Effects of Browsing Conditions and Visual Alert Design on Human Susceptibility to Deepfakes

Emilie L. Josephs, Camilo L. Fosco, and Aude Oliva

Abstract. The increasing reach of deepfakes raises practical questions about people's ability to detect false videos online. How vulnerable are people to deepfake videos? What technologies can help improve detection? Previous experiments that measure human deepfake detection historically omit a number of conditions that can exist in typical browsing conditions. Here, we operationalized four such conditions (low prevalence, brief presentation, low video quality, and divided attention), and found in a series of online experiments that all conditions lowered detection relative to baseline, suggesting that the current literature underestimates people's susceptibility to deepfakes. Next, we examined how AI assistance could be integrated into the human decision process. We found that a model that exposes deepfakes by amplifying artifacts increases detection rates, and also leads to higher rates of incorporating AI feedback and higher final confidence than text-based prompts. Overall, this suggests that visual indicators that cause distortions on fake videos may be effective at mitigating the impact of falsified video.

1 Introduction

Deepfakes are increasingly common, increasingly easy to create, and increasingly convincing. "Deepfake" is the colloquial term for an image, video, or audio clip that has been manipulated using deep learning techniques. While they are sometimes harmless, they can also be used for harm, ranging from fraud, impersonation, blackmail, nonconsensual intimate imagery, fake news, and political propaganda. Their propagation in the modern information landscape raises two major practical questions. How effective are deepfake videos at fooling human observers? What is the best way to warn viewers about deepfakes, and insulate them from the false information they contain?

Research to date suggests that humans are less susceptible to deepfake videos than images. Deepfake images have reached a point where they are indistinguishable from real images, and may even elicit higher levels of trust and social compliance (Nightingale and Farid 2022; Tucciarelli et al. 2022; Shen et al. 2021; Lago et al. 2021). In contrast, deepfake videos are still detectable at above chance levels (Groh et al. 2022; Köbis, Doležalová, and Soraperra 2021; Korshunov and Marcel 2021; Rossler et al. 2019; Lovato et al. 2022). While it is impossible to provide a single estimate of video deepfake detection

10.54501/jots.v2i2.144

rates, since studies differ in their design and stimuli, a survey of recent deepfake video detection studies suggest average detection rates in the 65%–75% range (Groh et al. 2022; Rossler et al. 2019; Boyd et al. 2022; Prasad et al. 2022).

However, it is likely that these studies overestimate a typical user's ability to detect fake video. Most recent experiments use settings that can inflate detection rates: participants are informed about deepfakes, responses are untimed, deepfakes are abundant, or video streaming quality is controlled. During real-word browsing, conditions are less ideal: deepfakes are rare, people are distracted, video quality is variable, and videos can be short. All of these conditions can impair the detection of even the most obvious signals (Rich et al. 2008; Prasad et al. 2022). To date, there is little understanding of how human deepfake detection varies under more ecologically valid search conditions. Here, we operationalize a range of detection conditions and examine how they affect deepfake detection in humans. To anticipate, we find that all conditions we tested reduced the detectability of deepfake videos.

Given that people are generally susceptible to fake images and videos, there is a need to develop methods for alerting and protecting users. Recent work has explored the effect of changing the motivational state of the viewer, using high-level interventions. However, these methods have not shown significant success: motivating participants by teaching them about the harms of deepfakes (Köbis, Doležalová, and Soraperra 2021), adding financial incentives (Köbis, Doležalová, and Soraperra 2021), or even eliciting emotional states (Groh et al. 2022) have all failed to affect detection rates.

A second, emerging direction is to supplement human users with additional information about the video from an independent source. Specifically, recent work has suggested human-AI teaming, where users are given access to a computer vision model that specializes in deepfake detection. Pairing humans with models is an emerging possibility because of rapid advances in deepfake detection by machine learning models (Boyd et al. 2022; Groh et al. 2022; Sohrawardi et al. 2020). Most models use computer vision methods, detecting signs of tampering such as pixel artifacts, anomalies in the biological signals of the video's subject, or inconsistencies in individual-specific features (Durall et al. 2019; L. Li et al. 2019; J. Li et al. 2019; Yang, Li, and Lyu 2019; H. Li et al. 2020; Li, Chang, and Lyu 2018; Ciftci, Demir, and Yin 2020; Matern, Riess, and Stamminger 2019; Boháček and Farid 2022; Agarwal et al. 2020; Haliassos et al. 2021; Cozzolino et al. 2021; Yang, Li, and Lyu 2019; Matern, Riess, and Stamminger 2019). Another kind of approach relies on authenticating a video based on its metadata, with proposals such as digital watermarking, blockchain-based tracking, and dataset fingerprinting (Qureshi, Megías, and Kuribayashi 2021; Neekhara et al. 2022; Chan et al. 2020; Alattar, Sharma, and Scriven 2020; Yu et al. 2021). In general, methods for automatically detecting deepfakes are under active development.

So far, however, teaming humans with AI assistants has met with mixed success. Studies of humans paired with deepfake detection models find surprisingly low willingness to incorporate AI feedback. In one study, participants who had access to model suggestions only updated their responses 24% of the time, and only changed their mind 12% of the time (Groh et al. 2022). Even when models are highly accurate, people embrace them only some of the time: Boyd et al. (2022) found that teaming humans with a model that is 90% accurate only yielded final human accuracy of 63%. Thus, current approaches to human-AI teaming for deepfake detection have significant unrealized potential.

How can we increase viewers' engagement with model suggestions? In traditional approaches, a model's predictions are communicated to the user using text. One direction for improvement may be to develop visual indicators that are more intuitive and compelling. Recent efforts tried showing users saliency maps of suspicious video

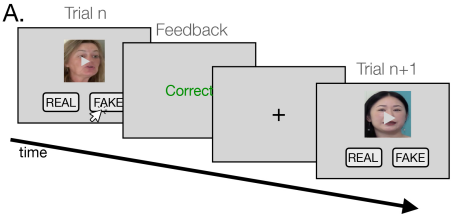
regions, but these did not improve engagement relative to text-based indicators (Boyd et al. 2022; Malolan, Parekh, and Kazi 2020). Here, we propose an alternate approach to visual indicator design, which relies on motion magnification to amplify artifacts in fake video (Fosco et al. 2022).

Artifact amplification (i.e., increasing the visibility of flaws and artifacts in fake videos) is promising for deepfake signaling for many reasons. First, it is a highly intuitive signal. It targets and amplifies the same information that humans instinctively use to make an unassisted judgment, like the naturalness of motion and the coherence of the faces (Mittal, Hegde, and Memon 2022). It also targets a visual domain that humans are particularly attuned to. Humans are exceptionally sensitive to the proportions of faces (Benson and Perrett 1991; Mauro and Kubovy 1992; Farah et al. 1998; Sinha et al. 2006) and are highly sensitive to unnatural faces (e.g., the uncanny valley effect; Mori 1970; Seyama and Nagayama 2007; MacDorman et al. 2009; Kagan et al. 1966). Second, it is practical: videos in the current online landscape are already loaded with text and icons (e.g., video playback controls, social media platform watermarks, news tickers and crawlers, closed captions), so additional text may not be very salient.

Here, we recruited subjects into online experiments to assess how different viewing conditions affect deepfake detection, and whether artifact amplification is an effective visual indicator. Participants performed a series of deepfake detection tasks for video deepfakes, where we varied the prevalence, duration, and streaming quality of deepfakes, as well as the user's cognitive load. Half of the participants performed the tasks without AI support, allowing us to isolate and measure the effect of different viewing conditions. The other half had AI support in the form of artifact amplification, allowing us to measure the behavioral impact of artifact magnification across viewing conditions. We also compared artifact magnification with a more conventional visual indicator design (text on the video). Participants performed a deepfake detection task where an AI provided feedback on their responses using either motion magnification or text, and we measured the final accuracy and subjective confidence for each condition.

Overall, we found that all viewing conditions we tested decreased deepfake detection rates relative to a baseline. In contrast, artifact magnification was highly effective at boosting deepfake detection and increased deepfake detection across all viewing conditions. Additionally, artifact amplification was more effective than text at encouraging users to incorporate the AI's suggestion, leading to higher accuracy and higher confidence, particularly for medium- and difficult-to-detect deepfakes. Altogether, this paper makes two broad contributions to the science of human deepfake detection: it advances our understanding of the risks they pose to humans, by charting the detectability of deepfakes across commonly encountered browsing conditions, and explores ways to mitigate this risk.

A.



B.

Condition	Description
Baseline	2AFC, 50% prevalence, unspeeded, feedback
Low Prevalence	Fake trials have 20% prevalence
Brief Presentation	2 second response time limit
Divided Attention	Concurrent digit counting task
Noisy Video	Jpeg compression noise

Figure 1: Methods, and conditions testing deepfake detection under typically encountered browser conditions. (A) Baseline procedure: participants viewed one video at a time, and indicated whether they thought it was real or fake. Feedback was given on each trial. (B) Summary of the conditions explored. The baseline procedure was modified to accommodate each condition; see Section 2.

2 Methods

2.1 Study 1: Detectability of deepfake Caricatures

Stimuli. Stimuli consisted of videos of single individuals. Videos were selected from the Deepfake Detection Challenge preview dataset (DFDC) (Dolhansky et al. 2019). Videos were preprocessed as described in Fosco et al. (2022): audio was removed, videos were cut into 12-second clips and cropped to show only one face. Each cropped clip measured 360x360 pixels, with a minimum of a 100px margin between the edge of frame and the face. Next, videos were selected for the experiment as follows.

First, we selected 300 real clips, by sampling 2–5 real video clips for different actors in the DFDCp. We sampled the videos to include variation in gender, race, body type, hair, age, and bearing. Next, we selected the corresponding deepfakes. The DFDCp features multiple deepfakes generated from each real video. For each of the real clips we selected, we retrieved all of the corresponding deepfakes, and filtered them for quality. Our goal was to match the quality of deepfakes in our study to the quality of deepfakes that would plausibly be shared online. Thus, we excluded any deepfake that contained artifacts for its whole duration; contained artifacts covering the whole face at any point in time; contained a momentary failure revealing the real face underneath; or had mismatches in gender, lighting, or skin tone from the underlying head and body. We additionally removed any deepfake that was indistinguishable from the real face it was generated from. For each real clip, we randomly selected one of the corresponding deepfake clips from the set that survived this filtering, yielding 300 real-fake pairs.

Finally, from each of the tampered videos, a Caricature was created using the CariNet approach (Fosco et al. 2022). CariNet is a semi-supervised framework that predicts which regions in a tampered video of a face contain artifacts that are salient to human observers, and selectively amplifies them using motion magnification. This causes faces with artifacts to ripple and warp throughout the video (see Figure 3A for example frames). These distorted videos are called Deepfake Caricatures. The effect of the transformation is most compelling when viewed in video form, so we have have provided a gallery of fake videos with the Caricature effect applied: https://camilofosco.com/deepfake_caricatures_website/gallery.html.

Thus, our video set contained a total of 900 videos: 300 real, 300 deepfakes, and 300 Caricatures. We calibrated our experiments to take a median of 15 minutes (in the Baseline condition; see below) by presenting only a subset of the videos to each individual subject. We divided the video set into subsets of 100 videos each: 50 real videos and 50 fake videos (fake videos within a subset could not be generated from real videos in the same subset).

Sample size estimation. Sample size was determined by estimating the number of participants required to gain a stable estimate of the detectability of a single video, based on pilot data. A pilot study was conducted with 40 participants per condition, using the Baseline procedure (see below). Following Strong and Alvarez (2019), we simulated sample sizes ranging between 10 and 200, by sampling 1,000 times with replacement from the pilot participants. For each sample size, we calculated the variance in average detectability for a given video, averaged over all videos. We identified the sample size at which this variance plateaus (n = 6), then doubled and rounded up to obtain a sample size of 15 subjects per video.

Participants. A total of 913 people participated across all sub-experiments in Study 1 (52% Female, 45% Male, 3% Not Reported). Participants were recruited from the Prolific online experiment platform (www.prolific.com). Participants were required to meet the following criteria: 95% approval rates or higher, a history of more than 500

tasks completed on the platform, and located in the US. Participants were recruited and compensated according to procedure approved by MIT's Committee on the Use of Humans as Experimental Subjects. Participants were paid an hourly rate of $11.25 per hour.

Design. There were five slightly different designs of detectability experiments, depending on the viewing condition being measured.

In the Baseline condition, participants viewed one video at a time and indicated whether they thought the video was real or fake. Responses were not time limited, and participants received feedback on every trial. There were 100 trials, divided into five blocks of 20. The experiment started with five exposure trials, in which a 12-second deepfake video was displayed, but no response was required. The experiment contained five attention check trials, which consisted of a video with the message "this is an attention check, please select 'REAL' " in capital letters. Real and fake videos were equally prevalent, and randomly intermixed.

In the Low Prevalence condition, the same design was used, with the exception that only 20% of the trials contained fake videos. This design follows typical low-prevalence studies in the visual search literature (Rich et al. 2008; Wolfe et al. 2007; Hout et al. 2015). The subset of deepfakes used here were randomly sampled from the full set.

In the Speeded Presentation condition, the Baseline condition was modified such that videos were only presented for 2 seconds. This value was selected by taking the median reaction time in the Baseline condition (2.8 s), then rounding down to yield a moderately challenging time limit. In order to enforce the time limit, the response screen replaced the video after the time limit.

In the Divided Attention condition, participants performed a concurrent digit-counting task. Similar tasks have been used in driving and automobile research to elicit multi-tasking and increase cognitive load (Yamani et al. 2018; Horrey, Wickens, and Consalus 2006). A nine-digit string was displayed on the video itself, one digit at a time, with a 0.45-second interval between digits. Participants were asked to count the number of odd digits in the string, which ranged from three to five (inclusive). Participants reported the number of digits after reporting their response for the video. Because this is a challenging task, the experiment was shortened to 50 trials.

The Noisy Video condition was identical to the Baseline condition, with the exception that the videos had been manipulated to mimic compression artifacts caused by lossy encoding. Similar to Rössler et al. (2018), videos were compressed using a constant rate factor of 40 (18 is considered perceptually lossless, 23–28 is considered acceptable), yielding blurring and aliasing.

Each of these viewing conditions had a deepfake version, or a Caricature version. These versions were presented in a between-subjects design, because we were concerned that including deepfakes and Caricatures in the same subject would cause criterion shifts. We took the following steps to reduce population effects: there were no outward differences between the deepfake and caricature versions until participants started the task, both versions of the task were released on the website at the same time, and condition assignment was simply determined by which link participants clicked.

Analysis. The following preregistered procedures were used for removing low-quality data: any participant who failed three or more attention check trials was removed and replaced, and any trial that took longer than 60 seconds was dropped. For the divided attention condition, we additionally dropped any subject performing lower than two standard deviations below the mean on the digit counting task, in order to ensure they were devoting sufficient attention to the number task.

Some participants had 100% accuracy rates, especially in the Caricatures condition. Thus, for calculating signal detection measures, we used the method proposed in Hautus (1995) to calculate sensitivity and criterion in case with extreme values (.5 is added to the count of Hits, False Alarms, Misses, and Correct Rejections to avoid a divide by zero).

Power analysis indicated that we had very high power, so we took a conservative approach to quantifying differences in hit rate, false alarm rate, sensitivity, and criterion from Baseline. Specifically, we used a Bonferroni correction where the alpha level was divided by the total number of statistical tests (16). Effect sizes were measured using Cohen's d_s for two-sample designs (Lakens 2013). To compute the statistical significance of the sensitivity difference within each condition, for which we preregistered a hypothesis, we used one-sided unpaired t-test, with a Bonferroni correction such that alpha level was divided by the total number of tests (one for each experimental condition, 5).

2.2 Study 2: Caricatures compared to text-based prompts

Stimuli. Stimuli for Study 2 were a subset of those in Study 1. Deepfakes from Study 1 were given a detectability score, which was the percent of time it was detected in the Baseline condition of Study 1, across all participants who viewed it. Overly easy and difficult deepfakes were discarded (detection rates below 20% or above 90%), and the remaining deepfakes were binned by difficulty: difficult (detected 0.21%–0.44% of the time), medium (0.44%–0.67%), and easy (0.67%–0.9%). Finally, 30 deepfakes were randomly selected from each bin to make up the stimulus set, as well as the corresponding Caricatures. To achieve 50% fake prevalence, and additional 90 real videos were randomly selected, with the caveat that they did not include the real videos corresponding to the selected deepfakes.

Participants. Participant recruitment followed the same criteria, ethical safeguards, and payment rates as Experiment 1. A total of 298 people participated (47% Female, 52% Male, 1% Not Reported).

Design. Our design was based on Groh et al. (2022). Participants saw one video at a time, and were asked to "Rate your certainty that this is a deepfake" using a slider, whose poles were labeled "100% confident REAL" and "100% confident FAKE." Next, they clicked a button labeled "Click here to see the model's guess," and were shown the prediction of the model as either a text-based prompt or a Caricature.

Text-based prompts were in the form "The model estimates that this video is REAL" (or "FAKE"). They were introduced with the text "Here is the output of the modeling procedure. As a reminder, our model works by estimating the likelihood that the video is fake, then labeling the videos it thinks are probably fake." In the Caricatures, this screen displayed the same video after it had been passed through the Caricature model, which distorts deepfakes but leaves real videos intact. The Caricatures were introduced with the text "Here is the output of the modeling procedure. As a reminder, our model works by estimating the likelihood that the video is fake, then distorting the videos it thinks are probably fake." Real-fake labels reflected the ground truth.

This experiment included real and fake videos at 50% prevalence, and there was no time pressure. Participants did not receive feedback on their accuracy. There were 60 trials total per participant, divided into blocks of 10. There were five randomly placed catch-trials, on which the text "This is an attention check, please set confidence to 100% Real" was displayed instead of the model prediction. Text-based and Caricature conditions were collected in a between-subjects manner.

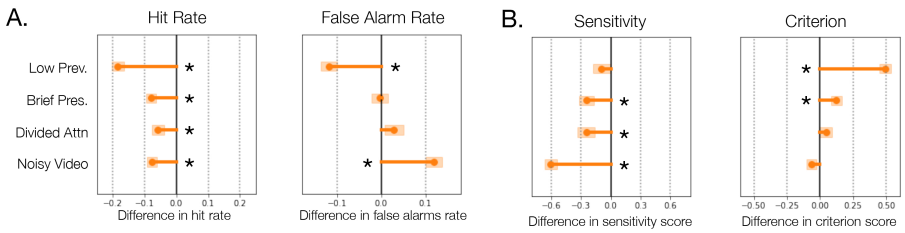


Figure 2: Results of testing deepfake detection under typically encountered browser conditions. (A) Changes in hit and false alarm rates relative to baseline. Light orange box indicates standard error of the mean, and stars indicate significance. Baseline hit and false alarm rates were 0.73 and 0.28 respectively. (B) Change in sensitivity and criterion from baseline. Baseline values were 1.29 and -0.01 respectively.

Data analysis. The following preregistered procedures were used for removing low-quality data: any participant who failed three or more attention checks was removed and replaced, and any trial that took longer than 60 seconds was dropped.

Four analyses were performed on this data, each using an ANOVA to test for a main effect of visual indicator type (text-based prompt or Caricature) or interaction between visual indicator type and difficulty on each of four measures of interest (if an interaction was present, main effects were not analyzed). There were two preplanned measures of interest (final confidence and final accuracy), for which we used standard p value of 0.05, and two post hoc measures (proportion of trials on which participants updated their responses, magnitude of response changes) for which we used a Bonferoni-corrected p value of 0.025.

Post hoc individual differences. Single subject averages were extracted for initial accuracy (i.e., accuracy before model input), initial confidence (i.e., confidence level before model input) and AI amenabilty (i.e., the proportion of trials in which a participant updated their responses following model feedback). Partial correlations for the individual difference measures were performed using the ppcor package in R.

3 Results

Study 1a: Typical browsing conditions reduce deepfake detection performance

Deepfake detection experiments with human participants typically present videos under conditions that are favorable for detection. However, these conditions do not reflect the reality of encountering a deepfake video while browsing the internet. How do deepfake detection rates change when we introduce conditions that can exist during an internet browsing session?

We tested the detectability of deepfakes across five different viewing conditions. The Baseline condition used settings similar to most deepfake detection studies to date: participants performed an untimed detection task, where they viewed one video at a time and reported whether it was real or fake (Figure 1A). Half of the videos were fake (i.e., 50% prevalence).

The remaining conditions operationalized some challenges that can arise in a typical browsing session. This included: (1) Low Prevalence, where only 20% of videos were deepfakes; (2) Brief Presentation, where participants were only shown two seconds of the video, mimicking a situation where only a brief clip of video is provided or attended to; (3) Divided Attention, where we simulated a case where users encounter video

engaging in a concurrent task (i.e., multitasking) by asking participants to perform a digit-counting task at the same time as the detection task; and (4) Noisy Video, where the video was blurred and degraded, mimicking situations where video quality is reduced due to compression and streaming limits (Prasad et al. 2022). These conditions targeted both endogenous factors under the control of the user and exogenous factors that may depend on the browsing session. These conditions were not designed to be perfect reproductions of what users might encounter during real-world browsing, but rather to isolate common viewing conditions, and test their effects in a controlled and measurable way.

Stimuli consisted of 360-by-360px videos of faces, selected and cropped from the Deepfake Detection Challenge dataset preview version (Dolhansky et al. 2019; see Methods). There were 180 participants per experiment, for a total of 900 participants. Results were analyzed in a signal detection framework (Abdi 2007; Batailler et al. 2022), with a particular focus on hit rates (i.e., proportion correct on target-present trials), since this is the most direct measure of how many deepfakes were correctly detected under different conditions. Additionally, this avoids ambiguity that can arise from only reporting overall accuracy rates: overall accuracy does not distinguish whether participants were good at identifying which videos were fake (i.e., correct on target-present trials) or at confirming which videos were real (i.e., correct on target-absent trials). This distinction is especially important in low-prevalence settings, when a high rate of correct rejection can make overall accuracy high, even in the face of a low hit rate.

Results for the detection experiments are reported in Figure 2 (and also in tabular format in the Appendices). In the Baseline condition, the average hit rate across participants was 73.3% (SEM (standard error of the mean): 1.13%), at the cost of a relatively high false alarm rate (28.1%). Crucially, the hit rate was reduced for all of the experimental conditions we examined. Low Prevalence suffered the most, with the average hit rate reduced to 54.8% (SEM: 1.58%), followed by Brief Presentation and Noisy Video (65.3% and 65.5%, SEM: 1.07% and 1.14%, respectively) and Divided Attention (67.4%, SEM: 1.29%). All differences were significant, with effect sizes in the medium to large range (Low Prevalence: $t(175) = 9.44$, $p < 0.001$, $d_s = 1.43$; Brief Presentation: $t(176) = 5.13$, $p < 0.001$, $d_s = 0.77$; Divided Attention: $t(176) = 3.39$, $p < 0.001$, $d_s = 0.51$; Noisy Video: $t(176) = 4.78$, $p < 0.001$, $d_s = 0.72$). Altogether, these results suggest that deepfake detection rates are sensitive to the conditions under which detection is occurring, and that detection is lowered when deepfakes are rare, when engagement with the video is short, when participants are distracted, or when the video is degraded.

What accounts for the reductions in hit rates? One source of the decrease could be reduced *sensitivity*, that is, a reduction in the sensory system's ability to detect the video artifacts under these viewing conditions. Another could be increased *criterion*, that is, a behavioral change causing a more conservative decision process and decreased willingness to say that a target (i.e., a deepfake) is present. Using signal detection theory, we measured sensitivity and criterion across conditions and found that the mechanism for the reduced hit rate differs across viewing conditions (Figure 2B). For Brief Presentation, Divided Attention, and Noisy Video, there is a significant decrease in sensitivity relative to Baseline (Brief Presentation: $t(176) = 3.21$, $p = 0.0016$, $d_s = 0.48$; Divided Attention: $t(176) = 3.04$, $p = 0.0027$, $d_s = 0.46$; Noisy Video: $t(176) = 9.27$, $p < 0.001$, $d_s = 1.39$; Low Prevalence: $t(175) = 1.14$, not significant) suggesting that the perceptual difference between real and fake videos is not as salient when people are rushed or distracted, or when the video quality is reduced. In contrast, for Low Prevalence, and to a lesser extent for Brief Presentation, there is a substantial increase in criterion (Low Prevalence: $t(175) = 11.0$, $p < 0.001$, $d_s = 1.66$; Brief Presentation: $t(176) = 3.22$, $p = 0.0015$, $d_s = 0.48$; Divided Attention: $t(176) = 1.01$, not significant; Noisy Video: $p = 1.69$, not

significant). This suggests that participants are biased to respond "REAL" when they are rushed or when fake videos are rare, and require more obvious signs of tampering in order to overcome this bias. Taken together, different viewing conditions affect different dimensions of detection performance.

So far we have discussed how individuals perform when confronted with real and fake videos, across a range of conditions. However, it is also valuable to understand the accuracy of group-level responses on individual videos. Certain fake news detection systems rely on crowdsourced ratings and explanations, but it is an open question whether consensus responses to deepfake videos are accurate and resilient to viewing conditions. In an exploratory analysis, we took a wisdom-of-crowds approach (Groh et al. 2022) to our data, and examined aggregate responses on individual videos. For each video, we took the majority response ("REAL" or "FAKE") on a given video as the "consensus response," and compared it to the ground truth. In the Baseline experiment, the consensus response was correct 84% of the time, slightly higher than previous results (74% and 80% in Groh et al. 2022).

Is this aggregate performance level resilient to commonly encountered browsing conditions? In keeping with the exploratory nature of this analysis, we only report effect sizes, using Cohen's *h*, for calculating effect sizes of proportions. In general, changes in viewing conditions had little effect on the accuracy of the consensus response, with accuracies of 82%, 82%, and 80% for Low Prevalence, Brief Presentation, and Divided Attention, respectively (Cohen's *h*: 0.05, 0.05, and 0.01, respectively). The only viewing condition to have an effect was Noisy Video, with a consensus performance of 72% (Cohen's *h*: 0.3, considered small to medium). This suggests that while individual performance is susceptible to changes in viewing conditions, aggregate measures are more resilient.

Study 1b: Artifact amplification increases deepfake detectability across viewing conditions

While human observers can still achieve above-chance success at detecting deepfake videos, this advantage may not last much longer. As deepfakes become more realistic, deepfake mitigation may rely on pairing the human user with a machine learning model. How will these models communicate their predictions to the human user? Here, we test the viability of using artifact amplification for indicating fake videos.

For these studies, we used a computer vision model that detects then amplifies artifacts in deepfake videos (Fosco et al. 2022). In this approach, the model generates a heat map predicting the locations of artifacts in the input video. In addition to training on large sets of deepfakes, the model is semi-supervised with human annotations of artifacts, so these heat maps identify artifacts that are salient to people as well as machines. This heatmap is used to guide the application of motion magnification to frames of the video, yielding distorted versions of deepfakes where the faces appear to ripple and warp. These distorted outputs are called Deepfake Caricatures (Figure 3A).

We first established whether making a Caricature of a deepfake improves their detectability. A separate pool of participants (n = 180 per study, for a total of 900) were recruited to perform detection tasks as above, except all fake videos had been subjected to the Caricatures transformation. To quantify how much Caricatures facilitate detection, we report the difference in average sensitivity between participants viewing plain deepfakes (using the above data) and those viewing Caricatures, for a given detection setting. Across all conditions, Caricatures led to a substantial increase in sensitivity ($t > 20$, $p < 1e^{-50}$) for all conditions; see Supplement for full statistical reporting). The hit rate was improved to 95.1% in the Baseline condition, and importantly, remained high across all other

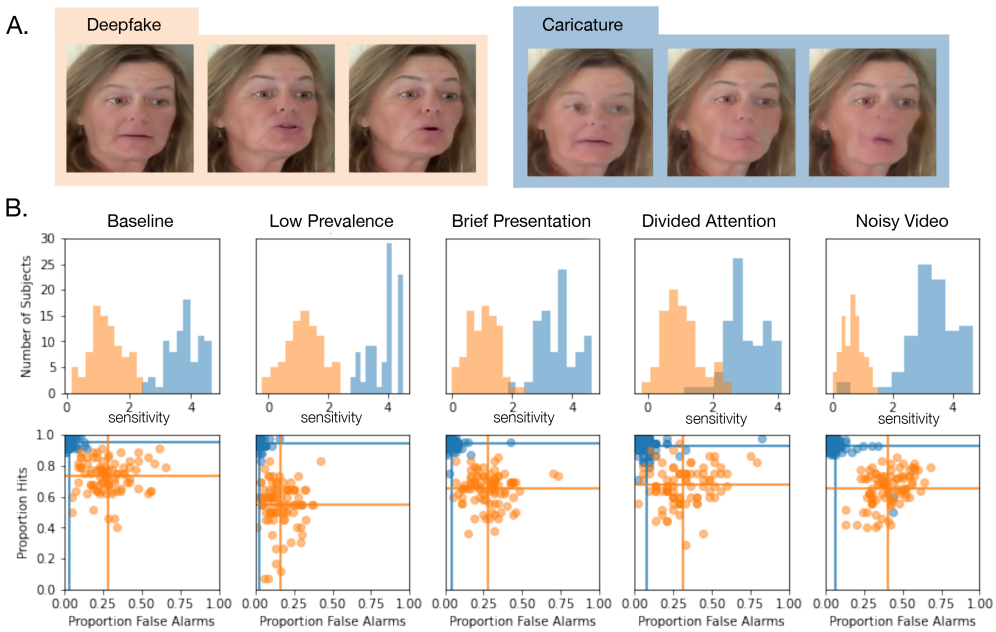


Figure 3: Detectability of Deepfake Caricatures. (A) Comparison between frames of a deepfake video and the same video with the Caricature transformation applied. The amplification of motion artifacts causes the faces in Caricatures to warp and distort. (B) Results: top row shows the distribution of average participant sensitivity under different viewing conditions. Orange denotes participants who saw plain deepfakes, and blue denotes those who saw Caricatures. The bottom row replots the data, explicitly showing average participant hit and false alarm rates for deepfakes and Caricatures across viewing conditions.

conditions, with hit rates of 94.9%, 94.4%, 92.6%, and 93.2% for Low Prevalence, Brief Presentation, Divided Attention, and Noisy Video, respectively (Figure 3B; see Supplement for all signal detection measures). In all cases, the distribution of sensitivity scores across participants in the Caricature condition had little to no overlap with the distribution for participants in the Deepfake condition, indicative of very large effect sizes (Cohen's d_s = 4.67, 4.79, 4.41, 3.41, and 4.57 for Baseline, Low Prevalence, Brief Presentation, Divided Attention, and Noisy Video, respectively).

Taken together, these results show that artifact amplification is highly effective at making fake video distinguishable from real, and that this improvement is present across a range of viewing conditions.

Study 2: Artifact amplification is more convincing than traditional text-based prompts

So far, we have shown that human observers are well below the ceiling at detecting deepfakes, and suggested that artifact amplification is an effective way to boost the detectability of fake videos. However, the perceptibility of a visual indicator is only one way to quantify how effective it is. A crucial second measure of an indicator's effectiveness is whether users find it convincing enough to accept the model's suggestion.

We next examined whether using text-based prompts versus artifact amplification changes the likelihood that a user would change their behavior based on the visual indicator. We measured the indicators' impact on behavior in three ways: (1) the user's

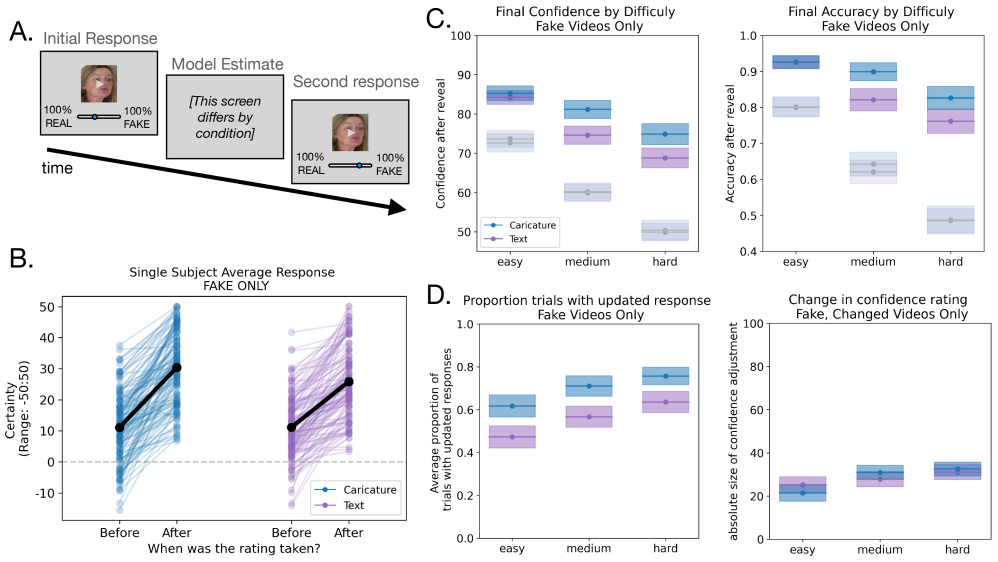


Figure 4: (A) Procedure: Participants view a video and make their responses using a slider. Next, they view model prediction. In the Text condition, this screen showed text that read "Our model estimates that this video is REAL" (or "FAKE"). In the Caricature condition, it showed the video with the Caricature procedure applied (this amplifies artifacts in fake videos, but leaves real videos intact). Then, participants could adjust their response. (B) Single subject confidence levels, before and after model input, for fake videos. Blue denotes Caricatures and purple denotes Text. (C) Average confidence and accuracy after model input, broken down by deepfake difficulty. Boxes show 95% CI, and lighter colors show responses before model input. (D) Which behavioral changes underlie increased confidence following Caricatures.

average degree of confidence in their final, model-assisted response; (2) the proportion of time users accepted the model's suggestion; and (3) the amount of confidence change participants reported on single trials.

Following Groh et al. (2022), participants were shown a video and provided an initial response on a slider ranging from "100% confident REAL" to "100% confident FAKE" (Figure 4A). Next, participants were shown a model prediction screen, where model predictions were conveyed either in text (e.g., "Our model estimates that this video is [REAL/FAKE]"), or by displaying a Caricature of the video. Since the Caricature model works by detecting and amplifying artifacts in fake videos, this has the effect of distorting fake videos and leaving real videos intact. Predictions in this stage were not generated by a real model, but instead reflected ground truth, yielding a "model" performance of 100% accuracy. This allowed us to estimate the indicator's impact on behavior in the best-case scenario of a perfect model, and to isolate the role of the visual indicator, since previous work has indicated that model accuracy has its own influence on behavior (Naujoks, Kiesel, and Neukum 2016; Sendelbach and Funk 2013; Yin, Wortman Vaughan, and Wallach 2019). After viewing the model prediction screen, participants were returned to the screen with the video and given the opportunity to update their response on the slider.

We hypothesized that the difference between visual indicators might be more pronounced for more difficult videos, which appear more convincingly real. Thus, we included videos at three levels of difficulty (operationalized as their overall detectability in the Baseline condition of the above experiments; see Methods and Materials). Deepfakes were present

at a 50% prevalence, and there was no time limit for responses. Since this experiment is concerned with how well the different methods convince users that a video is fake, we discuss only target-present trials, where the video is fake.

Figure 4B visualizes a high-level summary of participant's behavior on target-present trials, graphed as the average scores assigned for the video before and after participants viewed model feedback. The score was registered on a 100-point scale centered on 0, where 0 means that participants were unsure if the video was real or fake, 50 means they were sure the video was fake, and -50 means they were sure the video was real. A first observation is that there are large between-subject differences, in both deepfake detection ability (shown by spread of dots in the Before condition), and in people's willingness to change their behavior in response to the indicator(shown by the variety in slope between Before and After). Second, it is clear that both visual indicators are effective at changing people's judgments of the video's authenticity.

To quantify the difference between text-based indicators and Caricatures, we examined the difference in the average confidence on After trials across difficulty levels (Figure 4C). Overall, confidence decreased with video difficulty, suggesting that people are more susceptible to challenging deepfakes, even with high-quality model support. Crucially, this decrease is less pronounced in the Caricatures condition: while easy trials showed no difference in final confidence between the two methods, medium and hard trials showed higher subjective confidence for Caricatures (significant interaction: $F(2,798) = 3.51$, $p = 0.030$). Importantly, there was no difference between Text and Caricatures in participants' responses before model input (grey bars in Figure 4C). Thus, models that used artifact amplification to convey their prediction made people more confident in their model-supported decision compared to models that used text-based prompts.

Accuracy scores were also calculated, by transforming the continuous confidence score into a binary response. Positive scores were translated to a "Fake" response, negative scores were translated to a "Real" response, and the accuracy of these binary responses was assessed against the ground truth. We found that accuracy scores followed the same pattern as confidence scores (Figure 4C): accuracy of AI-assisted responses were not different between Caricatures or Text for easy trials, but Caricatures led to higher accuracy for medium and difficult trials (significant interaction, $F(2,798) = 4.63$, $p = 0.010$). This illustrates how models that elicit higher subjective confidence can increase detection outcomes, in cases where the model is highly accurate.

What change in user behavior underlies this increase in overall confidence? One possibility is that Caricatures increase the frequency with which users accept the model's suggestion. Another is that the frequency remains the same, but users experience larger changes in confidence in the Caricatures condition. We compared these options in a follow-up analysis (Figure 4D), and found that the average proportion of trials in which subjects changed their responses was higher for Caricatures, at all levels of difficulty (significant main effect: $F(1,798) = 17.28$, $p < 0.001$, no interaction effect). In contrast, there was no difference in the magnitude of the adjustment people made following Caricatures vs. text-based indicators (no main effect: $F(1,768) = 2.208813$, $p = 0.14$, no interaction effect; analyzing subset of trials where participants adjusted their responses). This suggests that the efficacy of Caricatures comes from their ability to make an impression more often, not necessarily from their ability to make a larger impression.

Post hoc individual differences

Given the large number of participants, and high inter-individual variation, we saw an opportunity for an exploratory analysis on the individual characteristics that might influence a user's willingness to incorporate AI feedback. We assigned each participant

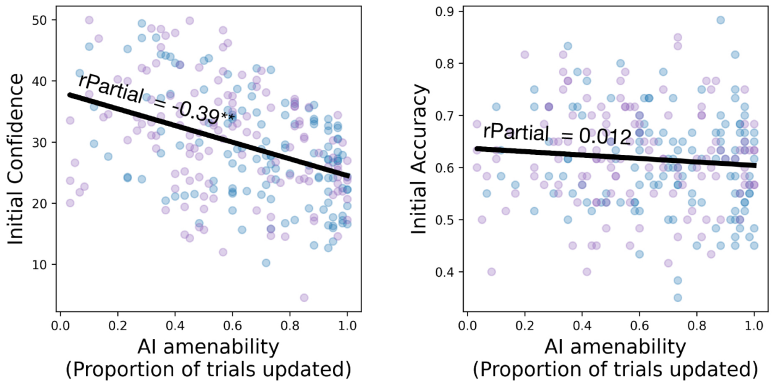


Figure 5: Exploratory analysis of behavioral measures that correlate with a participant's AI amenability, operationalized as the proportion of trials on which they updated their response based on model input. We examined its relationship with participant confidence (before model input) and participant accuracy (before model input). Since confidence and accuracy are related (VIF = 11.96), we used partial correlations. There was no difference in the trend between Text and Caricature trials, so all data were combined for this analysis, but condition colors are preserved in these figures for transparency. Stars indicate significance.

an AI amenability score, operationalized as the proportion of trials on which they adjusted their response following AI feedback, and examined what other behavioral measures were correlated.

One possibility is that individuals with low AI amenability were simply the individuals where were better at the task: if their initial answers tended to be more correct, there would be no reason to engage with AI assistance. Another possibility is that low-amenability participants were more confident in their initial response, and therefore less likely to adjust, even when the model indicates they were wrong. Of course, a participant's confidence depends partially on their perceived accuracy in the study. Indeed, the VIF (variance inflation factor) between a participant's initial accuracy (i.e., before model input) and their initial confidence was above 10, indicating high covariance (VIF = 11.96). Thus we used partial correlations to assess the relationship between AI amenability, and initial accuracy and confidence, respectively, while partialling out the dependence between initial accuracy and initial confidence.

Overall, results (Figure 5) indicate that there is a a small but significant negative relationship between a participant's initial confidence (independent of their actual accuracy) and their likelihood to accept the AI's suggestion (rpartial = -0.39, $p < 0.001$). In contrast, there was no relationship between a participant's unassisted accuracy and their AI amenability (rpartial = 0.012, $p = 0.84$). This suggests that another possible predictor of whether a user will successfully pair with an AI, independent of the visual design of the indicator, is how confident they feel about their ability to perform the task unassisted. Future work is required to quantify the contribution of this factor.

4 Discussion

Two pressing questions in today's media landscape are how susceptible people are to deepfakes, and how to mitigate the risks that deepfakes pose. Here, we advance our understanding of both of these issues. We show that deepfake detection rates are highly sensitive to the conditions under which they are viewed, and are negatively impacted by

many of the conditions present in typical browsing sessions. We also confirm previous findings that pairing a human observer with a machine learning model can increase detection rates, but illustrate how the design of the visual indicator supplied by the model can affect the quality of the collaboration.

One broad implication of these results is that the field of behavioral deepfake detection is overestimating people's detection rates. There is limited value in discussing precise detection accuracy values, since detection rates will change as the technology improves, and can depend on the type of deepfake used as stimuli (Rossler et al. 2019). Instead, we focus on how detection rates changed when the experimental conditions varied: detection rates were reduced relative to baseline in every condition we tested. These results suggest that misinformation mitigation researchers should increase their estimates of people's susceptibility to deepfakes, and the false messages they may convey.

Additionally, we found that the specific mechanisms of this reduction varied depending on the viewing conditions (e.g., increased criterion vs. reduced sensitivity). This suggests that different conditions have independent effects, which may stack when the conditions are combined. A number of additional conditions that would be present during a browsing session were not tested here, but could be expected to further impact performance. Videos are sometimes viewed while scrolling, and this motion may disguise motion artifacts in deepfakes. Videos are often embedded in text, or near other images and banners, which may add clutter to the visual display. Users outside of experimental setting may have their own reasons for engaging with videos, and may not be as vigilant for flaws in the videos. More work is required to understand the full range of conditions, and their individual (and combined) impacts on detection rates. More broadly, it is useful to name and test the variety of detection conditions that exist in typical browsing sessions, because they have implications for how we deploy warnings about deepfakes. Given this variability, visual indicators for deepfake signaling should be tested under a variety of conditions, to ensure that they remain useful and robust across all anticipated conditions.

Interestingly, our exploratory results suggest that crowd-consensus deepfake detection is much more robust to commonly encountered browsing conditions than individual detection. This suggests that techniques that rely on aggregate annotations can be successful even if individual judgments are less reliable. One example is the use of human annotation data to supervise deepfake detection models (Fosco et al. 2022; Boyd et al. 2022; Gupta et al. 2020). Another is the use of crowdsourcing as part of a real-time deepfake detection pipeline. Misinformation mitigation in some online communities relies on aggregating reports on content already in circulation from users themselves. The present results suggest that this approach could be useful for deepfakes in active circulation on platforms with wide and active user bases. More targeted research is require to confirm these exploratory results.

A second implication of this work is that human-AI teaming, while effective, has unmet potential, and that the choice of visual indicator can influence users' willingness to incorporate AI feedback. Our work speaks to previous studies where pairing humans with high-performing deepfake detection models achieved performance well below ceiling (Groh et al. 2022; Lai and Tan 2019; Boyd et al. 2022). These studies used model performance values that were high, but not perfect. This is ecological (no current model performs at 100% accuracy), but it introduces a confound when trying to assess how visual indicator design affects users' willingness to incorporate AI feedback: people are less likely to cooperate with models that have made errors in the past (Naujoks, Kiesel, and Neukum 2016; Sendelbach and Funk 2013; Yin, Wortman Vaughan, and Wallach 2019). Here, we fixed model performance at 100%, which serves two roles. First, it allows us to observe how visual indicator design can affect behavior without

confounds from source reliability effects. Second, it allows for observation of willingness to incorporate AI feedback in the best possible scenarios. Overall, even when paired with a perfect model, participants achieved performance well below ceiling (64.4% for Caricatures, aggregated across difficulty levels). This adds to the growing literature about a human acceptance gap in human-AI teaming for deepfake detection.

These results also make the case that visual indicator design is one factor in reducing this gap. We tested one particular kind of visual indicator, artifact amplification, and found that it is detectable under a variety of viewing conditions, and that it affects participants' subjective impression of the video more than traditional text-based indicators. This adds to previous results from our group showing that artifact amplification on deepfakes is effective at boosting detection for both high- and low-vigilance individuals, and that it is effective after as little as 500ms of exposure (Fosco et al. 2022). There are several possible reasons that artifact amplification is so effective: it increases the amount of motion in the video, which humans are very perceptually attuned to, and it increases the subjective impression of unnaturalness as the faces change shape over time. Future work is required to untangle these two contributions.

Overall, artifact amplification can be considered part of a broader family of distortion-based visual indicators. Such visual indicators rely on the conscious, deliberate distortion of an image to enable visual observation of an otherwise invisible signal (Le Ngo and Phan 2019; Śmieja et al. 2021). These have existed for some time across a number industrial and civil settings, as a visual aid in quality control applications. For example, motion amplification has been found useful for monitoring vibrations in iron pipes (Kupwade-Patil et al. 2020) and pedestrian bridges (Shang and Shen 2018), and for visualizing the deformation in wind turbine blades (Sarrafi et al. 2018) and antique structures (Fioriti et al. 2018). Motion and color amplification has even been proposed for facilitating the observation of subtle physiological signals like heart rate in infants (Wu et al. 2012; Balakrishnan, Durand, and Guttag 2013).

Deepfake mitigation measures have only recently begun to recognize the perceptual power of distortion. Some approaches actively inject human-invisible artifacts into real images or video, which cause any subsequent video manipulation to contain large and visible artifacts (Chen et al. 2021; Wang et al. 2022). We introduce a complementary, reactive approach, which identifies and amplifies distortions caused by the deepfake-generation pipeline itself. Distortion-based indicators could also be applied to deepfakes identified via metadata-based detection methods (Qureshi, Megías, and Kuribayashi 2021; Neekhara et al. 2022; Chan et al. 2020; Alattar, Sharma, and Scriven 2020; Yu et al. 2021), by injecting artifacts into the video stream. Crucially, we have empirically demonstrated the effectiveness of distortion-based visual indicators in deepfake mitigation, and this principle can be applied regardless of the method used to identify the deepfake.

There are some limitations to the present work. This work uses videos that have no sound. This matches many viewing contexts (e.g., GIFs, platforms set to mute by default), but does not generalize to all contexts (e.g., long-form interviews, news broadcasts), which can include additional information streams, such as the quality of the audio and the semantic content of the speech. Such contexts open up novel research directions, such as testing distortion-based indicators in the audio domain. Additionally, we assessed the effectiveness of the visual indicator based on the accuracy and confidence achieved by the participants, but this only captures responses in the moment, and does not give insight into downstream effects of different visual indicators, such as how they affect memory for the videos or memory for the information they contain.

This work also raises questions about the risks of distortion-based visual indicators

for deepfake signaling. While we show they are effective for signaling deepfakes in the moment, they may harm longer-term information literacy goals. If people only see distorted deepfakes, they may not learn what artifacts exist in unsignaled deepfakes. Widespread distribution of distortion-based visual indicators may also cause a criterion shift, where people become less sensitive to the subtle artifacts in regular deepfakes because they have become accustomed to more obvious visual distortions.

5 Conclusion

We demonstrate how conditions that exist during normal browsing can increase human susceptibility to deepfakes. However, we also demonstrate how human-centered principles can be applied to visual indicator design to increase their effectiveness. We leveraged people's natural sensitivity to distortions in faces by amplifying artifacts in videos, and found that this method of marking fake videos was more convincing than text-based alerts, and led to higher accuracy. More broadly, this paper demonstrates the promise of integrating knowledge about what perceptual tasks are easy and automatic for humans into the development of visual indicators.

References

Abdi, Hervé. 2007. "Signal detection theory (SDT)." *Encyclopedia of Measurement and Statistics,* 886–89. https://doi.org/10.4135/9781412983907.

Agarwal, Shruti, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. "Detecting deep-fake videos from appearance and behavior." In *2020 IEEE International Workshop on Information Forensics and Security (WIFS),* 1–6. IEEE. https://doi.org/10.1109/WIFS49906.2020.9360904.

Alattar, Adnan, Ravi Sharma, and John Scriven. 2020. "A system for mitigating the problem of deepfake news videos using watermarking." *Electronic Imaging* 32:1–10. https://doi.org/10.2352/ISSN.2470-1173.2020.4.MWSF-117.

Balakrishnan, Guha, Fredo Durand, and John Guttag. 2013. "Detecting pulse from head motions in video." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 3430–37. https://doi.org/10.1109/CVPR.2013.440.

Batailler, Cédric, Skylar M. Brannon, Paul E. Teas, and Bertram Gawronski. 2022. "A signal detection approach to understanding the identification of fake news." *Perspectives on Psychological Science* 17 (1): 78–98. https://doi.org/10.1177/1745691620986135.

Benson, Philip J., and David I. Perrett. 1991. "Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images." *European Journal of Cognitive Psychology* 3 (1): 105–35. https://doi.org/10.1080/09541449108406222.

Boháček, Matyáš, and Hany Farid. 2022. "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms." *Proceedings of the National Academy of Sciences* 119 (48): e2216035119. https://doi.org/10.1073/pnas.2216035119.

Boyd, Aidan, Patrick Tinsley, Kevin Bowyer, and Adam Czajka. 2022. "The Value of AI Guidance in Human Examination of Synthetically-Generated Faces." arXiv: 2208.10544 [cs.CV].

Chan, Christopher Chun Ki, Vimal Kumar, Steven Delaney, and Munkhjargal Gochoo. 2020. "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media." In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G),* 55–62. IEEE. https://doi.org/10.1109/AI4G50087.2020.9311067.

Chen, Zhikai, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. 2021. "Magdr: Mask-guided detection and reconstruction for defending deepfakes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 9014–23. https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_MagDR_Mask-Guided_Detection_and_Reconstruction_for_Defending_Deepfakes_CVPR_2021_paper.pdf.

Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. 2020. "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* ISSN: 1939-3539. https://doi.org/10.1109/tpami.2020.3009287.

Cozzolino, Davide, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. "ID-Reveal: Identity-aware deepfake video detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 15108–17. https://openaccess.thecvf.com/content/ICCV2021/papers/Cozzolino_ID-Reveal_Identity-Aware_DeepFake_Video_Detection_ICCV_2021_paper.pdf.

Dolhansky, Brian, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. "The Deepfake Detection Challenge (DFDC) Preview Dataset." arXiv: 1910.08854 [cs.CV].

Durall, Ricard, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2019. "Unmasking DeepFakes with simple Features." arXiv: 1911.00686 [cs.LG].

Farah, Martha J., Kevin D. Wilson, Maxwell Drain, and James N. Tanaka. 1998. "What is 'special' about face perception?" *Psychological Review* 105 (3): 482. https://doi.org/10.1037/0033-295X.105.3.482.

Fioriti, Vincenzo, Ivan Roselli, Angelo Tatì, Roberto Romano, and Gerardo De Canio. 2018. "Motion Magnification Analysis for structural monitoring of ancient constructions." *Measurement* 129:375–80. https://doi.org/10.1016/j.measurement.2018.07.055.

Fosco, Camilo, Emilie Josephs, Alex Andonian, Allen Lee, Xi Wang, and Aude Oliva. 2022. "Deepfake Caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines." arXiv: 2206.00535 [cs.CV].

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): e2110013119. https://doi.org/10.1073/pnas.2110013119.

Gupta, Parul, Komal Chugh, Abhinav Dhall, and Ramanathan Subramanian. 2020. "The eyes know it: FakeET-an eye-tracking database to understand deepfake perception." In *Proceedings of the 2020 International Conference on Multimodal Interaction,* 519–27. https://doi.org/10.1145/3382507.3418857.

Haliassos, Alexandros, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. "Lips don't lie: A generalisable and robust approach to face forgery detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 5039–49. June. https://openaccess.thecvf.com/content/CVPR2021/papers/Haliassos_Lips_Dont_Lie_A_Generalisable_and_Robust_Approach_To_Face_CVPR_2021_paper.pdf.

Hautus, Michael J. 1995. "Corrections for extreme proportions and their biasing effects on estimated values of d′." *Behavior Research Methods, Instruments, & Computers* 27:46–51. https://doi.org/10.3758/BF03203619.

Horrey, William J., Christopher D. Wickens, and Kyle P. Consalus. 2006. "Modeling drivers' visual attention allocation while interacting with in-vehicle technologies." *Journal of Experimental Psychology: Applied* 12 (2): 67. https://doi.org/10.1037/1076-898X.12.2.67.

Hout, Michael C., Stephen C. Walenchok, Stephen D. Goldinger, and Jeremy M. Wolfe. 2015. "Failures of perception in the low-prevalence effect: Evidence from active and passive visual search." *Journal of Experimental Psychology: Human Perception and Performance* 41 (4): 977. https://doi.org/10.3758/BF03203619.

Kagan, Jerome, Barbara A. Henker, Amy Hen-Tov, Janet Levine, and Michael Lewis. 1966. "Infants' differential reactions to familiar and distorted faces." *Child Development,* 519–32. https://doi.org/10.2307/1126676.

Köbis, Nils C., Barbora Doležalová, and Ivan Soraperra. 2021. "Fooled twice: People cannot detect deepfakes but think they can." *iScience* 24 (11): 103364. https://doi.org/10.1016/j.isci.2021.103364.

Korshunov, Pavel, and Sébastien Marcel. 2021. "Subjective and objective evaluation of deepfake videos." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2510–14. IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414258.

Kupwade-Patil, Kunal, Justin G. Chen, Murat Uzun, Denvid Lau, Maranda L. Johnston, Ao Zhou, Dirk Smit, and Oral Büyüköztürk. 2020. "Corrosion assessment of ductile iron pipes using high-speed camera technique: Microstructural validation." *NDT & E International* 116:102362. https://doi.org/10.1016/j.ndteint.2020.102362.

Lago, Federica, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. 2021. "More real than real: A study on human visual perception of synthetic faces." *IEEE Signal Processing Magazine* 39 (1): 109–16. https://doi.org/10.1109/MSP.2021.3120982.

Lai, Vivian, and Chenhao Tan. 2019. "On human predictions with explanations and predictions of machine learning models: A case study on deception detection." In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 29–38. https://doi.org/10.1145/3287560.3287590.

Lakens, Daniël. 2013. "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs." *Frontiers in Psychology* 4:863. https://doi.org/10.3389/fpsyg.2013.00863.

Le Ngo, Anh Cat, and Raphaël C.-W. Phan. 2019. "Seeing the invisible: Survey of video motion magnification and small motion analysis." *ACM Computing Surveys (CSUR)* 52 (6): 1–20. https://doi.org/10.1145/3355389.

Li, Haodong, Bin Li, Shunquan Tan, and Jiwu Huang. 2020. "Identification of deep network generated images using disparities in color components." *Signal Processing* 174:107616. https://doi.org/10.1016/j.sigpro.2020.107616.

Li, Jia, Tong Shen, Wei Zhang, Hui Ren, Dan Zeng, and Tao Mei. 2019. "Zooming into Face Forensics: A Pixel-level Analysis." arXiv: 1912.05790 [cs.CV].

Li, Lingzhi, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2019. "Face X-ray for More General Face Forgery Detection." arXiv: 1912.13458 [cs.CV].

Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. 2018. "In ictu oculi: Exposing AI generated fake videos by detecting eye blinking." In *2018 IEEE International Workshop on Information Forensics and Security (WIFS),* 1–7. IEEE. https://doi.org/10.1109/WIFS.2018.8630787.

Lovato, Juniper, Laurent Hébert-Dufresne, Jonathan St-Onge, Randall Harp, Gabriela Salazar Lopez, Sean P. Rogers, Ijaz Ul Haq, and Jeremiah Onaolapo. 2022. "Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks." arXiv: 2210.10026 [cs.SI].

MacDorman, Karl F., Robert D. Green, Chin-Chang Ho, and Clinton T. Koch. 2009. "Too real for comfort? Uncanny responses to computer generated faces." *Computers in Human Behavior* 25 (3): 695–710. 10.1016/j.chb.2008.12.026.

Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi. 2020. "Explainable deep-fake detection using visual interpretability methods." In *2020 3rd International Conference on Information and Computer Technologies (ICICT),* 289–93. IEEE. https://doi.org/10.1109/ICICT50521.2020.00051.

Matern, Falko, Christian Riess, and Marc Stamminger. 2019. “Exploiting visual artifacts to expose deepfakes and face manipulations.” In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW),* 83–92. IEEE. https://doi.org/10.1109/WACVW.2019.00020.

Mauro, Robert, and Michael Kubovy. 1992. “Caricature and face recognition.” *Memory & Cognition* 20 (4): 433–40. https://doi.org/10.3758/BF03210927.

Mittal, Govind, Chinmay Hegde, and Nasir Memon. 2022. “Gotcha: Real-Time Video Deepfake Detection via Challenge-Response.” arXiv: 2210.06186 [cs.CR].

Mori, Masahiro. 1970. “Bukimi no tani [The uncanny valley].” *Energy* 7:33. https://spectrum.ieee.org/the-uncanny-valley.

Naujoks, Frederik, Andrea Kiesel, and Alexandra Neukum. 2016. “Cooperative warning systems: The impact of false and unnecessary alarms on drivers’ compliance.” *Accident Analysis & Prevention* 97:162–75. https://doi.org/10.1016/j.aap.2016.09.009.

Neekhara, Paarth, Shehzeen Hussain, Xinqiao Zhang, Ke Huang, Julian McAuley, and Farinaz Koushanfar. 2022. “FaceSigns: Semi-Fragile Neural Watermarks for Media Authentication and Countering Deepfakes.” arXiv: 2204.01960 [cs.CV].

Nightingale, Sophie J., and Hany Farid. 2022. “AI-synthesized faces are indistinguishable from real faces and more trustworthy.” *Proceedings of the National Academy of Sciences* 119 (8): e2120481119. https://doi.org/10.1073/pnas.2120481119.

Prasad, Swaroop Shankar, Ofer Hadar, Thang Vu, and Ilia Polian. 2022. “Human vs. Automatic Detection of Deepfake Videos Over Noisy Channels.” In *2022 IEEE International Conference on Multimedia and Expo (ICME),* 1–6. IEEE. https://doi.org/10.1109/ICME52920.2022.9859954.

Qureshi, Amna, David Megías, and Minoru Kuribayashi. 2021. “Detecting deepfake videos using digital watermarking.” In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* 1786–93. IEEE. http://www.apsipa.org/proceedings/2021/pdfs/0001786.pdf.

Rich, Anina N., Melina A. Kunar, Michael J. Van Wert, Barbara Hidalgo-Sotelo, Todd S. Horowitz, and Jeremy M. Wolfe. 2008. “Why do we miss rare targets? Exploring the boundaries of the low prevalence effect.” *Journal of Vision* 8 (15): 15–15. https://doi.org/10.1167/8.15.15.

Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. “Faceforensics++: Learning to detect manipulated facial images.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 1–11. https://doi.org/10.1109/ICCV48922.2021.01418.

Rössler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. “FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces.” arXiv: 1803.09179 [cs.CV].

Sarrafi, Aral, Zhu Mao, Christopher Niezrecki, and Peyman Poozesh. 2018. “Vibration-based damage detection in wind turbine blades using Phase-based Motion Estimation and motion magnification.” *Journal of Sound and Vibration* 421:300–318. https://doi.org/10.1016/j.jsv.2018.01.050.

Sendelbach, Sue, and Marjorie Funk. 2013. “Alarm fatigue: a patient safety concern.” *AACN Advanced Critical Care* 24 (4): 378–86. https://doi.org/10.4037/NCI.0b013e3182a903f9.

Seyama, Jun'ichiro, and Ruth S. Nagayama. 2007. "The uncanny valley: Effect of realism on the impression of artificial human faces." *Presence* 16 (4): 337–51. https://doi.org/10.1162/pres.16.4.337.

Shang, Zhexiong, and Zhigang Shen. 2018. "Multi-point vibration measurement and mode magnification of civil structures using video-based motion processing." *Automation in Construction* 93:231–40. https://doi.org/10.1016/j.autcon.2018.05.025.

Shen, Bingyu, Brandon RichardWebster, Alice O'Toole, Kevin Bowyer, and Walter J. Scheirer. 2021. "A study of the human perception of synthetic faces." In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021),* 1–8. IEEE. https://doi.org/10.1109/FG52635.2021.9667066.

Sinha, Pawan, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. "Face recognition by humans: Nineteen results all computer vision researchers should know about." *Proceedings of the IEEE* 94 (11): 1948–62. https://doi.org/10.1109/JPROC.2006.884093.

Śmieja, Michał, Jarosław Mamala, Krzysztof Prażnowski, Tomasz Ciepliński, and Łukasz Szumilas. 2021. "Motion Magnification of Vibration Image in Estimation of Technical Object Condition-Review." *Sensors* 21 (19): 6572. https://www.mdpi.com/1424-8220/21/19/6572.

Sohrawardi, Saniat Javid, Sovantharith Seng, Akash Chintha, Bao Thai, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2020. "Defaking DeepFakes: Understanding journalists' needs for DeepFake detection." In *Proceedings of the Computation+ Journalism 2020 Conference, Northeastern University, Boston, MA, USA,* vol. 21. https://bpb-us-w2.wpmucdn.com/sites.northeastern.edu/dist/d/53/files/2019/11/CJ_2020_paper_64.pdf.

Strong, Roger W., and George Alvarez. 2019. "Using simulation and resampling to improve the statistical power and reproducibility of psychological research." *Psyarxiv,* https://doi.org/10.31234/osf.io/2bt6q.

Tucciarelli, Raffaele, Neza Vehar, Shamil Chandaria, and Manos Tsakiris. 2022. "On the realness of people who do not exist: The social processing of artificial faces." *iScience* 25 (12): 105441. https://doi.org/10.1016/j.isci.2022.105441.

Wang, Run, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. 2022. "Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations." arXiv: 2206.00477 [cs.CR].

Wolfe, Jeremy M., Todd S. Horowitz, Michael J. Van Wert, Naomi M. Kenner, Skyler S. Place, and Nour Kibbi. 2007. "Low target prevalence is a stubborn source of errors in visual search tasks." *Journal of Experimental Psychology: General* 136 (4): 623. https://doi.org/10.1037/0096-3445.136.4.623.

Wu, Hao-Yu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. 2012. "Eulerian video magnification for revealing subtle changes in the world." *ACM Transactions on Graphics (TOG)* 31 (4): 1–8. https://doi.org/10.1145/2185520.2185561.

Yamani, Yusuke, Pınar Bıçaksız, James Unverricht, and Siby Samuel. 2018. "Impact of information bandwidth of in-vehicle technologies on drivers' attention maintenance performance: A driving simulator study." *Transportation Research Part F: Traffic Psychology and Behaviour* 59:195–202. https://doi.org/10.1016/j.trf.2018.09.004.

Yang, Xin, Yuezun Li, and Siwei Lyu. 2019. "Exposing deep fakes using inconsistent head poses." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 8261–65. IEEE. https://doi.org/10.1109/ICASSP.2019.8683164.

Yin, Ming, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. "Understanding the effect of accuracy on trust in machine learning models." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–12. https://doi.org/10.1145/3290605.3300509.

Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2021. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data." In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 14448–57. https://doi.org/10.1109/ICCV48922.2021.01418.

Authors

Emilie Josephs (ejosephs@mit.edu) is a Postdoctoral Researcher at MIT CSAIL.

Camilo Fosco is a Graduate Researcher at MIT CSAIL.

Aude Oliva (oliva@mit.edu) is a Senior Research Scientist at MIT CSAIL.

Acknowledgements

We thank Jeremy Wolfe, Ruth Rosenholtz, Vasha Dutell, Emma Stewart, Ben Lahner, Alex Lascelles, Bowen Pan, and Alex Andonian for helpful comments and discussion throughout the process, and Alex Lascelles for help administering the experiments and proofreading the paper.

Data availability statement

Replication files are available on the Open Science Framework at https://osf.io/nbzc8/. The following experiments had preregistration posted on AsPredicted.org: Baseline, Low Prevalence, Brief Presentation, Divided Attention, Noisy Video (https://aspredicted.org/63L_W8X, https://aspredicted.org/VX6_TDZ, https://aspredicted.org/132_1CL, https://aspredicted.org/ZV1_2JY, https://aspredicted.org/VKK_W35). Preregistration documents were prepared for Study 2, but never posted online due to an oversight.

Funding statement

This work was funded by an NSF Secure and Trustworthy Cyberspace grant (CNS-2319025) to Aude Oliva, an Ignite grant from SystemsThatLearn@CSAIL to Aude Oliva and Camilo Fosco, an EECS MathWorks Fellowship to Camilo Fosco, and funding from Eric Yuan, CEO and founder of Zoom Video Communications.

Ethical standards

Participants were recruited and compensated according to procedures approved by MIT's Committee on the Use of Humans as Experimental Subjects, under IRB Protocol number 2205000642.

Keywords

Deepfakes; Misinformation mitigation; Human AI teaming; Decision support systems; Human factors.

Appendices

Appendix A: Signal detection measures for non-signaled deepfakes across viewing conditions

Viewing Condition	Hit Rate	False Alarm Rate	Sensitivity	Criterion
Baseline	0.73	0.28	1.29	−0.01
Low Prevalence	0.55	0.16	1.19	0.48
Brief Presentation	0.65	0.28	1.04	0.11
Divided Attention	0.67	0.31	1.03	0.03
Noisy Video	0.66	0.40	4.57	−0.08

Appendix B: Complete statistical reporting of sensitivity difference between non-signaled deepfakes, and deepfakes signaled using Caricatures

Viewing Condition	*t*	*p*	Effect Size
Baseline	30.08	2.50E-69	4.66
Low Prevalence	31.86	3.14E-75	4.79
Brief Presentation	29.62	1.00E-70	4.41
Divided Attention	22.73	1.56E-54	3.41
Noisy Video	22.73	9.84E-76	4.57

Appendix C: Signal detection measures for deepfakes signaled with Caricatures across viewing conditions

Viewing Condition	Hit Rate	False Alarm Rate	Sensitivity	Criterion
Baseline	0.95	0.03	3.75	0.13
Low Prevalence	0.95	0.02	3.85	0.22
Brief Presentation	0.94	0.04	3.07	0.10
Divided Attention	0.93	0.08	3.06	−0.009
Noisy Video	0.93	0.06	3.32	0.05

Figure 6 on the following page shows a graphical representation of differences from Baseline across viewing conditions for Caricatures. Light blue boxes indicates standard error of the mean, and stars indicate significance (post hoc Bonferroni corrected two-sample t-test). In contrast to non-signaled deepfakes, viewing conditions only reduce hit rates in one case, when the viewer is distracted.

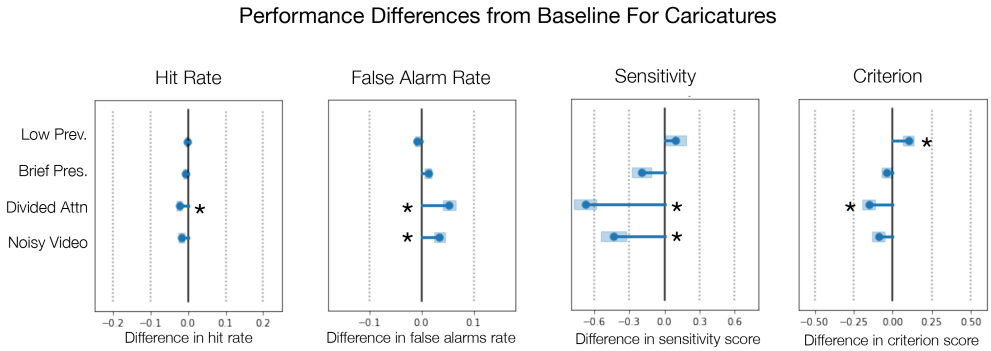

Performance Differences from Baseline For Caricatures
Hit Rate
False Alarm Rate
Sensitivity
Criterion
Low Prev.
Brief Pres.
Divided Attn
Noisy Video
−0.2 −0.1 0.0 0.1 0.2
Difference in hit rate
−0.1 0.0 0.1
Difference in false alarms rate
−0.6 −0.3 0.0 0.3 0.6
Difference in sensitivity score
−0.50 −0.25 0.00 0.25 0.50
Difference in criterion score

Figure 6