Is this real? Susceptibility to deepfakes in machines and humans

*Didem Pehlivanoglu[1,2], Mengdi Zhu[2,3], Jialong Zhen[1], Aude A. Gagnon-Roberge[1], Rebecca K. Kern[1], Damon Woodard[2,3], Brian S. Cahill[1], & Natalie C. Ebner[1,2,4]

[1]Department of Psychology, University of Florida, 945 Center Dr, Gainesville, FL 32603
[2]Florida Institute for National Security, University of Florida, 601 Gale Lemerand Dr, Gainesville, FL 32611, USA
[3]Department of Electrical and Computer Engineering, University of Florida, 968 Center Dr, Gainesville, FL 32603
[4]Center for Cognitive Aging and Memory, McKnight Brain Institute, University of Florida, 1149 Newell Dr, Gainesville, FL 32610, USA

***Corresponding Author:** Didem Pehlivanoglu (dpehlivanoglu@ufl.edu), Department of Psychology, University of Florida, 945 Center Dr., Gainesville, FL, 32611.

Acknowledgements

The authors have no conflict of interest to disclose and have complied with APA ethical standards in human subjects research. The full set of materials, de-identified data files, and analysis scripts are available on OSF repository (https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932).

Abstract (248 words)

Deepfakes are synthetic media created by deep-generative methods to fake a person's audio-visual representation. Growing sophistication of deepfake technology poses significant challenges for both machine learning (ML) algorithms and humans. Here we used real and deepfake static face images (Study 1) and dynamic videos (Study 2) *(i)* to investigate sources of misclassification errors in machines, *(ii)* to identify psychological mechanisms underlying detection performance in humans, and *(iii)* to compare humans and machines in their classification decision accuracy and confidence. Study 1 found that machines achieved excellent performance in classifying real and deepfake images, with good accuracy in feature classification. Humans, in contrast, experienced challenges in distinguishing between real and deepfake images. Their classification accuracy was at chance level, and this underperformance relative to machines was accompanied by a truth bias and low confidence for the detection of deepfake images. Using video stimuli, Study 2 found that performance of machines was near chance level, with poor feature classification. Further, the machines showed greater truth bias and low reduced decision confidence relative to humans who outperformed machines in the detection of video deepfakes. Finally, the study revealed that higher analytical thinking, lower positive affect, and greater internet skills were associated with better video deepfake detection in humans. Combined, findings across these two studies advance understanding of factors contributing to deepfake detection in both machines and humans and could inform intervention toward tackling the growing threat from deepfakes by identifying areas of particular benefit from human-AI collaboration to optimize deepfake detection.

Keywords: Deepfakes, Deception, Artificial Intelligence, Machine Learning, Confidence, Analytical Thinking, Truth Bias

1. Introduction

Significant advances in artificial intelligence (AI) have resulted in the use of sophisticated technology for producing manipulated media and has led to the emergence of *deepfakes* (for a review, see Nightingale & Wade, 2022). Deepfakes are algorithmic manipulations which are typically synthesized using generative adversarial networks, a type of machine learning that works by pitting two neural networks (a generator and a discriminator) against one another in an iterative back-and-forth process, to create any type of fake image, video, or audio (Tong et al., 2020). While deepfake technology presents numerous creative and entertainment possibilities for education, arts, and science, it also raises significant ethical, legal, and societal concerns, ranging from advertising to national security. In particular, deepfakes constitute a novel deception tactic to fake someone's entire audio-visual representation for spreading false information (Seow et al., 2022; Ternovski et al., 2022; Zhang, 2022) and are effectively harnessed in social engineering (Vaccari & Chadwick, 2020; Westerlund, 2019).

The growing presence of deepfakes to manipulate public opinion on social and news platforms (Fallis, 2021; Vaccari & Chadwick, 2020) has led to computer science and psychology research into investigating machine and human performance for detecting deepfakes (Groh et al., 2022; Karras et al., 2020; Montserrat et al., 2020; Nightingale & Farid, 2022). Importantly, these lines of research have been almost exclusively focused on deepfake detection performance, with factors that contribute to the ability to detect deepfakes in machines and humans still poorly understood. Further, this past research has mostly been conducted in isolation in each discipline and a direct comparison between machine and human performance has not been conducted yet. To fill these research gaps, here, we *(i)* identified sources of misclassification errors in machines, *(ii)* determined psychological mechanisms deepfake detection among humans, and *(iii)* directly contrasted machine and human performance in real and deepfake discernment; across two studies one employing static face images (Study 1) and one employing dynamic videos (Study 2). Next, we review work leading to these central research aims.

1.1. Machine Detection of Deepfakes

Deepfake images are typically synthesized using generative models that apply a replication process for the generation of new samples based on training data. Upon effective training, these models allow image synthesis, style transfer, and face-swapping. Among the multitude of existing generative models, generative adversarial networks (GANs; Goodfellow et al., 2014) are highly regarded for their ability to produce high-quality, high-resolution images. Detecting image deepfakes is a rapidly evolving field within computer vision and digital forensics. Generative models, however, often leave distinctive "fingerprints" on deepfakes, leading to the development of machine learning (ML) algorithms designed to identify and categorize such model-specific artifacts (Durall et al., 2019; Yu et al., 2021). In particular, Convolutional Neural Networks (CNNs) are trained on extensive datasets of real and fake images to learn distinguishing features. CNNs have demonstrated considerable promise in detecting deepfakes, with notable examples including ShallowNet (Tariq et al., 2018), ResNet-50 (Wang et al., 2020), Inception (Suratkar et al., 2020), Xception (Rössler et al., 2019; Suratkar et al., 2020), and MobileNet (Suratkar et al., 2020). CNN-based ML algorithms for image deepfake detection have reported accuracy rates between 83% and 100% (Afchar et al., 2018; Tolosana et al., 2020).

Video deepfakes often involve swapping of faces from source to target videos. Variational Autoencoders (VAEs; Kingma & Welling, 2014) are one of the methods frequently employed for face-swapping due to their proficiency in learning disentanglement within the data (Korshunova et al., 2017; Natsume et al., 2018). As for detecting video deepfakes, there are two primary methods. The first method involves CNNs, which follow a similar process as for image deepfake detection: each frame of an individual video from a training set of videos is processed by CNNs for final classification. CNNs have achieved video deepfake detection accuracies between 80% and 90% (Afchar et al., 2018; Sambhu & Canavan, 2020). Notably, the Xception network (Rössler et al., 2019) excels in learning complex data representations (i.e., face detection) and offers advantages such as efficiency and reduced susceptibility to overfitting. The second method leverages biometric and biological features for detection, as current deepfake technologies still struggle to accurately replicate such features. In particular, this approach includes analyzing facial features (Matern et al., 2019), eye blink patterns (Jung et al., 2020; Li et al., 2018), eye movements (Gupta et al., 2020), head poses (Yang et al., 2018), consistency of facial geometry (Tursman et al., 2020), facial expressions (Agarwal et al., 2020), lip syncing (Korshunov & Marcel, 2021), and biological signals from facial regions (e.g., photoplethysmography, head motion-based ballistocardiogram; Ciftci et al., 2020). Performance in distinguishing real videos from deepfakes using these feature-based ML algorithms ranges from 50% to 96%, but requires high-quality, high-resolution data for biometric feature extraction.

Growing evidence suggests variation in performance of different ML algorithms in detecting deepfake images and videos, but what contributes to this variation is not yet well understood. Going beyond existing work, here we determine factors that underlie machines' ability to spot real and deepfake material **(Aim 1)**. In particular, currently limited is understanding of how and why particular pieces of content get misclassified. Misclassifications are typically caused by biases within detection systems, such as training data bias, algorithmic bias, cultural and contextual bias, and/or performance discrepancies. We adopted a fine-grained approach by employing feature space analysis (Kulis, 2013) which allowed us to examine and compare two different ML algorithms regarding their classification accuracy of static face images (Study 1) and dynamic videos (Study 2). Analyzing feature detection performance will allow identification of misclassification sources by different ML algorithms, which will be crucial for improving the feature selection capacity of these algorithms.

1.2. Human Deepfake Detection

According to a recent national survey, a significant portion of Americans (63%) reported that made-up or altered images or videos create ‘‘a great deal of confusion’’ about the basic facts of current issues and events (Gottfried, 2019). Some studies employing static face images to determine discrimination ability between deepfake and real faces found that human performance was not better than chance (Miller et al., 2023; Nightingale & Farid, 2022; Rossi et al., 2023; Shen et al., 2021), with deepfake faces typically perceived as more real (Miller et al., 2023; Shen et al., 2021; Tucciarelli et al., 2022) and trustworthy (Nightingale & Farid, 2022) than real faces. Other studies demonstrated that while above chance, mean deepfake face detection accuracy in humans ranged from 60 to 65% only (Bray et al., 2023; Hulzebosch et al., 2020);

and this performance was accompanied by participants' overconfidence in their ability to detect deepfake faces (Bray et al., 2023; Miller et al., 2023). In brief, these findings reveal that humans are frequently fooled by deepfake faces and cannot reliably distinguish them from real faces.

There are also studies on human detection for dynamic video deepfakes, which vary widely regarding detection accuracy (from 58% to 89%; (Groh et al., 2022; Josephs et al., 2024; Köbis et al., 2021; Nas & de Kleijn, 2024; Somoray & Miller, 2023). For example, using deepfake videos pre-categorized based on subjective ratings of difficulty, one study found that participants detected “easy” video deepfakes with 71% accuracy whereas performance dropped to 25% for “very difficult” videos, indicating below chance-level performance for humans for high-quality deepfakes (Korshunov & Marcel, 2021). Another study found that overall detection accuracy for deepfake videos in humans was above chance (58%; Köbis et al., 2021). Thus, taken together, while human detection for static image deepfakes appears to be at chance, detection of at least some video deepfakes can be relatively good.

Currently less understood are individual differences in the ability to detect deepfakes. The limited research on this topic suggests that individuals with prior experience with histological images (i.e., microscopic images of tissues) were better able to distinguish between artificially generated and genuine histological samples than individuals without prior experience (Hartung et al., 2024). Also, somewhat counterintuitive, belief in conspiracy theories was positively correlated with deepfake video detection (Nas & de Kleijn, 2024). While informative, these studies have, however, failed to consider a larger spectrum of psychological factors that may contribute to deepfake detection ability. To fill this research gap, the current paper investigated interindividual differences in cognitive and socioemotional processing as well as experience and comfortability with the internet in their influence on deepfake detection accuracy in humans **(Aim 2)** for static face images (Study 1) and dynamic videos (Study 2). Investigation of these factors will inform the psychological mechanisms in deepfake detection, which can guide the development of interventions to reduce deception via deepfakes.

In particular, we assessed the following psychological variables:

Cognitive Processing. According to Dual-Process Theory(De Neys, 2012; Kahneman, 2011; Stanovich, 2009), individuals engage in two main routes of information processing: a quick, intuition-based route and a slow, deliberate route. While the intuition-based route leads to faster decision making, it is associated with low analytical reasoning and relies on cognitive heuristics. The slower route, in contrast, is associated with high analytical thinking and allows deliberation of information, often leading to less error-prone decision making. Indeed, research has consistently shown that individuals higher in analytical thinking were better at detecting misleading information (e.g., fake news; Pehlivanoglu et al., 2021, 2022; Pennycook & Rand, 2021). Further, need for cognition, which refers to the tendency to enjoy and engage in effortful and systematic thinking (Cacioppo et al., 1984), has been positively correlated with information seeking (Juric, 2017) and decision-making competence (Ding et al., 2020). Individuals with higher need for cognition are more willing to invest cognitive effort to solve demanding tasks and employ an elaborated information processing style instead of a heuristic processing style (Cacioppo et al., 1996; Verplanken et al., 1992). Also, individuals with higher need for cognition demonstrated greater skepticism toward information shared on social media (Tsfati & Cappella, 2003; Vraga & Tully, 2021). Based on this literature, we measured *analytical thinking* and need *for cognition* in their contributions to deepfake detection.

Socioemotional Processing. Affect has been shown to impact deception detection, though the direction of this effect is somewhat unclear (Ebner et al., 2020; Forgas & East, 2008; see also Ebner et al., 2023 for a summary). For example, individuals with greater feelings of sadness and distress (dysphoric mood) compared to non-dysphoric individuals were better at lie detection (Lane & DePaulo, 1999). Similarly, negative affect increased, while positive affect decreased, skepticism, deception detection, and ambiguity (Matovic et al., 2014; but see LaTour & LaTour, 2009). Further, heightened emotionality (in the form of both increased positive and negative affect) was associated with worse fake news detection (Martel et al., 2020). Additionally, interoceptive awareness, which reflects the ability to read one's inner bodily state (Bogaerts et al., 2022; Mehling et al., 2009), has been associated with deception detection (Gunderson & ten Brinke, 2022; Heemskerk et al., 2024; ten Brinke et al., 2019). Based on these findings, we measured *affect* and *interoceptive awareness* in their contributions to deepfake detection.

Experience and Comfortability with the Internet. Having relevant skills and experience with the internet and online materials may influence the ability to detect deception. For instance, time spent on social media was negatively correlated with believing fake news (Halpern et al., 2019) and positively with detection of deepfake videos (Nas & de Kleijn, 2024). Somewhat counterintuitive, however, one study found that self-reported IT affinity was not related to deepfake detection in both individuals with an IT background and non-professionals (Sütterlin et al., 2022). These previous studies, however, have not considered a broader set of internet and technology related skills that may contribute to the detection of deepfakes. Thus, going beyond existing literature, we measured self-reported *digital literacy* (i.e., internet skills) and *power usage* (i.e., mastery of technology use) in their contributions to deepfake detection.

1.3. Human vs. Machine Performance in Deepfake Detection

Currently, research from computer science on deepfake detection in machines is not well integrated with research on deepfake detection in humans. One exception is Groh et al. (2022) who directly compared human and machine performance and found comparable accuracy. Their study, however, only examined deepfake videos (not static images) and some videos involved familiar actors (i.e., political figures), which may have affected detection performance. Here we employed both static face images (Study 1) and dynamic videos (Study 2) of unfamiliar individuals and directly compared performance of the leading ML algorithm (i.e, the better performing ML algorithm among the two compared under Aim 1) with human performance **(Aim 3).** Findings from our work will provide insight into whether machines outperform humans in classification accuracy and confidence and increase knowledge about the nature of decision biases between machines and humans.

2. Study 1

2.1. Participants

Study 1 recruited 2,418 undergraduates through the Department of Psychology's SONA system. Of those, 680 participants were removed from analysis for the following reasons: 131 did not continue past consenting, 513 failed attention checks (e.g., *Please answer 2 to this question*), 27 had survey completion times 3 standard deviations greater than the group average, and 9 were older than 39 years. The final analysis sample comprised 1,738 participants (Age range: 18-39 years, *M* = 19.46, *SD* = 2.26; 81% female).

2.2. Measures

2.2.1. Image Rating Task. Participants were asked to rate the veracity of each of 200 faces on a scale from 1 (*Fake*) to 10 (*Real*). Each face was presented on the screen for at least 3 s to ensure sufficient processing time; beyond the 3s window, the task was self-paced.

Real images were 300 human face images randomly selected from the Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019), which contains 70,000 high-quality images (1024x1024 resolution) that vary in age, gender, ethnicity, and image background. The final set of real face images were crawled from Flickr, then aligned and cropped to ensure they contained only one face. For deepfake images, we randomly selected an additional set of 300 face images from the FFHQ dataset. These faces were then synthesized with a pre-trained styleGAN2 network released by NVIDIA (Karras et al., 2020). The styleGAN2 algorithm enables intuitive, scale-specific control of the synthesizing process via an automatically learned, unsupervised separation of high-level attributions (e.g., pose and identity when trained on human faces) and stochastic variation in the synthesized images (e.g., freckles, hair, accessories). For equal numbers of real vs. deepfake images by gender, deepfake images were first classified as male vs. female using a deep-learning based classification algorithm, then cross-validated via manual selection to exclude images with interference and/or warping artifacts.

We created three sets of 200 stimuli each by randomly selecting 100 real and 100 deepfake images from the larger pool we created, with face gender balanced within each set and image type (real vs. deepfake). Final image sets are achieved in the OSF repository (https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932). A third of participants, respectively, were assigned to view only one of the three sets to assure counterbalancing, with face presentation order within each set randomized.

2.2.2. Cognitive Reflection Test (CRT). Analytical thinking was assessed via the CRT (Frederick, 2005), which contains both numerical and logical propositions that have an intuitive and an analytical answer. For example, *"A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents."* Individuals who rely on intuition respond with the intuitive answer (10 cents), whereas individuals who rely on effortful thinking respond with the analytical answer (5 cents).

Validity of the CRT is affected by familiarity with the items (Haigh, 2016) as well as number of scale items (Toplak et al., 2014). Here, we used a 7-item version, which consisted of three items from Shenhav et al. (2012) and four items from Thomson and Oppenheimer (2016). An example item was: *"The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?".* Participants with high analytical thinking overcome the impulse to give the intuitive (incorrect) answer of *8 years old* and instead give the analytical (correct) answer of *4 years old.* We calculated sum scores across the 7 items, with higher CRT scores reflecting greater analytical thinking.

2.2.3. Need for Cognition (NFC). The NFC scale is a self-report questionnaire (Cacioppo & Petty, 1982) assessing how much an individual engages in and enjoys thinking or cognitively demanding tasks. We used a short version of the scale containing 18 items (Cacioppo et al., 1984). Each item consists of a statement, e.g. "*I would prefer complex to simpler problems*", and participants score themselves on a scale from 1 (*Extremely uncharacteristic*) to 5 (*Extremely characteristic*). We calculated the mean across all 18 items, with higher NFC scores reflecting greater need for cognition.

2.2.4. Positive and Negative Affect (PANAS). We administered the 20-item PANAS (Watson et al., 1988), an affect assessment that contains 20 items. We also included six additional items to capture hedonic balance (Röcke et al., 2009). For each item, participants were asked "*To what extent do you feel [emotion adjective] right now?*" and used a scale from 1 (*Very slightly or not at all*) to 5 (*Extremely*) to evaluate each adjective (e.g., *excited*, *happy*, *afraid*, *alert*; 13 positive and 13 negative adjectives). We calculated the mean across positive adjectives and negative adjectives, with higher scores reflecting more positive affect and more negative affect, respectively.

2.2.5. Multidimensional Assessment of Interoceptive Awareness Version 2 (MAIA-2). MAIA-2 (Mehling et al., 2018) is a 37-item self-report questionnaire that measures awareness of bodily sensations. The scale is composed of 8 subscales measuring different aspects of interoception (i.e., noticing, not-distracting, not-worrying, attention regulation, emotional awareness, self-regulation, and body listening). Each subscale has Likert-type items, with response options ranging from 0 (*Never*) to 5 (*Always*). Sample items are, "*When I am tense I notice where the tension is located in my body.", "I can notice an unpleasant body sensation without worrying about it.",* and "*I notice that my body feels different after a peaceful experience.*" We calculated the mean across all 37 items, with higher MAIA-2 scores reflecting greater interoceptive awareness.

2.2.6. Digital Literacy Scale (DLS). The DLS is a 21-item inventory that evaluates an individual's familiarity with computer and internet elements (Hargittai, 2009). The current study used a modified version, which updated the internet terms (Guess & Munger, 2023). For each item participants reported their level of understanding of various computer and internet elements (e.g., *phishing, tagging, selfie)* on a scale ranging from 1 (*No understanding*) to 5 (*Full understanding*). We calculated the mean across all 21 items, with higher scores reflecting greater understanding of digital media.

2.2.7. Power User Scale (PUS). The PUS (Sundar & Marathe, 2010) is a 12-item inventory that assesses mastery of information technology based on prior experience, expertise, and self-efficacy. The scale consists of two sub-scales, each with 6 items that were evaluated on a scale from -4 (*Strongly disagree*) to +4 (*Strongly agree*). One subscale captures low (e.g., "*I think most technological gadgets are complicated to use*") vs. high (e.g., "*I often find myself using many technological devices simultaneously"*) frequency of technology use. The other subscale captures low (e.g., "*I prefer to ask friends how to use any new technological gadget instead of trying to figure it out myself*") vs. high (e.g., "*I would feel lost without information technology*") comfortability with technology use. We calculated the mean across both subscales (all 12 items), with higher PUS scores reflecting greater power usage (i.e., expertise, experience, and efficacy in technology use).

2.3. Procedure

All procedures and measures were approved by the University of Florida Institutional Review Board (IRB# 202102022). Participants completed this study remotely through Qualtrics (https://www.qualtrics.com/). Prior to study enrollment, all participants consented electronically to participate. Participants then completed the Image Rating Task, CRT, NFC, MAIA-2, PANAS, DLS, PUS, and a brief demographic questionnaire, in this order. The study took approximately 100 mins and participants were reimbursed with SONA credits upon completion.

2.4. Analyses and Results

All de-identified datasets and analysis scripts used in Study 1 are available on the OSF repository (https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932).

2.4.1. Machine Performance

To measure how well a machine could detect image deepfakes, we chose two different ML algorithms, previously shown to be efficient in identifying specific artifacts existing in GAN-generated deepfake images (Mirsky & Lee, 2021; Verdoliva, 2020). The first approach applied a CNN (using a pre-trained ResNet-50 network, Wang et al., 2020); the second involved Frequency Domain Analysis (FDA; using a pre-trained Support Vector Machine; (Durall et al., 2019) to extract frequency characteristics from images to distinguish real and deepfake images. These ML algorithms generated predicted labels as outcome variable. Predicted labels can be either 0 = Deepfake face or 1 = Real face and reflect the classification of each face type. The CNN approach yielded 97% accuracy in distinguishing real and deepfake images, whereas the FDA approach resulted in 79% accuracy.

To understand the source of misclassification underlying image detection performance of these two ML algorithms, we used feature visualization. This approach compared feature detection of the two ML algorithms in their classification accuracy of static face images. Specifically, this technique involved reducing the dimensionality of features for 2D visualization of features using t-distributed stochastic neighbor embedding (t-SNE; van der Maaten et al., 2008). After applying feature analysis for both ML approaches, the decision boundary of the CNN approach (Figure 1A) was more clearly defined than that of the FDA approach (Figure 1B), resulting in higher classification accuracy of features in image deepfakes in the CNN than the FDA approach. One reason for the difference in performance could be that the CNN approach used persistent and distinctive features from both real and fake images, while the FDA approach only leveraged features within the deepfake images, causing features in real images to get misclassified.

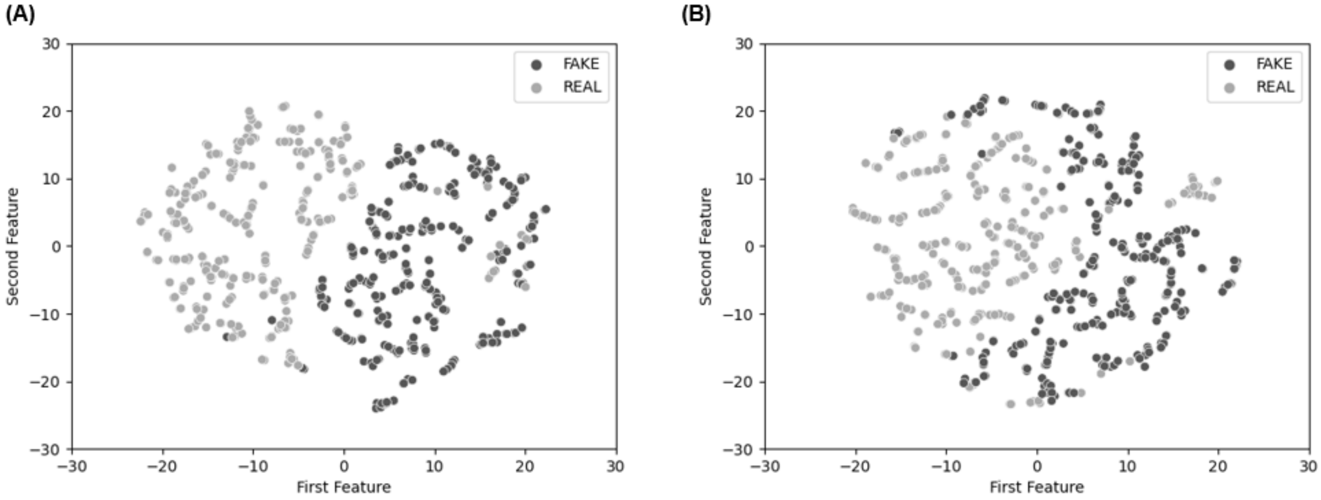

Figure 1. 2-D visualization of latent features from **(A)** Convolutional Neural Network (CNN) and **(B)** Frequency Domain Analysis (FDA). Real images are shown in gray, deepfake images in black.

2.4.2. Human Performance

To calculate how well humans could discriminate between deepfake and real face images, we computed the d-prime (d' in standard Signal Detection Theory; (Macmillan & Creelman, 1991). In this score, deepfake images were considered as "signal present". Given our 1 = Fake to 10 = Real response scale, ratings from 1 to 5 reflected 'hits' whereas ratings from 6 to 10 reflected 'misses' for deepfake images. For real face images, ratings from 1 to 5 reflected 'correct rejections' whereas ratings from 6 to 10 reflected 'false alarms'. Using the formula d' = z(H)-z(F), d' was calculated for each participant across all images, with higher d' indicating a participant's greater ability to discriminate between deepfake and real face images. The average d' score was close to 0 (Figure 2; *M* = -0.13, *SD* = 0.45), reflecting diminished ability to discriminate between deepfake and real face images in humans.

To examine the extent to which individual differences in psychological variables predicted discrimination ability we conducted a multiple linear regression model on d' as outcome variable. This model included the main effects of analytic thinking (CRT; continuous), need for cognition (NFC; continuous), positive and negative affect (PANAS, continuous), interoceptive awareness (MAIA-2; continuous), digital literacy (DLS; continuous), and power usage (PUS; continuous), with participant gender and age added as covariates. None of the individual difference measures predicted discrimination ability between real and deepfake images (all *F*s < 1.65, *p*s > .09).

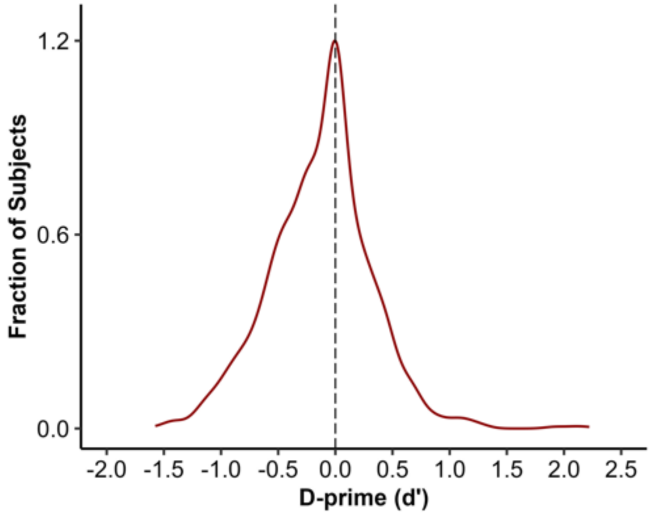

Figure 2. Distribution of discrimination ability (d' score) in humans. The dashed line reflects guessing (i.e., no discrimination between deepfake and real face images).

2.4.3. Machine vs. Human Performance

As noted, the CNN algorithm outperformed the FDA algorithm, yielding a 97% accuracy in the deepfake detection task. In comparison, humans performed poorer than the CNN, with overall accuracy in classifying face images remaining at 49%. Next, for both CNN algorithm and humans, we calculated the True Positive Rate (TPR), reflective of the prediction of real when a face image was real; and the True Negative Rate (TNR), reflective of the prediction of fake when a face image was a deepfake. Separate calculation of TPR and TNR for CNN and humans allowed us to determine whether accuracy was comparable for classifications of real and deepfake images or whether it was biased towards one or the other image type (e.g., whether accuracy was high for real images but low for deepfake images). The TPR for the CNN was 97% and the TNR was 97% (Figure 3A), indicating that this algorithm was equally successful in classifying real and deepfake images, with no detection bias towards one or the other image type. In contrast, the TPR for humans was 69%, whereas the TNR was only 29% (Figure 3B). This low TNR in humans was driven by a greater tendency to misclassify deepfake images as “real” (as reflected by a false positive rate (FPR) of 71% in Figure 3B), suggesting a truth bias in humans (i.e., a tendency to misclassify deepfakes as “real”).

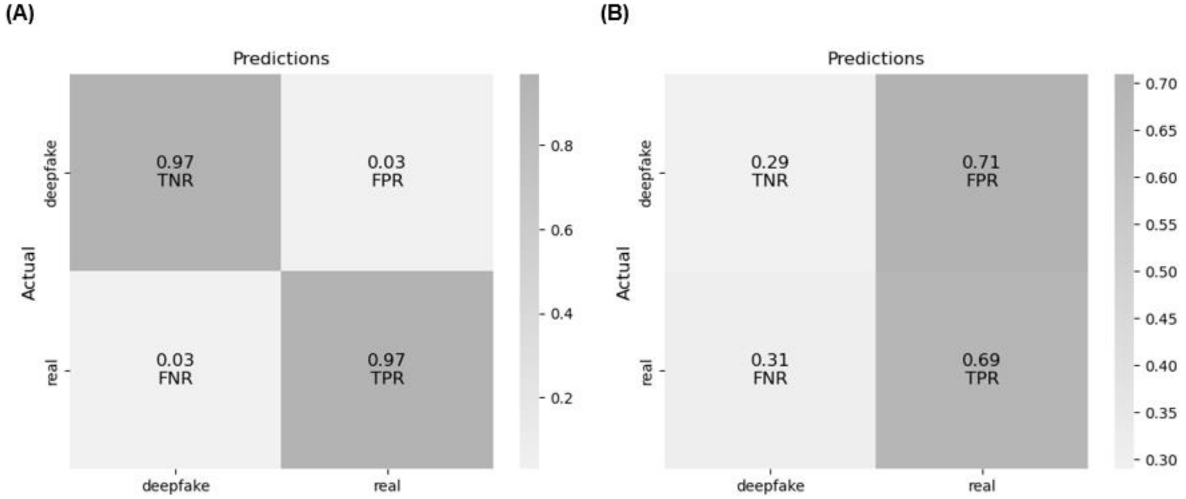

Figure 3. Confusion matrix indicating accuracy for **(A)** Convolutional Neural Network (CNN) and **(B)** humans. TNR = True Negative Rate (i.e., correctly classifying deepfake as “deepfake”); FNR = False Negative Rate (misclassifying real as “deepfake”); TPR = True Positive Rate (i.e., correctly classifying real as “real”); FPR = False Positive Rate (i.e., misclassifying deepfake as “real”). Colors in each confusion matrix serve as heatmap, with darker colors indicating higher values.

We also computed decision confidence scores in image classifications for the machine algorithm and humans. For the CNN algorithm, the confidence score was calculated using a probability score derived from the classification prediction. This score represents the confidence level of whether a given face classified as real was real on a scale from 0 (*Not confident at all)* to 1 (*Very confident)*. As shown in Figure 4A, approximately 45% of confidence scores for the CNN fell within 0 and 0.1, indicating high confidence in classifying deepfake images as deepfake. Correspondingly, approximately 45% of the confidence scores were within 0.9 and 1.0, indicating also high confidence in classifying real face images as real. That is, the machine was confident about its prediction. For decision confidence in humans, we re-coded image ratings from 1 (*Not confident at all)* to 10 (*Very confident),* with higher scores reflecting greater confidence. Humans showed greater confidence in classification of real (*M* = 6.86, *SD* = 1.68) than deepfake (*M* = 4.01, *SD* = 1.8) images (*t*(1737)= 35.12, *p* < .001, Cohen’s *d* = 1.64; Figure 4B), consistent with their higher accuracy for real than deepfake images.

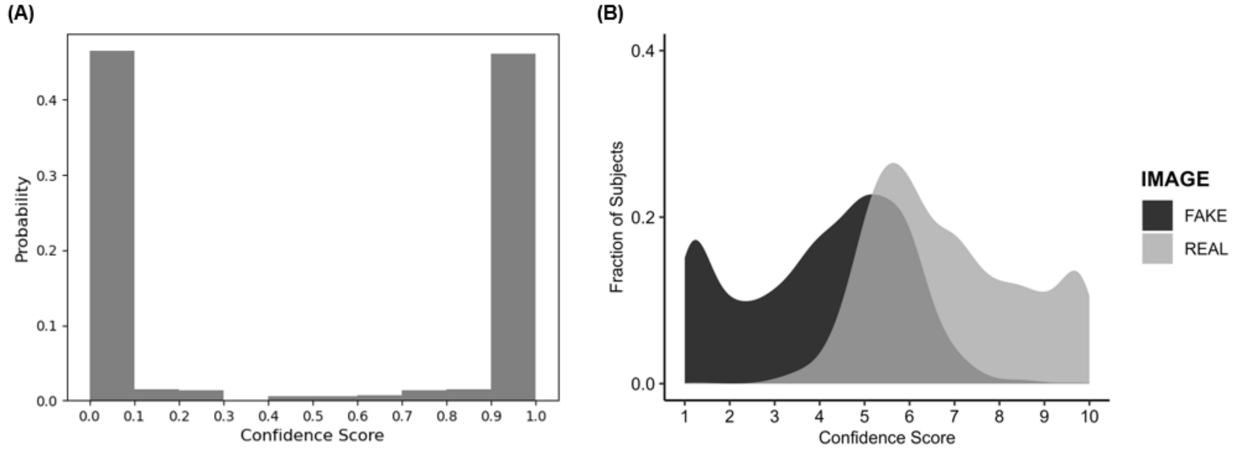

Figure 4. (A) Histogram of probability scores derived from Convolutional Neural Network (CNN) regarding image classification confidence. Scores from 0 to 0.1 reflect higher confidence for classification of deepfake images; scores from 0.9 to 1 reflect higher confidence for classification of real images. **(B)** Distribution of image classification confidence scores in humans. Real images are shown in gray; deepfake images in black. Higher confidence scores reflect greater classification confidence.

2.5. Summary and Brief Discussion of Study 1

In Study 1 we found that the CNN approach outperformed the FDA approach in detecting static real and deepfake images, as reflected in greater feature classification accuracy. In comparison, the ability of humans to discriminate between deepfake and real face images was rather poor (at chance level); and individual differences in cognitive and socioemotional processes as well as in the level of internet skills did not explain variability in detection performance for real or deepfake images. Furthermore, our direct comparison between machine and human performance revealed that the CNN algorithm outperformed humans by showing excellent prediction accuracy, with no decision bias and high classification confidence for both deepfake and real images. Humans' dramatic underperformance relative to the machine was coupled with a truth bias and low confidence for the classification of deepfake images.

In this first study, we addressed machine and human performance for deepfake images. Fast-developing AI advances, however, more and more confront us with dynamic deepfakes such as in videos in our everyday lives. Importantly, cues available in static vs. dynamic deepfakes differ in that videos often contain audio and visual input simultaneously, integrate behavioral (e.g., facial expressions, gestures) and non-behavioral (e.g., lighting, skin texture) features, and typically are more ecologically valid than static images. Thus, going beyond Study 1, Study 2 examined deepfake detection performance by employing videos *(i)* to investigate sources of misclassification errors in machines, *(ii)* to identify psychological mechanisms underlying detection performance in humans, and *(iii)* to compare humans and machines in their classification decision accuracy and confidence.

3. Study 2

3.1. Participants

Study 2 recruited 2,183 undergraduates from the Department of Psychology's SONA. Of those, 743 were removed from analysis for the following reasons: 142 did not continue the study after consenting, 560 failed attention checks, 28 had survey completion times 3 standard deviations greater than the group average, and 13 were older than 39 years. The final sample comprised 1,440 participants (Age range: 18-39 years, *M* = 19.94, *SD* = 2.27; 83% female).

3.2. Measures

3.2.1. Video Rating Task. Participants viewed 70 short videos of an individual discussing a topic (e.g., book presentations, video games, daily activities). At the end of each video clip, participants were asked to rate the veracity of the face shown in each of the videos (i.e., *"This Face was ____."*) on a scale from 100% *(Fake)* to 100% *(Real)*, with 50% reflecting just as likely real or fake. The presentation order of the videos was randomized, and beyond the 10-s video presentation, the task was self-paced.

Videos were obtained from the Deepfake Detection Challenge (DFDC) dataset (Dolhansky et al., 2020), which is a large-scale dataset containing over 100,000 videos, both real and deepfake, covering a variety of scenarios and individuals of diverse gender, age, and racial/ethnic backgrounds. Real videos were created by recording video clips of volunteers. Deepfake videos were generated by applying various manipulation techniques (e.g., face swapping, altering facial expressions, or audio swapping) to real videos.

We randomly selected an initial pool of 336 real and 322 deepfake videos. Each video was assessed on multiple criteria to ensure that *(i)* it had a landscape orientation, good sound quality and lighting, and had no text or written information embedded, *(ii)* there was only one person shown in each video, *(iii)* of unique identity (i.e., the same person was not shown in any of the other videos), *(iv)* the person was speaking by looking towards the camera without location change (e.g., walking), and *(v)* videos did not involve audio synthesis or replacement (i.e., audio swapped video) by checking lip syncing. In particular, the final set comprised 35 real and 35 fake videos, all trimmed to 10 s to ensure equal duration. To assure that detection performance was not confounded by audio in the videos, the same set of videos were muted to create non-audio video versions. For counterbalancing, approximately half of the participants (N=684) viewed the videos with audio and approximately the other half (N=756) viewed the muted versions, with videos presented in random order in each of these two stimuli lists. All videos are achieved under the OSF repository (https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932).

3.3. Procedure

Study procedures were approved by the University of Florida Institutional Review Board (IRB# 202102022). Identical to Study 1, participants consented electronically and completed the study remotely through Qualtrics (https://www.qualtrics.com/). Participants first completed the Video Rating Task, followed by the CRT, NFC, MAIA-2, PANAS, DLS, PUS, and a brief demographic questionnaire, in this order. The study took approximately 100 mins and participants were reimbursed with SONA credits upon completion.

3.4. Analyses and Results

All de-identified datasets and analysis scripts used in Study 2 are available on the OSF repository (https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932).

3.4.1. Machine Performance

To measure how well machines detect video deepfakes, we tested two different ML algorithms, known to be efficient in finding inconsistencies and manipulations on video frames. The first was FaceForensics (using a pre-trained Xception network, Rössler et al., 2019), and the second involved Recurrent Neural Network (RNN) (using the pre-trained network; (Güera & Delp, 2018) to identify inconsistencies of latent features from continuous frames. As in Study 1, predicted labels generated by the ML algorithms were either 0 = Deepfake or 1 = Real face and reflected the classification for each face type within the videos. FaceForensics yielded 51% accuracy in distinguishing real and deepfake videos, whereas RNN resulted in 39% accuracy.

To identify the source of misclassification, we applied the same feature visualization technique described in Study 1. Features were intertwined for both FaceForensics (Figure 5A) and RNN (Figure 5B), making it difficult to establish a clear decision boundary and resulting in poor classification of features for both ML algorithms. Of note, RNN misclassified features even more than FaceForensics, possibly because the RNN algorithm uses the entire frame to extract features whereas the FaceForensics model uses a frontal face frame only.

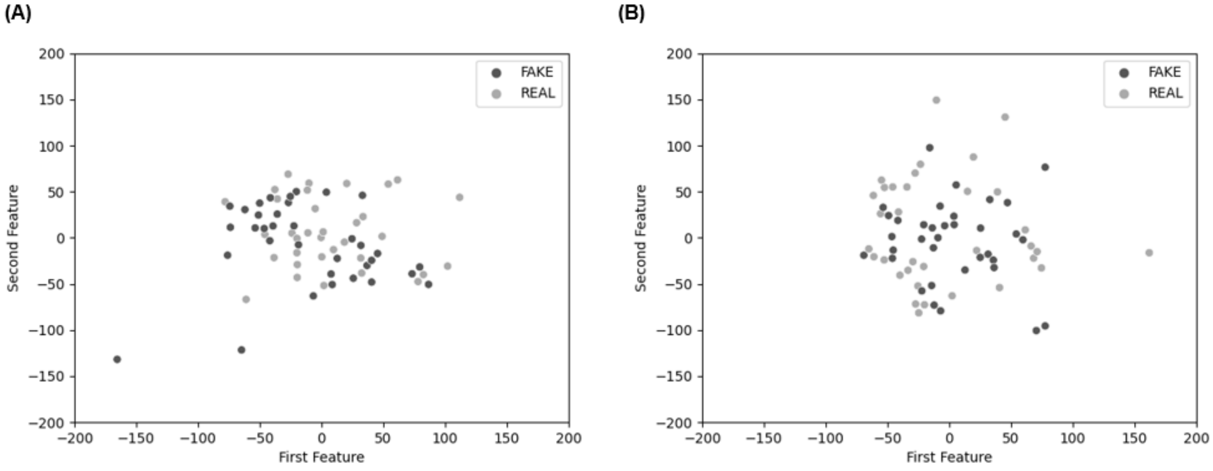

Figure 5. 2-D visualization of latent features from **(A)** FaceForensics and **(B)** Recurrent Neural Network (RNN). Real videos are shown in gray, deepfake videos in black.

3.4.2. Human Performance

As in Study 1, we computed discrimination ability, with higher d' indicating greater discriminate ability between deepfake and real videos. Deepfake videos were considered as 'signal present'. Given our 100% = Fake to 100% = Real response scale, ratings from 100% to 60% Fake reflected 'hits' whereas ratings from 100% to 60% Real reflected misses for deepfake videos. For real videos, ratings from 100% Real to 60% Real reflected 'correct rejections' whereas ratings from 100% Fake to 60% Fake reflected 'false alarms'. Responses of 50% were omitted from the analysis as they reflected that participants were undecided whether a video was real or fake (8% of the trials). The ability to discriminate between deepfake and real videos was relatively good in humans (Figure 6A; *M* = 0.86, *SD* = 0.65, Range= -0.91 to 3.50).

We also again conducted a multiple linear regression on d' for formal analysis by applying the identical analytical approach as described in Study 1. Greater ability to discern between deepfake and real videos was associated with higher analytical thinking (reflected by a significant main effect of CRT: *F* = 2.12, *p* = .03, Cohen's f^2 = 0.01; Figure 6B), lower positive affect (reflected by a significant main effect for PA: *F* = 2.35, *p* = .02, Cohen's f^2 = 0.01; Figure 6C), and greater power usage (reflected by a significant main effect for PUS: *F* = 2.86, *p* < .01, Cohen's f^2 = 0.02; Figure 6D). None of the other individual difference variables predicted discrimination ability (all *F*s < 1.29, *p*s > .19).

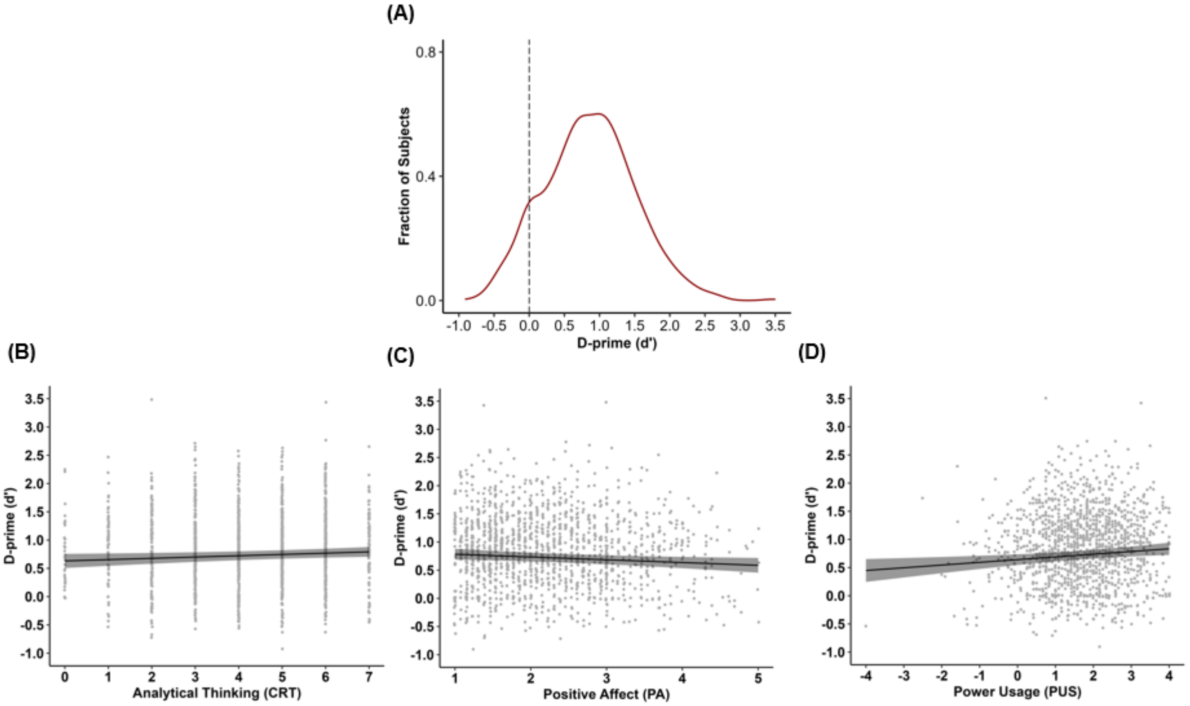

Figure 6. (A) Distribution of discrimination ability (d' score) in humans. The dashed line reflects guessing. Greater ability to discern between deepfake and real videos was associated with **(B)** higher analytical thinking, indexed by Cognitive Reflection Test (CRT) scores, **(C)** lower positive affect, indexed by Positive and Negative Affect Scale (PANAS) scores, and **(D)** greater power usage, indexed by Power User Scale (PUS) scores. Each dot represents a participant. Shaded areas around the regression lines reflect the 95% confidence interval.

3.4.3. Machine vs. Human Performance

As noted, FaceForensics outperformed RNN, yielding a 51% accuracy in deepfake video detection. Of note, however, and different from the results for static images, humans outperformed FaceForensics, with 64% overall accuracy in classifying videos. We again calculated the TPR and TNR. The TPR for FaceForensics was 63%, whereas the TNR was only 23% (Figure 7A). This low TNR for FaceForensics was driven by a greater tendency to misclassify deepfake videos as "real" (i.e., truth bias), as reflected by an FPR of 77% (Figure 7A). The TPR for humans was 76%, whereas the TNR was 51% (Figure 7B). While not as pronounced as in the machine, humans also were more likely to misclassify deepfake videos as "real", as reflected by an FPR of 49% (Figure 7B).

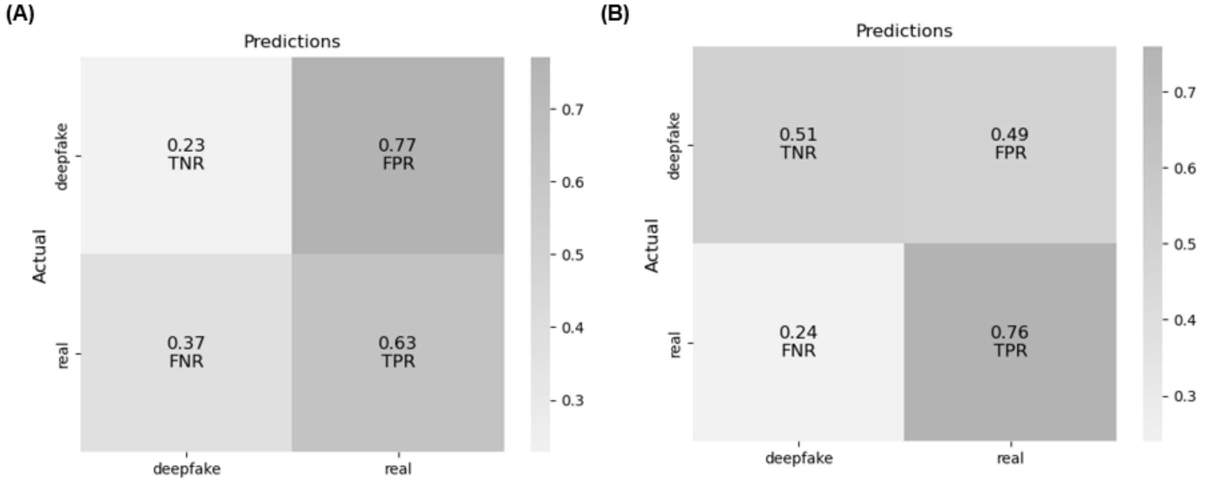

Figure 7. Confusion matrix indicating accuracy for **(A)** FaceForensics and **(B)** humans. TNR = True Negative Rate (i.e., correctly classifying deepfake as "deepfake"); FNR = False Negative Rate (misclassifying real as "deepfake"); TPR = True Positive Rate (i.e., correctly classifying real as "real"); FPR = False Positive Rate (i.e., misclassifying deepfake as "real"). Colors in each confusion matrix serve as heatmap, with darker colors indicating higher values.

Parallel to Study 1, we again computed decision confidence scores in video classifications for both the machine and humans. For FaceForensics (Figure 8A), only 27% of the confidence scores fell within the range of 0.9 to 1.0, while the remaining 63% were fairly evenly distributed across other bins. This pattern suggests that FaceForensics was uncertain about its decisions. For humans (Figure 8B), confidence in the classification of real videos (*M* = 7.13, *SD* = 1.26) was greater than confidence in the classification of deepfake videos (*M* = 5.6, *SD* = 1.5; *t*(1,439)= 23.98, *p* < .001, Cohen's *d* = 1.11), consistent with higher accuracy for real than deepfake videos.

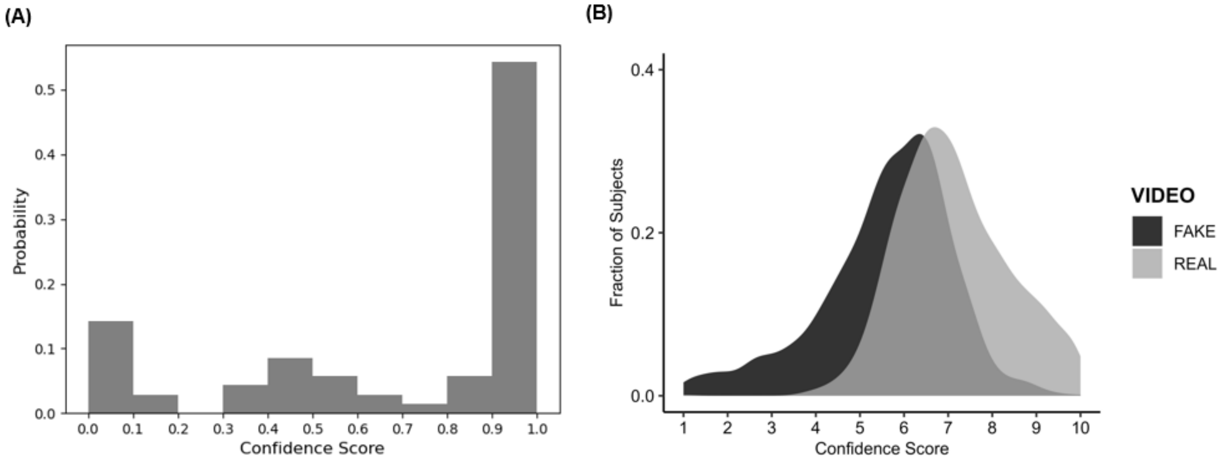

Figure 8. (A) Histogram of probability scores from FaceForensics for video classification confidence. Scores from 0 to 0.1 reflect higher confidence for classification of deepfake videos; scores from 0.9 to 1 reflect higher confidence for classification of real videos. **(B)** Distribution of

video classification confidence scores in humans. Real videos are shown in gray, deepfake videos in black. Higher confidence scores reflect greater confidence.

3.5. Summary and Brief Discussion of Study 2

In Study 2 we found that while the FaceForensics algorithm performed slightly better than the RRN algorithm at detecting real and deepfake videos, accuracy for FaceForensics was rather low (at chance level). We observed that classification accuracy of features in videos were intertwined and poor for both ML algorithms. In contrast, discrimination ability between deepfake and real videos in humans was rather good. Further, higher analytical thinking, less positive affect, and greater internet skills were associated with better discernment ability. Directly comparing machine and human performance furthermore showed that the overall classification accuracy of FaceForensics was lower than human performance, with this underperformance by the machine characterized by a truth bias and low classification confidence. A decision bias was less evident in humans, with decision confidence patterns in alignment with detection accuracy for real and deepfake videos.

4. General Discussion

With rapidly increasing sophistication of AI, deepfakes represent a serious challenge in today's society. They are being used to deceive and disseminate disinformation, undermining trust in media and institutions. While research on deepfake detection performance in both machines and humans is growing, the processes underlying deepfake detection ability are not well understood; and direct comparisons of machine vs. human performance are still rare. Here we identified sources of misclassification errors in machines, psychological mechanisms of discrimination ability in humans, and directly contrasted machine and human performance regarding classification accuracy and confidence for real and deepfake images (Study 1) and videos (Study 2). Across two studies, our data yielded three key findings: First, ML algorithms were overall more accurate and better at classifying features in real and deepfake images than videos. Second, humans outperformed the ML algorithm in deepfake video detection, but they experienced challenges in deepfake image detection, where they displayed a truth bias and low confidence. In turn, the ML algorithm's quite weak performance with videos was marked by a truth bias and low classification confidence. Third, we found that higher analytical thinking, lower positive affect, and more internet skills improved discernment of deepfake from real videos in humans. Collectively, these findings suggest that ML excels at detecting deepfake images (static input) but humans have an advantage in video detection (dynamic input). This differential pattern of findings highlights the need for collaboration between humans and AI to optimize the detection of deepfakes. Theoretical and practical implications of our novel findings are discussed next.

4.1. Machine Excels in Image Deepfake Detection but Experiences Challenges with Videos

CNN- and FDA ML algorithms achieved high detection accuracies of 97.17% and 79.33%, respectively, for image deepfakes (Study 1). Follow-up feature space analysis further demonstrated that CNN was more effective at discernment than FDA because features learned by this algorithm clustered more tightly for both deepfake and real mages. The scattering of features in FDA compared to the more tightly clustered features learned by CNN may have stemmed from their different approaches to feature selection. That is, CNN is trained to identify distinctive features from both deepfake and real images. FDA, in contrast, produces more dispersed features from real images because it relies on the Fourier transform to detect unique patterns specific to deepfake (but not real) images. This, in turn, leads to less accurate real image classification.

In contrast, video deepfake detection by FaceForensics and RNN algorithms (Study 2) achieved low accuracies of 51.43% and 38.57%, respectively. Follow-up feature space analysis for these algorithms revealed that both methods struggled with the identification of distinctive features that effectively differentiated between deepfake and real videos. Visualization of this performance pattern revealed that features from real and deepfake videos were entangled and indistinguishable, leading to classification error. It is also possible that the FaceForensics and RNN models were trained on data that did not match the characteristics of the test videos used in our study. As a result, the algorithms may have extracted irrelevant or "wrong" features, leading to incorrect classifications. This alternative explanation is somewhat supported by our finding of relatively better deepfake detection by FaceForensics than RNN as deepfake videos in the DFDC dataset are created using face-swapping, which is the technique that FaceForensics specializes in. The RNN algorithm, in contrast, analyzes the entire frame for feature extraction (Güera & Delp, 2018), which may have resulted in higher misclassifications for deepfake videos taken from the DFDC dataset, in which manipulations are present on the face regions only.

4.2. Machines Outperform in Image Detection but Humans Lead in Video Deepfake Detection

Our comparison of machine and human performance for classifying face images (Study 1) found that the CNN algorithm outperformed human detection ability, showing excellent accuracy without decision bias and maintaining high confidence. In contrast, humans performed significantly worse with overall accuracy at chance level. Humans also showed a truth bias in their decision criteria, reflected in a greater tendency to misclassify deepfake images as ‘real’ and this bias was accompanied by low deepfake image classification confidence. These findings suggest that sophisticated ML models can generate deepfake face images that are indistinguishable from real face images in the eye of human perceivers.

Regarding classification of deepfake videos (Study 2), however, we found that humans outperformed the FaceForensics algorithm in overall accuracy, and the machine also showed a truth bias and low classification confidence for deepfake videos. In contrast, humans’ greater accuracy and lower truth bias when classifying deepfake videos than images, and also relative to machine performance, suggests that rich perceptual cues in dynamic stimuli (e.g., motion and temporal consistency) facilitate deepfake detection in humans whereas ML algorithms were less able to benefit from such cues.

This differential pattern of findings for images vs. videos point out that humans and machines employ rather different processes in deepfake detection by highlighting the potential for a human-AI collaboration to optimize performance (e.g., by supporting human decision making with machine predictions and by feeding human-perceived cues/features to improve the algorithms’ predictions, Groh et al., 2022; Miller et al., 2023). Along these lines, future research could use two-alternative forced choice designs, where a deepfake face image is presented alongside its corresponding real face while recording eye movements of human perceivers. This approach would allow researchers to identify erroneous visual viewing patterns and attention to non-diagnostic cues in humans, and this could then be followed up with AI-guided eye tracking training, in which diagnostic features deemed as critical by ML for deepfake detection are targeted.

4.3. Higher Analytical Thinking, Lower Positive Affect, and Greater Internet Skills Predict Better Video Deepfake Detection

Results from Study 2 fill an important research gap in that they support that higher analytical thinking, less positive affect, and greater internet skills subserved better discernment of deepfake from real videos. Analytical thinking is a reliable predictor of fake news detection (Bago et al., 2020; Pehlivanoglu et al., 2021, 2022; Pennycook & Rand, 2019). Extending this work to deepfakes for the first time, findings from our study suggest that elaborative, relative to shallow, processing may foster attention to spot digital manipulations (e.g., face swapping) in video deepfakes. We also found that less positive affect was related with greater discernment between deepfake and real videos. This finding is in line with evidence that less positive affect enhances deliberative decision making (Schwarz & Clore, 2003) and deception detection (Matovic et al., 2014; but see Ebner et al., 2020). Finally, higher power usage was related to better ability to distinguish between deepfake and real videos. There is previous evidence showing that time spent on social media was linked to less susceptibility to fake news (Halpern et al., 2019) and deepfake videos (Nas & de Kleijn, 2024). Our measure on power usage went beyond previous research, which focused solely on time spent on social media, by considering and demonstrating the role of prior experience, expertise, and self-efficacy on video deepfake detection for the first time.

4.4. Conclusions

Across two studies, employing deepfake images and videos, and directly comparing humans and machines, here we found that ML algorithms have superior accuracy and better feature classification for real and deepfake images than videos. The machines' underperformance for videos is accompanied by a truth bias and low classification confidence for deepfake videos. We also found that humans outperformed ML algorithms in deepfake video detection; while they perform only at chance level on deepfake images, for which they display a truth bias and low decision confidence. We also provide first evidence that higher analytical thinking, less positive affect, and greater internet skills are conducive to better discernment between real and deepfake discernment videos. These findings combined importantly advance understanding of the processes involved in deepfake detection, delineating conditions under which human-machine collaboration may be particularly fruitful.

Funding Sources

This work was supported by the Department of Psychology, College of Liberal Arts and Science, University of Florida, the National Institute on Aging of the National Institutes of Health grants 1R01AG057764 and R01AG072658, and the Florida Department of Health Ed and Ethel Moore Alzheimer's Disease Research Program grant 22A10. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CRediT authorship contribution statement

Didem Pehlivanoglu: Conceptualization, Methodology, Resources, Data Curation, Formal Analysis, Visualization, Writing - Original Draft, Writing - Review & Editing, Supervision. **Mengdi Zhu:** Conceptualization, Methodology, Resources, Software, Data Curation, Formal analysis, Visualization, Writing - Original Draft, Writing - Review & Editing. **Jialong Zhen:** Conceptualization, Methodology, Data Curation, Formal analysis, Visualization, Writing - Review & Editing. **Aude A. Gagnon-Roberge:** Software, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing. **Rebecca K. Kern:** Software, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing. **Damon Woodard:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Funding acquisition. **Brian S. Cahill:** Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Supervision. **Natalie C. Ebner:** Conceptualization, Methodology, Resources, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition.

Data Availability

The full set of de-identified datasets, analysis scripts, and materials from Study 1 and 2 are available on OSF at https://osf.io/qhm3y/?view_only=bdc41a53bf7a4367bde6951372d9c932.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/WIFS.2018.8630761

Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S.-N. (2020). Detecting Deep-Fake Videos from Appearance and Behavior. *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. https://doi.org/10.1109/WIFS49906.2020.9360904

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bogaerts, K., Walentynowicz, M., Van Den Houte, M., Constantinou, E., & Van den Bergh, O. (2022). The Interoceptive Sensitivity and Attention Questionnaire: Evaluating Aspects of Self-Reported Interoception in Patients With Persistent Somatic Symptoms, Stress-Related Syndromes, and Healthy Controls. *Psychosomatic Medicine*, *84*(2), 251. https://doi.org/10.1097/PSY.0000000000001038

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect ‘deepfake’ images of human faces. *Journal of Cybersecurity*, *9*(1), tyad011. https://doi.org/10.1093/cybsec/tyad011

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

Ding, D., Chen, Y., Lai, J., Chen, X., Han, M., & Zhang, X. (2020). Belief Bias Effect in Older Adults: Roles of Working Memory and Need for Cognition. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02940

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The DeepFake Detection Challenge (DFDC) Dataset* (arXiv:2006.07397). arXiv. http://arxiv.org/abs/2006.07397

Durall, R., Keuper, M., Pfreundt, F.-J., & Keuper, J. (2019). *Unmasking DeepFakes with simple Features*. arXiv. https://doi.org/10.48550/ARXIV.1911.00686

Ebner, N. C., Ellis, D. M., Lin, T., Rocha, H. A., Yang, H., Dommaraju, S., Soliman, A., Woodard, D. L., Turner, G. R., Spreng, R. N., & Oliveira, D. S. (2020). Uncovering Susceptibility Risk to Online Deception in Aging. *The Journals of Gerontology: Series B*, *75*(3), 522–533. https://doi.org/10.1093/geronb/gby036

Ebner, N. C., Pehlivanoglu, D., & Shoenfelt, A. (2023). Financial Fraud and Deception in Aging. *Advances in Geriatric Medicine and Research*, *5*(3), e230007. https://doi.org/10.20900/agmr20230007

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, *34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, *44*(5), 1362–1367. https://doi.org/10.1016/j.jesp.2008.04.010

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks*.

Gottfried, J. (2019, June 14). About three-quarters of Americans favor steps to restrict altered videos and images. *Pew Research Center*. https://www.pewresearch.org/short-reads/2019/06/14/about-three-quarters-of-americans-favor-steps-to-restrict-altered-videos-and-images/

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (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

Güera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. https://doi.org/10.1109/AVSS.2018.8639163

Guess, A. M., & Munger, K. (2023). Digital literacy and online political behavior. *Political Science Research and Methods*, *11*(1), 110–128. https://doi.org/10.1017/psrm.2022.17

Gunderson, C. A., & ten Brinke, L. (2022). The Connection Between Deception Detection and Financial Exploitation of Older (vs. Young) Adults. *Journal of Applied Gerontology*, *41*(4), 940–944. https://doi.org/10.1177/07334648211049716

Gupta, P., Chugh, K., Dhall, A., & Subramanian, R. (2020). The eyes know it: FakeET- An Eye-tracking Database to Understand Deepfake Perception. *Proceedings of the 2020 International Conference on Multimodal Interaction*.

Haigh, M. (2016). Has the standard Cognitive Reflection Test become a victim of its own success? *Advances in Cognitive Psychology*, *12*(3), 145–149. https://doi.org/10.5709/acp-0193-5

Halpern, D., Valenzuela, S., Katz, J., & Orrego Miranda, J. (2019). *From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News* (pp. 217–232). https://doi.org/10.1007/978-3-030-21902-4_16

Hargittai, E. (2009). An Update on Survey Measures of Web-Oriented Digital Literacy. *Social Science Computer Review*, *27*(1), 130–137. https://doi.org/10.1177/0894439308318213

Hartung, J., Reuter, S., Kulow, V. A., Fähling, M., Spreckelsen, C., & Mrowka, R. (2024). *Experts fail to reliably detect AI-generated histological data*. https://doi.org/10.1101/2024.01.23.576647

Heemskerk, A., Lin, T., Pehlivanoglu, D., Hakim, Z., Valdes Hernandez, P. A., ten Brinke, L., Grilli, M. D., Wilson, R. C., Turner, G. R., Spreng, R. N., & Ebner, N. C. (2024). Interoceptive Accuracy Enhances Deception Detection in Older Adults. *The Journals of Gerontology: Series B*, gbae151. https://doi.org/10.1093/geronb/gbae151

Hulzebosch, N., Ibrahimi, S., & Worring, M. (2020). Detecting CNN-Generated Facial Images in Real-World Scenarios. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2729–2738. https://doi.org/10.1109/CVPRW50498.2020.00329

Josephs, E., Fosco, C., & Oliva, A. (2024). Effects of Browsing Conditions and Visual Alert Design on Human Susceptibility to Deepfakes. *Journal of Online Trust and Safety*, *2*(2), Article 2. https://doi.org/10.54501/jots.v2i2.144

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, *8*, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Juric, M. (2017, March 15). *The role of the need for cognition in the university students' reading behaviour* [Text]. University of Borås. https://informationr.net/ir/22-1/isic/isic1620.html

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and Improving the Image Quality of StyleGAN*.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *CoRR*, *abs/1312.6114*.

Köbis, N. C., Doležalová, B., & Soraperra, I. (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, P., & Marcel, S. (2021). Subjective and Objective Evaluation of Deepfake Videos. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2510–2514. https://doi.org/10.1109/ICASSP39728.2021.9414258

Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). Fast Face-Swap Using Convolutional Neural Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3697–3705. https://doi.org/10.1109/ICCV.2017.397

Kulis, B. (2013). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, *5*(4), 287–364. https://doi.org/10.1561/2200000019

Lane, J. D., & DePaulo, B. M. (1999). Completing Coyne’s Cycle: Dysphorics’ Ability to Detect Deception. *Journal of Research in Personality*, *33*(3), 311–329. https://doi.org/10.1006/jrpe.1999.2253

Li, Y., Chang, M.-C., & Lyu, S. (2018). *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking*. arXiv. https://doi.org/10.48550/ARXIV.1806.02877

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user’s guide* (pp. xv, 407). Cambridge University Press.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, *5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92. https://doi.org/10.1109/WACVW.2019.00020

Matovic, D., Koch, A. S., & Forgas, J. P. (2014). Can negative mood improve language understanding? Affective influences on the ability to detect ambiguous communication. *Journal of Experimental Social Psychology*, *52*, 44–49. https://doi.org/10.1016/j.jesp.2013.12.003

Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The Multidimensional Assessment of Interoceptive Awareness, Version 2 (MAIA-2). *PLOS ONE*, *13*(12), e0208034. https://doi.org/10.1371/journal.pone.0208034

Mehling, W. E., Gopisetty, V., Daubenmier, J., Price, C. J., Hecht, F. M., & Stewart, A. (2009). Body awareness: Construct and self-report measures. *PloS One*, *4*(5), e5614. https://doi.org/10.1371/journal.pone.0005614

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, *34*(12), 1390–1403. https://doi.org/10.1177/09567976231207095

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, *54*(1), 7:1-7:41. https://doi.org/10.1145/3425780

Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E. R., Yang, J., Guera, D., Zhu, F. M., & Delp, E. J. (2020). Deepfakes Detection with Automatic Face Weighting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2851–2859.

Nas, E., & de Kleijn, R. (2024). Conspiracy thinking and social media use are associated with ability to detect deepfakes. *Telematics and Informatics*, *87*, 102093. https://doi.org/10.1016/j.tele.2023.102093

Natsume, R., Yatagawa, T., & Morishima, S. (2018). *FSNet: An Identity-Aware Generative Model for Image-based Face Swapping*.

Nightingale, S. J., & Farid, H. (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

Nightingale, S. J., & Wade, K. A. (2022). Identifying and minimising the impact of fake visual media: Current and future directions. *Memory, Mind & Media*, *1*, e15. https://doi.org/10.1017/mem.2022.8

Pehlivanoglu, D., Lighthall, N. R., Lin, T., Chi, K. J., Polk, R., Perez, E., Cahill, B. S., & Ebner, N. C. (2022). Aging in an “infodemic”: The role of analytical reasoning, affect, and news consumption frequency on news veracity detection. *Journal of Experimental Psychology: Applied*, *28*(3), 468. https://doi.org/10.1037/xap0000426

Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, *6*(1), 24. https://doi.org/10.1186/s41235-021-00292-3

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Röcke, C., Li, S.-C., & Smith, J. (2009). Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults? *Psychology and Aging*, *24*(4), 863–878. https://doi.org/10.1037/a0016276

Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2023). *Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?* Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2023.017

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images* (arXiv:1901.08971). arXiv. https://doi.org/10.48550/arXiv.1901.08971

Sambhu, N., & Canavan, S. (2020). *Detecting Forged Facial Videos using convolutional neural network*. arXiv. https://doi.org/10.48550/ARXIV.2005.08344

Schwarz, N., & Clore, G. L. (2003). Mood as Information: 20 Years Later. *Psychological Inquiry*, *14*(3–4), 296–303. https://doi.org/10.1207/S15327965PLI1403&4_20

Seow, J. W., Lim, M. K., Phan, R. C. W., & Liu, J. K. (2022). A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, *513*, 351–371. https://doi.org/10.1016/j.neucom.2022.09.135

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). A Study of the Human Perception of Synthetic Faces. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8. https://doi.org/10.1109/FG52635.2021.9667066

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*, 107917. https://doi.org/10.1016/j.chb.2023.107917

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought* (pp. xv, 308). Yale University Press.

Sundar, S. S., & Marathe, S. S. (2010). Personalization versus Customization: The Importance of Agency, Privacy, and Power Usage. *Human Communication Research*, *36*(3), 298–322. https://doi.org/10.1111/j.1468-2958.2010.01377.x

Suratkar, S., Johnson, E., Variyambat, K., Panchal, M., & Kazi, F. (2020). Employing Transfer-Learning based CNN architectures to Enhance the Generalizability of Deepfake Detection. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–9. https://doi.org/10.1109/ICCCNT49239.2020.9225400

Sütterlin, S., Lugo, R. G., Ask, T. F., Veng, K., Eck, J., Fritschi, J., Özmen, M.-T., Bärreiter, B., & Knox, B. J. (2022). The Role of IT Background for Metacognitive Accuracy, Confidence and Overestimation of Deep Fake Recognition Skills. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition* (pp. 103–119). Springer International Publishing. https://doi.org/10.1007/978-3-031-05457-0_9

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting Both Machine and Human Created Fake Face Images In the Wild. *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, 81–87. https://doi.org/10.1145/3267357.3267367

ten Brinke, L., Lee, J. J., & Carney, D. R. (2019). Different physiological reactions when observing lies versus truths: Initial evidence and an intervention to enhance accuracy. *Journal of Personality and Social Psychology*, *117*(3), 560. https://doi.org/10.1037/pspi0000175

Ternovski, J., Kalla, J., & Aronow, P. (2022). The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments. *Journal of Online Trust and Safety*, *1*(2), Article 2. https://doi.org/10.54501/jots.v1i2.28

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Tong, X., Wang, L., Pan, X., & Wang, J. G. (2020). An Overview of Deepfake: The Sword of Damocles in AI. *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 265–273. https://doi.org/10.1109/CVIDL51233.2020.00-88

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Tsfati, Y., & Cappella, J. (2003). Do People Watch what they Do Not Trust? *Communication Research*, *30*, 504–529. https://doi.org/10.1177/0093650203253371

Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (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

Tursman, E., George, M., Kamara, S., & Tompkin, J. (2020). Towards Untrusted Social Video Verification to Combat Deepfakes via Face Geometry Consistency. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2784–2793. https://doi.org/10.1109/CVPRW50498.2020.00335

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910–932. IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2020.3002101

Verplanken, B., Hazenberg, P. T., & Palenéwen, G. R. (1992). Need for cognition and external information search effort. *Journal of Research in Personality*, *26*(2), 128–136. https://doi.org/10.1016/0092-6566(92)90049-A

Vraga, E. K., & Tully, M. (2021). News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society*, *24*(2), 150–166. https://doi.org/10.1080/1369118X.2019.1637445

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). *CNN-generated images are surprisingly easy to spot... For now*.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Yang, X., Li, Y., & Lyu, S. (2018). *Exposing Deep Fakes Using Inconsistent Head Poses*. arXiv. https://doi.org/10.48550/ARXIV.1811.00661

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A Survey on Deepfake Video Detection. *IET Biometrics*, *10*(6), 607–624. https://doi.org/10.1049/bme2.12031

Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, *81*(5), 6259–6276. https://doi.org/10.1007/s11042-021-11733-y