonah Peretti, who produced the first viral deepfake video in 2018 of Barack Obama telling the world that "President Trump is a complete and utter dipsh*t". In the correspondence below, he explained how the deepfake, created via a professional actor's expert impersonation and synthesized via face-swap, came to exist, emphasizing the need for a high-quality impersonator and post-production:

> "The idea was shaped by Jordan's ability to do a good Obama impersonation - so that part isn't fancy tech. Jordan is just better at impressions than other people making deep fakes and he did Obama as a character on Key & Peele.
>
> Then we worked with Jared who used a combination of deep fake software downloaded from Reddit and Adobe products we use to do video effects and post production work. It wasn't straightforward and required a combination of approaches and Jared's prodigious talents."

In collaboration with an industry partner and following the lessons from our correspondence with *Buzzfeed*, we produced a series of deepfake videos using target footage of 2020 presidential candidate and senator Elizabeth Warren and performances of a professional Elizabeth Warren impersonator. Warren's campaign disseminated a series of campaign video recordings of the senator in her home kitchen making personal thank-you calls to campaign donors and, in some cases, discussing policy matters and events during the campaign. We produced a series of videos performances of the impersonator in a similar kitchen performing several different sketches that each represented a potential plausible "scandal" for Warren.

Table C5: **Descriptions and Scripts of Scandal Performances**

Scandal Description	Title	Script
In-Party Incivility	LEAK: Elizabeth Warren calls Joe Biden "a piece of sh*t" and a pedophile in 2019 campaign call	"Why shouldn't you vote for Joe Biden in 2020? Well, I'll tell you why: because he's a sexist piece of shit who likes to grope young girls, that's why."
Out-Party Incivility	LEAK: Elizabeth Warren calls Donald Trump "a piece of sh*t" and a pedophile in 2019 campaign call	"Why shouldn't you vote for Donald Trump in 2020? Well, I'll tell you why: because he's a sexist piece of shit who likes to grope young girls, that's why."
Past Controversy (racialized comment)	LEAK: Elizabeth Warren re-claims Cherokee heritage in 2019 campaign call	"Well, you know, as someone who has Cherokee ancestry, who's proud of their Native heritage, I deeply identify with other indigenous people and people of color in this country and I will do everything I can to fight for you in Washington."
Novel Controversy (anti-LGBTQ comment)	LEAK: Elizabeth Warren admits she doesn't "endorse the LGBTQ lifestyle" in 2019 campaign call	"Well, as a Christian woman of faith, I don't personally support the LGBTQ lifestyle, but I will try to do what I can for marriage equality in Washington."
Political Insincerity	LEAK: Elizabeth Warren flips stance on student loan debt in 2019 campaign call	"Well, I know I've said that before, but I don't really think that eliminating student loan debt for anyone is fair or realistic."

To script these scandals, we carefully studied past controversial hot mic scandals of Democratic politicians as well as exact statements made by Warren in these campaign videos. We then scripted statements in Warren's natural tone and affliction that appeared plausible in our qualitative assessments of the Warren campaign, and also captured a diversity of scandal types from incivility to controversial speech to policy-based insincerity. As such, these statements are not meant to invoke extreme disbelief or incredulity, though testing the credulity threshold of deepfake scandals in a principled manner could the subject of future research.

Table C5 describes the content of the final performances selected for our experiment. We

used the audio from these sketches for the audio condition and the video plus audio for the parody skit. We then performed the procedure to create a face-swap deepfake to produce the final deepfake video treatments, one for each selected scandal performance.

C.2 Face-swap algorithm

Deepfakes that swap the face of a target (e.g. President Barack Obama) with an actor (e.g. Hollywood actor Jordan Peele) – dubbed face-swaps in Figure 1 – are synthesized via a particular class of artificial neural networks called Adversarial Autoencoders (Makhzani et al., 2015).

The deepfaker's task is to train two autoencoders to accurately represent (encode) the two respective faces in a latent space and accurately reconstruct (decode) them as images. Let $\mathbf{X}_{\text{target}}$ denote a set of facial images of the target and $\mathbf{X}_{\text{actor}}$ denote a set of facial images of the actor. Denoting $\mathcal{G}_{\text{target}}$ as the function for the target autoencoder and $\mathcal{G}_{\text{actor}}$ as the function for the actor autoencoder, the networks are structured as $\mathcal{G}_{\text{target}}(x) = \delta_{\text{target}}\{\pi(x)\}$ and $\mathcal{G}_{\text{actor}}(x') = \delta_{\text{actor}}\{\pi(x')\}$ where π is an encoder subnetwork, δ_{target} and δ_{actor} are the decoder subnetworks for the target and actor respectively, and $x \in \mathbf{X}_{\text{target}}, x' \in \mathbf{X}_{\text{actor}}$. Both autoencoders share an encoder function π which discover a common latent representation for the targets' and actors' faces; separate decoders are charged with realistically reconstructing the input faces. The objective function to be optimized is:

$$\min_{\substack{\pi, \\ \delta_{\text{target}}, \\ \delta_{\text{actor}}}} \mathbb{E}_{x \sim \mathbf{X}_{\text{target}}}\Big[||\delta_{\text{target}}\{\pi(x)\} - x||^2\Big] + \mathbb{E}_{x' \sim \mathbf{X}_{\text{actor}}}\Big[||\delta_{\text{actor}}\{\pi(x')\} - x'||^2\Big] \tag{1}$$

To produce a *face-swap* deepfake given a audiovisual performance of the actor with respective facial image frames $\mathbf{Y}_{\text{actor}} = \Big[y_1, \ldots, y_N\Big]$, we input the frames into the trained target autoencoder which outputs $\mathbf{Y}_{\text{actor}} = \Big[\delta_{\text{target}}\{\pi(y_1)\}, \ldots, \delta_{\text{target}}\{\pi(y_N)\}\Big]$ that can be recombined with the audio of the actor's performance.

To maximize the realism of outputs created from actor inputs fed to the target autoencoder, we trained a third discriminator neural network $\mathcal{D}$ which aims to accurately classify the latent representations of images as belonging to either the target or actor. The final adversial objective is given as:

$$\begin{aligned}\max_{\mathcal{D}} \min_{\substack{\pi, \\ \delta_{\text{target}}, \\ \delta_{\text{actor}}}} & \mathbb{E}_{x \sim \mathbf{X}_{\text{target}}}\Big[||\delta_{\text{target}}\{\pi(x)\} - x||^2\Big] + \mathbb{E}_{x' \sim \mathbf{X}_{\text{actor}}}\Big[||\delta_{\text{actor}}\{\pi(x')\} - x'||^2\Big] \\ + \quad & \mathbb{E}_{x'' \sim \mathbf{X}}\Big[||\mathcal{D}\{\pi(x'')\} - \mathbf{1}\{x'' \in \mathbf{X}_{\text{actor}}\}||^2\Big]\end{aligned} \tag{2}$$

Optimization of this objective function can be performed via alternating iterative updating of the two networks' weights using stochastic gradient descent. After sufficient rounds of training, the target autoencoder can accurately reproduce the target's face using images of

only the actor's face and is thus able to effectively 'fool' the discriminator.

Finally, we reduced the resolution and bit-rate of our stimuli. This increases realism in two ways: (1) by masking any artifacts of the visual alterations of each face-swap and (2) by credibly presenting each video as a 'leaked' mobile phone recording. Both attributes are representative of deepfakes in current circulation, according to our qualitative assessment of the observable population described in Appendix A.

D Stimuli in Detection Experiment

This section provides sceenshots of the videos used in the detection experiment. All subject are assigned a mix of videos in which there are either no deepfakes (i.e., all displayed videos are of real media), a low proportion of deepfakes, or a high proportion of deepfakes. Each of these three conditions employs eight videos. While the order in which videos are presented varies within these conditions, the videos within each condition are fixed across subjects. The choice and frequency of targets (e.g., Donald Trump, Joe Biden, Barack Obama) and lip-syncs vs. face-swaps was informed by our best knowledge of the distribution of these videos on the public Internet (see Appendix Section A).

Subjects assigned to the no-fake condition saw real videos D7a through D7h. Subjects in the low-fake condition saw fake videos D8a and D8b, and real videos D7c, D7d, D7e, D7g, D7h, and D7i. Subjects in the high-fake condition saw fake videos D8a, D8b, D8c, D8d, D8e, D8f and real videos D7b and D7g.

Heterogeneity in detection performance at the clip level (both for the entire pool and across subgroups) can be found in Section J.

E Ethical Considerations

We highlight the ethical considerations pursuant to a study that uses stimuli which are expected to be uniquely deceptive.

First, in addition to the subjects randomly assigned to a debrief in the middle of the survey, we extensively debrief all subjects at the completion of the survey. This debrief goes beyond the standard description of study procedures. We require respondents to type out the following phrase, depending on which experimental arm they were assigned to:

> The [video/audio/text] about Elizabeth Warren is false.

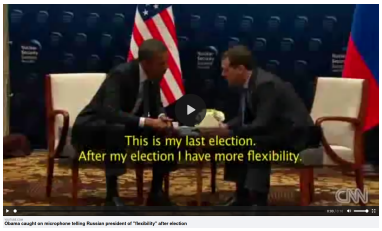
Second, to minimize the risk of influencing the proximate election, we opted to make a deepfake of high-profile 2020 Democratic Presidential candidate who was not ultimately selected as the nominee. Elizabeth Warren is a salient politician, making our experiment more ecologically valid than one with a low-profile or hypothetical politician, but she is slated for re-election until 2024. We selected a female candidate because women are more likely

(a) ***Donald Trump ("soup" press conference gaffe).*** Following national demonstrations in the summer of 2020, President Donald Trump decries protestors weaponizing cans of soup against police officers in a soon-to-be viral press conference clip (Blum, 2020).

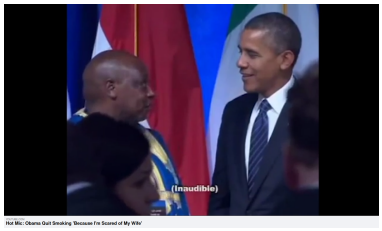
(b) ***Joe Biden (town hall 'push-up contest' gaffe).*** After a heated exchange, Democratic presidential candidate Joe Biden challenges a combative voter at a town hall to a push-up contest.

(c) ***Joe Biden (stutter gaffe).*** A video compilation of Joe Biden stuttering in various campaign appearances.

(d) ***Donald Trump (COVID-19 precautions announcement).*** In a public address from the White House, President Trump urges Americans to take personal precautions to avoid COVID-19.



(e) ***Barack Obama (Russian president hot mic).*** President Barack Obama is caught on a hot mic telling Russian President Dmitry Medvedev of "more flexibility" following his "last election" to negotiate on the issue of missile defense; an exchange that critics suggested revealed a lack of concern about re-election and lack of diplomatic transparency criticized (Goodman, 2012).



(f) ***Barack Obama (smoking hot mic).*** President Barack Obama is caught on a hot mic to a U.N. National Assembly attendee saying that he quit smoking because "I'm scared of my wife".

(g) ***Elizabeth Warren (Instagram beer gaffe).*** Democratic primary candidate Elizabeth Warren furnishes a beer on an livestream video broadcasted on Instagram, a moment criticized as inauthentic and pandering by news media (Zimmer, 2019).

(h) ***Elizabeth Warren (post-debate hot mic).*** Democratic primary candidate Elizabeth Warren confronts fellow candidate Bernie Sanders on live television for "calling me a liar on national TV".

(i) ***Donald Trump (Apple press conference gaffe).*** During an on-camera White House event, President Donald Trump mistakenly calls Apple CEO Tim Cook "Tim Apple" in a clip to go viral soon after (Rupar, 2019).

Figure D7: **Authentic Videos in Detection Task Experiment**

(a) ***Donald Trump (fake AIDS cure announcement).*** In a campaign rally speech, President Donald Trump announces that under his administration, scientists have found a cure to AIDS.

(b) ***Barack Obama (fake news announcement).*** In a White House address, President Barack Obama stresses the importance of relying on trusted news sources.

(c) ***Bernie Sanders (fake debate).*** In a televised presidential town hall event, Democratic primary candidate Bernie Sanders recalls marching for civil rights in Selma, Alabama.

(d) ***Boris Johnson (fake Brexit announcement).*** Sitting Prime Minister Boris Johnson announces that in order to "rise above the divide" on Brexit, he will endorse opposition party leader Jeremy Corbyn in the upcoming U.K. general election.

(e) ***Donald Trump (fake resignation announcement).*** In a White House address, President Donald Trump notes the American public's disappointment in his leadership and announces his resignation before the 2020 election, citing a need to "put the interests of America first".

(f) ***Hillary Clinton (fake debate).*** In a televised debate, 2016 Democratic presidential candidate Hillary Clinton labels opponent Donald Trump's tax plan as only benefiting the 1%.

Figure D8: **Deepfake Videos in Detection Task Experiment**

to be the targets of non-political deepfakes, and we specifically test for whether pre-existing prejudice against women among subjects changes the effect of the deepfake. Two of the treatments do refer to Presidential nominees Trump and Biden, but since they are otherwise identical, any effects they produce would be offset.

Third, we carefully weigh the risks to subjects against the potential risks that may be averted with the knowledge gained through our experiment. The potential long-term consequences of exposure to a single piece of media are minimal. That is, participants are unlikely to change their political behavior as a response to treatment, given our extensive debrief. Given that we have no experimental evidence either way, it is at least as likely that our experiment will *benefit* subjects as cause harm. The experiment gives subjects experience detecting fake media, followed up by the debrief which contains feedback and information about how the deepfake process works. Given the importance and seeming inevitability of more deepfakes in the future, and the uncertainty around their effects, we argue that academics in fact have an "obligation to experiment" (Ko, Mou and Matias, 2016). We believe that improved understanding of how deepfakes function and evidence from our low-cost interventions will in fact serve to prevent real-world harms from deepfakes in the future.

Fourth, a similar argument applies to the knowledge we generate from the perspective of policy-makers, journalists, and election administrators (Agarwal et al., 2019). More specifically, our study can inform future legislation or platform policies designed to minimize the threat posed by this technology.[13]

Finally, we were very concerned our experiment inadvertently contribution to the supply of online misinformation and tried to minimize this risk to the greatest possible extent. We formatted the source in the survey so as to make it impossible to download videos (the videos were not clickable, for example). Moreover, we have searched extensively for each of the deepfakes that we created (with both text and image searches). To the best of our knowledge, these preventative measures appear to have worked. We can find no evidence that we have contributed to the supply of misinformation with our study.

F Sample Description

Our survey experiment was fielded to a nationally representative sample on the Lucid survey research platform to a total of 17,501 subjects launched in two waves between September 29th 2020 and October 29th 2020. Of this 17,501, only 5,724 subjects successfully completed the survey experiment or passed a series of quality checks. One of these quality checks was a battery of randomly dispersed attention checks in response to a recently-publicized issue with in-attention among survey respondents during this period as documented in Aronow

[13]See `SB 6513` introduced in the Washington state legislature at the time of writing, intended to restrict the use of deepfake audio or visual media in campaigns for elective office.

et al. (2020). Additionally, we imposed a series of "technology checks," namely that the subjects be able to watch and listen to a video. In addition, 629 respondents failed front-end pre-treatment attention checks: namely, they entered gender or age values that did not match up (or come to close to matching up) with respondent demographic characteristics provided by Lucid. We coded these respondents as "low-quality" respondents which we drop in our analyses as a robustness measure. As expected by Aronow et al. (2020), results largely hold across the two cohorts, but nearly all coefficient estimates are slightly diminished for the low-quality cohort.

Table F6 compares our sample's demographic traits to the demographic traits in the most recent Current Population Survey (CPS) – in particular, traits like education, age, and household income that are hypothesized to have correlations with deepfake deception and affective appeal (by their correlation with digital literacy, internet usage, and political knowledge) as well race, gender and ethnicity which are correlates of partisanship, another predictor of our measured behavioral responses. To adjust for remaining disrepancies, we generate post-stratification weights via raking to match the CPS marginal population totals. We perform weighted regression in our analyses as a robustness measure to guard against measurement error from possible demographic skews.

Table F6: **Sample Demographics and Representativeness after Post-stratification**

		CPS	Unweighted Sample	Weighted Sample
Education	<High school	10.95%	0.94%	2.62%
	High school	47.14%	29.25%	45.26%
	College	30.3%	47.22%	36.37%
	Postgraduate	11.61%	21.93%	15.75%
Age	18-24	10.42%	5.49%	8.32%
	25-34	13.88%	12.81%	15%
	35-44	12.58%	17.24%	17.04%
	45-64	25.76%	31.43%	34.31%
	65+	15.81%	33%	25.33%
Household Income	<$25k	19.11%	29.98%	22.4%
	$100k-$150k	14.95%	6.13%	10.71%
	>$150k	15.47%	4.89%	10.39%
	$25k-$49k	20.79%	21.77%	23.1%
	$50k-$74k	17.2%	15.67%	18.85%
	$75k-$99k	12.48%	19.37%	14.54%
Gender	Female	51.25%	65.83%	55.91%
	Male	48.75%	33.67%	44.09%
Race	Asian	5.42%	3.93%	4.57%
	Black	10.28%	5.63%	8.09%
	Other	4.18%	3.67%	3.74%
	White	80.12%	86.06%	83.6%
Hispanic	Yes	14.66%	5.08%	8.47%
	No	85.34%	94.16%	91.53%

Notes: Weights are constructed via Iterative Proportional Fitting to match sample marginal totals to CPS marginal totals on displayed demographic traits. Weights in the final column used for all analyses in paper.

G Pre-Registration

Our pre-analysis plan containing all pre-registered hypotheses can be found at `https://osf.io/yh53p` (Barari, Lucas and Munger, 2020). For all models, unless otherwise noted or displayed, controls include age group, education, 3 point party ID, cognitive reflection, political knowledge, internet usage, and an indicator for mobile (vs. desktop) exposure. The reference stimuli for all analyses of the incidental exposure experiment is `video`. Reference category for environment in the detection task experiment is high-fake. Cognitive reflection, political knowledge, ambivalent sexism, and internet usage are all re-scaled to [0,1]. Unless otherwise noted, analyses exclude respondents who receive information prior to the incidental exposure experiment, however results (effect magnitudes, statistical significance) are substantively similar in all cases with their inclusion. As pre-registered, all p-values are "step-up" adjusted to $p \cdot r/K$ where r denotes the rank of the unadjusted p-value amongst K total estimated p-values (Benjamini-Hochberg procedure). Analyses do not additionally adjust for respondent wave, for brevity, though we find that including respondent wave as either an interaction term or a linear term does not change our results.

G.1 Divergences from pre-registration

Though our pre-analysis plan was specified in great detail and faithfully executed following our experiment, the analyses presented in this paper differ for a number of reasons – none of which relate to whether the results were favorable to our priors. For the purposes of research transparency, we believe it is nevertheless crucial to share our pre-analysis plan (and the outcomes for each specification), however here we briefly explain the conceptual, logistical, or methodological reasons for divergences or omissions from the PAP.

First, for presentational reasons, we organized our paper around three broad research questions rather than individual outcomes (e.g., "belief", "affect") or treatments (e.g., "information provision") as is structured in the pre-analysis plan. For similar reasons, we organized results for $\mathbf{H}_{3b}$ into Table G25, rather than its own table.

Second, we pre-registered a comparison to the `skit` condition in all hypotheses pertaining to the incidental exposure experiment. However, due to methodological concerns about differential item functioning (i.e. differing meanings of credibility for a skit vs. purportedly real video) pointed out to us by a reviewer, we omit such invalid comparisons to the skit in the forthcoming tables.

Third, for space and scope concerns, we choose to omit analyses pertaining to the trust outcome ($\mathbf{H}_{3a}$).

Fourth, as a conceptual correction, we refer to the outcome called "belief" or "deception" in the pre-analysis plan ($\mathbf{H}_{1,4,5,6,7}$) instead as "credbility" in the main text.

Fifth, we omit any conclusions from tests that perform a three-way interaction between partisanship, stimuli condition, and cognitive reflection ($\mathbf{H}_{6a}$,$\mathbf{H}_{6b}$) due to a lack of statistical power in our observed sample (see Appendix Section H).

G.2 Pre-registered analyses

Table G7: **Models of Credibility Confidence in Incidental Exposure Experiment**

	Confidence that clipping was credible [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.08	0.14	0.11	0.10	0.13	0.14	0.17*
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)
Text	−0.02	−0.11	−0.04	−0.02	−0.14	−0.03	−0.13
	(0.09)	(0.09)	(0.10)	(0.09)	(0.08)	(0.10)	(0.10)
On Mobile				0.06	0.15	0.22**	0.24*
				(0.09)	(0.09)	(0.10)	(0.10)
Age 65+				0.12	0.17*	0.09	0.13
				(0.08)	(0.08)	(0.09)	(0.09)
High School				−0.45	−0.53*	−0.56	−0.63**
				(0.41)	(0.24)	(0.51)	(0.30)
College				−0.42	−0.48	−0.50	−0.54
				(0.41)	(0.24)	(0.50)	(0.30)
Postgrad				−0.57	−0.54	−0.61	−0.51
				(0.41)	(0.25)	(0.51)	(0.31)
Independent PID					0.19*	0.16	0.30
					(0.11)	(0.13)	(0.13)
Republican PID					0.55***	0.61***	0.70***
					(0.08)	(0.09)	(0.09)
C.R.				−0.04	0.01	−0.03	0.12
				(0.15)	(0.15)	(0.16)	(0.16)
Male				−0.07	−0.02	0.02	0.09
				(0.08)	(0.07)	(0.08)	(0.08)
Political Knowledge				0.30	0.24	0.16	0.17
				(0.18)	(0.17)	(0.21)	(0.20)
Internet Usage				0.64	0.93***	0.67	0.74*
				(0.33)	(0.33)	(0.39)	(0.38)
Ambivalent Sexism				0.14***	0.03	0.02	−0.03
				(0.04)	(0.04)	(0.05)	(0.05)
Constant	3.40***	3.43***	3.46***	2.61***	2.50***	2.82***	2.85***
	(0.06)	(0.06)	(0.07)	(0.55)	(0.43)	(0.65)	(0.50)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	1,345	1,345	968	1,345	1,345	968	968
R^2	0.001	0.01	0.003	0.02	0.06	0.07	0.09
Adjusted R^2	−0.0002	0.005	0.0005	0.01	0.05	0.05	0.07

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G8: **Models of Binarized Credibility Confidence of Scandal Clipping in Incidental Exposure Experiment**

	Somewhat/strongly confident clipping was credible						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.01	0.06	0.02	0.02	0.05	0.04	0.08*
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Text	−0.04	−0.04	−0.04	−0.04	−0.06*	−0.04	−0.03
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
On Mobile				0.08**	0.11***	0.14***	0.17***
				(0.03)	(0.03)	(0.04)	(0.04)
Age 65+				0.06*	0.09**	0.07*	0.08*
				(0.03)	(0.03)	(0.03)	(0.04)
High School				−0.24	−0.26***	−0.24	−0.28**
				(0.16)	(0.09)	(0.20)	(0.12)
College				−0.21	−0.22	−0.20	−0.22
				(0.16)	(0.09)	(0.20)	(0.12)
Postgrad				−0.22	−0.20*	−0.22	−0.17
				(0.16)	(0.10)	(0.20)	(0.12)
Independent PID				0.02	0.03	0.02	0.05
				(0.04)	(0.04)	(0.05)	(0.05)
Republican PID				0.20***	0.19***	0.22***	0.25***
				(0.03)	(0.03)	(0.04)	(0.04)
C.R.				−0.04	−0.04	−0.04	−0.001
				(0.06)	(0.06)	(0.07)	(0.07)
Male				0.01	0.02	0.03	0.04
				(0.03)	(0.03)	(0.03)	(0.03)
Political Knowledge				0.18**	0.23***	0.15	0.23***
				(0.07)	(0.07)	(0.08)	(0.08)
Internet Usage				0.29**	0.41***	0.33**	0.35**
				(0.13)	(0.13)	(0.15)	(0.15)
Ambivalent Sexism				0.02	0.03	0.02	0.005
				(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.47***	0.47***	0.48***	0.08	−0.09	0.07	−0.01
	(0.02)	(0.02)	(0.03)	(0.21)	(0.17)	(0.26)	(0.20)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	1,345	1,345	968	1,345	1,345	968	968
R^2	0.002	0.01	0.003	0.06	0.08	0.07	0.10
Adjusted R^2	0.001	0.01	0.001	0.05	0.07	0.06	0.09

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G9: **Models of Scandal Target Affect in Incidental Exposure Experiment**

	Elizabeth Warren Affect Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Video	−4.77***	−4.51**	−1.78	−3.75***	−4.53***	−2.88	−5.03***
	(1.64)	(1.63)	(1.97)	(1.29)	(1.31)	(1.53)	(1.55)
Audio	−2.05	−3.45*	−0.45	−3.57***	−4.94***	−3.03*	−5.10***
	(1.60)	(1.60)	(1.91)	(1.27)	(1.29)	(1.49)	(1.53)
Text	−1.85	−1.17	−0.35	−2.76**	−1.46	−1.83	−0.68
	(1.61)	(1.61)	(1.92)	(1.27)	(1.30)	(1.50)	(1.54)
Skit	−2.96	−3.38**	−1.13	−3.57***	−4.05***	−2.95*	−4.52***
	(1.60)	(1.59)	(1.91)	(1.27)	(1.29)	(1.49)	(1.51)
Attack Ad	−4.49***	−4.84***	−3.35*	−4.20***	−4.04**	−4.18**	−5.28***
	(1.61)	(1.58)	(1.92)	(1.27)	(1.27)	(1.50)	(1.51)
Information				0.65	0.58	0.28	0.46
				(0.73)	(0.75)	(0.86)	(0.87)
On Mobile				−2.46**	−1.84	−3.54***	−3.21**
				(0.91)	(0.93)	(1.08)	(1.11)
Age 65+				−4.50***	−5.21***	−5.49***	−5.79***
				(0.81)	(0.86)	(0.94)	(0.99)
High School				−0.56	−1.56	−1.98	−2.68
				(3.86)	(2.40)	(4.52)	(2.80)
College				1.69	1.68	0.27	−0.16
				(3.85)	(2.44)	(4.52)	(2.86)
Postgrad				11.68**	14.29***	9.68	12.18***
				(3.90)	(2.56)	(4.58)	(3.01)
Independent PID				−27.29***	−27.06***	−26.08***	−26.41***
				(1.22)	(1.23)	(1.44)	(1.44)
Republican PID				−40.03***	−37.95***	−40.76***	−39.27***
				(0.84)	(0.84)	(0.98)	(0.99)
C.R.				−1.27	−1.84	−0.92	−1.21
				(1.62)	(1.66)	(1.84)	(1.88)
Male				0.98	−0.09	1.16	0.47
				(0.81)	(0.79)	(0.95)	(0.93)
Political Knowledge				−1.86	−1.66	0.31	−0.18
				(1.90)	(1.85)	(2.27)	(2.24)
Internet Usage				5.35	4.41	3.20	1.68
				(3.43)	(3.48)	(4.15)	(4.20)
Ambivalent Sexism				−3.87***	−3.28***	−4.09***	−3.99***
				(0.47)	(0.48)	(0.56)	(0.57)
Constant	45.95***	45.35***	43.91***	72.67***	72.10***	75.41***	77.97***
	(1.15)	(1.13)	(1.39)	(5.40)	(4.49)	(6.39)	(5.31)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	5,472	5,472	3,872	5,471	5,471	3,871	3,871
R^2	0.002	0.003	0.001	0.38	0.36	0.40	0.38
Adjusted R^2	0.001	0.002	−0.0002	0.38	0.35	0.39	0.38

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$
Reference category for clip type is `control`.

Table G10: **Models of Information Provision and Media Trust in Incidental Exposure Experiment**

	Dependent variable:						
	Trust in Media (Combined Index)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Information	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	2.27***	2.30***	2.27***	2.30***	2.31***	2.30***	2.31***
	(0.02)	(0.02)	(0.02)	(0.17)	(0.14)	(0.17)	(0.14)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
Observations	2,542	2,542	2,542	2,542	2,542	2,542	2,542
R^2	0.001	0.001	0.001	0.19	0.24	0.19	0.24
Adjusted R^2	0.0005	0.001	0.0005	0.19	0.24	0.19	0.24
Note:	*Notes:* · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$ Respondents in `skit` and `ad` conditions are excluded.						

Table G11: **Models of Information Provision and Media Trust Across Sources in Incidental Exposure Experiment**

	Trust in...							
	Offline Media		Online Media		Social Media		Combined Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Information	−0.01	−0.02	−0.05	−0.05	−0.05	−0.05	−0.04	−0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
Constant	2.64***	2.96	2.31***	1.99	1.87***	1.96	2.27***	2.30***
	(0.02)	(0.22)	(0.02)	(0.21)	(0.02)	(0.22)	(0.02)	(0.17)
Controls?		✓		✓		✓		✓
Observations	2,542	2,542	2,542	2,542	2,542	2,542	2,542	2,542
R^2	0.0000	0.16	0.001	0.14	0.001	0.18	0.001	0.19
Adjusted R^2	−0.0003	0.16	0.001	0.13	0.001	0.17	0.0005	0.19
Note:	· $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$ Respondents in `skit` and `ad` conditions are excluded.							

Table G12: **Models of Deepfake Exposure, Credibility, and Media Trust Across Sources in Incidental Exposure Experiment**

	Trust in Media (Combined Index)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Credible	−0.11***	−0.07***	−0.12***	−0.08***	−0.06*	−0.05***	−0.14***
	(0.02)	(0.02)	(0.04)	(0.02)	(0.02)	(0.02)	(0.04)
Video	−0.32***	−0.28**	−0.09*	−0.44***	−0.22	−0.32**	−0.19***
	(0.11)	(0.10)	(0.05)	(0.10)	(0.12)	(0.12)	(0.05)
Credible x Video	0.08**	0.07**	0.11	0.10***	0.05	0.07	0.19*
	(0.03)	(0.03)	(0.07)	(0.03)	(0.03)	(0.03)	(0.07)
Constant	2.66***	2.53***	2.36***	2.51***	2.39***	2.53***	2.34***
	(0.06)	(0.25)	(0.25)	(0.20)	(0.31)	(0.24)	(0.20)
Weighted?				✓		✓	✓
Low-Quality Dropped?					✓	✓	✓
Controls?		✓	✓	✓	✓	✓	✓
Credibility Binarized?			✓				✓
N	1,343	1,343	1,343	1,343	966	966	1,343
R^2	0.03	0.24	0.23	0.29	0.23	0.29	0.28
Adjusted R^2	0.03	0.23	0.22	0.28	0.22	0.28	0.27

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$
Respondents in `skit` and `ad` conditions are excluded.

Table G13: **Models of Deepfake Exposure, Credibility, and Media Trust Across Sources in Incidental Exposure Experiment**

	Trust in...							
	Offline Media		Online Media		Social Media		Combined Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Credibility	−0.08***	−0.15***	−0.06***	−0.10*	−0.08***	−0.11*	−0.07***	−0.12
	(0.02)	(0.05)	(0.02)	(0.05)	(0.02)	(0.05)	(0.02)	(0.04)
Video	−0.22	−0.05	−0.26	−0.10	−0.37	−0.13	−0.28***	−0.09
	(0.13)	(0.06)	(0.12)	(0.06)	(0.13)	(0.06)	(0.10)	(0.05)
Credibility x Video	0.06	0.07	0.07	0.15	0.08**	0.10	0.07**	0.11
	(0.03)	(0.09)	(0.03)	(0.09)	(0.03)	(0.09)	(0.03)	(0.07)
Constant	2.96***	2.79***	2.21***	2.08***	2.41***	2.22***	2.53***	2.36
	(0.32)	(0.32)	(0.31)	(0.30)	(0.32)	(0.32)	(0.25)	(0.25)
Controls?	✓	✓	✓	✓	✓	✓	✓	✓
Credibility Binarized?		✓		✓		✓		✓
Observations	1,343	1,343	1,343	1,343	1,343	1,343	1,343	1,343
R^2	0.21	0.20	0.16	0.16	0.21	0.20	0.24	0.23
Adjusted R^2	0.20	0.19	0.15	0.15	0.20	0.20	0.23	0.22

Note: *Notes:* · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$
Respondents in `skit` and `ad` conditions are excluded.

Table G14: **Models of Information Provision and Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Confidence that clipping was credible [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Information	−0.35***	−0.30***	−0.40***	−0.38***	−0.36***	−0.40***	−0.36***
	(0.09)	(0.09)	(0.11)	(0.09)	(0.09)	(0.10)	(0.10)
Audio	0.08	0.14	0.11	0.11	0.15*	0.15	0.21*
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)
Text	−0.02	−0.11	−0.04	−0.01	−0.13	−0.02	−0.12
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)
Info x Audio	0.09	−0.06	0.04	0.10	−0.004	0.02	−0.12
	(0.12)	(0.12)	(0.15)	(0.12)	(0.12)	(0.14)	(0.14)
Info x Text	0.20	0.24	0.16	0.25**	0.32**	0.18	0.23
	(0.12)	(0.13)	(0.15)	(0.12)	(0.12)	(0.14)	(0.14)
Constant	3.40***	3.43***	3.46***	2.90***	3.14***	2.99***	3.32***
	(0.06)	(0.06)	(0.07)	(0.37)	(0.30)	(0.46)	(0.37)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.01	0.01	0.02	0.06	0.06	0.08	0.08
Adjusted R^2	0.01	0.01	0.02	0.05	0.05	0.07	0.07

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G15: **Models of Information Provision and Binarized Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Somewhat/strongly confident clipping was credible						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Information	−0.11**	−0.09**	−0.12**	−0.12***	−0.11***	−0.12***	−0.10**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Audio	0.01	0.06*	0.02	0.02	0.06	0.04	0.09
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Text	−0.04	−0.04	−0.04	−0.04	−0.05	−0.04	−0.03
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Info x Audio	0.02	−0.04	−0.01	0.03	−0.02	−0.02	−0.07
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)
Info x Text	0.09	0.08	0.07	0.11**	0.12*	0.09	0.07
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
Constant	0.47***	0.47***	0.48***	0.14	0.16	0.14	0.24
	(0.02)	(0.02)	(0.03)	(0.14)	(0.11)	(0.18)	(0.14)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.01	0.01	0.01	0.05	0.05	0.06	0.06
Adjusted R^2	0.005	0.01	0.01	0.04	0.04	0.05	0.05

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G16: **Models of Cognitive Reflection and Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Confidence that clipping was credible [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.09	0.08	0.13	0.12	0.13	0.19	0.21
	(0.11)	(0.11)	(0.12)	(0.10)	(0.10)	(0.12)	(0.12)
Text	0.11	−0.03	0.04	0.11	−0.03	0.02	−0.16
	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.11)	(0.11)
C.R.	−0.09	−0.10	−0.12	−0.10	−0.08	−0.11	−0.06
	(0.20)	(0.19)	(0.23)	(0.19)	(0.19)	(0.22)	(0.21)
C.R. x Audio	0.10	0.07	0.01	0.13	0.01	−0.08	−0.22
	(0.27)	(0.27)	(0.31)	(0.27)	(0.27)	(0.30)	(0.30)
C.R. x Text	−0.11	0.11	0.001	0.01	0.16	0.14	0.48
	(0.27)	(0.27)	(0.31)	(0.26)	(0.26)	(0.30)	(0.30)
Constant	3.25***	3.32***	3.29***	2.86***	3.16***	3.04***	3.41***
	(0.08)	(0.07)	(0.09)	(0.37)	(0.29)	(0.45)	(0.37)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.002	0.001	0.002	0.06	0.05	0.08	0.07
Adjusted R^2	0.0001	−0.001	−0.001	0.05	0.05	0.07	0.06

Notes: $^{\cdot}$ $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G17: **Models of Cognitive Reflection and Binarized Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Somewhat/strongly confident clipping was credible						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	−0.01	0.003	−0.01	−0.002	0.02	0.01	0.05
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
Text	−0.003	−0.02	−0.01	−0.004	−0.03	−0.02	−0.06
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
C.R.	−0.10	−0.12	−0.11	−0.11	−0.13	−0.11	−0.12
	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)	(0.08)
C.R. x Audio	0.11	0.11	0.09	0.11	0.09	0.06	0.02
	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.12)	(0.12)
C.R. x Text	0.01	0.08	0.03	0.05	0.11	0.07	0.22*
	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.12)	(0.12)
Constant	0.45***	0.46***	0.45***	0.15	0.18	0.17	0.29*
	(0.03)	(0.03)	(0.03)	(0.14)	(0.11)	(0.17)	(0.14)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.002	0.002	0.002	0.04	0.04	0.06	0.06
Adjusted R^2	−0.0001	0.001	−0.001	0.04	0.04	0.05	0.05

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G18: **Models of Partisan Group Identity and Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Confidence that clipping was credible [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	0.50	0.29*	0.54***	0.48***	0.32*	0.51**	0.29
	(0.15)	(0.15)	(0.17)	(0.15)	(0.15)	(0.17)	(0.17)
Audio	0.14	0.09	0.17	0.15	0.15	0.20	0.22
	(0.14)	(0.14)	(0.15)	(0.14)	(0.14)	(0.15)	(0.15)
Text	0.09	−0.16	−0.0001	0.10	−0.12	0.01	−0.31
	(0.13)	(0.14)	(0.15)	(0.13)	(0.13)	(0.15)	(0.15)
C.R.	0.005	−0.18	−0.02	0.01	−0.09	−0.001	−0.09
	(0.25)	(0.25)	(0.29)	(0.25)	(0.25)	(0.29)	(0.28)
Repub x Audio	−0.09	0.001	−0.04	−0.09	−0.05	−0.05	−0.01
	(0.21)	(0.21)	(0.24)	(0.21)	(0.21)	(0.24)	(0.24)
Repub x Text	0.05	0.31	0.07	−0.003	0.21	0.02	0.37
	(0.21)	(0.21)	(0.24)	(0.21)	(0.21)	(0.23)	(0.23)
C.R. x Repub	−0.23	0.15	−0.23	−0.24	0.06	−0.28	0.09
	(0.39)	(0.38)	(0.46)	(0.39)	(0.38)	(0.46)	(0.43)
Audio x C.R.	−0.06	−0.07	−0.18	−0.04	−0.14	−0.22	−0.41
	(0.35)	(0.36)	(0.40)	(0.35)	(0.36)	(0.40)	(0.40)
Text x C.R.	−0.17	0.19	0.004	−0.14	0.12	0.05	0.54
	(0.34)	(0.35)	(0.39)	(0.34)	(0.35)	(0.39)	(0.40)
Repub x Audio x C.R.	0.45	0.34	0.33	0.44	0.40	0.36	0.48
	(0.54)	(0.55)	(0.62)	(0.54)	(0.54)	(0.62)	(0.61)
Repub x Text x C.R.	0.37	−0.07	0.20	0.45	0.14	0.28	−0.10
	(0.54)	(0.54)	(0.62)	(0.54)	(0.53)	(0.61)	(0.60)
Constant	3.03***	3.19***	3.06***	2.93***	3.29***	3.09***	3.60***
	(0.10)	(0.10)	(0.11)	(0.37)	(0.30)	(0.46)	(0.37)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.04	0.04	0.04	0.06	0.05	0.07	0.07
Adjusted R^2	0.03	0.03	0.04	0.05	0.05	0.06	0.06

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$
PID is pooled to Republican/Not Republican for brevity. PID interacted with C.R. to test possible mechanism of motivated reasoning (pre-registered), although, as a reviewer pointed out, this is not a sufficient test of a motivated reasoning mechanism by itself.

Table G19: **Models of Partisan Group Identity and Binarized Credibility Confidence of Clipping in Incidental Exposure Experiment**

	Somewhat/strongly confident clipping was credible						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	0.14	0.10	0.14**	0.14**	0.11*	0.15*	0.09
	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)
Audio	−0.01	0.03	−0.02	−0.01	0.05	−0.01	0.07
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
Text	−0.04	−0.08	−0.06	−0.03	−0.07	−0.05	−0.13
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
C.R.	−0.05	−0.11	−0.07	−0.06	−0.09	−0.07	−0.11
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.11)
Repub x Audio	0.02	−0.05	0.05	0.02	−0.06	0.03	−0.05
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)
Repub x Text	0.07	0.12	0.09	0.06	0.10	0.07	0.16
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)
C.R. x Repub	−0.11	−0.03	−0.08	−0.12	−0.07	−0.11	−0.003
	(0.15)	(0.15)	(0.18)	(0.15)	(0.15)	(0.18)	(0.17)
Audio x C.R.	0.06	−0.01	0.03	0.06	−0.04	0.01	−0.15
	(0.13)	(0.14)	(0.15)	(0.13)	(0.14)	(0.15)	(0.16)
Text x C.R.	−0.02	0.07	0.04	−0.01	0.06	0.04	0.23
	(0.13)	(0.13)	(0.15)	(0.13)	(0.13)	(0.15)	(0.15)
Repub x Audio x C.R.	0.13	0.29	0.10	0.14	0.31	0.13	0.39
	(0.21)	(0.21)	(0.24)	(0.20)	(0.21)	(0.24)	(0.24)
Repub x Text x C.R.	0.16	0.07	0.06	0.19	0.11	0.11	−0.01
	(0.21)	(0.20)	(0.24)	(0.21)	(0.20)	(0.24)	(0.23)
Constant	0.39***	0.42***	0.39***	0.17	0.23	0.19	0.36**
	(0.04)	(0.04)	(0.04)	(0.14)	(0.12)	(0.18)	(0.15)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.03	0.03	0.04	0.05	0.05	0.06	0.06
Adjusted R^2	0.03	0.03	0.03	0.04	0.04	0.05	0.05

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G20: **Models of Partisan Group Identity and Scandal Target Affect in Incidental Exposure Experiment**

	Elizabeth Warren Feeling Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	−34.13***	−30.48***	−38.34***	−32.08***	−28.72***	−36.56***	−35.22***
	(3.28)	(3.33)	(3.78)	(3.19)	(3.24)	(3.68)	(3.81)
Video	−7.34*	−9.58***	−7.75	−8.15***	−8.85**	−8.98**	−11.94***
	(3.11)	(3.22)	(3.60)	(3.02)	(3.12)	(3.50)	(3.67)
Audio	−8.50**	−8.26**	−8.52*	−9.19***	−7.42**	−9.56***	−9.89**
	(3.09)	(3.19)	(3.54)	(3.00)	(3.10)	(3.45)	(3.65)
Text	−1.38	0.73	−2.58	−2.25	1.51	−3.76	−0.35
	(2.98)	(3.17)	(3.50)	(2.90)	(3.08)	(3.41)	(3.69)
Skit	−6.69*	−7.94**	−6.55	−7.39*	−7.43**	−8.32*	−9.95**
	(3.02)	(3.13)	(3.50)	(2.93)	(3.03)	(3.40)	(3.62)
Attack Ad	−4.65	−4.47	−5.74	−5.49*	−4.61	−6.97*	−8.59
	(3.11)	(3.20)	(3.59)	(3.02)	(3.11)	(3.49)	(3.68)
C.R.	1.81	−0.67	−1.21	−3.65	−3.17	−6.86	−6.67
	(5.78)	(5.92)	(6.72)	(5.63)	(5.76)	(6.56)	(6.87)
Repub x Video	2.46	4.05	9.12	3.19	3.32	9.74*	12.05**
	(4.70)	(4.77)	(5.37)	(4.57)	(4.62)	(5.23)	(5.30)
Repub x Audio	4.26	4.42	4.98	3.72	3.01	4.30	5.40
	(4.67)	(4.79)	(5.32)	(4.54)	(4.64)	(5.17)	(5.34)
Repub x Text	−3.16	−7.25	1.63	−2.09	−6.15	2.96	−0.14
	(4.58)	(4.70)	(5.19)	(4.45)	(4.56)	(5.04)	(5.24)
Repub x Skit	−0.83	0.19	0.17	−0.45	−0.74	1.68	2.60
	(4.55)	(4.66)	(5.15)	(4.42)	(4.52)	(5.01)	(5.18)
Repub x Ad	0.52	−3.72	1.85	2.17	−1.41	4.51	3.79
	(4.63)	(4.62)	(5.30)	(4.51)	(4.48)	(5.15)	(5.18)
C.R. x Repub	−15.24	−12.89	−9.24	−14.02	−12.72	−8.73	−7.82
	(8.81)	(8.79)	(10.22)	(8.56)	(8.53)	(9.93)	(10.04)
Video x C.R.	6.67	14.75	10.29	8.53	12.05	11.75	13.77
	(8.12)	(8.33)	(9.37)	(7.90)	(8.09)	(9.11)	(9.32)
Audio x C.R.	12.00	5.74	14.42	12.26	2.43	14.81	8.48
	(7.91)	(8.30)	(9.11)	(7.70)	(8.06)	(8.87)	(9.40)
Text x C.R.	−6.67	−5.95	−0.73	−5.64	−7.70	−0.16	−3.39
	(7.63)	(8.11)	(8.83)	(7.43)	(7.89)	(8.59)	(9.24)
Skit x C.R.	6.12	6.53	9.83	7.62	5.31	12.22	9.65
	(7.93)	(8.10)	(9.16)	(7.71)	(7.87)	(8.90)	(9.21)
Ad x C.R.	1.90	4.81	3.87	3.79	7.12	5.66	8.73
	(7.91)	(8.11)	(9.02)	(7.70)	(7.87)	(8.78)	(9.11)
Repub x Video x C.R.	2.61	−7.37	−11.11	3.03	−3.45	−8.29	−10.98
	(12.43)	(12.40)	(14.38)	(12.09)	(12.03)	(13.98)	(13.80)
Repub x Audio x C.R.	0.51	1.03	1.24	2.69	4.23	4.24	2.18
	(12.18)	(12.55)	(13.86)	(11.84)	(12.17)	(13.47)	(13.94)
Repub x Text x C.R.	14.80	15.44	3.87	13.49	12.89	3.51	5.56
	(12.18)	(12.29)	(13.79)	(11.84)	(11.95)	(13.40)	(13.77)
Repub x Skit x C.R.	12.16	11.10	9.70	12.26	15.20	9.85	14.61
	(12.18)	(12.31)	(13.84)	(11.85)	(11.94)	(13.46)	(13.65)
Repub x Ad x C.R.	−4.21	−3.90	−2.75	−6.00	−9.37	−4.47	−6.39
	(12.33)	(12.27)	(14.03)	(11.98)	(11.90)	(13.63)	(13.71)
Constant	61.66***	60.53***	62.56***	65.27***	64.38***	68.28***	71.53***
	(2.21)	(2.32)	(2.61)	(5.93)	(5.07)	(6.95)	(5.99)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	5,472	5,472	3,872	5,471	5,471	3,871	3,871
R^2	0.29	0.25	0.31	0.33	0.30	0.35	0.33
Adjusted R^2	0.28	0.25	0.30	0.32	0.30	0.34	0.32

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G21: **Models of Ambivalent Sexism and Credibility Confidence in Scandal Clipping in Incidental Exposure Experiment**

	Confidence that clipping was credible [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	0.24***	0.17***	0.30***	0.15***	0.09	0.19***	0.10
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
Audio	0.50	0.25	0.57*	0.53	0.30	0.56*	0.36
	(0.21)	(0.22)	(0.26)	(0.21)	(0.22)	(0.25)	(0.26)
Text	0.17	−0.04	0.29	0.26	0.05	0.37	0.11
	(0.22)	(0.22)	(0.26)	(0.21)	(0.22)	(0.25)	(0.26)
A.S. x Audio	−0.13	−0.05	−0.15	−0.13	−0.06	−0.14	−0.07
	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)	(0.09)
A.S. x Text	−0.03	0.02	−0.09	−0.05	−0.01	−0.11	−0.04
	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)	(0.09)
Constant	2.54***	2.81***	2.39***	2.69***	3.08***	2.78***	3.25***
	(0.16)	(0.16)	(0.19)	(0.38)	(0.31)	(0.47)	(0.39)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.02	0.01	0.02	0.06	0.06	0.08	0.07
Adjusted R^2	0.02	0.01	0.02	0.05	0.05	0.07	0.06

Notes: $^{\cdot}$ $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G22: **Models of Ambivalent Sexism and Binarized Credibility Confidence in Scandal Clipping in Incidental Exposure Experiment**

	Somewhat/strongly confident clipping was credible						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	0.06***	0.05	0.08***	0.03	0.03	0.04	0.04
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Audio	0.09	0.10	0.14	0.09	0.11	0.13	0.16
	(0.08)	(0.08)	(0.10)	(0.08)	(0.08)	(0.10)	(0.10)
Text	−0.02	−0.03	0.04	0.002	−0.02	0.06	0.03
	(0.08)	(0.09)	(0.10)	(0.08)	(0.08)	(0.10)	(0.10)
A.S. x Audio	−0.02	−0.02	−0.04	−0.02	−0.02	−0.04	−0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
A.S. x Text	0.01	0.01	−0.01	0.004	0.01	−0.02	−0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Constant	0.25***	0.28***	0.19***	0.12	0.15	0.10	0.22
	(0.06)	(0.06)	(0.07)	(0.15)	(0.12)	(0.18)	(0.15)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	2,701	2,701	1,906	2,701	2,701	1,906	1,906
R^2	0.01	0.01	0.01	0.04	0.04	0.06	0.06
Adjusted R^2	0.01	0.01	0.01	0.04	0.04	0.05	0.05

Notes: $\cdot$ $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G23: **Models of Ambivalent Sexism and Scandal Target Affect in Incidental Exposure Experiment**

	Elizabeth Warren Feeling Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	−12.48***	−10.45***	−13.42***	−5.43***	−5.33***	−5.88***	−6.40***
	(1.31)	(1.30)	(1.64)	(1.10)	(1.10)	(1.36)	(1.36)
Video	−6.36	−8.97	−4.31	−5.31	−9.39*	−5.09	−11.49*
	(5.49)	(5.61)	(6.72)	(4.53)	(4.68)	(5.51)	(5.62)
Audio	−9.99*	−9.01	−10.70	−11.27**	−9.92*	−12.25**	−10.76
	(5.35)	(5.45)	(6.55)	(4.42)	(4.55)	(5.36)	(5.50)
Text	−2.92	1.16	−2.33	−7.76	−3.06	−7.44	−5.03
	(5.39)	(5.51)	(6.56)	(4.46)	(4.61)	(5.37)	(5.54)
Skit	−3.24	−10.05	−1.13	−5.82	−14.24**	−5.19	−15.82***
	(5.39)	(5.47)	(6.57)	(4.45)	(4.57)	(5.38)	(5.48)
A.S. x Video	0.73	1.58	0.97	0.56	1.69	0.81	2.32
	(1.86)	(1.88)	(2.28)	(1.53)	(1.57)	(1.87)	(1.90)
A.S. x Audio	2.83	2.04	3.53	2.73	1.75	3.29*	2.01
	(1.82)	(1.82)	(2.24)	(1.51)	(1.52)	(1.84)	(1.85)
A.S. x Text	0.31	−0.76	0.66	1.77	0.58	1.99	1.57
	(1.85)	(1.84)	(2.24)	(1.52)	(1.54)	(1.83)	(1.86)
A.S. x Skit	0.10	2.26	−0.04	0.79	3.58**	0.80	4.02
	(1.84)	(1.84)	(2.24)	(1.52)	(1.54)	(1.83)	(1.86)
Constant	81.03***	75.28***	81.77***	77.20***	79.41***	80.00***	85.04***
	(3.85)	(3.88)	(4.81)	(6.50)	(5.61)	(7.84)	(6.70)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	4,555	4,555	3,213	4,555	4,555	3,213	3,213
R^2	0.08	0.06	0.09	0.38	0.35	0.40	0.37
Adjusted R^2	0.08	0.06	0.09	0.38	0.35	0.39	0.37

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G24: **Predictors of Detection Task Accuracy**

	Deepfake Detection Accuracy (% Correctly Classified)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		0.25***	0.24***	0.24***	0.22***	0.24***
		(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Accuracy Prime	−0.002		−0.005	−0.003	−0.005	−0.004
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Debrief			0.01	0.01	0.005	0.003
			(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Information			−0.01	−0.002	−0.01	−0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			0.16***	0.17***	0.16***	0.16***
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			−0.02	−0.05	−0.02	−0.07*
			(0.03)	(0.03)	(0.04)	(0.04)
Low-fake Env.			0.03***	0.04***	0.04***	0.06***
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.04***	0.04***	0.05***	0.06***
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.02**	0.01	0.01	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.02	0.02	0.05	0.05
			(0.03)	(0.02)	(0.04)	(0.02)
College			0.03	0.04	0.05	0.06**
			(0.03)	(0.02)	(0.04)	(0.02)
Postgrad			0.01	0.01	0.03	0.03
			(0.03)	(0.02)	(0.04)	(0.03)
C.R.			0.06***	0.07***	0.06***	0.07***
			(0.02)	(0.02)	(0.02)	(0.02)
C.R. x Republican			−0.05	−0.05	−0.06	−0.04
			(0.03)	(0.03)	(0.03)	(0.03)
Ambivalent Sexism			0.001	−0.001	0.003	0.003
			(0.004)	(0.004)	(0.005)	(0.005)
Republican			0.08***	0.07***	0.09***	0.07***
			(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.57	0.35***	0.14	0.15***	0.14**	0.16***
	(0.005)	(0.02)	(0.05)	(0.04)	(0.06)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,445	5,445	5,444	5,444	3,846	3,846
R^2	0.0000	0.02	0.08	0.08	0.08	0.09
Adjusted R^2	−0.0002	0.02	0.08	0.08	0.07	0.08

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G25: **Predictors of Detection Task False Positive Rate (FPR)**

	Detection FPR (% Real Videos Classified as Deepfakes)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		−0.03	−0.01	−0.03	−0.01	−0.03
		(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Accuracy Prime	−0.001		0.01	−0.0004	0.01	0.004
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Debrief			−0.03***	−0.03***	−0.03***	−0.03***
			(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Information			0.01	0.01	0.01	0.004
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			−0.12***	−0.11***	−0.10***	−0.10***
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			−0.01	−0.01	0.01	0.01
			(0.03)	(0.03)	(0.04)	(0.04)
Low-fake Env.			0.03***	0.02**	0.02**	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.03***	0.03***	0.03***	0.03***
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.01	−0.001	0.02*	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.001	0.003	−0.004	0.002
			(0.03)	(0.02)	(0.04)	(0.02)
College			0.01	0.02	0.0000	0.004
			(0.03)	(0.02)	(0.04)	(0.03)
Postgrad			0.04	0.04	0.03	0.04
			(0.04)	(0.02)	(0.04)	(0.03)
C.R.			−0.06***	−0.08***	−0.05**	−0.05*
			(0.02)	(0.02)	(0.02)	(0.02)
C.R. x Republican			0.04	0.06	0.04	0.05
			(0.03)	(0.03)	(0.03)	(0.03)
Ambivalent Sexism			0.01	0.01*	0.003	−0.001
			(0.004)	(0.004)	(0.01)	(0.01)
Republican			−0.06***	−0.07***	−0.07***	−0.06***
			(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.22***	0.24***	0.31***	0.34***	0.30***	0.32***
	(0.005)	(0.02)	(0.05)	(0.04)	(0.06)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,443	5,443	5,442	5,442	3,844	3,844
R^2	0.0000	0.0004	0.03	0.03	0.03	0.02
Adjusted R^2	−0.0002	0.0002	0.03	0.03	0.03	0.02

Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

Table G26: **Predictors of Detection Task False Negative Rate (FNR)**

	Detection FNR (% Deepfakes Classified as Real Videos)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		−0.03	−0.03	−0.01	−0.02	−0.003
		(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Accuracy Prime	−0.01		−0.01	−0.01	−0.02	−0.03***
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Debrief			0.01	0.01	0.01	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Exp 1 Information			−0.003	0.002	−0.01	−0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			−0.03	−0.06**	−0.03	−0.04
			(0.02)	(0.02)	(0.03)	(0.03)
Internet Usage			0.12**	0.08	0.12*	0.10**
			(0.04)	(0.04)	(0.05)	(0.05)
Low-fake Env.			0.01	0.001	0.01	−0.01
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.02	0.02*	0.03	0.03
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.001	0.01	0.004	0.01
			(0.05)	(0.03)	(0.05)	(0.03)
High School			0.02	0.03	0.03	0.03
			(0.05)	(0.03)	(0.05)	(0.03)
College			0.08	0.12***	0.07	0.10**
			(0.05)	(0.03)	(0.05)	(0.03)
Postgrad			−0.04	−0.03	−0.04	−0.02
			(0.02)	(0.03)	(0.03)	(0.03)
Republican			0.09**	0.08*	0.06	0.02
			(0.04)	(0.04)	(0.04)	(0.04)
C.R.			0.02***	0.02***	0.01	0.005
			(0.01)	(0.01)	(0.01)	(0.01)
Ambivalent Sexism			−0.04**	−0.03**	−0.03	−0.01
			(0.01)	(0.02)	(0.02)	(0.02)
C.R. x Republican	0.34***	0.35***	0.20***	0.23***	0.22***	0.24***
	(0.01)	(0.03)	(0.07)	(0.06)	(0.08)	(0.07)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	3,654	3,654	3,654	3,654	2,557	2,557
R^2	0.0003	0.0002	0.02	0.03	0.02	0.02
Adjusted R^2	0.0000	−0.0001	0.02	0.02	0.01	0.01

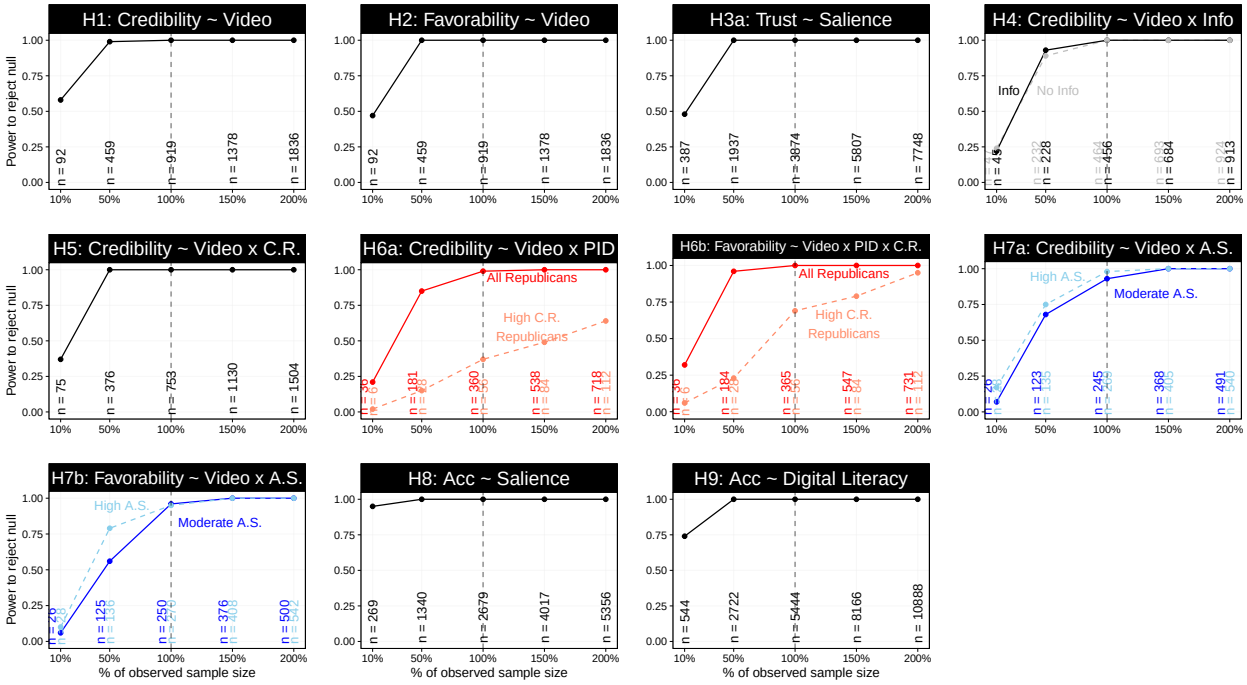
Notes: · $p \cdot r/K < .1$ * $p \cdot r/K < .05$ ** $p \cdot r/K < .01$ *** $p \cdot r/K < .001$

H Power Analyses

Null results like the ones we have reported in Figure 2 may be the inadvertent consequence of an under-powered experiment. However, Figure H9 demonstrates that, with only a few unsurprising exceptions, all of our pre-registered hypotheses (which include our topline findings for **RQ1–3**) are powered at 90% or higher at our observed sample size to detect effects at our stated equivalence bounds (±0.5 standard deviations of each outcome).

To compute the statistical power for each hypothesis, we re-sample the relevant units to test the hypothesis at different sample sizes: 10%, 50%, 100%, 150%, and 200% of the actual number of units from our experiment used to report effects in the main text. For each sample size, we then simulate effects equivalent to the observed upper equivalence bound (±0.5 s.d.) in the re-sampled data for the particular outcome. We conduct our procedure for estimating effects (in this case, the simplest univariate regression that we pre-registered for each hypothesis), collect whether we reject the null hypothesis ($\alpha = 0.01$), and repeat this process 1000 times at each sample size.

Figure H9: **Statistical Power for Pre-Registered Analyses**



Notes: Each observed sample size for which power is computed (1,000 simulations) is denoted both the number of units sampled into the experiment (text label "$n =$") and the corresponding % out of the observed number of units for the relevant subgroup(s) for that hypothesis (horizontal axis). For simplicity and for a more conservative power computation, all hypotheses involving the video stimuli are tested using only the `text` and `video` conditions. For interaction effects, power is computed using the regression specification for individual subgroups (some omitted for brevity).

Most notably Figure H9 shows that we are not well-powered to detect the effects of the `video` condition in Experiment 1 on high-cognition Republicans' credibility perceptions ($\mathbf{H}_{6a}$) and candidate favorability ($\mathbf{H}_{6b}$). Hence, we exclude any inferences about the three-way interaction between partisanship, cognitive reflection and stimuli condition in Experiment 1.

Although this particular power analysis was not conducted ahead of our experiment, we believe it is equally or more informative an ex-ante power analysis. This is because we must evaluate our statistical power *given the observed data and the observed effect size*, removing the researcher discretion used to speculate plausible effect sizes (potentially favorable to our hypotheses) and compute power across simulated data that are subject to modeling assumptions.

I Robustness Checks

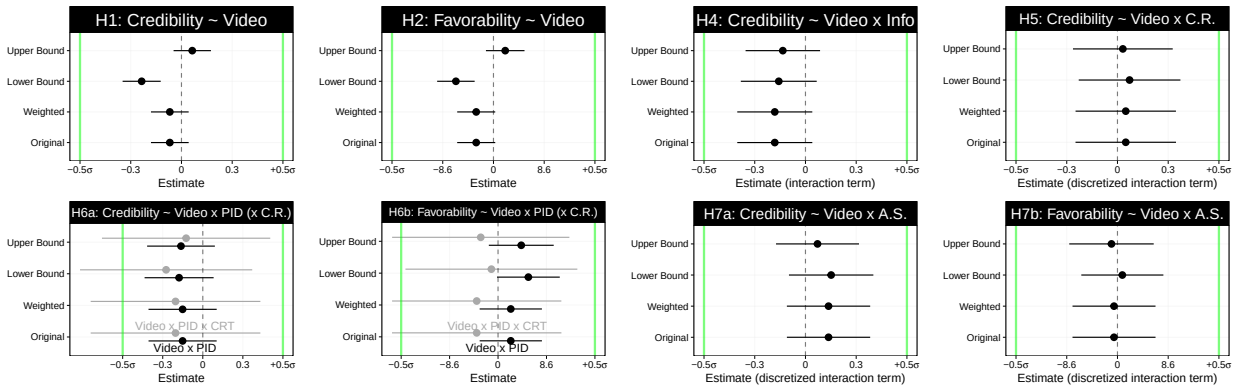
Unless otherwise stated, all bars resembling confidence intervals are 95% confidence intervals.

I.1 Attrition checks

A careful reader[14] might raise concern about the imbalance across stimulus conditions in Experiment 1 (Table 3): the `video` condition has only 872 participants whereas all other conditions have ≈950. This might indicate differential attrition, which could potentially bias our reported treatment effects.

We find no evidence that this imbalance is due to a randomization failure or differential attrition. After contacting our survey provider, Lucid, we confirmed that there was no failure of the complete randomization function within the survey flow. Moreover, the differences in these cells does not survive statistical significance from a Chi-square test or the slightly more conservative exact Binomial test. Assignment to the `video` cell also does not significantly predict whether respondents complete the survey, although it does predict the length of time taken to complete the survey (which is expected, given the inherently more time-consuming nature of loading and viewing a video relative to a headline).

Figure I10: **Manski Bounds and Reweighted Estimates for Pre-Registered Analyses**



Notes: The vertical axis shows our original estimates for each hypothesis (numbered according to our pre-registration), re-weighted estimates after weighting each unit by the inverse proportion of units in its' Experiment 1 treatment condition, and lower and upper extreme Manski bounds (Horowitz and Manski, 2000) of the estimates. Lines correspond to 95% confidence intervals. All estimates are computed using a univariate regression model specific for each hypothesis. The horizontal axis shows the value of the cofficient of interest on its original scale with the stated equivalence bounds for the outcome in green (±0.5 standard deviations). Grey lines correspond to hypotheses that have been determined to be under-powered (see Section H).

To err on the side of caution, we re-evaluated our hypotheses after re-adjusting the data with extreme value bounds (Horowitz and Manski, 2000) – possible to due the finite range of

[14]We thank an anonymous reviewer for taking this care.

all outcomes – and re-adjusting the data with inverse weights by the Experiment 1 treatment cell. Figure I10 shows the results. Indeed estimates from the Manski bounds do flip signs and, taking the point estimates from each bound together to form the complete bound, do include zero as is typically expected (Horowitz and Manski, 2000). Notably, however, no hypothesis that previously failed to reach substantive significance (according to our equivalence bounds) does so in this most conservative imputation of the missing `video` condition outcomes.

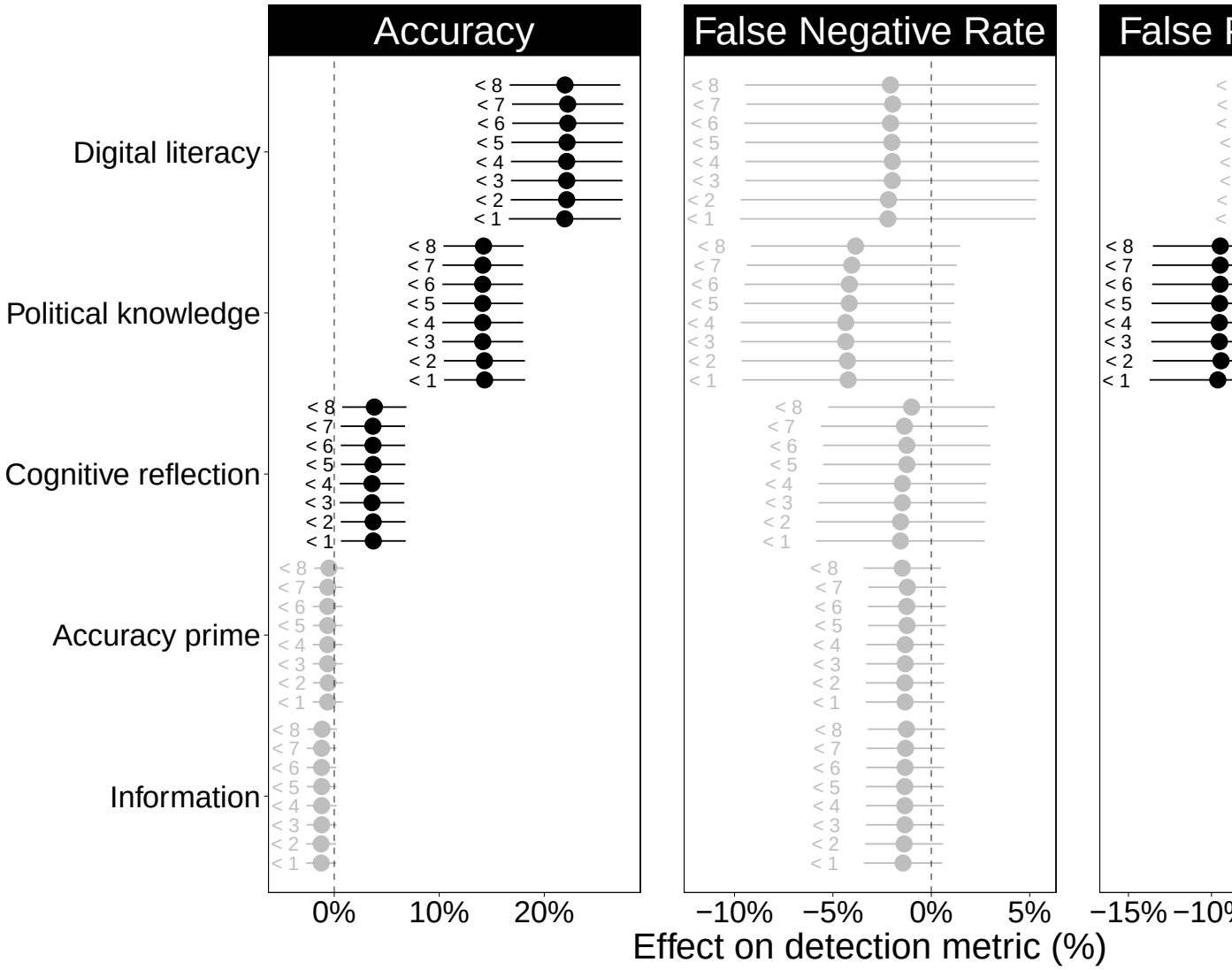
I.2 Nonresponse checks

Some respondents failed to enter a detection for all videos in Experiment 2. However, this number is relatively small and has no bearing on our findings in Figure 4. As Figure I11 shows, removing non-responding subjects does not substantially change our reported estimates.

I.3 Placebo checks

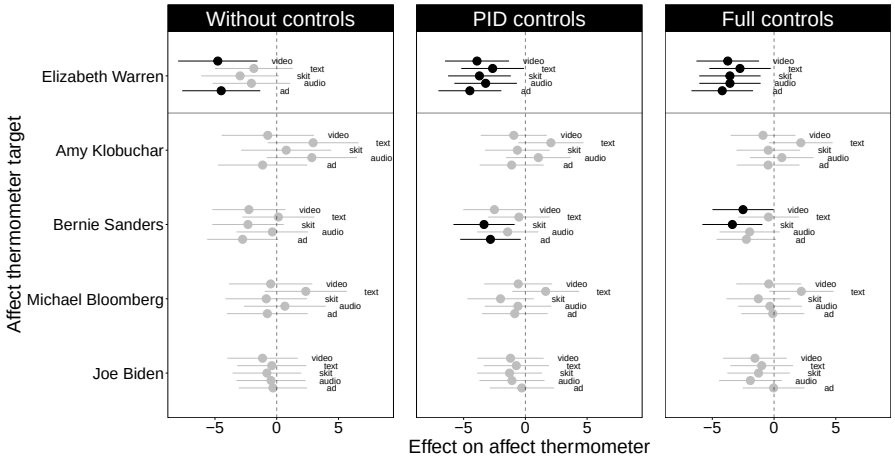
Next we show the results of two placebo tests in order to validate that the treatment stimuli manipulated attitudes towards Elizabeth Warren as intended. Figure I12 shows that exposure to the scandal stimuli of Elizabeth Warren reduced Warren's favorability across media conditions as intended. Figure I13 shows that ambivalent sexism predicted lower favorability for Warren (although it predicted lower favorability for other male Democratic candidates as well), but not for Joe Biden.

Figure I11: **Sensitivity of Predictors of Detection Experiment Performance to Non-Response Thresholding**



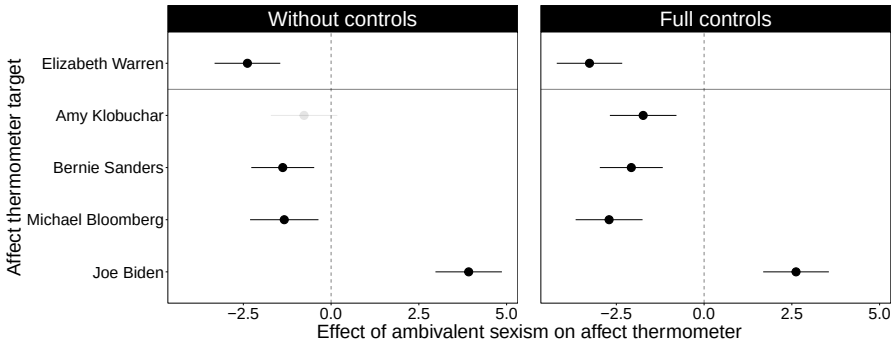
Notes: Each estimate is the effect (95% confidence interval) of the corresponding predictor estimated from a model with full controls (see tables for detection task results) excluding respondents with $< x$ number of videos completed in the detection task.

Figure I12: **Clip Type and Affect Towards Placebo Targets in Incidental Exposure Experiment**



Notes: Shown are other candidates who ran in the 2020 Democratic primary for whom we selected clips to mask our deepfake in the incidental exposure experiment.

Figure I13: **Ambivalent Sexism and Affect Towards Placebo Targets in Incidental Exposure Experiment**



Notes: Shown are other candidates who ran in the 2020 Democratic primary.

J Exploratory Analyses

This section produces additional results bolstering findings in the main text that are not pre-registered (i.e. exploratory). Unless otherwise stated, all bars resembling confidence intervals are 95% confidence intervals.

Figure J14: **Baseline Comparisons of Credibility and Affective Response of Video in Incidental Exposure Experiment**

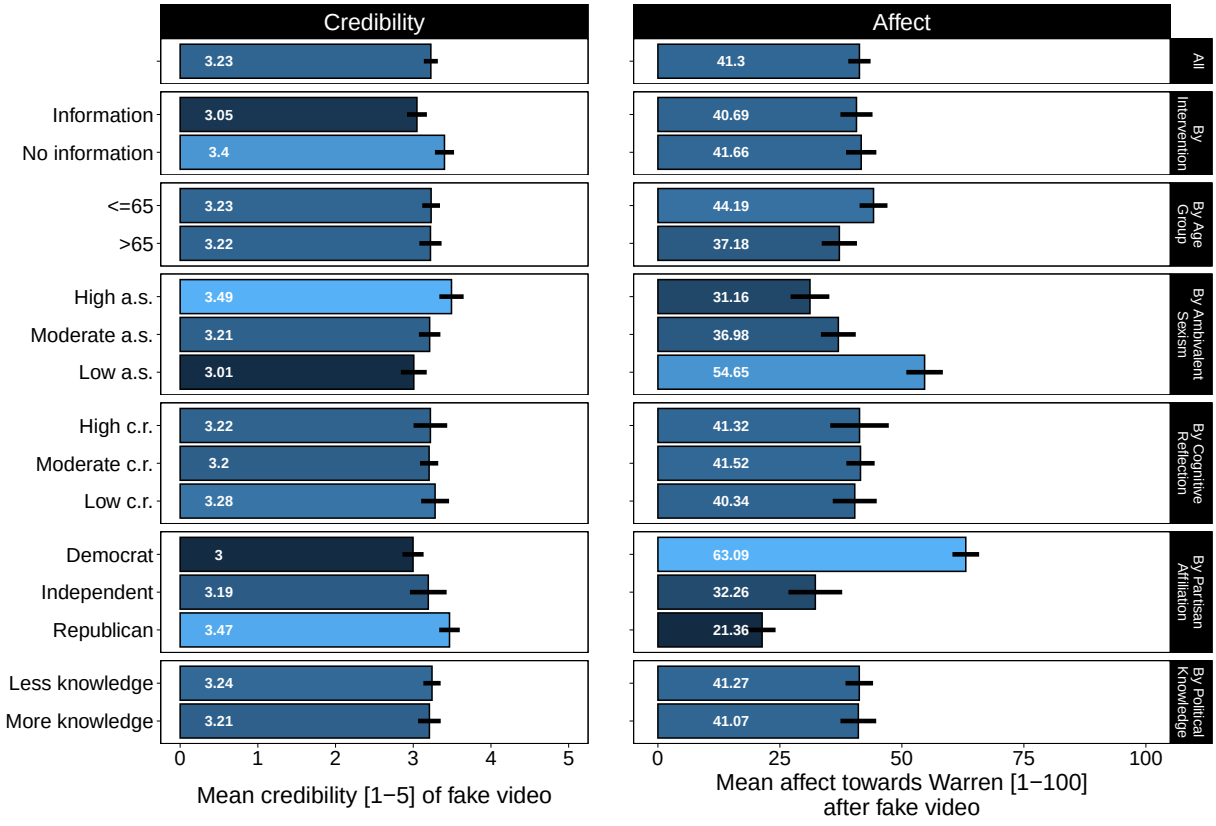
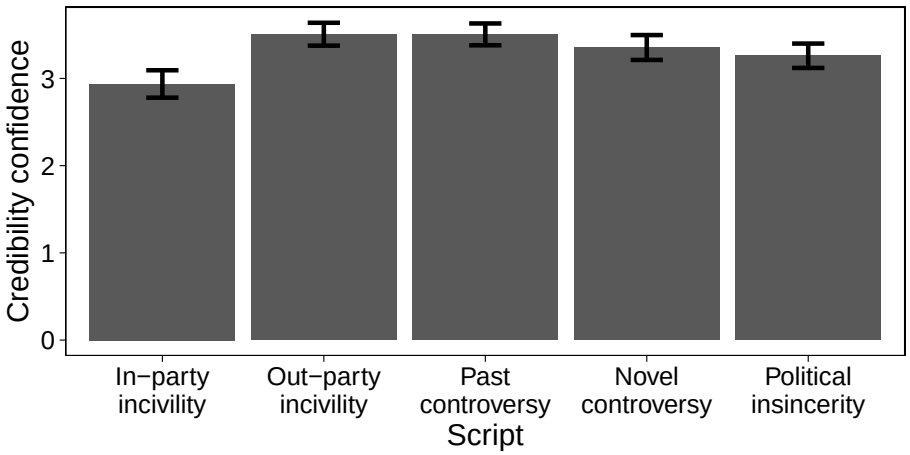
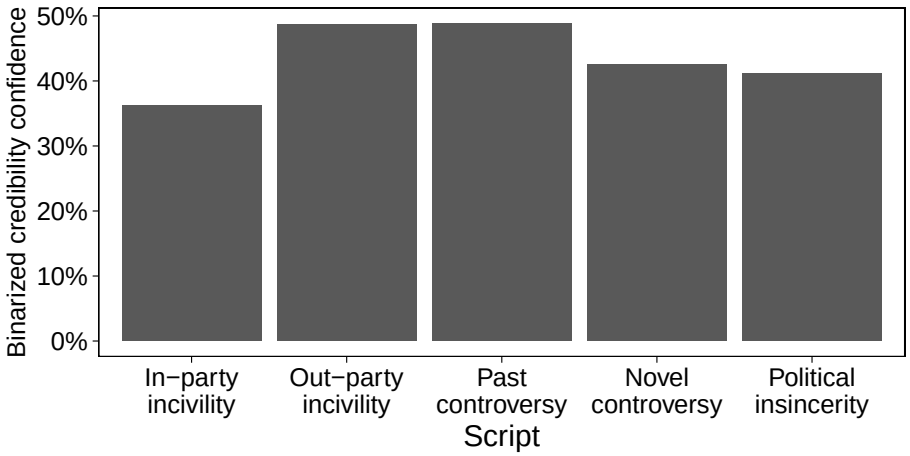


Figure J15: **Heterogeneity in Incidental Exposure Credibility by Scandal Script**

(a) **Categorical Response**



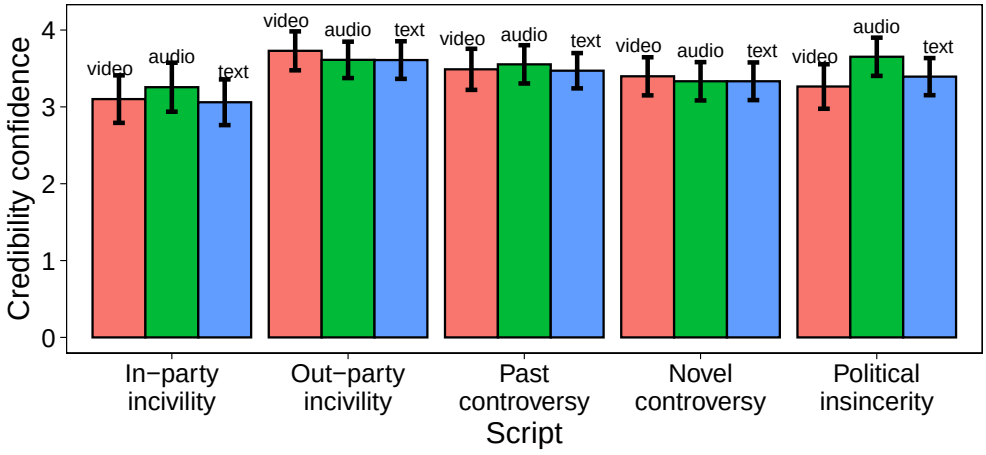
(b) **Binarized Response**



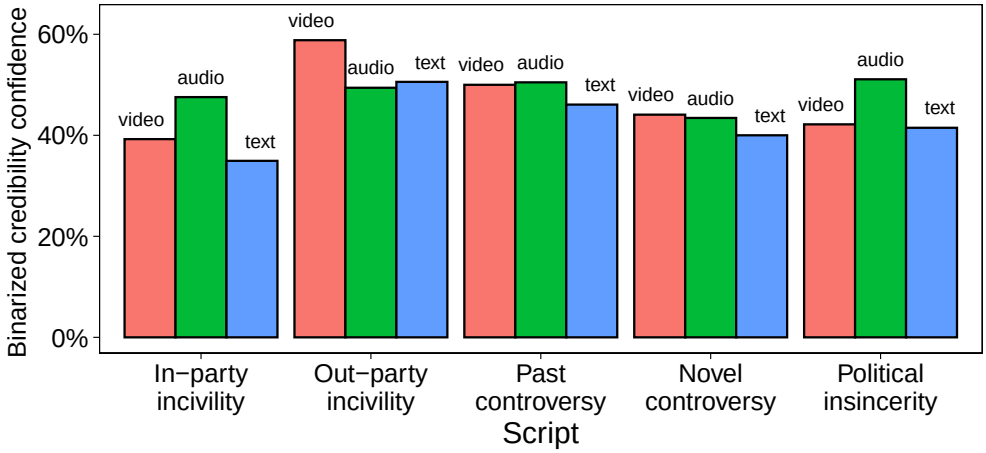
Notes: Results from the subset of respondents exposed to a scandal stimuli, not assigned an information treatment, and who provided a response to our deception question (n=1848). Bars in (a) indicate a 95% confidence interval around the mean credibility response.

Figure J16: **Heterogeneity in Incidental Exposure Credibility by Scandal Script and Medium**

(a) **Categorical Response**

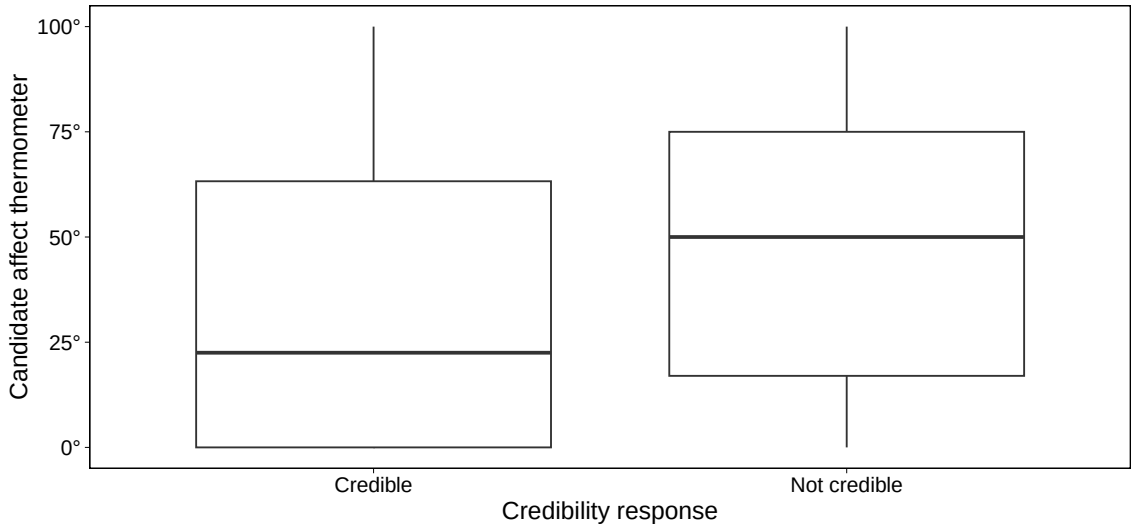


(b) **Binarized Response**



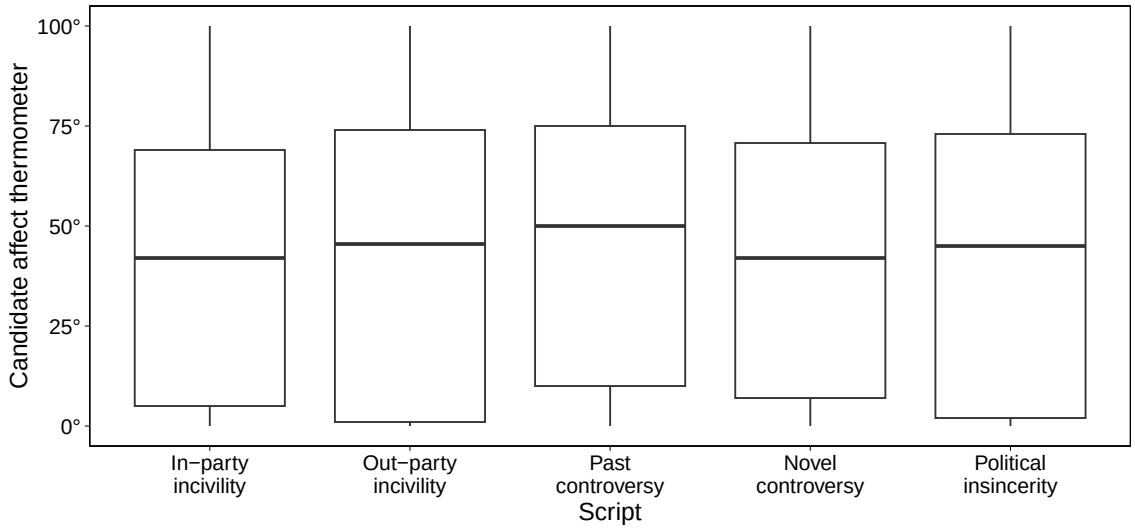
Notes: Results from the subset of respondents exposed to a scandal stimuli, not assigned an information treatment, and who provided a response to our credibility question (n=1,848). To reduce the number of experimental cells, only three of five scripts were used for the `skit` stimuli. Bars in (a) indicate a 95% confidence interval around the mean credibility response.

Figure J17: **Heterogeneity in Incidental Exposure Affect By Credibility**



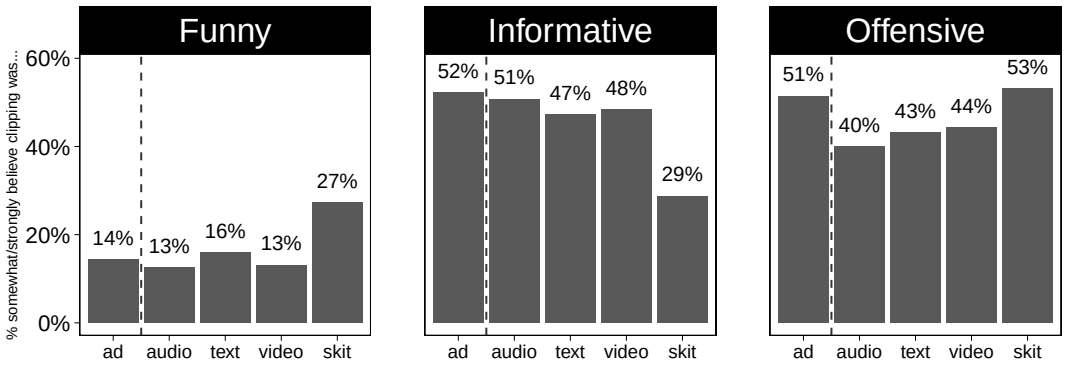
Notes: Results from the subset of respondents not assigned an information treatment, and who provided a response to our affect thermometer and credibility questions (n=2,070). Group means differ significantly ($t = -11.64,\ p < 0.001$). We emphasize that this difference cannot be interpreted causally.

Figure J18: **Heterogeneity in Incidental Exposure Affect By Scandal Script**



Notes: Results from the subset of respondents not assigned an information treatment, not assigned to control (no stimulus) and who provided a response to our affect thermometer questions (n=1,829). Relative to the "In-party incivility" script, only the "Novel controversy" script predicts a significantly higher favorability level but a much weaker level of significance than the rest of our study ($t = 1.697,\ p = 0.08$). We emphasize that this difference cannot be interpreted causally.

Figure J19: **Other Affective Responses to First-Stage Clip**

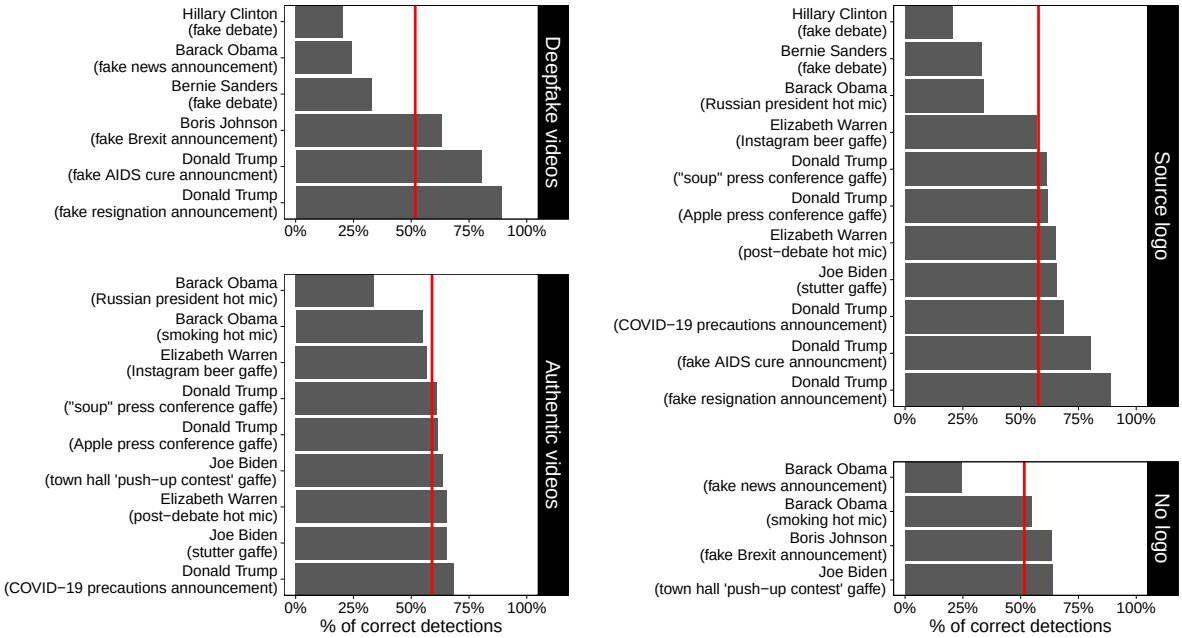


Notes: Responses evaluating whether clip was "funny," "informative," or "offensive" were solicited alongside belief that clip was not fake or doctored. The attack ad condition excluded since it is not a directly comparable clip of the scandal.

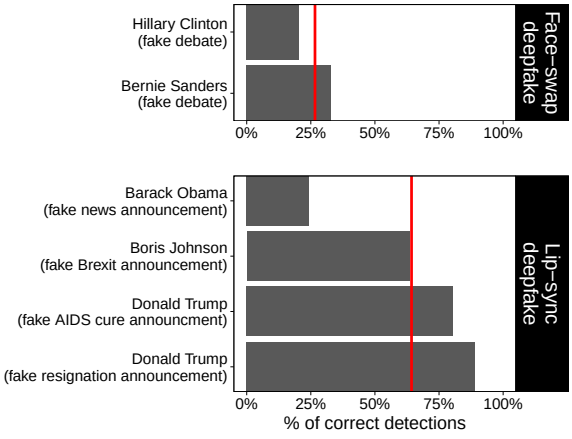
Figure J20: **Detection Task Performance for Specific Clips**

(a) **By Authenticity**

(b) **By Source Cue**

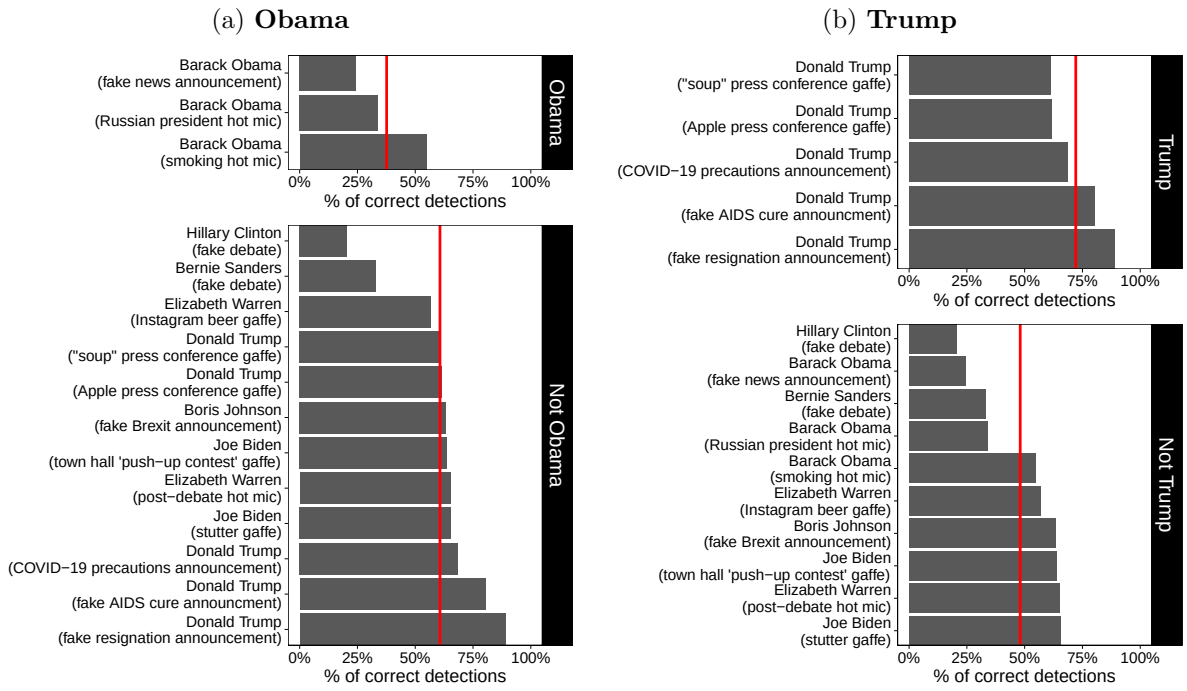


(c) **By Deepfake Technology**



Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. Fake clips are detected less well than real clips, but this difference (Δ) is not significant according to a t-test ($\Delta = -7.20\%, t = 0.57, p = 0.58$). Clips without source outlet logos are detected less well than clips with source logos, but this difference is also not significant ($\Delta = -6.03\%, t = 0.53, p = 0.61$).

Figure J21: **Detection Task Performance for Specific Clips by Target (Obama vs. Trump)**

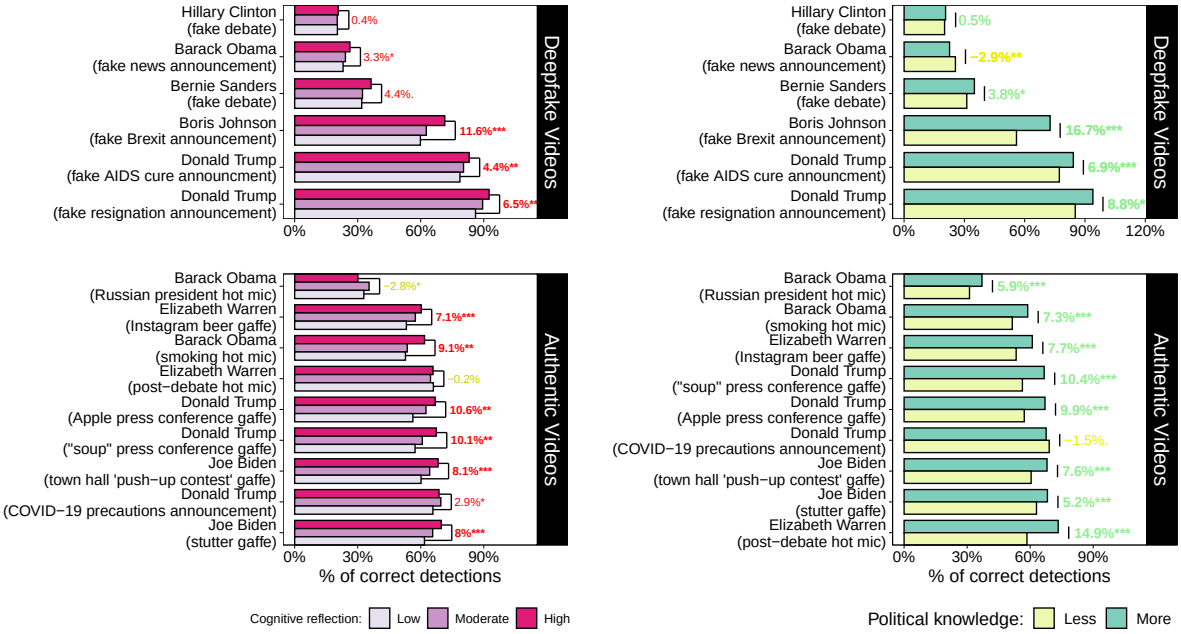


Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. Fake clips are detected less well than real clips, but this difference (Δ) is not significant according to a t-test ($\Delta = -7.20\%, t = 0.57, p = 0.58$). Clips without source outlet logos are detected less well than clips with source logos, but this difference is also not significant ($\Delta = -6.03\%, t = 0.53, p = 0.61$).

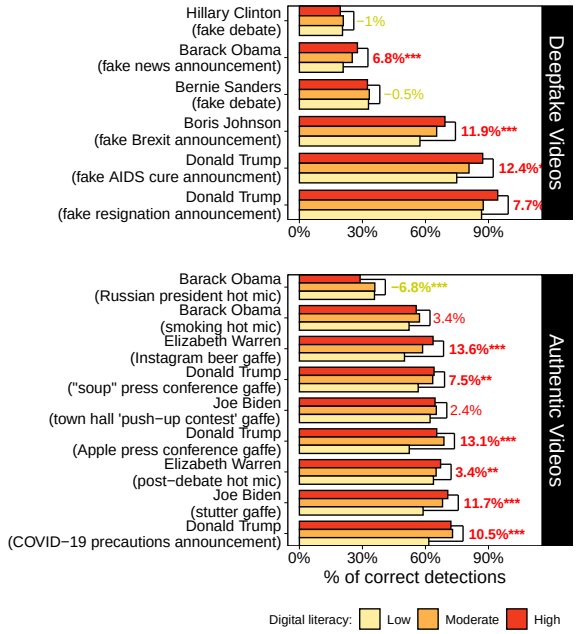
Figure J22: **Detection Task Performance for Specific Clips by Subgroup**

(a) **By Cognitive Reflection**

(b) **By Political Knowledge**

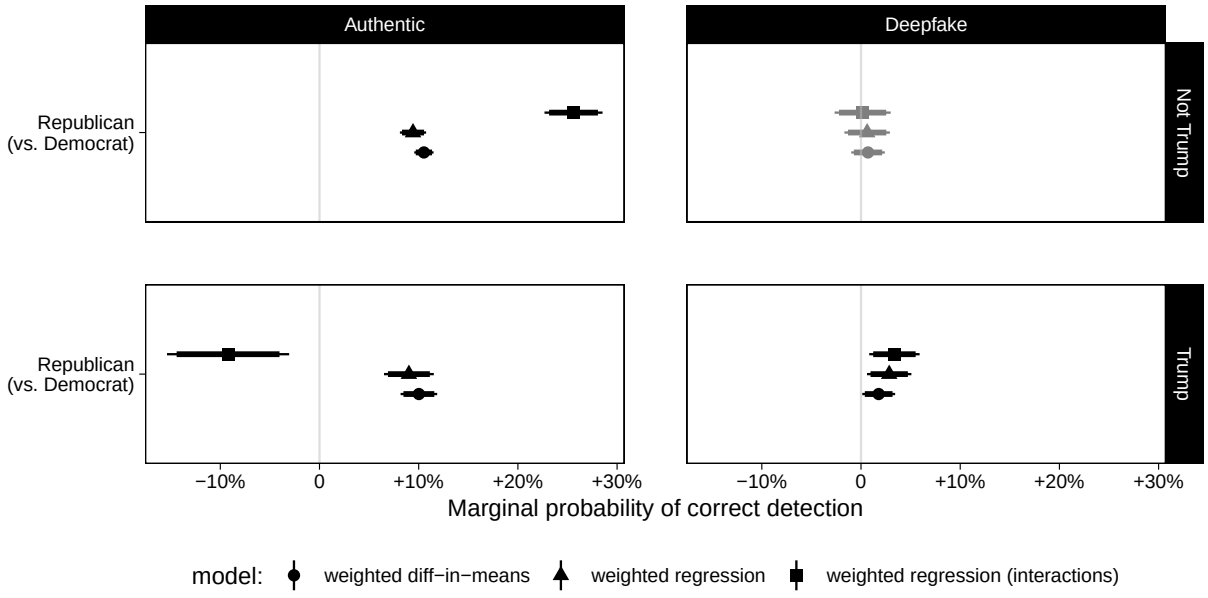


(c) **By Digital Literacy**



Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. Cognitive reflection and digital literacy categories constructed as equal-sized quartiles.

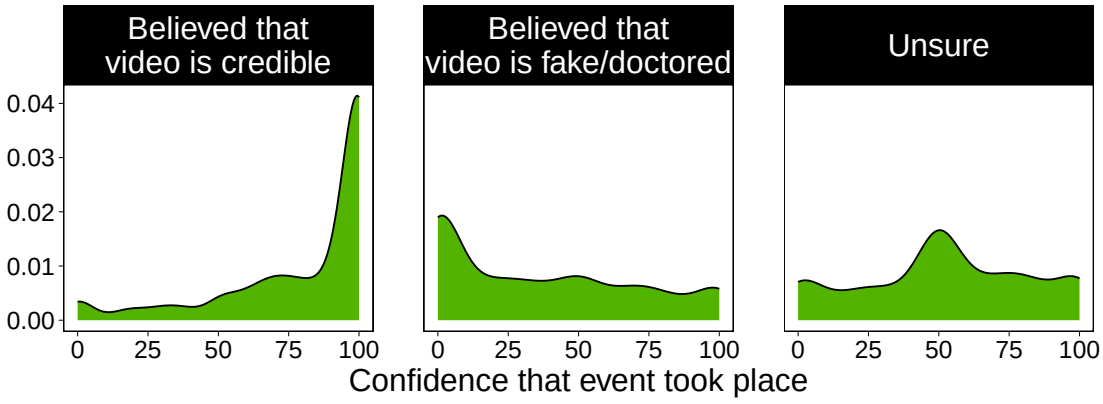
Figure J23: **Comparison of Partisan Group Identity Effects on Detection Task Performance**



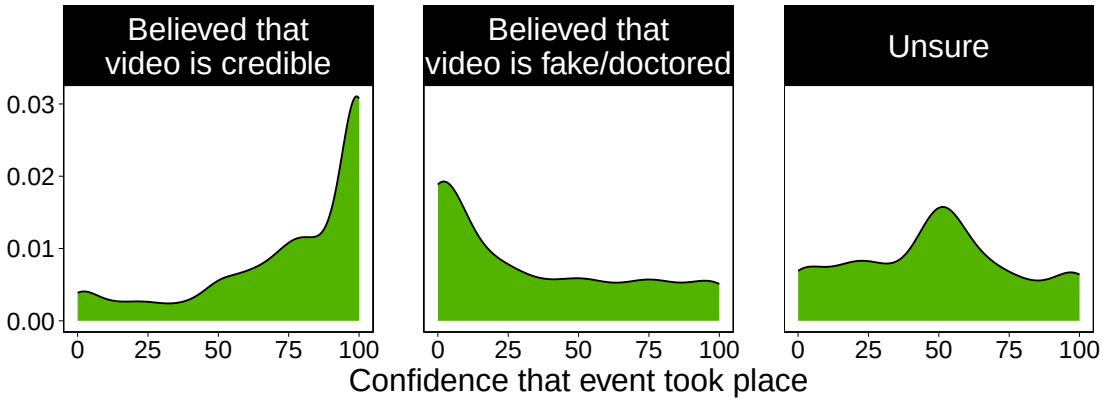
Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. The vertical axis denotes partisan group identity of the respondent for which effects are computed (Republican) relative to a baseline group (Democrat). All models incorporate weights from post-stratification described in Appendix F. The regression controls for the characteristics described at the start of this section as well as whether the clip contains a source logo or not. For simplicity, Ordinary Least Squares linear regression is used to estimate marginal probabilities rather than a binary outcome regression model, though results look similar from a logistic regression model. The interaction model fits an interaction term of partisanship with an indicator for particular clip within each cell.

Figure J24: **Relationship Between Credibility of Video Clip and Credibility of Event**

(a) **Barack Obama (Russian president hot mic)**



(b) **Donald Trump (COVID-19 precautions announcement)**



Notes: The vertical axis denotes density in each plot. Clips (a) and (b) are real videos. Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. A variety of regression specifications estimate large, robust and statistically significant positive relationship between a respondent's belief in the video's authenticity and confidence in the depicted event's occurrence.

K Survey Measures

In this section we list the survey measures that we use. The subsection titles refer to boxes in Figure B6, which more abstractly demonstrates the survey flow.

K.1 Pre-Exposure Questionnaire

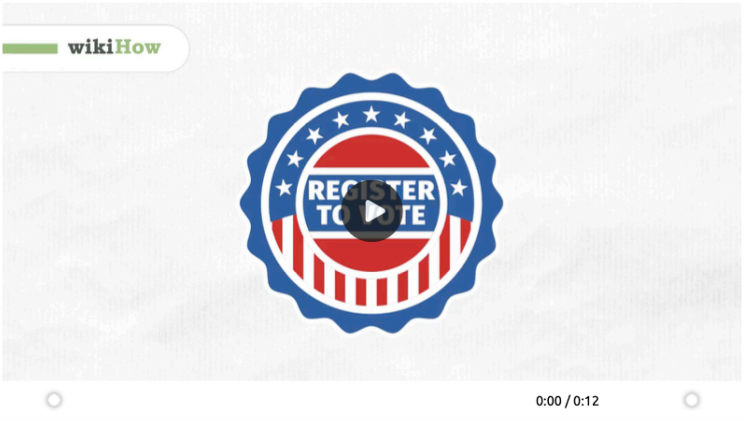
> Note for readers: In this section, we performed a short attention check, and asked a question to confirm that the participant was able to watch and listen to video. We then asked a series of pre-experiment demographic questions, which we use to test for the reported heterogeneities.

For our research, careful attention to survey questions is critical! To show that you are paying attention please select "I have a question."

○ I understand
○ I do not understand
○ I have a question

People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both "extremely interested" and "slightly interested."

○ Extremely interested
○ Very interested
○ Moderately interested
○ Slightly interested
○ Not interested at all



Watch the video above and answer the questions below.

In the video above, which of the following describes the speaker's characterization of the process of registering to vote in most US states?

○ Quick and easy
○ Swift and speedy
○ Slow and arduous
○ Undemocratic and illegal
○ Important and necessary

The name of the video's producer is first displayed in the top left corner, then the bottom right. Who produced the video?

○ wikiHow
○ Buzzfeed
○ The New York Times
○ Vice Media

Before we get started, we'd like to learn a little bit about your background as well as your opinions and knowledge on a few different topics. Please answer them *truthfully*.

How old are you?

What is your gender?

○ Male
○ Female
○ Other

Below is a series of statements concerning **men and women** and their relationships in contemporary society. Please indicate the degree to which you agree or disagree with each statement.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Women complain too often about being discriminated against.	○	○	○	○	○
Most women interpret innocent remarks or acts as being sexist.	○	○	○	○	○
Women should remain level-headed and calm in difficult situations.	○	○	○	○	○
Women should display superior moral virtue and judgment compared to men.	○	○	○	○	○
Women are too often uncivil, abrasive, or shrill in difficult situations.	○	○	○	○	○

What is the highest level of education you've completed?

○ Have not finished high school
○ High school
○ College
○ Postgraduate degree

How often do you use the Internet?

○ Pretty much all the time
○ Several times a day
○ About once a day
○ 3 to 6 days a week
○ 1 to 2 days a week
○ Every few weeks
○ Less often

How often do you use Facebook?

○ Pretty much all the time
○ Several times a day
○ About once a day
○ 3 to 6 days a week
○ 1 to 2 days a week
○ Every few weeks
○ Less often

Generally speaking, do you usually think of yourself as a Democrat, a Republican, or an independent?

○ Democrat
○ Independent
○ Republican

Please answer the following questions, thinking through the answers carefully and *without* consulting external sources.

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How many cents does the ball cost? Please enter in the format of $X.XX.

If it takes 5 machines 5 minutes to make 5 widgets, how many minutes would it take 100 machines to make 100 widgets?

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake?

What is Saturday Night Live?

○ A gameshow
○ A sitcom
○ A late-night sketch comedy show
○ A televised dance competition

Who is the current Speaker of the US House of Representatives?

○ John Boehner
○ Mike Pence
○ Mitch McConnell
○ Nancy Pelosi

What is Medicare?

○ A program run by the US federal government to pay for old people's health care
○ A program run by state governments to provide health care to poor people
○ A private health insurance plan sold to individuals in all 50 states
○ A private, non-profit organization that runs free health clinics

Which political party has a majority in the U.S. House of Representatives?

○ Democrats
○ Republicans
○ I don't know

Which political party has a majority in the U.S. Senate?

○ Democrats
○ Republicans
○ I don't know

How much of a majority is required for the US Senate and US House to override a presidential veto?

○ One-half
○ Two-thirds
○ Three-fifths
○ Three-fourths

Please select the color blue.

○ red
○ green
○ blue
○ orange

As of today do you lean more to the Republican Party or more to the Democratic Party?

○ Democrat
○ Republican
○ Neither

How often do you read news stories online?

○ Several times a day
○ About once a day
○ 3 to 6 days a week
○ 1 to 2 days a week
○ Every few weeks
○ Less often

How often do you read news stories offline (in the newspaper, printed news magazines)?

○ Several times a day
○ About once a day
○ 3 to 6 days a week
○ 1 to 2 days a week
○ Every few weeks
○ Less often

K.2 No Information/Information About Deepfakes

> Note for readers: This section displays our informational intervention. Subjects assigned to the "Information About Deepfakes" condition saw the normal and italicized text, whereas those assigned to the "No Information" condition saw only the normal text.

We're going show you a series of media clippings about candidates in the Democratic primary for the 2020 U.S. Presidential Election. These will consist of a mix of text headlines, audio clippings, and video clippings - similar to the experience of looking at a news site or scrolling through Facebook or Twitter.

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.

Please take your time to fully read, watch, or listen to each piece.

K.3 Newsfeed

> Note for readers: As described in Section 2.1, to create a natural environment for media consumption, we surround the experimentally manipulated media exposure with four media clips, two before and two after. In this section of the

survey, subjects were sequentially exposed to five separate pages, each of which contained a separate piece of media. The third page contained one of the experimental conditions enumerated in Table 3. The first, second, fourth, and fifth pages were constant across all subjects, and each contained true but potentially scandalous stories. The first page was a video of Biden, the second a text report on Klobuchar, the fourth an audio recording of Michael Bloomberg, and the fifth a skit of Larry David playing Bernie Sanders (but with a subheading which clearly denotes that this is a skit).

There was a page break between each of these five pages, all of which are shown below.

First Page:

YOUTUBE.COM
Joe Biden tells auto plant worker he's "full of sh" and shushes aid**

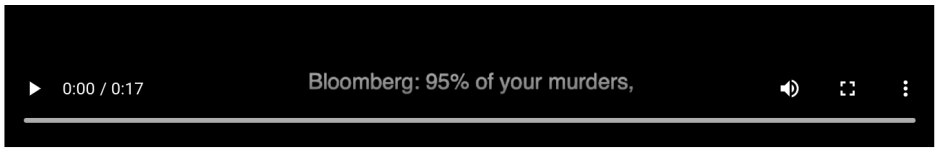
Second Page:

Amy Klobuchar on Budget Cuts: "This President Lacks Empathy"

Democratic primary candidate Amy Klobuchar says that the President's proposed cuts to Medicare and Medicaid shows he lacks empathy because when he boasts about the economy, "most Americans aren't feeling that prosperity".

Third Page: **EMBED PRIMARY MANIPULATION: EXPOSURE TO 1 OF THE 6 CONDITIONS DEPICTED IN TABLE 3.**

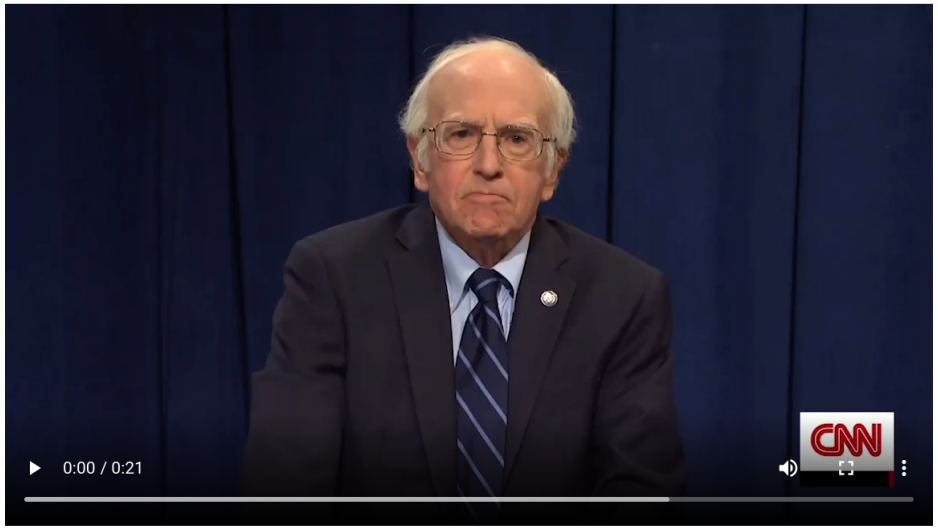
Fourth Page:



YOUTUBE.COM

Leaked audio: Michael Bloomberg defends racial profiling and stop-and-frisk policing in 2015

Fifth Page:



YOUTUBE.COM

Larry David as Bernie Sanders in DNC Debate

Post-Exposure Questionnaire

> Note for readers: We next measured out primary outcomes, first by asking about each of the five media to which the subject was exposed was offensive, funny, fake or doctored, and informative to voters. We then measured a feeling thermometer for each of the five politicians in the media clips, and then asked an additional attention check ('Please select the number 2'). We then ask our primary measures of digital literacy, political knowledge, and media trust.

To what extent do you think that the clipping of Joe Biden telling an auto plant worker he's "full of sh**" ...

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
is offensive	○	○	○	○	○
is funny	○	○	○	○	○
is fake or doctored	○	○	○	○	○
is informative for voters	○	○	○	○	○

To what extent do you think that the clipping of Amy Klobuchar saying that President Trump lacks empathy because of his budget cuts ...

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
is offensive	○	○	○	○	○
is funny	○	○	○	○	○
is fake or doctored	○	○	○	○	○
is informative for voters	○	○	○	○	○

To what extent do you think that the clipping of Elizabeth Warren [TEXT DESCRIPTION OF THE PARTICULAR WARREN CLIP SEEN BY SUBJECT (SEE TABLE C5 FOR A COMPLETE LIST)] ...

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
is offensive	○	○	○	○	○
is funny	○	○	○	○	○
is fake or doctored	○	○	○	○	○
is informative for voters	○	○	○	○	○

To what extent do you think that the clipping of Michael Bloomberg defending racial profiling and stop-and-frisk policing in 2015 ...

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
is offensive	○	○	○	○	○
is funny	○	○	○	○	○
is fake or doctored	○	○	○	○	○
is informative for voters	○	○	○	○	○

To what extent do you think that the clipping of Bernie Sanders announcing his 2020 campaign slogan as "let's Bern this place to the ground" ...

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
is offensive	○	○	○	○	○
is funny	○	○	○	○	○
is fake or doctored	○	○	○	○	○
is informative for voters	○	○	○	○	○

We'd like you to rate how you feel towards each of the following candidates from the 2020 Democratic primary on a scale of 0 to 100. Zero means very unfavorable and 100 means very favorable. Fifty means you do not feel favorable or unfavorable. How would you rate your feeling toward each candidate?

	Very unfavorable										Very Favorable
	0	10	20	30	40	50	60	70	80	90	100
Joe Biden	○	○	○	○	○	○	○	○	○	○	○
Amy Klobuchar	○	○	○	○	○	○	○	○	○	○	○
Elizabeth Warren	○	○	○	○	○	○	○	○	○	○	○
Michael Bloomberg	○	○	○	○	○	○	○	○	○	○	○
Bernie Sanders	○	○	○	○	○	○	○	○	○	○	○

Please select the number 2.

○ 1
○ 2
○ 3
○ 4

How familiar are you with the following computer and Internet-related items? Please choose a number between 1 and 5 where 1 represents "no understanding" and 5 represents "full understanding" of the item.

	1: No understanding	2	3	4	5: Full understanding
Hashtag	○	○	○	○	○
App	○	○	○	○	○
Smartphone	○	○	○	○	○
Fitibly	○	○	○	○	○
Selfie	○	○	○	○	○
Tablet	○	○	○	○	○
PDF	○	○	○	○	○

Who is the current female senator from Massachusetts?

○ Ed Markey
○ Elizabeth Warren
○ Joni Ernst
○ Susan Collins
○ I don't know

Who is the current Prime Minister of the U.K.?

○ Jeremy Corbyn
○ Theresa May
○ Boris Johnson
○ Keir Starmer
○ I don't know

In general, how much trust and confidence do you have in **offline media** — such as newspapers, T.V. and radio — when it comes to reporting the news fully, accurately, and fairly — a great deal, a fair amount, not very much, or none at all?

○ A great deal
○ A fair amount
○ Not very much
○ None at all

In general, how much trust and confidence do you have in **online-only media** — such as blogs and online-only news websites — when it comes to reporting the news fully, accurately, and fairly — a great deal, a fair amount, not very much, or none at all?

○ A great deal
○ A fair amount
○ Not very much
○ None at all

In general, how much trust and confidence do you have in **social media** — such as Facebook or Twitter — when it comes to covering the news fully, accurately, and fairly — a great deal, a fair amount, not very much, or none at all?

○ A great deal
○ A fair amount
○ Not very much
○ None at all

K.4 Exposure Debrief

Note for readers: All subjects were debriefed. Subjects were randomized to be debriefed before or after the Detection experiment. In the debrief, subjects were again shown the Warren media to which they were exposed, then were told that it was fabricated, and could not advance to the next page until typing "The video about Elizabeth Warren is false."

K.5 Accuracy Prime

> Note for readers: Before the detection experiment, all subjects saw the normal text, and those assigned to the accuracy prime also saw the italicized text.

Now we're going to show you a series of videos of politicians. Some of these may have been digitally manipulated to depict someone saying something they did not actually say in the original clip. We'd like you to identify which of these you think are fake/doctored, and which are real.

Sometimes when people watch political videos, they get angry or excited based on what is being shown, rather than taking the time to stop and think about whether it's conveying accurate information. Democracy works best when people take the time to consider the accuracy of what they see.

K.6 Detection

Subjects then participated in the detection experiment. Section D fully enumerates these conditions, so we point interested readers to that section and do not duplicate it here.

Recall that subjects were shown either zero, two, or six doctored videos in this experiment. After completing the experiment, subjects assigned to zero doctored videos were told that "None of the media on the previous page was doctored or fake." Subjects assigned to see either two or six doctored videos could not advance to the next page until typing "Two/Six videos did not take place as depicted" and were given a chance to review the doctored videos to which they were exposed.

Appendix References

Abram, Cleo. 2020. "The most urgent threat of deepfakes isn't politics. It's porn." *Vox* .
URL: https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kriste

Adamic, Lada and Bernardo Huberman. 2000. "Power-Law Distribution of the World Wide Web." *science* 287(5461):2115–2115.

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*. pp. 38–45.

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

Aronow, P., Josh Kalla, Lilla Orr and John Ternovsk. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020." *Working Paper* .
URL: https://osf.io/preprints/socarxiv/8sbe4

Barari, Soubhik, Christopher Lucas and Kevin Munger. 2020. "Pre-Analysis Plan: An Experiment on the Effect of Political Deepfakes on Beliefs and Attitudes.".

Blum, Jeremy. 2020. "Trump Rant About 'Anarchist' Protesters Wielding Deadly 'Cans Of Soup' Goes Viral." *Huffington Post* .
URL: https://www.huffpost.com/entry/trump-deadly-cans-of-soup_n_5f4fbcc6c5b69eb5c0379f01

Davis, Raina. 2020. "Technology Factsheet: Deepfakes." *Policy Brief* .
URL: https://www.belfercenter.org/publication/technology-factsheet-deepfakes

Goodman, J. David. 2012. "Microphone Catches a Candid Obama." *New York Times* .
URL: https://www.nytimes.com/2012/03/27/us/politics/obama-caught-on-microphone-telling-medvedev-of-flexibility.html

Horowitz, Joel L and Charles F Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95(449):77–84.

Ko, Allan, Merry Mou and Nathan Matias. 2016. "The Obligation to Experiment." *Medium* .

Krook, Mona Lena and Juliana Restrepo Sanín. 2020. "The Cost of Doing Politics? Analyzing Violence and Harassment Against Female Politicians." *Perspectives on Politics* 18(3):740–755.

Lewis, Rebecca. 2018. "Alternative Influence: Broadcasting the Reactionary Right on YouTube." *Data & Society* 18.

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow and Brendan Frey. 2015. "Adversarial Autoencoders." *Working Paper* .

Rupar, Aaron. 2019. "Trump's bizarre "Tim/Apple" tweet is a reminder the president refuses to own tiny mistakes." *Vox* .
URL: https://www.vox.com/2019/3/11/18259996/trump-tim-cook-apple-tweet-time-and-words

Suwajanakorn, Supasorn, Steven M Seitz and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing Obama: Learning Lip Sync From Audio." *ACM Transactions on Graphics (TOG)* 36(4):1–13.

Tammekänd, Johannes, John Thomas and Kristjan Peterson. 2020. "Deepfakes 2020: The Tipping Point.".

Westerlund, Mika. 2019. "The Emergence of Deepfake Technology: A Review." *Technology Innovation Management Review* 9(11).

Zimmer, Ben. 2019. "Elizabeth Warren and the Down-to-Earth Trap." *The Atlantic* .
URL: https://www.theatlantic.com/entertainment/archive/2019/01/why-elizabeth-warrens-beer-moment-fell-flat/579544/