Human detection of political deepfakes across transcripts, audio, and video

Matthew Groh\textsuperscript{a,*}, Aruna Sankaranarayanan\textsuperscript{a}, Andrew Lippman\textsuperscript{a}, and Rosalind Picard\textsuperscript{a}

\textsuperscript{a}Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
\textsuperscript{*}To whom correspondence should be addressed. E-mail: groh@mit.edu

ABSTRACT

Recent advances in technology for hyper-realistic visual effects provoke the concern that deepfake videos of political speeches will soon be visually indistinguishable from authentic video recordings. The conventional wisdom in communications research predicts people will fall for fake news more often when the same version of a story is presented as a video rather than text. Here, we evaluate how accurately 41,822 participants distinguish real political speeches from fabrications in an experiment where speeches are randomized to appear as permutations of text, audio, and video. We find access to audio and visual communication modalities improve participants’ accuracy. Here, human judgment relies more on what is said, the audio-visual cues, than what is said, the speech content. However, we find that reflective reasoning moderates the degree to which participants consider visual information: low performance on the Cognitive Reflection Test is associated with an over-reliance on what is said.

Introduction

Recent advances in technology for algorithmically applying hyper-realistic manipulations to video are simultaneously enabling new forms of interpersonal communication and posing a threat to traditional standards of evidence and trust in media\textsuperscript{1–7}. In the last few years, computer scientists have trained machine learning models to generate photorealistic images of people who do not exist\textsuperscript{8–16}, inpaint people out of images\textsuperscript{11,12}, clone voices based on a few samples\textsuperscript{13,14}, modulate the lip movements of people in videos to make them appear to say something they have not said\textsuperscript{15,18}, and create fake videos based on simple text prompts\textsuperscript{17}. The synthetic videos’ false appearance of indexicality – the presence of a direct relationship between the photographed scene and reality\textsuperscript{18,19} – has the potential to lead people to believe video-based messages that they otherwise would not have believed if the messages were communicated via text. This potential influence is particularly concerning because research demonstrates that videos, especially videos of an injustice, elicit more engagement and emotional reactions (e.g., anger, sympathy) than text descriptions displaying the same information\textsuperscript{20–22} (although, see ref.\textsuperscript{23}). Moreover, visual misinformation is common on social media\textsuperscript{24} and the emotional and motivational influences of visual communication have been attributed to why fake, viral videos have provoked mob-violence\textsuperscript{25,26}. While people are more likely to believe a real event occurred after watching a video of the event than reading a description of the event\textsuperscript{27}, an open question remains: Does visual communication relative to text increase the believability of fabricated events?

The realism heuristic\textsuperscript{26,28} predicts “people are more likely to trust audiovisual modality [relative to text] because its content has a higher resemblance to the real world.” This prediction is relevant for many deepfake videos\textsuperscript{29} and suggests fabricated video would be more believable than fabricated text conditional on the absence of obvious perceptual distortions. Yet there exists little direct empirical evidence for this heuristic applied to algorithmically manipulated video. In an experiment using three fake videos as stimuli, researchers found that stories presented as videos are perceived as more credible than stories presented as text or read aloud in audio form\textsuperscript{26}. In contrast, in an experiment showing 6 political deepfake videos (videos manipulated by artificial intelligence to make someone say something they did not say) and 9 non-manipulated videos, researchers did not find differences between truth discernment rates in video, audio, and text\textsuperscript{30}. Perhaps some of the experiments’ participants did not take the videos’ “indexicality” as evidence of authenticity because participants were aware of how easily such videos could be manipulated. Alternatively, some participants may have noticed perceptual distortions in the videos, which would naturally lead one to believe the video has been manipulated. The mixed evidence on how communication modalities mediate people’s ability to discern fabricated content may be due to the small samples of stimuli in media effects research\textsuperscript{31}. In related work on how fake images can be persuasive and difficult to distinguish from real images: research finds people rarely question the authenticity of images even when primed\textsuperscript{12}, images can increase the credibility of disinformation\textsuperscript{33}, and images of synthetic faces produced by StyleGAN2\textsuperscript{9} are indistinguishable from the original photos on which the StyleGAN2 algorithm was trained\textsuperscript{34}. Moreover, research shows that non-probative and uninformative photos can lead people to believe false claims\textsuperscript{35}, lead people to believe they know more than they actually know\textsuperscript{36}, promote “truthiness” by creating illusory truth effects\textsuperscript{37,38}, which can
lead people to believe falsehoods they previously knew to be falsehoods. When it comes to ostensibly probative videos of political speeches, the question whether people are more likely to believe an event occurred because they saw it as opposed to only read about it remains open.

In fact, today’s algorithmically generated deepfakes are not yet consistently indistinguishable from real videos. On a sample of 166 videos from the largest publicly available dataset of deepfake videos to date, people are significantly better than chance but far from perfect at discerning whether an unknown actor’s face has been visually manipulated by a deepfake algorithm. This finding is significant because it demonstrates that people can identify deepfake videos from real videos based solely on visual cues. However, some videos are more difficult than others to distinguish due to blurry, dark, or grainy visual features. On a subset of 11 of the 166 videos, researchers do not find that people can detect deepfakes better than chance. In another experiment with 25 deepfake videos and 4 real videos but only 94 participants, researchers found that the overall discernment accuracy is 51% and a media literacy training increases discernment accuracy by 24 percentage points for participants assigned to the training relative to the control group. In experiments examining how people react to deepfake videos of politicians, researchers find people are more likely to feel uncertain than misled after viewing a deepfake of Barack Obama and people consider a deepfake of a Dutch politician significantly less credible than the real video from which it was adapted and the deepfake video is not more persuasive than the text alone. In the experiment examining the fabricated video of a Dutch politician, some respondents explained their credibility judgements by indicating audio-visual cues of how the message was communicated (e.g., unnatural mouth movements); others indicated inconsistency in the content of the message itself (e.g., contextually unrealistic speeches).

People’s capacity to identify multimedia manipulations raises questions: how do various kinds of fabricated media (e.g., audio and video of fake political speeches) alter the perceived credibility of misinformation, how do audience characteristics (e.g., reflective reasoning) moderate media effects, and how does the source and content of a message interact with the fabricated sources people engage with and share on social media?

In this paper, we evaluate discernment across 32 political speeches by two well-known politicians. We present these speeches to participants via the 7 possible permutations of 3 digital media communication modalities: text, audio, and video. Based on 46,713 responses from 3,317 individuals who participated in a pre-registered (and an additional 387,274 responses from 38,510 participants who participated after the pre-registration window) cross-randomized experiment, we examine ordinary people’s performance at discerning political speeches randomized to appear in each of the following seven conditions: a transcript, an audio clip, a silent video, audio with subtitles, silent video with subtitles, video with audio, and video with audio and subtitles. By randomly assigning political speeches to these permutations of text, audio, and video modalities and asking participants to discern truth from falsehood, this experiment is designed to disentangle the degree to which participants attend to and consider the content of what is said and the audio-visual cues as to how it is said. In addition, we evaluate these disentangled components across message types (speeches that are either concordant or discordant with the general public’s perception of a speaker’s political identity) and audience characteristics (reflective reasoning as measured by the Cognitive Reflection Test (CRT), which is a robust test for measuring an individual’s tendency towards reflecting on questions before answering that helps explain why people fall for fakes news and is strongly associated with the reliability of news sources people engage with and share on social media.

\(^1\)The pre-registered analysis is available at [https://aspredicted.org/VFZ_6HE](https://aspredicted.org/VFZ_6HE). We continued collecting responses from participants who found our experiment organically after the pre-registered cut-off date for data collection passed, and our final sample includes 432,987 responses from 41,822 participants who passed the attention check. Our results are robust to include or exclusion of participants who participate after the pre-registered cut-off date.
Figure 1. Mean identification accuracy across the 32 silent videos (with no subtitles) to illustrate the heterogeneity in how difficult the visual deepfake manipulations are to detect. There are 8 fabricated videos and 10 non-fabricated videos out of the 32 on which participants identify less than 67% of the time. There are 2 fabricated videos accurately identified in more than 80% of observations. The 95% confidence interval range is less than 1% for all silent videos.

Results

Participants

A total of 73,236 individuals participated in the experiment. We used the Prolific platform to recruit 554 individuals from the United States who completed 16,699 trials. In addition to the recruited participants, 5,106 individuals (76% of whom visited from outside the United States) participated in the experiment during the pre-registration window from March 4, 2021 to June 1, 2021. These participants found the website organically and completed 44,461 trials. Between June 1, 2021 and July 1, 2022, an additional 67,576 individuals (70% of whom visited from outside the United States) completed 566,343 trials. We focus our analysis on 41,822 participants: the 509 of 554 recruited participants and 41,313 of 72,682 non-recruited participants who passed the attention check where we presented an obvious deepfake and explicitly instructed participants to respond that the video is a deepfake with 100% confidence.

The sample of 509 recruited participants is balanced across political identities (in this experiment failure on the attention check does not correlate with political identities); 257 recruited participants self-report as Democrats, and the other 252 recruited participants self-report as Republicans. We do not find demographic differences in recruited participants who passed the attention check. We did not collect data for recruited participants who failed the attention check, but we did collect data for the non-recruited participants who failed the attention check. In the Supplementary Information, we demonstrate that the main results are robust to including participants who failed the attention check and robust to including or excluding participants who participated after June 1, 2021.

Many but not all participants responded to all 32 speeches; 482 recruited participants and 6,374 non-recruited participants viewed all 32 speeches. Before the experiment began, participants in the recruited cohort (but not the non-recruited cohort) responded to a baseline survey that included questions on political preferences, trust in media and politics, and the three questions from the CRT.

Discernment Performance across Communication Modalities

We begin by examining how frequently participants correctly identified the stimuli as fabricated or not. Across all 224 stimuli, recruited and non-recruited participants correctly identified the stimuli in 75% and 70% of observations, respectively.

We find the fabricated political speech transcripts and visual deepfake manipulations are difficult for participants to discern. The proportion of people who accurately identified fabrications from authentic text varied by stimuli. Across the 32 text transcripts, the least accurately identified transcript is identified correctly in 26% of trials, the most accurately identified one is identified correctly in 67% of trials, and the median accurately identified one is identified correctly in 43% of trials. Similarly, the range for accurate identification across the 32 silent videos (silent videos refers to only silent videos and not silent videos with subtitles) is 37% to 86% with a median of 66%. There are 8 out of 16 fabricated silent videos and 10 out of 16 non-fabricated silent videos that participants accurately identify less than than 67% of the time. Figure 1 illustrates the proportion of participants who accurately distinguish between authentic and fabricated for the 32 silent videos.

In contrast, we find audio clips are easier to discern than text transcripts or silent videos. On the audio clips with no subtitles, the accurate identification ranges from 56% to 86% with a median of 79%.
Figure 2 presents participants’ weighted accuracy, confidence, perceived fabrications in fabricated speeches, perceived fabrications in non-fabricated speeches, and response duration across modality conditions. Weighted accuracy indicates participants’ accuracy weighted by confidence (e.g., if a participant responded “82% confidence this is fabricated” and the participant is correct, then the participant is assigned a weighted accuracy score of 82, and otherwise, if the participant is incorrect, then the participant would be assigned a weighted accuracy score of 18). Confidence indicates participants’ self-reported level of confidence which ranges from 50 (just as likely as chance) to 100 (full certainty). Perceived fabrications in fabricated and non-fabricated speeches is defined as a participant indicating a 51% or higher confidence that a stimulus is fabricated. Response time is measured in seconds and winsorized at the 99th percentile to control for time response outliers, which are an artifact of participants who return to the experiment after an extended time.

Figure 2. The mean and distribution of (a) weighted accuracy, (b) confidence, (c) perceived fabrications in fabricated speeches, (d) perceived fabrications in real speeches, and (e) response time are plotted for each of the seven modality conditions. The black lines indicate the 95% confidence interval of the true mean and the gray dots indicate each of the 32 speeches. Figure 2b plots confidence on a scale that ranges from a minimum of 50% confidence (just as likely as chance) to 100% confidence (full confidence). Figure 2e plots response time winsorized at the 99th percentile to control for time response outliers, which are an artifact of participants who return to the experiment after an extended time.

We evaluate the marginal effect of each condition on participants’ weighted accuracy (and additional outcomes) via an ordinary least squares regression with standard robust errors clustered at the participant level following Abadie et al (2017)². We find both recruited and non-recruited participants’ accuracy increase as political speeches are presented with video and audio modalities. In this regression, which is also presented in column 1 of Table 1 in the Appendix, the dependent variable is weighted accuracy and the independent variables are binary indicators for assignment to communication modalities. Recruited participants’ accuracy is 58% ($p < 0.001$) on transcripts, 7% ($p < 0.001$) higher on silent videos, 9% ($p < 0.001$) higher on silent videos with subtitles, 19% ($p < 0.001$) higher on audio clips and audio clips with subtitles, and 25% ($p < 0.001$) higher on videos with audio and videos with audio and subtitles.²

²All p-values reported in this paper are generated by linear regression with robust standard errors clustered at the participant level unless otherwise noted.
higher on audio clips and audio clips with subtitles, and 28-29% \( (p < 0.001) \) higher on videos with audio and videos with audio and subtitles. Overall, participants are better at identifying whether an event actually happened when watching videos or listening to audio than reading transcripts.

In contrast to the high variability in participants’ accuracy across speeches and modality conditions, participants’ confidence is less variable. On text transcripts, participants’ mean confidence is 79%. Speeches presented via video and audio increase participants’ confidence relative to text transcripts by 9% and 11% \( (p < 0.001) \) independently, respectively, and 15% together \( (p < 0.001) \). We find small effects of learning over time; for every stimulus seen, participants’ accuracy increases by 0.27% \( (p = 0.001) \) and participants’ confidence increases by 0.03% \( (p = 0.006) \), which means that on average accuracy increased by 8.64% and confidence increase by 0.96% from the first stimulus seen to the last one seen.

As participants have access to additional communication modalities, participants’ weighted accuracy, confidence, discernment of fabricated speeches, and discernment of real speeches increase on average. However, we do not find any significant, marginal effects of subtitles on any of the dependent variables for modality conditions that already include audio. The median response time across all stimuli was 27 seconds, which is 6 seconds longer than the average video length. The median response time for the silent, subtitled videos is 34 seconds, which is slightly longer than the response time for all other modality conditions. Across all 7 modality conditions, the median response time for fabricated stimuli is shorter than the median response time for non-fabricated stimuli; fabricated text, video, and audio have 3.8 seconds \( (p < 0.001) \), 5.6 seconds \( (p < 0.001) \), and 3.5 seconds \( (p < 0.001) \) shorter response times than their non-fabricated counterparts.

Based on this experiment’s large sample size of 432,987 observations by participants who passed the attention check, the 224 stimuli in this experiment have a mean of 1,932 observations each. This large sample size per stimulus provides high statistical power to individually evaluate whether participants are discerning stimuli more accurately than chance. Specifically, using 1,933 observations provides over 99% statistical power to detect a 15 percentage point increase beyond chance at the \( p < 0.05 \) threshold. We evaluate the degree to which participants’ discernment surpasses random chance by running a binomial test on responses to each stimulus within a modality condition and applying a Bonferroni correction\(^2\), which means multiplying each \( p \)-value by 32 (the number of speeches per modality condition) to correct for multiple hypothesis testing.

After applying this correction for multiple hypothesis testing, we find participants’ discernment is statistically significantly better than chance \( (p < 0.05) \) on 5 of 32 text transcripts and 26 of the 32 silent videos. In particular, participants are no better than chance \( (p < 0.05) \) on 4 of the 16 non-fabricated, silent videos and 2 of the 16 fabricated, silent videos. In other words, we have high statistical power, and we do not find evidence that participants are better than chance on 6 of the silent videos and 27 of the 32 text transcripts.

When the information from the political speech transcript and video are combined in the silent, subtitled videos, we find participants discern better than chance \( (p < 0.05) \) on all 16 of 16 fabricated, silent videos with subtitles and 9 of 16 non-fabricated, silent videos with subtitles. Likewise, the addition of audio significantly increases discernment rates; in all modality conditions with audio, participants discern better than chance \( (p < 0.05) \) on between 31 to 32 of the 32 political speeches.

Figure 2c and Figure 2d show the distributions of discernment rates across modality conditions for fabricated and real videos. Similarly to Figure 2a and Figure 2b, these plots show that regardless of whether the stimuli are fabricated or not, the addition of audio or video is associated with an increase in participants’ discernment. However, we find slight differences in response bias: participants tend to identify text transcripts as real and the rest of the modalities as fabricated more often than random chance would suggest. For participants who did not select “Just as likely real or fabricated,” participants respond that text transcripts and silent videos are fabricated in 44% \( (p < 0.001) \) and 53% \( (p = 0.002) \) of trials, respectively, while participants respond that the other 5 modality conditions are fabricated in 55% to 57% of trials \( (p < 0.001) \) (see Figure 9 for the percent of participants guessing a video is fabricated over the number of speeches a participant has seen). Participants selected “Just as likely real or fabricated” in 21% of text transcripts, 7% of silent videos, 6% of silent videos with subtitles, 6% of audio, 5% of audio with subtitles, and 3% of video and audio with or without subtitles.

In Figure 3, we present participants’ marginal accuracy on transcripts, silent videos, and video with audio relative to silent, subtitled videos for each of the 32 speeches. Figure 3a reveals that participants are mostly less accurate on text transcripts than silent, subtitled videos. Likewise, Figure 3c shows participants are consistently more accurate on videos with audio than silent, subtitled videos. In contrast, Figure 3b illustrates heterogeneity in participants’ performance with and without subtitles. In the following section, we examine this heterogeneity along two dimensions: whether the video is fabricated or not and whether the speech content is considered discordant with the politician’s identity or not.

**Heterogeneous Moderating Effects of Discordant Messages**

We evaluate how discordant messages influence participants’ discernment by examining the interactions between discordance and modality conditions in the linear regressions on participants’ weighted accuracy presented in Table 2 and Table 3 in the Appendix. We limit this analysis to recruited participants for two reasons: first, recruited participants are all from the United
We find that participants' performance on the CRT moderates participants' discernment accuracy. In this analysis, the CRT we find participants are 5.0 percentage points (\(p = 0.002\)) less accurate on the discordant silent, subtitled videos than the same silent videos without subtitles. When considering all 32 fabricated and real speeches together (see column 1 of Table 2 in the Appendix), we find participants are 4.7 percentage points (\(p = 0.002\)) more accurate on silent, subtitled videos than the same videos without subtitles. However, we find participants are 5.0 percentage points (\(p = 0.018\)) less accurate on the discordant silent, subtitled videos than the same silent videos without subtitles. In other words, the addition of subtitles reduces discernment accuracy for political speeches that are discordant with the general public’s perception of what politicians would say.

In order to further evaluate this effect, we consider fabricated videos and non-fabricated videos separately in columns 2 and 3 in Table 2 in the Appendix. We find the negative effect of discordance on subtitled videos is driven by participants’ discernment of non-fabricated videos. We find participants are 6.8 percentage points (\(p = 0.021\)) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. In contrast, we do not find a statistically significant difference (\(p = 0.341\)) between participants’ performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles. The negative effects of subtitles on non-fabricated yet discordant silent videos indicates the content of a message can change how participants weigh visual information.

The heterogeneous effects of subtitles on the discernment of silent videos is robust to our specification of discordance. In Table 3 in the Appendix, we consider the same regressions as Table 2 in the Appendix except we replace the binary variable indicating discordance with a continuous variable for how discordant the speech is with the speaker based on the independent survey with 84 participants on how well the political speeches match either politicians’ political views. The regressions in columns 2 and 3 of Table 3 in the Appendix present qualitatively similar results as Table 2 in the Appendix. When we consider discordance based on the public’s perceived discordance, we find participants are 4.2 percentage points (\(p = 0.003\)) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. Likewise, we do not find a statistically significant difference (\(p = .751\)) between participants’ performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles.

Heterogeneous Moderating Effects of the Cognitive Reflection Test (CRT)

We find that participants’ performance on the CRT moderates participants’ discernment accuracy. In this analysis, the CRT score is a continuous variable ranging from 0 to 3 with 124 participants answering none correctly and 109, 122, and 154 participants answering 1, 2, and 3 questions correctly, respectively. For every question that participants answer correctly, the absolute value of the difference in accuracy between the silent, subtitled video and the modality condition to which it is being compared. The legend indicates whether the video shows the politician expressing political views concordant or discordant with his expected political ideology and “F” and “NF” refer to fabricated and not fabricated, respectively.

States while the majority of non-recruited participants visited the website from outside the United States and it is unclear how familiar non-recruited participants are with United States politicians’ viewpoints; second, we also evaluate these effects with respect to CRT performance, which we only collected for recruited participants.

When considering all 32 fabricated and real speeches together (see column 1 of Table 2 in the Appendix), we find participants are 4.7 percentage points (\(p = 0.002\)) more accurate on silent, subtitled videos than the same videos without subtitles. However, we find participants are 5.0 percentage points (\(p = 0.018\)) less accurate on the discordant silent, subtitled videos than the same silent videos without subtitles. In other words, the addition of subtitles reduces discernment accuracy for political speeches that are discordant with the general public’s perception of what politicians would say.

In order to further evaluate this effect, we consider fabricated videos and non-fabricated videos separately in columns 2 and 3 in Table 2 in the Appendix. We find the negative effect of discordance on subtitled videos is driven by participants’ discernment of non-fabricated videos. We find participants are 6.8 percentage points (\(p = 0.021\)) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. In contrast, we do not find a statistically significant difference (\(p = 0.341\)) between participants’ performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles. The negative effects of subtitles on non-fabricated yet discordant silent videos indicates the content of a message can change how participants weigh visual information.

The heterogeneous effects of subtitles on the discernment of silent videos is robust to our specification of discordance. In Table 3 in the Appendix, we consider the same regressions as Table 2 in the Appendix except we replace the binary variable indicating discordance with a continuous variable for how discordant the speech is with the speaker based on the independent survey with 84 participants on how well the political speeches match either politicians’ political views. The regressions in columns 2 and 3 of Table 3 in the Appendix present qualitatively similar results as Table 2 in the Appendix. When we consider discordance based on the public’s perceived discordance, we find participants are 4.2 percentage points (\(p = 0.003\)) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. Likewise, we do not find a statistically significant difference (\(p = .751\)) between participants’ performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles.

Heterogeneous Moderating Effects of the Cognitive Reflection Test (CRT)

We find that participants’ performance on the CRT moderates participants’ discernment accuracy. In this analysis, the CRT score is a continuous variable ranging from 0 to 3 with 124 participants answering none correctly and 109, 122, and 154 participants answering 1, 2, and 3 questions correctly, respectively. For every question that participants answer correctly, the absolute value of the difference in accuracy between the silent, subtitled video and the modality condition to which it is being compared. The legend indicates whether the video shows the politician expressing political views concordant or discordant with his expected political ideology and “F” and “NF” refer to fabricated and not fabricated, respectively.

Heterogeneous Moderating Effects of the Cognitive Reflection Test (CRT)

We find that participants’ performance on the CRT moderates participants’ discernment accuracy. In this analysis, the CRT score is a continuous variable ranging from 0 to 3 with 124 participants answering none correctly and 109, 122, and 154 participants answering 1, 2, and 3 questions correctly, respectively. For every question that participants answer correctly, the absolute value of the difference in accuracy between the silent, subtitled video and the modality condition to which it is being compared. The legend indicates whether the video shows the politician expressing political views concordant or discordant with his expected political ideology and “F” and “NF” refer to fabricated and not fabricated, respectively.

Heterogeneous Moderating Effects of the Cognitive Reflection Test (CRT)

We find that participants’ performance on the CRT moderates participants’ discernment accuracy. In this analysis, the CRT score is a continuous variable ranging from 0 to 3 with 124 participants answering none correctly and 109, 122, and 154 participants answering 1, 2, and 3 questions correctly, respectively. For every question that participants answer correctly, the absolute value of the difference in accuracy between the silent, subtitled video and the modality condition to which it is being compared. The legend indicates whether the video shows the politician expressing political views concordant or discordant with his expected political ideology and “F” and “NF” refer to fabricated and not fabricated, respectively.
(p < 0.001), which means that participants are that much less accurate on non-fabricated discordant silent, subtitled videos than the same silent videos without subtitles while holding all else constant. The interaction between “CRT Score,” “Discordant,” and “Silent Subtitled Video” is 6.3 percentage points (p = 0.011), which means for each correct response to the CRT, participants are 6.3 percentage points more accurate at identifying discordant silent, subtitled videos while holding all else constant. This means that participants who answered all 3 CRT items correctly would be 18.9 percentage points (p = 0.011) more accurate on discordant silent, subtitled videos than participants who failed to answer any CRT item correctly. This improvement by 18.9 percentage points for answering all CRT items correctly cancels out the 17.5 percentage point decrease associated with discordant silent, subtitled videos compared to the same silent videos without subtitles. In other words, perfect performance on the CRT moderates the negative effects of discordant content such that participants are considering visual information and discerning just as accurately on silent subtitled videos as the same silent videos without subtitles. These results are qualitatively similar when we replace the binary variable for discordance with the continuous variable for discordance in Table 3.

Figure 4. Average treatment effect of assignment to modality conditions and their interaction with discordant speeches and participants’ performance on the Cognitive Reflection Test. The error bars represent 95% confidence intervals. The regressions on which this plot is based are presented in columns 5 and 6 of Table 2 in the Appendix.

Discussion
This work provides evidence, via a randomized experiment with 224 authentic and fabricated stimuli and 41,822 participants, that visual and auditory communication modalities increase people’s ability to distinguish authentic political speeches from fabricated political speeches. In particular, we provide corroborating evidence to the conventional wisdom around the “seeing is believing” narrative (the realism heuristic that suggests people will tend to trust video over text and recent results showing people “are more likely to believe an event occurred when it is presented in video versus textual form”) in the context of authentic speeches: people are significantly more accurate at identifying authentic speeches as authentic when the speeches include audio and visual modalities as opposed to only text. However, these results add considerable nuance to the seeing is believing narrative when considering fabricated speeches: people are significantly more accurate at identifying fabricated speeches as fabricated when the speeches include audio and visual modalities as opposed to only text. In other words, we find participants are significantly more accurate at distinguishing between authentic and fabricated political videos than transcripts.

These results are based on an experiment with a stimuli set that is much larger than most stimuli sets for the psychology of media effects research, and deepfake detection, but it is important to add a caveat that we focused on a single context, political speeches, and algorithm, the deepfake lip-syncing wav2lip algorithm, which is very effective at manipulating a person who is facing forward and already speaking into a convincing fake video. While we present evidence that adds considerable nuance to the media effects literature on communication modalities, future work may consider additional nuances by exploring heterogeneity based on other kinds of deepfake manipulations like face swapping and head puppetry, contexts that require more sophistication to produce a convincing fake (e.g. where a person is moving, turning their head, and interacting with other people), and who is being manipulated.
These results cannot simply be explained by the deepfake manipulations being too obvious or unrealistic. On silent videos without subtitles, we find participants are only 64% accurate at identifying manipulations (see Figure 2c and Figure 2d for the distribution of people guessing stimuli are fabricated across the seven modalities). Moreover, we find participants do not perform better than chance in nearly half of the silent videos. Participants’ performance on the silent videos is relevant to the quality of the deepfake manipulations because it avoids the confounding from the speech content and the audio. The participants’ low performance on silent videos offers evidence that visual artefacts and inconsistencies created by the lip syncing deepfake manipulations are not readily apparent to most people, and as such, these videos represent a reasonable stimuli set for examining how well people can distinguish real from fake videos.

People distinguish authentic from fabricated videos based on perceptual cues from video and audio and considerations about the content (e.g., the degree to which what is said matches participants’ expectations of what the speaker would say, which is known as the expectancy violation heuristic). With the message content alone, participants are only slightly better than random guessing at 57% accuracy on average. With perceptual information from video and the message content via subtitles, participants are slightly more accurate (and more confident) at 66% accuracy on average, and with information from both video and audio, participants are even more accurate (and more confident) at 82% accuracy on average. Our finding that participants are more accurate at distinguishing between real and fabricated on audio than silent video with subtitles aligns with the social psychology literature demonstrating people tend to rely on auditory information more than visual information for both discerning sincerity and ascribing authorship of a script to a human (as opposed to a computer). Overall, the experiment’s results show that as participants have access to more information via audio and video, they are better able to distinguish whether a political speech has been fabricated.

However, we find one notable exception to the result that more information leads to higher accuracy in distinguishing fabricated speeches from authentic ones: political speeches that conflict with the public’s perspective of what a politician would say are harder to discern in silent, subtitled videos than the same silent videos without subtitles. This effect on discordant speeches (but not concordant speeches) is not driven by subtitles distracting participants. We do not find any evidence of any effect on subtitles when audio is included. Instead, the heterogeneous effects of concordant and discordant speech content are a consequence of how participants handle cognitive dissonance and balance the consideration of perceptual and content-based information. We find that these effects are driven by responses to non-fabricated videos and are moderated by deliberative, reflective thinking as measured by the CRT.

Fabricated videos differ from non-fabricated videos in how people can discern their authenticity. Fabricated videos involve visual manipulations, which can sometimes be explicitly identified (e.g., a glitch, a flicker, or mechanical and otherwise out of place lip movement). If someone finds a suspicious visual artefact, then that individual can be quite confident the video has been fabricated. In contrast, non-fabricated videos have not been visually manipulated. As a result, there is no single bit of information to signify fabrication or authenticity. Furthermore, we find people take on average 2.5s to 3.8s longer to provide a response to non-fabricated speeches than fabricated speeches. If someone cannot find a visual distortion, then that individual cannot be perfectly certain that the video has or has not been fabricated; for example, the video may have been fabricated without any perceptible distortion, or perhaps, the individual has yet to find the subtle visual distortion. This asymmetry between assessing fabricated and non-fabricated speeches exacerbates the “liar’s dividend” where the general possibility that speeches can be fabricated calls into question whether any speech is fabricated and thus enables “liars to avoid accountability for things that are in fact true.”

Clear articulation of the precise state-of-the-art algorithms and associated contexts in which audio-visual content can be fabricated to be indistinguishable from the real thing can help inform how people assess the content they consume and reduce the effects of the “liar’s dividend.”

We find that participants’ performance on the CRT moderates the effects of subtitles on the discernment accuracy of silent videos. In particular, participants who correctly answered all three CRT items show no difference in discernment rates of discordant silent, subtitled videos relative to the same silent videos without subtitles. But, for every CRT item that participants incorrectly answer, participants are 6.3 percentage points less accurate on real discordant silent, subtitled videos than the same silent videos without subtitles. In other words, reflective thinking moderates how participants balance what is said (the content of the speech) with how it is said (visual information). Our results show that the least reflective participants tend to rely on the expectancy violation heuristic and discount visual information more than the most reflective participants.

Unlike for videos and transcripts, we cannot disentangle the content and perceptual information for audio modalities. Nevertheless, we find that the interaction between discordant speeches and any audio condition is negative after controlling for the level effects of discordance and any audio. This suggests that discordant media not only impair the incorporation of visual cues but may also impair attention to and incorporation of auditory cues as well.

The danger of fabricated videos may not be the average algorithmically produced deepfake but rather a single, highly polished, and extremely convincing video. For example, hyper-realistic deepfakes like the Tom Cruise deepfakes on Tiktok (see https://www.tiktok.com/@deeptomcruise) are produced by visual effects artists using both artificial intelligence algorithms and video editing software. While these hyper-realistic deepfakes may still contain manipulation artifacts (e.g.,
unattached earlobes that do not match Tom Cruise’s attached earlobes), future work on the psychology of multimedia misinformation may consider hyper-realistic videos produced by visual effects studios in addition to algorithmically manipulated videos.

Political deepfakes are most dangerous when people are least expecting information to be manipulated, and this experiment on multimedia truth discernment does not match the ecological realities that people typically face when confronted with fake news. In this experiment, 50% of content is fake, and we explicitly inform participants of this base rate. In today’s media ecosystem, fake news is relatively rare: less than a fraction of a percent of news is fake news. As such, this experiment is useful to study how people discern multimedia information when attending to questions of accuracy, but it is less useful in understanding how people will share misinformation they read on social media. People are generally highly accurate in discerning the veracity of news headlines yet share fake news headlines because their attention is not focused on accuracy. On social media, video-based misinformation will often be designed to incorporate characteristics (e.g., fear, disgust, surprise, novelty) that divert people’s focus from accuracy and make content go viral. Given that multimedia misinformation may be both easier to discern and more frequently shared on social media than text-based media, more research needs to be done to understand how people allocate attention while browsing the Internet. Recent research shows that educational material on common misinformation techniques can improve people’s ability to discern trustworthy from untrustworthy videos. Finally, discernment – how accurately people discern misinformation – is different than belief – how much people report they believe misinformation. It is possible (though quite peculiar) that someone could be highly accurate at discerning truth from falsehood while also tending to believe the fabricated content and not believe the true content. For example, research on fake news headlines and articles finds that people are better at discerning news concordant with their political leanings than discordant news while also believing discordant news more often than discordant news.

The finding that videos of political speeches are easier to distinguish as authentic or fabricated than text transcripts highlights the need to re-introduce and explain the oft-forgotten second half of the “seeing is believing” adage. In 1732, the old English adage appears as: “Seeing is believing but feeling is the truth.” Here, “feeling” does not refer to emotion but rather direct experience. Since the advent of photography, society has generally understood that what we see in a photograph is not always the truth and further assessment is often necessary. In this experiment, we asked participants to identify fabricated and non-fabricated snippets we show are real and half are fabricated.” Figure 5 in the Supplementary Information section presents a screenshot of the user interface, which shows participants were instructed to move a slider to report their confidence from 50% to 100% that a stimulus is fabricated (or 50% to 100% that a stimulus is not fabricated). After each response, we informed participants of the user interface, which shows participants were instructed to move a slider to report their confidence from 50% to 100% that a stimulus is fabricated (or 50% to 100% that a stimulus is not fabricated). Finally, these findings offer insights into political communication and communication theory more generally; there is more to how humans form beliefs than the “seeing is believing” narrative would suggest because people are paying attention to both what is said and how something is said.

Methods

Virtual Experiment Website

We hosted multimedia stimuli – transcripts, audio, and video of fabricated and authentic political speeches – on a custom designed website called Detect Fakes. In the experiment, we asked participants to identify fabricated and non-fabricated stimuli. After collecting informed consent and presenting participants with instructions, we show participants a short political speech and ask “Did [Joseph Biden/Donald Trump] say that?” followed by “Please [read/listen/watch] this [transcript/audio clip/video] from [Joseph Biden/Donald Trump] and share how confident you are that it is fabricated. Remember half the media snippets we show are real and half are fabricated.” Figure 5 in the Supplementary Information section presents a screenshot of the user interface, which shows participants were instructed to move a slider to report their confidence from 50% to 100% that a stimulus is fabricated (or 50% to 100% that a stimulus is not fabricated). After each response, we informed participants whether the stimulus was actually fabricated and then presented participants with another stimulus selected at random until participants viewed all 32 stimuli or decided to leave the experiment. Each participant began the experiment with an attention check stimulus.

Detect Fakes is currently hosted at https://detectfakes.media.mit.edu/.
Multimedia Stimuli
The multimedia stimuli are drawn from the Presidential Deepfake Dataset (PDD)\textsuperscript{93}, which is made up of 32 videos showing two United States presidents making political speeches. Half the videos are authentic videos that have not been altered by a deepfake algorithm. The other half have been fabricated to make the politicians appear to say something that they have not said. The fabricated videos were produced by writing a fabricated script, recording professional voice-actors reading the script, and applying a deepfake lip-syncing algorithm\textsuperscript{15} to real videos of Joseph Biden and Donald Trump to make it appear as if the politicians actually gave such a fabricated speech. The mean duration of the videos is 21 seconds and all videos are recorded at 30 frames per second and have a resolution of 854 by 480 pixels. The PDD is balanced across three dimensions: (1) videos that have and have not been fabricated, (2) videos of Joseph Biden and Donald Trump, and (3) videos of the two politicians making concordant and discordant speeches with what the general public believes are the politicians’ political views.

In this experiment, we transform each of the original videos from the PDD into 7 different forms of media: a transcript, an audio clip, a silent video, audio with subtitles, silent video with subtitles, video with audio, and video with audio and subtitles. As a result, there are 7 modality conditions, 32 unique speeches, and 224 unique stimuli. On the experiment website, the transcript appears as HTML text and the six other forms of media content appear in a video player. The audio clip shows a black screen in the video player and the audio clip with subtitles shows a black screen with subtitles at the bottom. Each participant encounters each political speech in only one modality.

Concordance and Discordance Validation
In order to validate the concordance and discordance of speeches, we conducted an independent survey where 84 participants who passed an attention check rated each of the 32 transcripts for how well the political speeches match either politicians’ political views. Participants were instructed “For each statement, we want you to rank how closely the statement matches your understanding of President Joseph Biden or President Donald Trump’s political views” and asked to provide a judgment on a 5-point Likert scale from “Strongly Disagree” (-2) to “Strongly Agree” (2) that “This statement matches President [Joseph Biden’s/Donald Trump’s] political viewpoint: [statement].” Participants’ responses confirm that speeches designed to be concordant and discordant with the two politicians views were indeed concordant and discordant with the average participants’ perception of the politicians’ views. The Z-values of participants responses to concordant and discordant speeches are -0.25 and 0.21, respectively, and this difference is statistically significant with $p < 0.001$ based on a T-Test. In Table 2, the “Discordant (Binary)” variable refers to the categories as outlined in the PDD, and in Table 3 the “Discordant (Continuous)” variable refers to these Z-values.

Randomization
We randomly assigned the order in which the 32 unique political speeches are presented to participants and each political speech is randomly assigned to one of the seven conditions. By randomly assigning the order of political speeches and the modality condition in which speeches were presented, we can identify the causal impact of media modality on participants’ ability to discern misinformation.

Consent and Ethics
This research complies with all relevant ethical regulations and the Massachusetts Institute of Technology’s Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. This study’s exemption identification number is E-3105. All participants are informed that “Detect Fakes is an MIT research project. All guesses will be collected for research purposes. All data for research were collected anonymously. For questions, please contact detectfakes@mit.edu. If you are under 18 years old, you need consent from your parents to use Detect Fakes.” Most participants arrived at the website via organic links on the Internet. For participants recruited from Prolific, we compensated participants at a rate of $9.78 an hour and provided bonus payments of $5 to the top 1% of participants. Before beginning the experiment, all participants from Prolific were also provided a research statement, “The findings of this study are being used to shape science. It is very important that you honestly follow the instructions requested of you on this task, which should take a total of 15 minutes. Check the box below based on your promise:” with two options, “I promise to do the tasks with honesty and integrity, trying to do them uninterrupted with focus for the next 15 minutes.” or “I cannot promise this at this time.” Participants who responded that they could not do this at this time were re-directed to the end of the experiment.

Data and Code Availability
The datasets and code generated and analyzed during the current study are available in our public Github repository, https://github.com/mattgroh/fabricated-political-speeches (the Github repository will be set to public upon peer-reviewed publication). All PDD videos are available on Youtube with links provided in the Presidential Deepfakes Dataset paper\textsuperscript{93}.
Acknowledgements
The authors would like to acknowledge funding from MIT Media Lab member companies, thank JL Cauvin and Austin Nasso (@jlcauvin and @austinnasso on Tiktok) for providing voice impressions, and thank the following individuals for helpful feedback: David Rand, Gordon Pennycook, Rahul Bhui, Yunhao (Jerry) Zhang, Ziv Epstein, and members of the Affective Computing lab at the MIT Media Lab and the Human Cooperation lab at MIT Sloan School of Management.

Author Contributions
M.G. conceived the experiments, A.S. developed the synthetic media and conducted the experiments, A.S. and M.G. analyzed the results, M.G. wrote the manuscript, and A.S., A.L., M.G., and R.P. reviewed and edited the manuscript.

Competing Interests
The authors declare no competing interests.

References
1. Hancock, J. T. & Bailenson, J. N. The social impact of deepfakes. Cyberpsychology, Behav. Soc. Netw. 24, 149–152 (2021).
2. Chesney, B. & Citron, D. Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 107, 1753 (2019).
3. Paris, B. & Donovan, J. Deepfakes and cheap fakes. United States Am. Data & Soc. (2019).
4. Leibowicz, C., McGregor, S. & Ovadya, A. The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media. arXiv:2102.06109 [cs] (2021). Accessed 2021-07-07.
5. Agarwal, S. et al. Protecting world leaders against deep fakes. In CVPR workshops, vol. 1 (2019).
6. Pataranutaporn, P. et al. Ai-generated characters for supporting personalized learning and well-being. Nat. Mach. Intell. 3, 1013–1022 (2021).
7. Guess, A. M. & Lyons, B. A. Misinformation, disinformation, and online propaganda. Soc. media democracy: The state field, prospects for reform 10–33 (2020).
8. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410 (2019).
9. Karras, T. et al. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8110–8119 (2020).
10. Nichol, A. et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021).
11. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M. & Rahwan, I. Human detection of machine-manipulated media. Commun. ACM 64, 40–47 (2021).
12. Suvorov, R. et al. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2149–2159 (2022).
13. Arik, S. O., Chen, J., Peng, K., Ping, W. & Zhou, Y. Neural voice cloning with a few samples. arXiv preprint arXiv:1802.06006 (2018). Accessed 2021-07-07.
14. Luong, H.-T. & Yamagishi, J. Nautilus: a versatile voice cloning system. IEEE/ACM Transactions on Audio, Speech, Lang. Process. 28, 2967–2981 (2020).
15. Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P. & Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, MM ’20, 484–492 (Association for Computing Machinery, New York, NY , USA, 2020).
16. Lahiri, A., Kwatra, V., Frueh, C., Lewis, J. & Bregler, C. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2755–2764 (2021).
17. Hong, W., Ding, M., Zheng, W., Liu, X. & Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, DOI: 10.48550/ARXIV.2205.15868 (2022).
18. Peirce, C. S. *Peirce on signs: Writings on semiotic* (UNC Press Books, 1991).
19. Messaris, P. & Abraham, L. The role of images in framing news stories. In *Framing public life*, 231–242 (Routledge, 2001).
20. Glasford, D. E. Seeing is believing: communication modality, anger, and support for action on behalf of out-groups. *J. Appl. Soc. Psychol.* 43, 2223–2230 (2013).
21. Yadav, A. *et al.* If a picture is worth a thousand words is video worth a million? differences in affective and cognitive processing of video and text cases. *J. Comput. High. Educ.* 23, 15–37 (2011).
22. Appiah, O. Rich media, poor media: The impact of audio/video vs. text/picture testimonial ads on browsers’ evaluations of commercial web sites and online products. *J. Curr. Issues & Res. Advert.* 28, 73–86 (2006).
23. Powell, T. E., Boomgaard, H. G., De Swert, K. & de Vreese, C. H. Video killed the news article? comparing multimodal framing effects in news videos and articles. *J. broadcasting & electronic media* 62, 578–596 (2018).
24. Garimella, K. & Eckles, D. Images and misinformation in political groups: Evidence from whatsapp in india. *Harv. Kennedy Sch. Misinformation Rev.* (2020).
25. Goel, V., Raj, S. & Ravichandran, P. How whatsapp leads mobs to murder in india. *The New York Times* 18 (2018).
26. Sundar, S. S., Molina, M. D. & Cho, E. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *J. Comput. Commun.* 26, 301–319 (2021).
27. Wittenberg, C., Tappin, B. M., Berinsky, A. J. & Rand, D. G. The (minimal) persuasive advantage of political video over text. *Proc. Natl. Acad. Sci.* 118 (2021).
28. Sundar, S. S. The main model: A heuristic approach to understanding technology effects on credibility. *Digit. Media, Youth, Credibil.* (2008).
29. Hancock, J. T., Naaman, M. & Levy, K. Ai-mediated communication: definition, research agenda, and ethical considerations. *J. Comput. Commun.* 25, 89–100 (2020).
30. Barari, S., Lucas, C. & Munger, K. Political Deepfake Videos Misinform the Public, But No More than Other Fake Media. preprint, Open Science Framework (2021). DOI: 10.31219/osf.io/cdfh3. Accessed 2021-01-14.
31. Reeves, B., Yeykelis, L. & Cummings, J. J. The use of media in media psychology. *Media Psychol.* 19, 49–71 (2016).
32. Kasra, M., Shen, C. & O’Brien, J. F. Seeing is believing: How people fail to identify fake images on the web. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*, 1–6 (2018).
33. Hameleers, M., Powell, T. E., Van Der Meer, T. G. & Bos, L. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Polit. Commun.* 37, 281–301 (2020).
34. Nightingale, S. J. & Farid, H. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci.* 119 (2022).
35. Cardwell, B. A., Henkel, L. A., Garry, M., Newman, E. J. & Foster, J. L. Nonprobative photos rapidly lead people to believe claims about their own (and other people’s) pasts. *Mem. & cognition* 44, 883–896 (2016).
36. Cardwell, B. A., Lindsay, D. S., Förster, K. & Garry, M. Uninformative photos can increase people’s perceived knowledge of complicated processes. *J. Appl. Res. Mem. Cogn.* 6, 244–252 (2017).
37. Newman, E. J., Jalbert, M. C., Schwarz, N. & Ly, D. P. Truthiness, the illusory truth effect, and the role of need for cognition. *Conscious. Cogn.* 78, 102866 (2020).
38. Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J. & Lindsay, D. S. Nonprobative photographs (or words) inflate truthiness. *Psychon. Bull. & Rev.* 19, 969–974 (2012).
39. Fazio, L. K., Brashier, N. M., Payne, B. K. & Marsh, E. J. Knowledge does not protect against illusory truth. *J. Exp. Psychol. Gen.* 144, 993 (2015).
40. Ecker, U. K. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* 1, 13–29 (2022).
41. Dolhansky, B. *et al.* The DeepFake Detection Challenge (DFDC) Dataset. *arXiv:2006.07397 [cs]* (2020). Accessed 2021-05-15.
42. Groh, M., Epstein, Z., Firestone, C. & Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci.* 119 (2022).
43. Köbis, N., Doležalová, B. & Soraperra, I. Fooled twice–people cannot detect deepfakes but think they can. *iScience* **24** (2021).

44. Tahir, R. *et al.* Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16 (2021).

45. Vaccari, C. & Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media+ Soc.* **6**, 2056305120903408 (2020).

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

47. Hameleers, M., van der Meer, T. G. & Dobber, T. You won’t believe what they just said! the effects of political deepfakes embedded as vox populi on social media. *Soc. Media+ Soc.* **8**, 20563051221116346 (2022).

48. Lee, E.-J. & Shin, S. Y. Mediated misinformation: Questions answered, more questions to ask. *Am. Behav. Sci.* **65**, 259–276 (2021).

49. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* **116**, 2521–2526 (2019).

50. Austin, E. W. & Dong, Q. Source v. content effects on judgments of news believability. *Journalism quarterly* **71**, 973–983 (1994).

51. Shen, C. *et al.* Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society* **21**, 438–463 (2019).

52. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harv. Kennedy Sch. Misinformation Rev.* **1** (2020).

53. Jakesch, M., Koren, M., Evtushenko, A. & Naaman, M. The role of source, headline and expressive responding in political news evaluation. *Headl. Expressive Responding Polit. News Eval. (December 5, 2018)* (2018).

54. Nadarevic, L., Reber, R., Helmecke, A. J. & Köse, D. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. *Cogn. Res. Princ. Implic.* **5**, 1–16 (2020).

55. Kim, A., Moravec, P. L. & Dennis, A. R. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *J. Manag. Inf. Syst.* **36**, 931–968 (2019).

56. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).

57. Appel, M. & Prietzel, F. The detection of political deepfakes. *J. Comput. Commun.* **27**, zmac008 (2022).

58. Arechar, A. A. *et al.* Understanding and reducing online misinformation across 16 countries on six continents. *PsyArXiv* (2022). Accessed 2022-02-22.

59. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends cognitive sciences* (2021).

60. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).

61. Dan, V. *et al.* Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Commun. Q.* **98**, 641–664 (2021).

62. Calo, R., Coward, C., Spiro, E. S., Starbird, K. & West, J. D. How do you solve a problem like misinformation? *Sci. advances* **7**, eabn0481 (2021).

63. Frederick, S. Cognitive reflection and decision making. *J. Econ. perspectives* **19**, 25–42 (2005).

64. Pennycook, G., Cheyne, J. A., Koehler, D. J. & Fugelsang, J. A. Is the cognitive reflection test a measure of both reflection and intuition? *Behav. Res. Methods* **48**, 341–348 (2016).

65. Stagnaro, M., Pennycook, G. & Rand, D. G. Performance on the cognitive reflection test is stable across time. *Stagnaro, MN, Pennycook, G., & Rand, DG (2018) Perform. on Cogn. Reflect. Test is stable across time. Judgm. Decis. Mak.* **13**, 260–267 (2018).

66. Bialek, M. & Pennycook, G. The cognitive reflection test is robust to multiple exposures. *Behav. research methods* **50**, 1953–1959 (2018).

67. Pennycook, G. & Rand, D. G. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *J. personality* **88**, 185–200 (2020).
68. Mosleh, M., Pennycook, G., Arechar, A. A. & Rand, D. G. Cognitive reflection correlates with behavior on twitter. Nat. communications 12, 1–10 (2021).

69. Palan, S. & Schitter, C. Prolific.ac—A subject pool for online experiments. J. Behav. Exp. Finance 17, 22–27, DOI: 10.1016/j.bef.2017.12.004 (2018).

70. Berinsky, A. J., Margolis, M. F. & Sances, M. W. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. Am. J. Polit. Sci. 58, 739–753 (2014).

71. Abadie, A., Athey, S., Imbens, G. & Wooldridge, J. When Should You Adjust Standard Errors for Clustering? arXiv:1710.02926 [econ, math, stat] (2017). Accessed 2021-12-15.

72. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Super. di Scienze Econ. e Commerciali di Firenze 8, 3–62 (1936).

73. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. Nat. Hum. Behav. 5, 980–989, DOI: 10.1038/s41562-021-01143-3 (2021).

74. Metzger, M. J., Flanagin, A. J. & Medders, R. B. Social and heuristic approaches to credibility evaluation online. J. communication 60, 413–439 (2010).

75. Barasch, A., Schroeder, J., Zev Berman, J. & Small, D. Cues to sincerity: How people assess and convey sincerity in language. ACR North Am. Adv. (2018).

76. Schroeder, J. & Epley, N. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. J. Exp. Psychol. Gen. 145, 1427 (2016).

77. Ternovski, J., Kalla, J. & Aronow, P. M. Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments. preprint, Open Science Framework (2021). DOI: 10.31219/osf.io/dta97.

78. Agarwal, S. & Farid, H. Detecting deep-fake videos from aural and oral dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 981–989 (2021).

79. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. Sci. Adv. 6, eaa5359 (2020).

80. Watts, D. J., Rothschild, D. M. & Mobius, M. Measuring the news and its impact on democracy. Proc. Natl. Acad. Sci. 118 (2021).

81. Pennycook, G. et al. Shifting attention to accuracy can reduce misinformation online. Nature 592, 590–595 (2021).

82. Lazer, D. Studying human attention on the internet. Proc. Natl. Acad. Sci. 117, 21–22 (2020).

83. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. Sci. Adv. 8, eabo6254 (2022).

84. Fuller, T. Gnomologia: adagies and proverbs; wise sentences and witty sayings, ancient and modern, foreign and British, vol. 1 (B. Barker, 1732).

85. Messaris, P. Visual persuasion: The role of images in advertising (Sage, 1997).

86. Farid, H. Digital doctoring: how to tell the real from the fake. Significance 3, 162–166 (2006).

87. King, D. The commissar vanishes: The falsification of photographs and art in Stalin’s Russia (Metropolitan Books New York, 1997).

88. Lai, V. & Tan, C. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. Proc. Conf. on Fairness, Accountability, Transpar. 29–38, DOI: 10.1145/3287560.3287590 (2019). ArXiv: 1811.07901.

89. Agarwal, S. et al. Watch those words: Video falsification detection using word-conditioned facial motion. arXiv preprint arXiv:2112.10936 (2021). Accessed 2022-02-14.

90. Sankaranarayanan, A., Groh, M., Picard, R. & Lippman, A. The presidential deepfakes dataset. In Proceedings of the ALoFAI Workshop at the International Joint Conference on Artificial Intelligence (2021).
Did Joseph Biden say that?
Please watch and listen to this video from Joseph Biden and share how confident you are that it is fabricated. Remember half the media snippets we show are real and half are fabricated.

☐ I have seen/heard/read this before.

Just as likely real or fabricated
Submit Next

Please play the video before submitting. You have rated 1 of 32 text, audio, and video examples.

Figure 5. Screenshot of experimental user interface.
Figure 6. Image from Sankaranarayan et al. (2021) showing the first frame at the 10 second mark of each of the 32 videos in the Presidential Deepfake Dataset, which is where the stimuli from this experiment are drawn.

Figure 7. Robustness check for Figure 2 based on the subset of 18 videos on which participants identified accurately in fewer than 67% of observations. The mean and distribution of (a) weighted accuracy, (b) confidence, (c) perceived fabrications in fabricated speeches, (d) perceived fabrications in real speeches, and (e) response time are plotted for each of the seven modality conditions. The black lines indicate the 95% confidence interval of the true mean and the gray dots indicate each of the 18 speeches. Figure 7b plots confidence on a scale that ranges from a minimum of 50% confidence (just as likely as chance) to 100% confidence (full confidence). Figure 7e plots response time windsorized at the 99th percentile to control for time response outliers, which are an artifact of participants who return to the experiment after an extended time.
Table 1. Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The “Transcript” condition is held out and represented by the constant term. Column (1) shows recruited participants, column (2) shows recruited participants on “difficult” videos with lower than 67% accurate identification, column (3) shows pre-registered non-recruited participants, column (4) shows non-recruited participants after the pre-registration window, column (5) shows all non-recruited participants, column (6) shows non-recruited participants on “difficult” videos, column (7) shows non-recruited participants including participants who fail the attention check, column (8) shows non-recruited participants including participants who fail the attention check on “difficult videos.”

| Condition                  | (1)   | (2)   | (3)    | (4)    | (5)    | (6)    | (7)    | (8)    |
|----------------------------|-------|-------|--------|--------|--------|--------|--------|--------|
| Constant                   | 57.61 | 56.19 | 52.70  | 53.14  | 53.11  | 52.09  | 52.77  | 51.76  |
|                            | (0.83)| (1.11)| (0.58) | (0.15) | (0.15) | (0.19) | (0.13) | (0.17) |
| Silent Video               | 6.58  | 4.52  | 11.70  | 12.82  | 12.74  | 6.83   | 11.97  | 6.45   |
|                            | (1.19)| (1.55)| (0.86) | (0.23) | (0.22) | (0.30) | (0.20) | (0.26) |
| Silent Video with Subtitles| 8.80  | 5.80  | 12.45  | 12.82  | 8.16   | 11.95  | 7.57   |
|                            | (1.11)| (1.47)| (0.87) | (0.23) | (0.22) | (0.29) | (0.19) | (0.25) |
| Audio                      | 19.48 | 18.52 | 20.45  | 20.66  | 19.32  | 19.30  | 18.07  |
|                            | (1.17)| (1.50)| (0.84) | (0.23) | (0.22) | (0.29) | (0.19) | (0.25) |
| Audio with Subtitles       | 19.41 | 18.76 | 19.71  | 20.89  | 19.56  | 19.25  | 18.13  |
|                            | (1.05)| (1.46)| (0.89) | (0.23) | (0.22) | (0.29) | (0.19) | (0.25) |
| Video with Audio           | 25.12 | 24.27 | 27.97  | 28.54  | 25.57  | 26.81  | 24.10  |
|                            | (1.09)| (1.46)| (0.80) | (0.21) | (0.21) | (0.28) | (0.18) | (0.24) |
| Video with Audio and Subtitles| 24.75| 22.87 | 27.42  | 27.82  | 27.79  | 24.96  | 26.18  | 23.50  |
|                            | (1.08)| (1.46)| (0.81) | (0.22) | (0.21) | (0.28) | (0.19) | (0.25) |

Number of Individuals: 509 509 2807 38510 41313 37688 58344 52428
Number of Speeches: 32 18 32 32 32 18 32 18
Observations: 16,086 9,053 29,627 387,274 416,901 234,899 537,936 303,264
$R^2$: 0.07 0.07 0.06 0.07 0.06 0.06 0.06 0.05

Note: *p<0.05; **p<0.01; ***p<0.001

Figure 8. Distribution of media truth discernment scores following Pennycook and Rand (2019) where the score is positively associated with accuracy at distinguishing fabricated media from authentic media.
Figure 9. Percent of participants who respond that the speech is fabricated across modalities and the number of videos seen.
## Table 2

| Dependent variable: Weighted Accuracy | All (1) | Fabricated (2) | Not Fabricated (3) | All (4) | Fabricated (5) | Not Fabricated (6) |
|--------------------------------------|--------|----------------|-------------------|--------|----------------|-------------------|
| Constant (Silent Video)              | 61.97*** | 57.97*** | 66.17*** | 57.28*** | 53.94*** | 61.22*** |
| Transcript                           | -3.92* | -9.24*** | 0.84 | -0.85 | -10.71** | 6.34 |
| Subtitled Silent Video               | 4.67** | 6.20** | 2.88 | 8.67** | 4.92 | 11.75** |
| Any Audio                            | 18.22*** | 23.73*** | 12.49*** | 19.06*** | 22.23*** | 15.31*** |
| Discordant (Binary)                  | 4.57** | 8.09*** | 0.85 | 6.64* | 8.97* | 3.72 |
| Discordant (Binary) * Transcript     | -5.46* | -2.68 | -7.97** | -10.18** | -4.67 | -13.45** |
| Discordant (Binary) * Subtitled Silent Video | -5.02* | -2.91 | -6.81* | -10.86** | -3.56 | -17.52*** |
| Discordant (Binary) * Any Audio      | -5.42** | -6.54** | -4.20 | -7.30* | -7.71 | -6.30 |
| CRT Score                            | 2.90** | 2.56 | 2.97* | 2.56 | 2.56 |
| CRT Score * Transcript               | -1.90 | 0.71 | -3.32 |
| CRT Score * Subtitled Silent Video   | -2.48 | 0.72 | -5.34** |
| CRT Score * Any Audio                | -0.48 | 0.90 | -1.63 |
| CRT Score * Discordant (Binary)      | -1.20 | -0.51 | -1.63 |
| CRT Score * Discordant (Binary) * Transcript | 2.83 | 1.36 | 3.23 |
| CRT Score * Discordant (Binary) * Subtitled Silent Video | 3.46 | 0.29 | 6.34* |
| CRT Score * Discordant (Binary) * Any Audio | 1.08 | 0.67 | 1.15 |

Note: *p<0.05; **p<0.01; ***p<0.001

Table 2. Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The “Silent Video” condition is held out and represented by the constant term. The "Discordance (Binary)" variable is defined in Sankaranarayanan et al. (2021) by whether the speaker’s political views are discordant with the speech content.
|                   | All (1) | Fabricated (2) | Not Fabricated (3) | All (4) | Fabricated (5) | Not Fabricated (6) |
|-------------------|---------|----------------|-------------------|---------|----------------|-------------------|
| Constant (Silent Video) | 64.19*** | 61.64*** | 66.76*** | 60.60*** | 58.03*** | 63.48*** |
|                   | (0.88)  | (1.29)        | (1.19)           | (1.55)  | (2.24)        | (2.27)           |
| Transcript       | -6.62*** | -12.21***    | -5.47*** | -6.02**  | -14.99*** | -3.19            |
|                   | (1.19)  | (1.84)        | (1.54)           | (2.11)  | (3.19)        | (2.81)           |
| Subtitled Silent Video | 2.22*  | 4.78**        | -1.52           | 3.26    | 3.05          | 1.15             |
|                   | (1.05)  | (1.61)        | (1.61)           | (1.87)  | (2.76)        | (3.10)           |
| Any Audio        | 15.58*** | 20.49***      | 9.35*** | 15.40*** | 18.34*** | 10.95***         |
|                   | (0.89)  | (1.42)        | (1.21)           | (1.53)  | (2.54)        | (2.27)           |
| Discordance (Continuous) | 0.29    | 0.99          | 0.73            | 0.93    | 1.30          | 1.74             |
|                   | (0.76)  | (1.25)        | (0.97)           | (1.32)  | (2.11)        | (1.74)           |
| Discordance (Continuous) * Transcript | -2.80** | 7.07***       | -8.99*** | -4.43*   | 6.79*   | -9.92***         |
|                   | (1.07)  | (1.61)        | (1.41)           | (2.00)  | (2.92)        | (2.52)           |
| Discordance (Continuous) * Subtitled Silent Video | -1.36  | 0.56          | -4.19** | -4.13*   | 0.73    | -8.51***         |
|                   | (1.10)  | (1.77)        | (1.39)           | (1.97)  | (3.22)        | (2.44)           |
| Discordance (Continuous) * Any Audio | -0.94  | 0.46          | -4.55*** | -1.59    | 0.52    | -4.91**          |
|                   | (0.80)  | (1.29)        | (1.10)           | (1.41)  | (2.20)        | (1.89)           |
| CRT Score        | 2.26**  | 2.32*         | 2.03            |
|                   | (0.76)  | (1.11)        | (1.06)           |
| CRT Score * Transcript | -0.43  | 1.55          | -1.42           |
|                   | (1.03)  | (1.55)        | (1.35)           |
| CRT Score * Subtitled Silent Video | -0.71  | 0.96          | -1.66           |
|                   | (0.91)  | (1.35)        | (1.46)           |
| CRT Score * Any Audio | 0.09   | 1.28          | -0.98           |
|                   | (0.75)  | (1.22)        | (1.05)           |
| CRT Score * Discordance (Continuous) | -0.36  | -0.17         | -0.58           |
|                   | (0.65)  | (1.06)        | (0.87)           |
| CRT Score * Discordance (Continuous) * Transcript | 0.96   | 0.28          | 0.51            |
|                   | (0.96)  | (1.39)        | (1.26)           |
| CRT Score * Discordance (Continuous) * Subtitled Silent Video | 1.62   | -0.16         | 2.51*           |
|                   | (0.97)  | (1.59)        | (1.23)           |
| CRT Score * Discordance (Continuous) * Any Audio | 0.34   | -0.07         | 0.16            |
|                   | (0.69)  | (1.10)        | (0.96)           |
| Number of Individuals | 509     | 507           | 509             | 509     | 507           | 509              |
| Number of Speeches | 32      | 16            | 16              | 32      | 16            | 16               |
| Observations     | 16,086  | 8,042         | 8,044           | 16,086  | 8,042         | 8,044            |
| \( R^2 \)        | 0.07    | 0.12          | 0.04            | 0.07    | 0.13          | 0.04             |

**Table 3.** Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The “Silent Video” condition is held out and represented by the constant term. The "Discordance (Continuous)" variable is computed by calculating the z-transformation of participants’ mean response on a 5-point Likert scale for how well a speech aligns with the public’s perception of the politicians’ viewpoints.