Deepfake detection by human crowds, machines, and machine-informed crowds

Matthew Groh[a,1], Ziv Epstein[a], Chaz Firestone[b], and Rosalind Picard[a]

[a]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218

Edited by Thomas Albright, Salk Institute for Biological Studies, La Jolla, CA; received May 29, 2021; accepted November 25, 2021

The recent emergence of machine-manipulated media raises an important societal question: How can we know whether a video that we watch is real or fake? In two online studies with 15,016 participants, we present authentic videos and deepfakes and ask participants to identify which is which. We compare the performance of ordinary human observers with the leading computer vision deepfake detection model and find them similarly accurate, while making different kinds of mistakes. Together, participants with access to the model's prediction are more accurate than either alone, but inaccurate model predictions often decrease participants' accuracy. To probe the relative strengths and weaknesses of humans and machines as detectors of deepfakes, we examine human and machine performance across video-level features, and we evaluate the impact of preregistered randomized interventions on deepfake detection. We find that manipulations designed to disrupt visual processing of faces hinder human participants' performance while mostly not affecting the model's performance, suggesting a role for specialized cognitive capacities in explaining human deepfake detection performance.

misinformation | artificial intelligence | forensic science | wisdom of crowds | facial recognition

How do we tell the difference between the genuine and the artificial? The emergence of deepfakes—videos that have been manipulated by neural network models to either swap one individual's face for another or alter the individual's face to make them appear to say something they have not said—presents challenges both for individuals and for society at large. Whereas a video of an individual performing an action or making a statement has long been one of the strongest pieces of evidence that the relevant event actually occurred, deepfakes undermine this gold standard, with potentially alarming consequences (1–4).

How should we best meet this new challenge of evaluating the authenticity of a video? One approach is to build automated deepfake detection systems that analyze videos and attempt to classify their authenticity. Recent advances in training neural networks for computer vision reveal that algorithms are capable of surpassing the performance of human experts in some complex strategy games (5, 6) and medical diagnoses (7, 8), so we might expect algorithms to be similarly capable of outperforming people at deepfake detection. Indeed, such computational methods often surpass human performance in detecting physical implausibility cues (9), such as geometric inconsistencies of shadows, reflections, and distortions of perspective (10–12). Similarly, face recognition algorithms often outperform forensic examiners (who are significantly better than ordinary people) at identifying whether pairs of face images show the same or different people (13). This focus on automating the analysis of visual content has advantages over certain methods from traditional digital media forensics, which often rely on image metadata (14) that are not available for many of today's most concerning deepfakes, which typically appear first on social media platforms stripped of such metadata (15, 16). Moreover, metadata from an individual's decision to share on social media may not be a reliable predictor of media's veracity because social media tends to focus people's attention on factors other than truth and accuracy (17, 18).

The artificial intelligence (AI) approach to classifying videos as real or fake focuses on developing large datasets and training computer vision algorithms on these datasets (19–31). The largest open-source dataset is the Deepfake Detection Challenge (DFDC) dataset, which consists of 23,654 original videos showing 960 consenting individuals and 104,500 corresponding deepfakes produced from the original videos. The first frames of both a deepfake and original video from this dataset appear in Fig. 1. The deepfakes examined here contain only visual manipulations produced using seven synthetic techniques: two deepfake autoencoders, a neural network face swap model (32), the neural talking heads (NTH) model (33), the face swapping generative adversarial network (FSGAN) for reenactment and inpainting (34), StyleGAN for generating synthetic faces (35), and sharpening refinement on blended faces (31). Unlike viral deepfake videos of politicians and other famous people, the videos from the competition have minimal context: They are all 10-s videos depicting unknown actors making uncontroversial statements in nondescript locations. As such, the cues for discerning real from fake can be based only on visual cues and not auditory cues or background knowledge of an individual or the topic they are discussing. In a contest run from 2019 to 2020, The Partnership for AI, in collaboration with large companies including Facebook, Microsoft, and Amazon, offered $1,000,000 in prize money to the most accurate deepfake detection models on the DFDC holdout set via Kaggle, a machine learning competition website. A total of 2,116 teams submitted computer vision models to the competition, and the leading model achieved an accuracy score of 65% on the 4,000

Significance

The recent emergence of deepfake videos raises theoretical and practical questions. Are humans or the leading machine learning model more capable of detecting algorithmic visual manipulations of videos? How should content moderation systems be designed to detect and flag video-based misinformation? We present data showing that ordinary humans perform in the range of the leading machine learning model on a large set of minimal context videos. While we find that a system integrating human and model predictions is more accurate than either humans or the model alone, we show inaccurate model predictions often lead humans to incorrectly update their responses. Finally, we demonstrate that specialized face processing and the ability to consider context may specially equip humans for deepfake detection.

Author contributions: M.G., Z.E., C.F., and R.P. designed research; M.G., Z.E., C.F., and R.P. performed research; M.G. analyzed data; and M.G., Z.E., C.F., and R.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: groh@mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2110013119/-/DCSupplemental.

Published December 28, 2021.

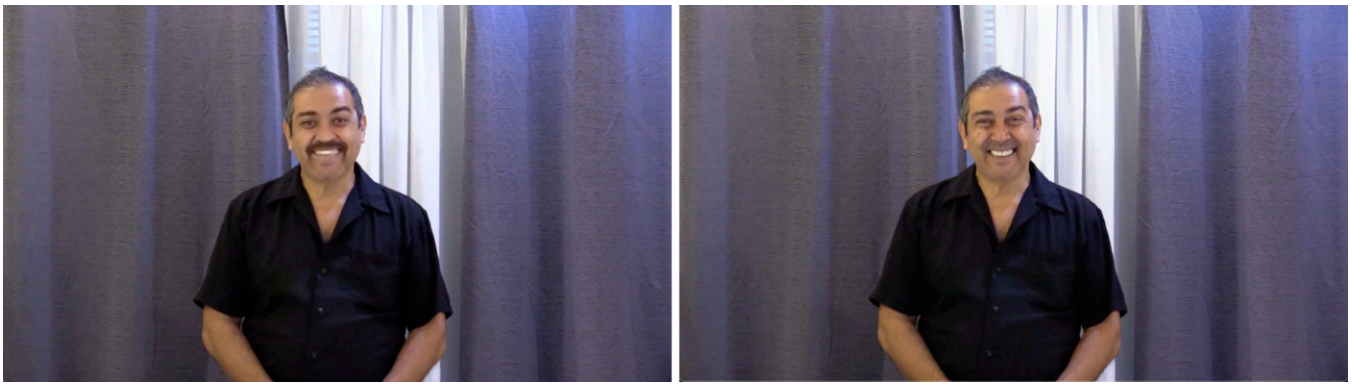

Fig. 1. One of these two images is the first frame of a deepfake from experiment 1; the other is the first frame of the original, authentic video from which the deepfake was created. Experiment 1 asked whether participants can tell which is which, using a two-alternative forced-choice paradigm (i.e., selecting which of two video clips is a deepfake). Experiment 2 presented a single video and asked participants for their confidence the video is a deepfake or not. (*Left*) The deepfake; the man was not mustachioed at the time of filming. (*Right*) Authentic.

videos in the holdout data, which consisted of half deepfake and half real videos (31, 36). While there are many proposed techniques for algorithmically detecting fakes (including affective computing approaches like examining heart rate and breathing rate (37) and looking for emotion-congruent speech and facial expressions) (38, 39), the most accurate computer vision model in the contest (40) focused on locating faces in a sample of static frames using multitask cascaded convolutional neural networks (41), conducting feature encoding based on EfficientNet B-7 (42), and training the model with a variety of transformations inspired by albumentations (43) and grid mask (44). Based on this model outperforming 2,115 other models to win significant prize money in a widely publicized competition on the largest dataset of deepfakes ever produced, we refer to this winning model as the "leading model" for detecting deepfakes to date.

The rules of the competition strictly forbid human-in-the-loop approaches, which leaves open questions surrounding how well human–AI collaborative systems would perform at discerning between manipulated and authentic videos. In this paper, we address the following questions: How accurately do individuals detect deepfakes? Is there a "wisdom of the crowds" (45, 46) effect when averaging participants' responses for each video? How does individual performance compare with the wisdom of the crowds, and how do these performances compare to the leading model's performance? Does access to the model's predictions and certainty levels help or hinder participants' discernment? And, what explains variation in human and machine performance; specifically, what is the role of video-level characteristics, can emotional priming influence participants' performance at detecting deepfakes, and does specialized processing of faces play a role in human and machine deepfake detection?

Crowdsourcing and averaging individuals' responses are promising and practical solutions for handling the scale of misinformation that would be otherwise overwhelming for an individual expert. Recent empirical research finds that averaged responses of ordinary people are on par with third-party fact-checkers for both factual claims in articles (47) and overall accuracy of content from URL domain names (48, 49). In order to comprehensively compare humans to the leading AI model and evaluate collective intelligence against its artificial counterpart, we need to conduct two comparisons: How often do individuals outperform the model, and how often does crowd wisdom outperform the model's prediction?

While a machine will consistently predict the same result for the same input, human judgment depends on a range of factors, including emotion. Recent research in social psychology suggests that negative emotions can reduce gullibility (50, 51), which could perhaps improve individual's discernment of videos. In particular, anger has been shown to reduce depth of thought by promoting reliance on stereotypes and previously held beliefs (52). Moreover, priming people with emotion has been demonstrated to both increase and decrease people's gullibility, depending on the category of emotion (53), and hinder people's ability to discern real from fake news (54). The role of emotion in deepfake detection is of practical concern because people share misinformation, especially political misinformation, because of its novelty and emotional content (55). While a detailed examination of emotions as potential mechanisms to explain deepfake detection performance is outside the scope of this paper, we have included a preregistered randomized experiment to evaluate whether experimentally elicited anger impairs participants' performance in detecting deepfakes.

Based on research demonstrating humans' expert visual processing of faces, we may expect humans to perform quite well at identifying the synthetic face manipulations in deepfake videos. Research in visual neuroscience and perceptual psychology has shown that the human visual system is equipped with mechanisms dedicated for face perception (56). For example, there is a region of the brain specialized for processing faces (57). Human infants show sensitivity to faces even before being exposed to them (58, 59), and adults are less accurate at recognizing faces when images are inverted or contain misaligned parts (60–62). The human visual system is faster and more efficient at locating human faces than other objects, including objects with illusory faces (63). Whether human visual recognition of faces is an innate ability or a learned expertise through experience, visual processing of faces appears to proceed holistically for the vast majority of people (64, 65). In order to examine specialized processing of faces as a potential mechanism explaining deepfake detection performance, we include a randomized experiment where we obstruct specialized face processing by inverting, misaligning, and occluding videos.

In order to answer questions about human and machine performance at deepfake detection, we designed and developed a website called Detect Fakes where anyone on the internet could view deepfake videos sampled from the DFDC and see for themselves how difficult (or easy) it is to discern deepfakes from real videos. On this website, we conducted two randomized experiments to evaluate participants' ability to discern real videos from deepfakes and examine cognitive mechanisms explaining human and machine performance at detecting fake videos. We present a screenshot of the user interface of these two experiments in *SI Appendix*, Fig. S4. In the first experiment, we presented a two-alternative forced choice design where a deepfake video is presented alongside its corresponding real video. In the second experiment, we presented participants with

a single video design and asked them to share how confident (from 50 to 100% in one percentage point increments) they are that the video is a deepfake (or is not a deepfake). In this single video framework, we present participants with the option to update their confidence after seeing the model's predicted likelihood that a video is a deepfake. By doing so, we evaluate how machine predictions affect human decision-making. In both experiments, we embedded randomized interventions to evaluate whether incidental emotion (emotion unrelated to the task at hand) or obstruction of specialized processing of faces influence participants' performance.

Results

Experiment 1: Two-Alternative Forced Choice (n = 5,524). In experiment 1, 5,524 individuals found our website organically and participated in 26,820 trials. The 56 pairs of videos in experiment 1 were sampled from the DFDC training dataset because the experiment was conducted before the holdout videos for the DFDC dataset were publicly released. As such, we compare participants' performance in experiment 1 to the overall performance of the leading model. We leave a direct comparison of participant and model performance for experiment 2, which focuses on performance across holdout videos.

Individual vs. machine. As stated in our pre-analysis plan for Experiment 1, we examined the accuracy of all participants who saw at least 10 pairs of videos, for a total of 882 participants.* Eighty-two percent of participants outperform the leading model, which achieves 65% accuracy on the holdout dataset (36). Half of the stimuli set (28 of 56 pairs of videos) was identified correctly by over 83% of participants, 16 pairs of videos were identified correctly by between 65% and 83% of participants, and 12 pairs of videos were identified correctly by less than 65% of participants. Out of these 12 pairs of videos, 3 pairs of videos were identified correctly by less than 50% of participants. Fig. 2*A* presents the distribution of participants' performance in experiment 1 (in blue in the second column) next to the model's overall performance (in black in the first column).

We do not find any evidence that participants improve in their ability to detect these videos within the first 10 videos seen ($P = 0.112$) (all P values reported in this paper are generated by linear regression with robust SEs clustered on participants unless otherwise stated). On average, participants took 42 s to respond to each pair of videos, and we find that, for every additional 10 s participants take to respond, participants' accuracy decreased by 1.1 percentage points ($P < 0.001$). We embedded three randomized experiments in experiment 1 to evaluate the roles of specialized processing of faces, time for reflection, and emotion elicitation. We find participants are 5.6 percentage points less accurate ($P = 0.004$) at detecting pairs of inverted videos than pairs of upright videos. In contrast, we do not find statistically significant effects of the additional time for reflection intervention or this particular emotion elicitation intervention. The custom emotion elicitation intervention in this first experiment did not have a statistically significant influence on participants' self-reported emotions, which suggests the custom emotion elicitation experiment did not work here. We provide additional details on the interventions in experiment 1 in *SI Appendix*.

Experiment 2: Single-Video Design (n = 9,492). In experiment 2, 9,492 individuals participated: 304 individuals were recruited from Prolific and completed 6,390 trials; 9,188 individuals found our website organically and completed 67,647 trials.† In the recruited cohort, all but three participants viewed 20 videos. In the nonrecruited cohort, over half of participants viewed seven videos, and the 90th percentile participant viewed 17 videos. The website instructed participants, about videos, that "half are deepfakes, half are not." After viewing each video, participants move a slider to report their response ranging from "100% confidence this is NOT a DeepFake" to "100% confidence this is a DeepFake" in 1% increments with "just as likely a DeepFake as not" in the middle (see *SI Appendix*, Fig. S4 for a screenshot of the user interface). Participants can never make a selection with less than 50% confidence; the slider's default position is in the center (at the "just as likely a DeepFake as not" position); one increment to the right becomes "51% confidence this is a DeepFake," and one increment to the left becomes "51% confidence this is NOT a DeepFake." The stimuli in experiment 2 consist of 50 videos randomly sampled from the competition holdout dataset (half deepfake and half nonmanipulated), 4 videos of Kim Jung-un and Vladimir Putin, including one deepfake and one nonmanipulated video of each leader, and a deepfake attention check video.

In experiment 2, we define the accuracy score as the participant's response between zero and one normalized for correctness, which is the participant's response if correct, or one minus the participant's response if incorrect. For example, if a participant responded "82% confidence this is a DeepFake" and the participant is correct, then the participant is assigned an accuracy score of 0.82. If the participant is incorrect, then the participant would be assigned an accuracy score of 0.18. We define accurate identification as an accuracy score greater than 0.5.

Participants' and the leading model's performance on deepfake detection depends on the population of participants, the population of videos, how performance is measured at the individual or collective level, and whether videos are presented side by side or by themselves. In some cases, we find a machine advantage, and, in others, we find a human advantage. The rest of *Results* examines individual participant performance compared with the leading model, participants' collective performance compared with the leading model, participants' collective performance when participants have access to the model's predictions, variations in human and machine performance across videos, and randomized experiments designed to evaluate the role of emotional priming and specialized visual processing of faces.

Individual vs. machine. For participants who pass the attention check, recruited participants accurately identified deepfakes from the randomly sampled holdout videos in 66% of attempts, while the nonrecruited participants accurately identified videos in 69% of trials (or 72% of attempts when limiting the analysis to nonrecruited participants who saw at least 10 videos). In comparison, the leading model accurately identified deepfakes on 80% of the sampled videos, which is significantly better than the 65% accuracy rate this model achieves on the full holdout dataset of 4,000 videos (36).

In a direct comparison of performance, 13% of recruited participants, 27% of nonrecruited participants who saw at least 10 videos, and 37% of nonrecruited participants who saw fewer than 10 videos outperform the model. Fig. 2*A* presents the distribution of participants' accuracy on the sampled holdout videos (in teal for recruited participants and gold for nonrecruited participants). Relative to the leading model, participants are less accurate at identifying deepfakes than they are at identifying real videos. Recruited participants accurately identify deepfakes as deepfakes in 57% of attempts compared to the leading model identifying deepfakes as deepfakes in 84% of videos, while both recruited participants and the leading model identify real videos as real videos at nearly same rate (75% of participants' observations and 76% of videos). Recruited participants predicted the sampled holdout videos were real (57% of observation) considerably more often than fake (38% of observations), while the computer vision

*Pre-analysis plan for nonrecruited participants in experiment 1 is available at https://aspredicted.org/g6497.pdf.

†Pre-analysis plan for recruited individuals participating in experiment 2 is available at https://aspredicted.org/hj6wb.pdf.

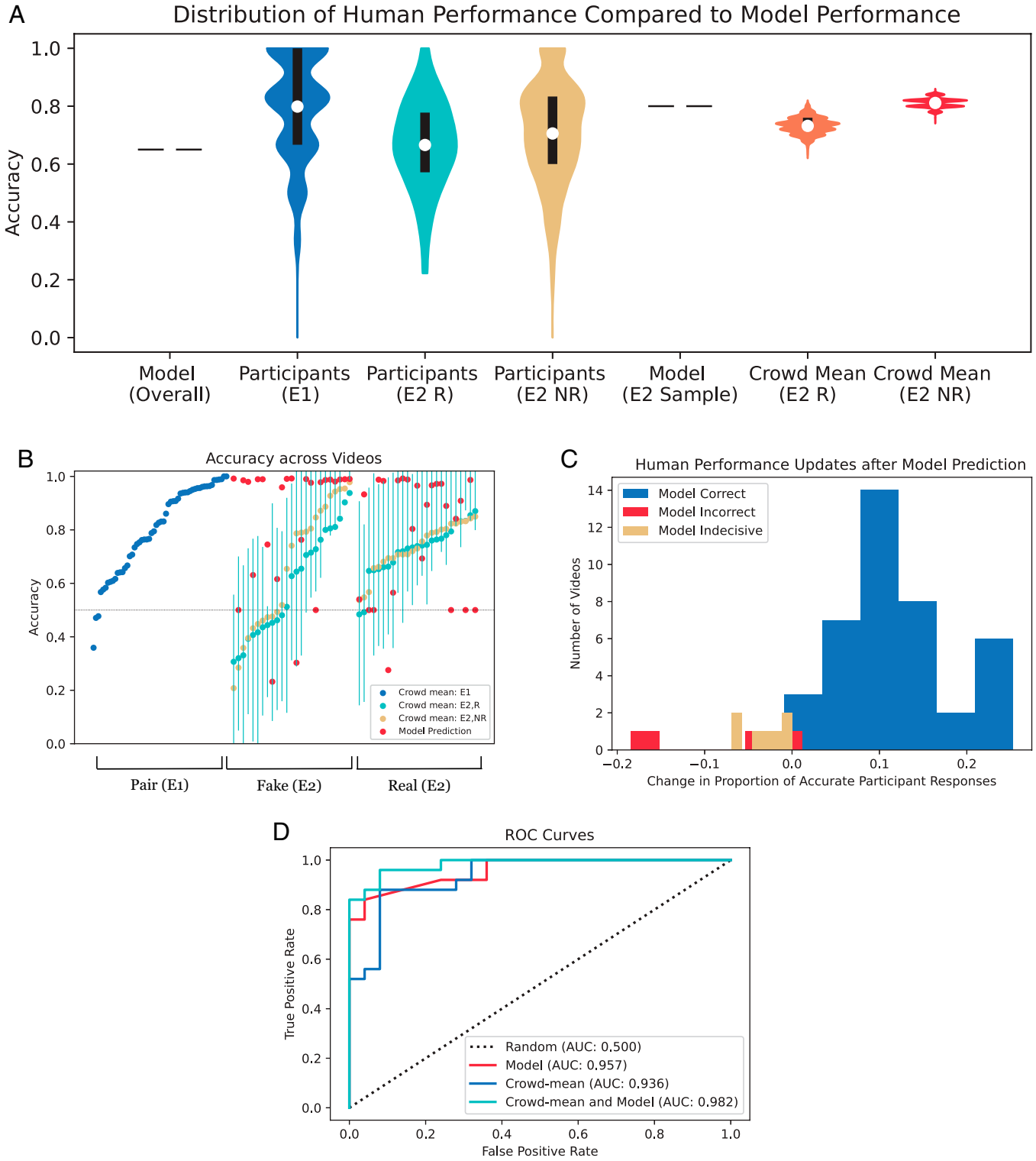

Fig. 2. (*A*) The distribution of participant performance across experiments compared to the model's performance via violin plots where the white dots indicate the mean and the black bars indicate the interquartile range. R, recruited participants; NR, nonrecruited participants; E1, experiment 1; E2, experiment 2. In experiment 1 (two-alternative forced choice), accuracy is defined as identifying a deepfake from a pair of videos correctly. In experiment 2 (single-video design), accurate identification is defined as responding with the correct answer with more than 50% confidence. The model's performance represents a single observation in each instance, and, as such, we present the model's performance as a horizontal black line with a white dot in the middle. The crowd mean distributions are obtained by bootstrapping CIs based on 1,000 randomly drawn samples that are each half of the total observations. (*B*) A scatter plot of the model's accuracy and the distribution of participants' accuracy scores for each video. The *x* axis is an index of the videos, and it is ordered by experiment, true class of each video, and participant's average accuracy. The teal lines represent the interquartile range of recruited participants' responses. (*C*) The distribution of changes in recruited participants' accuracy after updating their response based on whether the model's prediction is correct, incorrect, or indecisive. (*D*) The receiver operator characteristic curves of computer performance, recruited participants' collective performance, and recruited participants' collective performance with the model's decision support across the 50 DFDC videos in experiment 2.

model predicted videos were real (44% of observations) barely more frequently than fake (42% of observations). In 5% of recruited participant observations and 14% of computer vision model observations, the prediction was a 50–50 split between real and fake. We report confusion matrices for each treatment condition in *SI Appendix*, Tables S3–S7.

On the additional set of videos of political leaders, participants outperform the leading model. Specifically, 60% of recruited participants and 68% of nonrecruited participants who saw at least 10 videos outperform the model on these videos. For the deepfake videos of Kim Jong-un, Vladimir Putin, and the attention check, the state-of-the-art computer vision model outputs a 2%, 8%, and 1% probability score that each respective video is a deepfake, which is both confident and inaccurate.

We do not find any evidence that participants' overall accuracy changes as participants view more videos ($P = 0.433$). However, we find that, for every additional video seen by recruited participants, they are 0.9% ($P < 0.001$) more likely to report any video as a deepfake. This corresponds to performing about 18% better at detecting deepfakes and 18% worse at identifying real videos by the last video.

Recruited participants spent a median duration of 22 s before submitting their initial guess and a median duration of 3 s adjusting (or not adjusting) their initial guess when prompted with the model's predicted likelihood. Nonrecruited participants spend a similar amount of time. For both sets of participants, we find that, for every 10 additional seconds of participant response time, participants' accuracy decreases by one percentage point ($P < 0.001$).

Crowd wisdom vs. machine. The crowd mean, participants' responses averaged per video, is on par with the leading model performance on the sampled holdout videos. For recruited participants, the crowd mean accurately identifies 74% of videos. For nonrecruited participants, the crowd mean accurately identifies 80% of videos. For the 1,879 nonrecruited participants who saw at least 10 videos, the crowd mean is 86% accurate. In comparison, the leading model accurately identifies 80% of videos.

In Fig. 2*B*, we compare statistics on participants' accuracy (the mean and interquartile range) with the model's predictions for each video. In *SI Appendix*, Table S2, we present the mean accuracy of recruited and nonrecruited participants and the computer vision model for all videos. There are two videos (both deepfakes) on which both the crowd mean and the leading model are at or below the 50% threshold. One of these videos (video 7837) is extremely blurry, while the other video (video 4555) is filmed from a low angle, and the actress's glasses show significant glare.

There are 8 videos on which the crowd mean is accurate but the model is at or below the 50% threshold and another 5 to 13 videos on which the model is accurate but the crowd mean (depending on how the population selected) is below the 50% threshold.

Human–AI collaboration. In addition to comparing individual and collective performance to the leading model's performance, we examined how an AI model could complement human performance. After participants submitted their initial response for how confident they are that a video is or is not a deepfake in experiment 2, we revealed the likelihood that the video is a deepfake—as predicted by the leading model—and gave participants a chance to update their response. After taking into account the model's prediction, participants updated their confidence in 24% of trials (crossing the 50% threshold for accurate identification in 12% of trials). By updating their responses, recruited participants' accurate identification increased from 66 to 73% of observations ($P < 0.001$ based on a Student's t test). Fig. 2*C* presents the distribution of changes in overall participant accuracy for the 50 videos sampled from the DFDC. For the 40 videos upon which the model accurately identifies the video as a deepfake or not, participants updated their responses to be, on average, 10.4% more accurate at identification than before seeing the model's prediction. For the remaining 10 videos on which the model made an incorrect or equivocal prediction, participants updated their responses to be, on average, 2.7% less accurate at identification than before seeing the model's predictions. In the most extreme example of incorrect updating, the model predicted a 28% likelihood the video was a deepfake when it was indeed a deepfake, and participants updated their responses to be, on average, 18% less accurate at identifying the deepfake. This particular video (video 7837) is quite blurry, and, perhaps, participants changed their responses because it's very difficult to discern manipulations in low-quality video.

For the additional deepfake videos of Kim Jung-un and Vladimir Putin that are not included in the overall analysis, the model predicted a 2% and 8% likelihood, respectively, that the video was a deepfake. This prediction is not only incorrect but confidently so, which led participants to update their responses such that participants' accurate identification dropped from 56 to 34% on the Kim Jung-un deepfake and 70 to 55% on the Vladimir Putin deepfake.

In Fig. 2*D*, the receiver operating characteristic (ROC) curve of the leading model is plotted alongside the ROC curves of the crowd mean and the crowd mean responses where participants have access to the model's prediction for each video. While the model has a slightly higher area under the receiver operating characteristic curve (AUC) score of 0.957 relative to the crowd mean's AUC score of 0.936, either the model or the crowd mean could be considered to perform better, depending on the acceptable false positive rate. However, the crowd mean response after seeing the model's predictions strictly outperforms both the crowd mean and the leading model. When we condition the ROC analysis on confidence following methods for estimating the reliability of eyewitness identifications (66), we find that medium- and high-confidence responses outperform low-confidence responses by a large degree. We define low confidence as responses between 33.5 and 66.5, medium confidence as responses between 17 and 33.5 or 66.5 and 83, and high confidence as responses between 0 and 17 or 83 and 100. *SI Appendix*, Fig. S2 ROC curves present a visual comparison of model performance to low-, medium-, and high-confidence responses from participants, which reveals medium- and high- (but not low-) confidence responses can outperform the model's predictions, depending on the acceptable false positive rate.

Video features correlated with accuracy. Given the heterogeneity in both participants' and the leading model's performance on videos, we extend the analysis of performance across seven video-level features: graininess, blurriness, darkness, presence of a flickering face, presence of two people, presence of a floating distraction, and the presence of an individual with dark skin. These video-level characteristics were hand-labeled by the research team. On the 14 videos that are either grainy, blurry, or dark, the crowd wisdom of recruited participants is correct on 8 videos, while the crowd wisdom of nonrecruited participants and the model is correct on 10 videos. When we examine the 36 videos that are neither grainy, blurry, nor extremely dark, the crowd wisdom of recruited participants is correct on 29 out of 36 videos, the crowd wisdom of nonrecruited participants is correct on 32 out of 36 videos, and the model is correct on 30 of 36 videos. The presence of a flickering face is associated with an increase in recruited participants' accuracy rates by 24.2 percentage points ($P < 0.001$) and an increase in the model's accuracy rates by 21.7 percentage points ($P = 0.170$) in detecting a deepfake. The presence of two people in a video instead of a single person is associated with an overall increase in recruited participants' accuracy rates by 7.6% ($P < 0.001$) and a 21.9% decrease ($P = 0.023$) by the model in identifying real videos. The presence of a floating distraction is associated with a decrease in recruited participants' accuracy rates on real videos of 3.5%

($P = 0.034$) and an increase in recruited participants' accuracy rates on fake videos of 11.3% ($P < 0.001$). In 12 of 50 videos, at least one person in the video has dark skin (precisely defined as skin classified as type 5 or 6 on the Fitzpatrick scoring system, which is a classification system developed for dermatology that computer vision researchers have used to examine the context of skin color) (67). We find that the presence of an individual with dark skin in the video is associated with a decrease in recruited participants' accuracy by 8.8% ($P < 0.001$) and a decrease in the model's accuracy by 12.0% ($P = 0.192$). In order to control for these seven comparisons conducted simultaneously, we can apply a Bonferroni correction of 1/7 to the standard statistical significance thresholds (e.g., a *P* value threshold of 0.01 becomes 0.0014). Based on this correction, the influence of a flickering face, two people in the same video, floating distractions, and the presence of an individual with dark skin continue to be statistically significant for participants if the original *P* value threshold is chosen as 0.01.

Randomized experiments for evaluating emotion priming and specialized face processing. Within experiment 2, we embedded two randomized experiments to examine potential cognitive mechanisms underpinning how humans discern between real and fake videos. Specifically, we examine an affective intervention designed to elicit anger based on a well-established intervention (68) (*SI Appendix*, Fig. S3) and a perceptual intervention designed to obstruct specialized processing of faces via inversion (videos presented upside down), misalignment (videos presented with actors' faces horizontally split), and occlusion (videos presented with a black bar over the actors' eyes).

We present results of the anger elicitation intervention in columns 1 to 3 in Table 1. We do not find statistically significant effects ($P = 0.280$) of the anger elicitation intervention on overall accuracy. However, in our preregistered follow-up analysis limiting the dataset to real videos, we find that participants who were assigned to the anger elicitation treatment underperformed control participants by 5.2 percentage points ($P = 0.032$). In other words, participants in the anger elicitation treatment are more 5.2 percentage points more likely than participants in the control group to make a false positive identification that a real video is a deepfake. Notice here that the floor is not 0% accuracy but rather 50% accuracy (i.e., chance responding); the maximum effect of the anger elicitation treatment is 21.6 percentage points (71.6 from the constant term in column 2 of Table 1 minus 50), so a 5.2 percentage point reduction represents an effect that is 24.1% of the maximum possible effects under these conditions.

SI Appendix, Fig. S3 presents accuracy and confidence scores by treatment assignment to visually reveal the heterogeneous effect of anger elicitation on how participants discern between real and fake videos. When we examine the relationship between assignment to the anger elicitation group and how confidently participants guess, we do not find a statistically significant effect ($P = 0.347$). When we examine real videos and the relationship between anger elicitation and updating predictions after seeing what the model would predict, we find that participants assigned to the anger elicitation group are 3.7% ($P = 0.100$) more likely to change their guess to a correct answer than participants assigned to the control group. As a result, we do not find statistically significant effects of anger elicitation on accuracy after participants update their response ($P = 0.246$).

We present results of the perceptual obstruction intervention in columns 1 to 6 in Table 1. We find statistically significant effects of all three specialized processing obstructions on participants' ability to accurately identify deepfakes from authentic videos. The overall effects—reported in column 1 of Table 1—are all statistically significant and range from a decrease of 4.3 percentage points in accuracy for the inversion treatment ($P = 0.002$), to a decrease of 4.4 percentage points in accuracy for the eye occlusion treatment ($P = 0.004$), to a decrease of 6.3 percentage points for the misalignment treatment ($P < 0.001$) on a base rate of 65.5% accuracy when controlling for video fixed effects. In addition, we find that inverting the videos decreases participants' reported confidence scores (absolute distance in guesses from the 50–50 selection) by two percentage points ($P = 0.002$), but we do not find similar decreases in reported confidence on videos with misalignment or occlusion transformations.

In the sample of recruited participants, the specialized face processing obstructions have different effects depending on whether the videos are manipulated or not. When we limit the analysis to the algorithmically manipulated deepfakes (column 3 of Table 1), we do not find statistically significant effects on the inversion treatment ($P = 0.638$), but we do find that the misalignment and eye occlusion treatments show a decrease by 7.7 ($P = 0.002$) and 6.3 ($P = 0.008$) percentage points, respectively, relative to the control videos. In contrast, when we limit the analysis to the other half of videos that have not been manipulated (column 2 of Table 1), we do not find statistically significant effects for misalignment ($P = 0.075$) or eye occlusion interventions ($P = 0.263$), but we find participants' accuracy on inverted authentic videos is 9.1 percentage points lower than when the videos are upright ($P < 0.001$).

The experimental results on nonrecruited participants provide a replication and robustness check for the results on the recruited participants. The results from the nonrecruited participants were not preregistered because we weren't expecting many people to continue visiting our website organically. In fact, 9,188 visitors participated in the single video design between November 2020 and January 2021. In columns 4 through 6 in Table 1, we present the linear regressions results of the specialized face processing obstruction interventions on nonrecruited participants' accuracy. Similar to the results for the recruited sample, we find statistically significant effects ($P < 0.001$) of all three obstruction interventions on ability to accurately discern deepfakes from authentic videos. The number of observations in the nonrecruited sample is over 16 times larger than the number of observations in the recruited sample. Likewise, the number of participants is 33 times larger. These numbers differ because the number of videos seen by participants in the nonrecruited sample varied depending on participants' interest. With a larger sample size, we see statistically significant and negative effects of obstructions on all videos ranging from a 4 percentage point drop in accuracy from the eye occlusion intervention ($P < 0.001$) to a 7 percentage point drop on accuracy from the misalignment intervention ($P < 0.001$). We also find all three treatments decrease participants' confidence scores by one-half to one percentage point ($P < 0.001$).

In the nonrecruited sample, each of the 50 videos were viewed by between 945 and 1,168 participants. We run separate linear regressions for each video and find statistically significant and negative effects at the 1% significance level for inversion in 24 videos, misalignment in 15 videos, and occlusion in 20 videos. Furthermore, we find at least one of these specialized processing obstructions is negative and statistically significant at the 1% significance for 29 of the 50 videos.

In the sample of videos of political leaders, the specialized face processing obstructions had a significant effect on participants' ability to accurately identify the Vladimir Putin deepfake as a deepfake. The misalignment obstruction leads to a drop in accuracy of 20.7 percentage points ($P = 0.001$). Likewise, the occlusion obstruction leads to a drop of 10.3 percentage points ($P = 0.002$), and the inversion obstruction leads to a drop of 5.2 percentage points ($P = 0.072$).

In columns 7 through 9 in Table 1, we present the linear regression results of the specialized face processing obstruction interventions on the model's predictions and find the computer vision model is affected by one specialized face processing obstruction but not the other two. We find the computer vision model's predictive accuracy drops by 12.1 percentage points on

Table 1. Treatment effects of interventions on accuracy

	Dependent variable: Accuracy								
	Recruited			Nonrecruited			Computer		
	All	Real	Fake	All	Real	Fake	All	Real	Fake
Constant	0.655***	0.716***	0.567***	0.679***	0.700***	0.632***	0.813***	0.786***	0.841***
	(0.009)	(0.014)	(0.015)	(0.002)	(0.003)	(0.003)	(0.030)	(0.040)	(0.044)
Inversion	−0.043***	−0.091***	0.010	−0.053***	−0.080***	−0.027***	−0.121***	−0.110*	−0.132**
	(0.014)	(0.021)	(0.021)	(0.004)	(0.006)	(0.006)	(0.042)	(0.056)	(0.063)
Misalignment	−0.061***	−0.042*	−0.077***	−0.070***	−0.056***	−0.084***	0.011	0.000	0.021
	(0.016)	(0.024)	(0.025)	(0.005)	(0.007)	(0.007)	(0.042)	(0.056)	(0.063)
Eye occlusion	−0.044***	−0.023	−0.063***	−0.040***	−0.035***	−0.043***	−0.003	−0.007	0.001
	(0.015)	(0.021)	(0.024)	(0.004)	(0.006)	(0.006)	(0.042)	(0.056)	(0.063)
Anger	−0.020	−0.052**	0.012						
	(0.014)	(0.024)	(0.021)						
Number of participants	229	229	229	7,563	6,368	6,670	0	0	0
Number of guesses (real)	2,349	1,514	835	27,446	18,524	8,922	81	76	5
Number of guesses (deepfake)	1,707	549	1,158	22,766	6,316	16,450	87	7	80
Number of guesses (50–50)	180	68	112	3,713	1,726	1,987	32	17	15
Number of unique videos	50	25	25	50	25	25	50	25	25
Observations	4,236	2,131	2,105	53,925	26,566	27,359	200	100	100
R^2	0.180	0.069	0.225	0.185	0.057	0.273	0.062	0.054	0.073
Adjusted R^2	0.170	0.056	0.215	0.184	0.057	0.272	0.048	0.025	0.044
Residual SE	0.340	0.329	0.350	0.349	0.350	0.346	0.210	0.198	0.222
F statistic	288.686***	164.804***	169.388***	3,687.143***	2,150.874***	4,525.903***	4.337***	1.841	2.514*

Linear regressions on participant data include video fixed effects with Eicker–Huber–White SEs clustered at the participant level. * $P<0.1$; ** $P<0.05$; *** $P<0.01$.

the inverted videos ($P = 0.005$). We do not find a statistically significant difference in accuracy between either the control and misalignment sets of videos ($P = 0.800$) or the control and occlusion sets of videos ($P = 0.944$).

Discussion

How do ordinary human observers compare with the leading deepfake detection models? Our results are at odds with the commonly held view in media forensics that ordinary people have extremely limited ability to detect media manipulations. Past work in the cognitive science of media forensics has demonstrated that people are not good at perceiving and reasoning about shadow, reflection, and other physical implausibility cues (9–12). On first glance, deepfakes and other algorithmically generated images of people (e.g., images generated by StyleGAN) look quite realistic (35). But we show that deepfake algorithms generate artifacts that are perceptible to ordinary people, which may be partially explained by human's specialized visual processing of faces. In contrast to recent research showing that ordinary people quickly learn to detect AI-generated absences in photos (69), we do not find evidence that participants improve in their ability to detect deepfakes.

By showing participants videos of unknown individuals making uncontroversial statements, we focused the truth discernment task specifically on visual perception. The lack of additional context creates a level playing field for a reasonably fair comparison of human and machine vision, because humans cannot also reason about contextual, conceptual clues in these videos (70). In the two-alternative forced-choice paradigm of experiment 1, 82% of participants respond with higher accuracy than the leading model. In the more challenging single-video framework in experiment 2, participants still perform really well, and we find that between 13% and 37% of ordinary people outperform the leading deepfake detection model. When we aggregate participants' responses in experiment 2, we find that collective intelligence, as measured by the crowd mean, is just as accurate as the model's prediction.

In the extension of the experiment to videos of well-known political leaders (Vladimir Putin and Kim Jong-un), participants significantly outperform the leading model, which is likely explained by participants' ability to go beyond visual perception of faces. Unlike the 50 sample holdout videos, participants could critically contemplate the authenticity of the video of the political leader. For example, participants might have considered whether Vladimir Putin or Kim Jong-un speak English, whether they actually sound like they do in the video, and whether such a well-known political figure would say such a thing. Not only do the majority of participants identify the deepfake status of videos of political leaders correctly, but the computer vision model is confident in its wrong predictions. Perhaps the model failed because it was trained on face-swapping deepfake manipulations as opposed to synthetic lip-syncing manipulations. What the evidence shows is that today's leading model does not generalize well to stylistically different videos than the videos on which it has been trained, whereas human deepfake detection abilities do generalize across these different contexts.

The model's predictions helped participants improve their accuracy overall, but whether a participant's accuracy increased depended on whether the model accurately identified the video as a deepfake or not. Participants often made significant adjustments based on the model's predictions, and inaccurate or equivocal model predictions led participants astray in 8 of 10 instances. Moreover, the model's incorrect assessment of the political leader deepfake videos is associated with a decrease in participant accuracy, which is in line with recent empirical research that shows deepfake warnings do not improve discernment of political videos (71). Likewise, these results mirror other recent research revealing human–AI collaborative decision-making does not necessarily lead to more accurate results than either humans or AI alone (72–76).

Videos are heterogeneous, high-dimensional media, and, as a result, participants were accurate on some videos on which the leading model failed, and vice versa. In line with recent research examining perceptual differences between authentic and deepfake videos (77), we identified seven salient dimensions

across the 50 sampled holdout videos to evaluate differences in how participants and the leading model discern authenticity: We find that the leading model performs slightly better than participants on low-quality videos that were categorized as grainy, blurry, and very dark. This differential performance suggests that the model is picking up on low-level details that participants appear to ignore. On the other hand, we find that both recruited and nonrecruited participants attain similar accuracies to the model on standard quality videos. Both participants and the model are quite adept at picking up on flickering faces. The model has trouble discerning between real and deepfake videos when two actors appear in the video, while participants have no trouble in this context. This suggests that the model may be vulnerable to changes in context whereas participants are more robust to varying context. With respect to visual distractions, we find that distractions are associated with participants identifying videos more often as deepfakes. While we showed recruited participants examples of distraction videos that should not be reported as deepfakes, and we explicitly described these distractions in the instructions as not necessarily characteristic of deepfakes, we imagine the results concerning the distraction videos may possibly reflect confusion by the participants. Nonetheless, all reported results are robust to the exclusion of distraction videos. In light of recent research showing intersectional disparities in accuracy of commercial facial recognition software (67) and the impact of race on credibility with deepfakes (78), we examine accuracy on the videos with dark-skin actors. Participants and the leading model are both less accurate on videos with dark-skin actors, but, as we reported in *Results*, we only find a statistically significant difference in participants' performance, not the model's performance.

In experiment 2, we find some evidence for our preregistered hypothesis that anger would impair participants' ability to identify manipulated media. When we elicit incidental anger (i.e., anger unrelated to the task at hand), participants' accuracy at identifying real videos decreases, a pattern that held across almost all videos (see *SI Appendix*, Fig. S3 where participants assigned to the anger elicitation underperform participants assigned to the control in 22 out of 25 real videos, and see *Limitations*). The negative and heterogeneous effect of incidental anger on the discernment of real (but not fake) videos may be related to the negative and heterogeneous effect of emotion priming on accuracy ratings of fake (but not real) news headlines (54). Drawing on Martel et al., 2020, one potential explanation for the negative effects of anger elicitation on the discernment of authentic but not deepfake videos is emotion leading to an overreliance on intuition; in this experiment, if a participant sees something that looks like a deepfake manipulation, then she is unlikely to think the video is real, but, if a participant does not see something that looks like a deepfake manipulation, then he might think he's simply unable to spot the detailed manipulation and may respond based on his intuition that a video is fake rather than whether he clearly saw a manipulation or not.

Both experiments 1 and 2 provide support for the claim that specialized processing of faces helps people discern authenticity in visual media. In particular, we show that three visual obstructions designed to hinder specialized processing—inversion, misalignment, and partial occlusion—decrease participants' accuracy. In contrast to human visual processing, we find only inversion and not misalignment or partial occlusion change the model's performance. While the computer vision model is robust to misalignment and occlusion, this robustness may be a bug—the model overfitting to the training data—rather than a feature. Future research should explore whether specialized processing in computer vision models for deepfake detection enables better generalization to new contexts.

Limitations

We evaluated human and machine performance on 167 videos (84 deepfake and 83 authentic videos) across experiments 1 and 2. While these videos represent a balanced group of individuals across demographic dimensions and a variety of deepfake models, only the two political deepfake videos include lip-syncing manipulations, which are some of the most commonly used models for producing political deepfakes (31, 79–81). Moreover, we do not specifically recruit expert fact-checkers or expert media forensic analysts, and, as such, our results only generalize to the performance of ordinary people. Our comparison of untrained participants' predictions to the predictions of the leading computer vision model is limited to the best performance in 2020. If current trends continue as we expect they would, computer vision detection models will continue to improve (and possibly incorporate more human-like specialized processing of faces to better generalize across contexts), just as the realism of synthetic media generation algorithms will continue to improve. As a consequence, society will require more than just visual-based classification algorithms to protect against the potentially harmful threats that deepfakes pose (27).

The minimal context videos used here may not resemble the most problematic deepfakes, because the videos here show unknown people saying noncontroversial things in nondescript settings. On one hand, this minimal context makes the human participants' performance all the more impressive because such videos are missing many of the contextual cues they might normally use to discern authentic videos from deepfakes. On the other hand, perhaps videos designed to deceive are stylistically very different from the videos from the sampled holdout. As such, persuasive, manipulated video is important to consider in future research. The role of persuasion in synthetic media is beginning to be explored across varying media modalities (82, 83), but it is not the central focus of this paper. Instead, we ask how well the human visual processing system can detect the visual manipulations characteristic of deepfakes. We limit the bulk of our evaluation to uncontroversial videos of unknown actors to focus on the visual component of truth discernment. We begin to examine more realistic examples based on four videos of political leaders, but a larger sample size and further experimentation is necessary before making conclusions about how people judge the authenticity of political deepfakes. Furthermore, there is still much to learn about how AI systems and ordinary people can incorporate all the other information beyond facial features to make accurate judgments about a video's authenticity.

In this experiment, half of the videos were real and half are deepfakes. This is useful for comparing human and machine performance, but this base rate of deepfakes does not reflect the base rate of misinformation in today's media ecosystems (84). In 2021, less than a fraction of a percent of news was misinformation (85). Future experiments might consider examining people's ability to identify deepfakes when they do not have foreknowledge of the base rate of deepfakes. Moreover, an experiment embedded in a social media ecosystem could further identify how well people identify deepfakes within an ecologically valid context where people have access to contextual information such as who shared the video and how many others have shared or commented on the video. Ultimately, there are many ways to discern between real and fake videos, and visual perception should be considered as one tool in a user's toolkit for truth discernment.

We also considered how incidental emotions (i.e., emotions unrelated to the task at hand) affect participants' discernment of real and fake videos. Here, our two experiments found different results, and so we do not draw firm conclusions about the role of emotion on deepfake detection. In experiment 1, the custom emotion elicitation interventions did not significantly alter deepfake detection performance—although it also did not

significantly alter self-reported emotions, making it unclear how much to read into the lack of effects on performance. The results from experiment 2, although statistically significant by conventional standards, were near the cutoff for statistical significance for authentic videos and not statistically significant for deepfake videos. As such, future research could further explore the role of emotions in deepfake detection by running experiments with larger samples, examining additional emotions, ensuring effective elicitation, and focusing on integral emotions (emotions elicited directly from the stimuli). Recent research shows that inferences from feelings are context sensitive, and incidental emotions may be more likely to lead individuals astray in judgment tasks than integral emotions (86).

Implications

Relative to today's leading computer vision model, groups of individuals are just as accurate or more accurate, depending on which videos are considered. Participants and the model perform equally well on standard resolution, visual-only deepfake manipulations. Participants perform better on the four political videos and the attention check video, while the computer vision model performs slightly better on blurry, grainy, and very dark videos. The model's poor performance on both deepfakes of world leaders and videos with two people instead of one suggests that the model may not generalize well to stylistically different videos than the videos on which it has been trained. Humans have no problem with this kind of generalization, and, as a consequence, social media content moderation of video-based misinformation is likely to be more accurate when performed by teams of people than today's leading algorithm. As such, future research in crowd-based deepfake detection may consider how to most effectively aggregate wisdom of the crowds to improve discernment accuracy beyond the crowd mean [e.g., using algorithms such as the surprisingly popular answer (87) and revealed confidence (88)].

Sociotechnical systems may benefit from the combination of AI and crowd wisdom, but decision support tools for content moderation must be carefully designed to appropriately weigh human and model predictions. The confidently wrong predictions of the model on out-of-sample videos reveals the leading model is not ready to replace humans in detecting real-world deepfakes. Moreover, decision support tools can be counterproductive to accurate identification, as evidenced by the many instances in which participants saw incorrect predictions from the model and subsequently adjusted their predictions to be less accurate. Instead of solely informing people on the likelihood that a model is a deepfake, crowd wisdom could likely benefit from more explainable AI. Given that the leading model was more accurate at detecting certain classes of videos while humans were better at other classes, a future human–AI collaborative system might include additional information on video subtypes and how humans and machines perform across these subtypes. For example, video-level qualities (e.g., blurry, grainy, dark, specialized obstruction, stylistic similarities to training set, or other components upon which human and machine performance tends to diverge) and individual-level qualities could be factored into the interface and information presented by a human–AI collaborative system. By presenting model predictions alongside this information, it is possible humans could develop a better sense for confronting conflicting model predictions and deciding between second-guessing their own judgments and overriding the model's prediction. Machine-informed crowd wisdom can be a promising approach to deepfake detection and other classification tasks more generally where human and machine classification performance is heterogeneous on subtypes of the data.

Specialized visual processing of faces helps humans discern between real and deepfake videos. In future instances when humans are tasked with deepfake detection, it is important to consider whether a video has been manipulated in such a way as to reduce specialized processing. Moreover, given the usefulness of specialized processing of faces for humans in detecting deepfakes, it is possible that computer vision models for deepfake detection may find use in incorporating (and/or learning) such specialized processing (89).

Visual cues will continue to be helpful in deepfake detection, but, ultimately, identifying authentic video can involve much more than visual processing. When attempting to discern the truth from a lie, people rely on the available context, their knowledge of the world, their ability to critically reason, and their capacity to learn and update their beliefs. Similarly, the future of deepfake detection by both humans and machines should consider not only the perceptual clues but the greater context of a video and whether its message resembles an ordinary lie.

Methods

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-2070. All participants are informed that "Detect Fakes is an MIT research project. All guesses will be collected for research purposes. All data for research is collected anonymously. For questions, please contact detectfakes@mit.edu. If you are under 18 years old, you need consent from your parents to use Deep Fakes." Most participants arrived at the website via organic links on the internet. For recruited participants, we compensated each individual at a rate of $7.28 an hour and provided bonus payments of 20% to the top 10% of participants. Before beginning the experiment, all recruited participants 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 redirected to the end of the experiment.

We hosted the experiment on a website called Detect Fakes at https://detectfakes.media.mit.edu/. *SI Appendix*, Fig. S4 presents a screenshot of the user interface for both experiments 1 and 2. The rest of the methods are described in *SI Appendix*.

Data Availability. The datasets and code generated and analyzed during the current study are available in our public GitHub repository, https://github.com/mattgroh/cognitive-science-detecting-deepfakes. All DFDC videos are available at https://dfdc.ai/ (31), and the five non-DFDC videos are available in our public GitHub repository, https://github.com/mattgroh/cognitive-science-detecting-deepfakes.

ACKNOWLEDGMENTS. We acknowledge funding from MIT Media Lab member companies, thank Alicia Guo for excellent research assistance, and thank the following communities for helpful feedback: the Affective Computing lab, the Perception and Mind lab, Human Cooperation lab, and the moderator and participants at the Human and AI Decision-Making panel at the CODE2020 conference.

1. D. M. J. Lazer *et al.*, The science of fake news. *Science* **359**, 1094–1096 (2018).
2. B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. Law Rev.* **107**, 1753 (2019).
3. B. Paris, J. Donovan, Deepfakes and cheapfakes: The manipulation of audio and visual evidence. *Data and Society*, 18 September 2019. https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf. Accessed 15 May 2021.
4. C. Leibowicz, S. McGregor, A. Ovadya, The deepfake detection dilemma: A multi-stakeholder exploration of adversarial dynamics in synthetic media. arXiv [Preprint] (2021). https://arxiv.org/abs/2012.06109 (Accessed 15 May 2021).
5. D. Silver *et al.*, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
6. D. Silver *et al.*, Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
7. A. Esteva *et al.*, Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
8. S. M. McKinney *et al.*, International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

9. H. Farid, M. J. Bravo, "Image forensic analyses that elude the human visual system" in *Media Forensics and Security II* (International Society for Optics and Photonics, 2010), vol. 7541, p. 754106.
10. S. J. Nightingale, K. A. Wade, H. Farid, D. G. Watson, Can people detect errors in shadows and reflections? *Atten. Percept. Psychophys.* **81**, 2917–2943 (2019).
11. S. J. Nightingale, K. A. Wade, D. G. Watson, Can people identify original and manipulated photos of real-world scenes? *Cogn. Res. Princ. Implic.* **2**, 30 (2017).
12. M. Kasra, C. Shen, J. F. O'Brien, "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* (Association for Computing Machinery, 2018), pp. 1–6.
13. P. J. Phillips *et al.*, Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6171–6176 (2018).
14. H. Farid, *Fake Photos* (MIT Press, 2019).
15. A. M. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
16. S. Lyu, "Deepfake detection: Current challenges and next steps" in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (Institute of Electrical and Electronics Engineers, 2020), pp. 1–6.
17. G. Pennycook *et al.*, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
18. G. Pennycook, D. G. Rand, The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
19. P. Korshunov, S. Marcel, Deepfakes : A new threat to face recognition? Assessment and detection. arXiv [Preprint] (2018). https://arxiv.org/abs/1812.08685 (Accessed 15 May 2021).
20. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 3207–3216.
21. X. Yang, Y. Li, S. Lyu, "Exposing deep fakes using inconsistent head poses" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019), pp. 8261–8265.
22. A. Rössler *et al.*, FaceForensics: A large-scale video dataset for forgery detection in human faces. arXiv [Preprint] (2018). https://arxiv.org/abs/1803.09179 (Accessed 15 May 2021).
23. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: "Learning to detect manipulated facial images" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (IEEE, 2019), pp. 1–11.
24. L. Jiang, R. Li, W. Wu, C. Qian, C. Change Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE, 2020)*, pp. 2889–2898.
25. S. Agarwal, H. Farid, O. Fried, M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (Institute of Electrical and Electronics Engineers, 2020), pp. 660–661.
26. F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, "Detection of gan-generated fake images over social networks" in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (Institute of Electrical and Electronics Engineers, 2018), pp. 384–389.
27. Y. Mirsky, W. Lee, The creation and detection of deepfakes. *ACM Comput. Survey (Lond.)* **54**, 7 (2021).
28. L. Verdoliva, Media forensics and deepfakes: An overview. *IEEE J. Selected Topics Signal Process.* **14**, 910–932 (2020).
29. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* **64**, 131–148 (2020).
30. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. C. Ferrer, The Deepfake Detection Challenge (DFDC) preview dataset. arXiv [Preprint] (2019). https://arxiv.org/abs/1910.08854. Accessed 15 May 2021.
31. B. Dolhansky *et al.*, The deepfake detection challenge dataset. arXiv [Preprint] (2020). https://arxiv.org/abs/2006.07397 (Accessed 15 May 2021).
32. D. Huang, F. De La Torre, Facial action transfer with personalized bilinear regression in *European Conference on Computer Vision* (Springer, 2012), pp. 144–158.
33. E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Institute of Electrical and Electronics Engineers, 2019), pp. 9459–9468.
34. Y. Nirkin, Y. Keller, T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Institute of Electrical and Electronics Engineers, 2019), pp. 7184–7193.
35. T. Karras, S. Laine, T. Aila, "A style-based generator architecture for generative adversarial networks" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, 2019), pp. 4401–4410.
36. B. Dolhansky *et al.*, Deepfake detection challenge results: An open initiative to advance AI (2020). Accessed 27 January 2021.
37. H. Qi *et al.*, "DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms" in *Proceedings of the 28th ACM International Conference on Multimedia* (Association for Computing Machinery, 2020), pp. 4318–4327.
38. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues" in *Proceedings of the 28th ACM International Conference on Multimedia 2020* (Association for Computing Machinery, 2020), pp. 2823–2832.
39. S. Agarwal, H. Farid, T. El-Gaaly, S. Lim, "Detecting deep-fake videos from appearance and behavior" in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)* (IEEE, 2020), pp. 1–6.
40. S. Seferbekov, Deepfake detection challenge submission. https://github.com/selimsef/dfdc_deepfake_challenge. Accessed 15 January 2021.
41. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**, 1499–1503 (2016).
42. M. Tan, Q. Le, "EfficientNet : Rethinking model scaling for convolutional neural networks" in *International Conference on Machine Learning (PMLR)*, K. Chaudhuri, R. Salakhutdinov, Eds. (PMLR, 2019), pp. 6105–6114.
43. A. Buslaev *et al.*, Albumentations: Fast and flexible image augmentations. *Information (Basel)* **11**, 125 (2020).
44. P. Chen, S. Liu, H. Zhao, J. Jia, GridMask data augmentation. arXiv [Preprint] (2020). https://arxiv.org/abs/2001.04086 (Accessed 15 May 2021).
45. F. Galton, Vox populi (the wisdom of crowds). *Nature* **75**, 450–451 (1907).
46. J. Surowiecki, The Wisdom of Crowds (Anchor, 2005).
47. J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).
48. Z. Epstein, G. Pennycook, D. Rand, "Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2020), pp. 1–11.
49. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2521–2526 (2019).
50. N. M. Brashier, E. J. Marsh, Judging truth. *Annu. Rev. Psychol.* **71**, 499–515 (2020).
51. J. P. Forgas, R. East, On being happy and gullible: Mood effects on skepticism and the detection of deception. *J. Exp. Soc. Psychol.* **44**, 1362–1367 (2008).
52. G. Clore *et al.*, "Affective feelings as feedback: Some cognitive consequences" in *Theories of Mood and Cognition: A User's Handbook*, L. L. Martin, G. L. Clore, Eds. (L. Erlbaum, 2001), pp. 27–62.
53. J. P. Forgas, Happy believers and sad skeptics? Affective influences on gullibility. *Curr. Dir. Psychol. Sci.* **28**, 306–313 (2019).
54. C. Martel, G. Pennycook, D. G. Rand, Reliance on emotion promotes belief in fake news. *Cogn. Res. Princ. Implic.* **5**, 47 (2020).
55. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
56. P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE* **94**, 1948–1962 (2006).
57. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
58. C. C. Goren, M. Sarty, P. Y. Wu, Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics* **56**, 544–549 (1975).
59. V. M. Reid *et al.*, The human fetus preferentially engages with face-like visual stimuli. *Curr. Biol.* **27**, 1825–1828.e3 (2017).
60. R. K. Yin, Looking at upside-down faces. *J. Exp. Psychol.* **81**, 141–145 (1969).
61. G. Rhodes, S. Brake, A. P. Atkinson, What's lost in inverted faces? *Cognition* **47**, 25–57 (1993).
62. J. J. Richler, O. S. Cheung, I. Gauthier, Holistic processing predicts face recognition. *Psychol. Sci.* **22**, 464–471 (2011).
63. R. T. Keys, J. Taubert, S. G. Wardle, A visual search advantage for illusory faces in objects. *Attention Perception Psychophysics* **83**, 1942–1953 (2021).
64. J. J. Richler, I. Gauthier, A meta-analysis and review of holistic face processing. *Psychol. Bull.* **140**, 1281–1302 (2014).
65. A. W. Young, A. M. Burton, Are we face experts? *Trends Cogn. Sci.* **22**, 100–110 (2018).
66. J. T. Wixted, L. Mickes, J. C. Dunn, S. E. Clark, W. Wells, Estimating the reliability of eyewitness identifications from police lineups. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 304–309 (2016).
67. J. Buolamwini, T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification" in *Conference on Fairness, Accountability and Transparency (PMLR)*, S. A. Friedler, C. Wilson, Eds. (PMLR, 2018), pp. 77–91.
68. D. A. Small, J. S. Lerner, Emotional policy: Personal sadness and anger shape judgments about a welfare case. *Polit. Psychol.* **29**, 149–168 (2008).
69. M. Groh, Ziv Epstein, Nick Obradovich, Manuel Cebrian, Iyad Rahwan, Human detection of machine-manipulated media. *Commun. ACM* **64**, 40–47 (2021).
70. C. Firestone, Performance vs. competence in human-machine comparisons. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 26562–26571 (2020).
71. J. Ternovski, J. Kalla, P. M. Aronow, Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. OSF Preprints [Preprint] (2021). https://doi.org/10.31219/osf.io/dta97 (Accessed 15 May 2021).
72. M. Vaccaro, J. Waldo, The effects of mixing machine learning and human judgment. *Commun. ACM* **62**, 104–110 (2019).
73. P. Tschandl *et al.*, Human-computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
74. S. Gaube *et al.*, Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
75. A. Abeliuk, D. M. Benjamin, F. Morstatter, A. Galstyan, Quantifying machine influence over human forecasters. *Sci. Rep.* **10**, 15940 (2020).
76. M. Jacobs *et al.*, How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. *Transl. Psychiatry* **11**, 108 (2021).
77. L. Wöhler, M. Zembaty, S. Castillo, M. Magnor, "Towards understanding perceptual differences between genuine and face-swapped videos" in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2021).

78. K. Haut *et al.*, "Could you become more credible by being White? Assessing impact of race on credibility with deepfakes". arXiv [Preprint] (2021). https://arxiv.org/abs/2102.08054 (Accessed 15 May 2021).
79. S. Agarwal *et al.*, "Protecting world leaders against deep fakes" in *CVPR Workshops* (2019), pp. 38–45.
80. S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.* **36**, 1–13 (2017).
81. K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, C. V. Jawahar, "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*, pp. 484–492. 2020.
82. C. Wittenberg, A. Berinsky, J. Zong, D. G. Rand, The (minimal) persuasive advantage of political video over text. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2114388118 (2021).
83. T. Dobber, N. Metoui, D. Trilling, N. Helberger, C. de Vreese, Do (microtargeted) deepfakes have real effects on political attitudes? (2020).
84. J. Allen, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, "Evaluating the fake news problem at the scale of the information ecosystem." *Science Advances* 6, no. 14 (2020): eaay3539.
85. D. J. Watts, D. M. Rothschild, M. Mobius, Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1912443118 (2021).
86. N. Schwarz, "Feelings-as-information theory" in *Handbook of Theories of Social Psychology*, P. A. M. Van Lange, A. W. Kruglanski, E. T. Higgins, Eds. (Sage, 2011), vol. 1, pp. 289–308.
87. D. Prelec, H. S. Seung, J. McCoy, A solution to the single-question crowd wisdom problem. *Nature* **541**, 532–535 (2017).
88. Y. Zhang, "Identify experts through revealed confidence: Application to wisdom of crowds." SSRN [Preprint] (2020). 3739192 (2020). https://doi.org/10.2139/ssrn.3739192 (Accessed 15 May 2021).
89. A. Farzmahdi, K. Rajaei, M. Ghodrati, R. Ebrahimpour, S. M. Khaligh-Razavi, A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Sci. Rep.* **6**, 25025 (2016).
90. M. Groh, Z. Epstein, C. Firestone, R. Picard, Replication data for "Deepfake detection by crowds, machines, and machine-informed crowds." Github. https://github.com/mattgroh/cognitive-science-detecting-deepfakes. Deposited 29 January 2021.