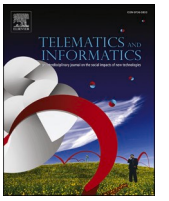
Telematics and Informatics 87 (2024) 102093

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele



Conspiracy thinking and social media use are associated with ability to detect deepfakes

Ewout Nas, Roy de Kleijn *

Leiden University, Cognitive Psychology Unit, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

ARTICLE INFO

Keywords:
Deepfake videos
Deepfakes
Fake news
Classification

ABSTRACT

Deepfake videos are highly realistic manipulated videos that are often difficult to distinguish from authentic videos. The technology rapidly evolves, making deepfake videos increasingly realistic. Most research on deepfake videos is focused on the algorithmic detection of deepfakes. Less is known about the human recognition of deepfake videos. The aim of the current study was to investigate the predictors of human performance at recognizing deepfake videos. Our findings show that humans perform better at recognizing deepfake videos of familiar persons compared to deepfakes of unfamiliar persons. Next, our findings show a positive relationship between time spent on social media and deepfake detection performance, as well as conspiracy thinking and deepfake detection performance. No relationships were found between age and gender and deepfake detection performance.

1. Introduction

Deepfake videos, also known as *deepfakes,* are highly realistic manipulated videos generated with artificial intelligence (Westerlund, 2019). Deepfake technology enables people to swap faces of individuals on other videos, making the individual say and do anything they want, making them a public concern (Maras and Alexandrou, 2019; Tolosana et al., 2020). The technology brings several potential applications, such as sabotaging an individual, falsifying media, or creating fake evidence. The development of deepfake technology is worrying due to it being easy to use, and highly realistic, making it difficult to detect (Fletcher, 2018).

The creation of deepfakes involves a process that starts with feeding a neural network a large amount of footage of two individuals (Westerlund, 2019). The architecture used is known as a *generative adversarial network* (GAN), which is a combination of a generative and a discriminative model (Goodfellow et al., 2014). The network learns how to mimic a person's facial form, facial expressions, mannerisms, and voice. Following this, the network swaps the face of a person into the face of another person, creating a fake video (Westerlund, 2019). Given that the network has to be trained with a large amount of data (footage of individuals), famous persons are the most likely targets of deepfake technology, as a lot of footage of them is readily available online (Westerlund, 2019). Until recently, deepfake technology was not accessible to the broad public, but due to the improvement of algorithms and hardware such as CPUs and GPUs the technology is now largely available to individuals (Maras and Alexandrou, 2019; Westerlund, 2019).

While deepfake technology has several commercial applications, e.g. computer-generated imagery in movies, dubbing advertisements to other languages while keeping the original voice, or creating virtual fitting rooms for webshop clients (Maras and Alexandrou, 2019; Westerlund, 2019), the technology also brings undesirable opportunities such as political sabotage (Fletcher,

* Corresponding author.
E-mail addresses: e.nas@hva.nl (E. Nas), kleijnrde@fsw.leidenuniv.nl (R. de Kleijn).

https://doi.org/10.1016/j.tele.2023.102093
Received 16 October 2023; Received in revised form 28 November 2023; Accepted 21 December 2023
Available online 27 December 2023
0736-5853/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2018). Deepfake technology can destabilize society and undermine a sense of social reality and the legal system by blurring the line between what is real and what is fake (Yadlin-Segal and Oppenheim, 2 2021.). One of the biggest threats may be that deepfakes can be used to create *fake news* (Westerlund, 2019). Fake news is defined as fabricated information that mimics news media content that is promoted on social media to mislead the audience for ideological or financial gain (Lazer et al., 2018).

1.1. Fake news

Allcott and Gentzkow (2017) studied the role of social media and fake news in the 2016 elections in the United States, describing the evolution of news publishing. In the 19th century, only a limited number of newspapers was available. Newspapers wrote for a big audience, were relatively objective and had strong requirements for quality control. In the current day and age, social media plays a big role in news sharing. This allows not only newspapers but also individuals to publish news for a wide audience to view. When individuals post articles and opinions on social media, there is no filtering, fact-checking or editorial judgment, making it easy to publish fake news. Additionally, due to the use of social media algorithms social media users are presented with posts that mostly align with their own opinions and interests, causing echo chambers. The combination of the ability to post articles without a filter and the presence of social media echo chambers enhances the sharing of fake news, potentially misleading millions of people (Allcott and Gentzkow, 2017).

To gain insight into how widespread fake news is, Allcott and Gentzkow (2017) analyzed a database of 156 fake news articles during the month around the 2016 U.S. elections. A conservative estimate was that the average American adult encountered one to three fake news articles from this database. Finally, Allcott and Gentzkow (2017) studied who are more likely to correctly identify true versus fake news articles. Three positive correlations were found: people who spend more time on consuming media, people with a higher education and people with a higher age have more accurate beliefs about news. No significant relationship was found between the use of social media and accurate beliefs about news (Allcott and Gentzkow, 2017).

Keersmaecker and Roets (2017) studied the role of cognitive ability on the impact of fake news. It was found that people change their attitudes when they find out that they are based on false information. However, people with lower cognitive ability adjust their attitudes to a lesser extent, and their attitudes remain more biased. This means that the effects of presenting people with incorrect information cannot simply be undone by pointing the incorrectness of this information afterwards. Social media platforms might be able to provide a solution to this problem, as algorithms can be used to detect fake news (Monti et al., 2019). This enables the platform to generate an automatic warning for fake news posts, making a user aware that a post might be fake news before the user reads the post.

Another factor that likely plays a role in fake news belief is conspiracy belief (Halpern et al., 2019). Conspiracy theories are defined as explanations for important events involving secret plots by powerful groups (Goertzel, 1994). Conspiracy theories are attractive for people with monological views, since it provides an easy explanation for any new phenomenon that threatens their belief system. There are various reasons people believe in conspiracy theories (Goreis and Voracek, 2019). Common traits from conspiracy theorists are feeling disconnected from society, being unhappy with circumstances, and not having a feeling of being in control of their life. Halpern et al. (2019) studied the relationship between conspiracy mentality and fake news, and suggest that people with a higher conspiracy mentality are more likely to believe in fake news.

Although social media plays a big role in the distribution of fake news, the use of social media use is negatively correlated with fake news belief, suggesting that people who spend more time on social media have developed more awareness of fake news (Guess et al., 2019; Halpern et al., 2019). Guess et al. (2019) found that Facebook users over 65 years old shared almost seven times as many articles from fake news websites, compared to the youngest group of this study, people between 18 and 29 years old. This effect held after correction for education, political ideology and overall posting activity. A potential explanation for this effect is cognitive deficit. Abilities such as episodic memory and abstract reasoning start declining when adults are in their 20s and 30s (Brashier and Schacter, 2020; Salthouse, 2009). Older adults tend to successfully categorize true and false headlines at the first glance but may start believing fake news after being exposed to it repeatedly. Another contributor could be that social networks of older adults shrink, making them lose peripheral social partners. This can lead to a misplaced trust, thinking that content shared by friends and family must be true. Furthermore, interpersonal trust increases with age (Poulin and Haase, 2015). Lastly, according to Krumsvik et al. (2016), there is a negative relationship between age and digital competence. Therefore, people of an older age might be less competent at recognizing deepfake videos.

1.2. Recognizing manipulated pictures and videos

The rise of deepfake technology brings more concerns than fake news on its own (Korshunov and Marcel, 2018). As open source deepfake generation software is broadly accessible to the public (e.g. DeepFaceLab), large amounts of realistic deepfake videos circulate on social media, making it important to study how humans perform at recognizing deepfake videos and what characteristics play a role in this performance. Several studies have been conducted on the human recognition of manipulated pictures and deepfake videos.

Nightingale et al. (2017) studied the human recognition of manipulated photos. It was found that the ability to detect manipulated photos is extremely limited, with 62% accuracy. People find it hard to detect manipulations in photos of the real world, even when a manipulated picture is physically implausible. Ostrovsky et al. (2005) found that the human visual system lacks sensitivity to inconsistent lighting in a picture, and inconsistent lighting could be an indicator of deepfake videos. Interestingly, no evidence was found that having interest in photography or having beliefs about the extent of manipulated photos in society is associated with

improved recognition accuracy.

Shahid et al. (2022) presented a qualitative study on the perception of deepfake videos by Indian social media users. 36 participants were presented with three deepfake videos and one authentic video. Only one participant assessed all videos as real or deepfake correctly. Key findings were that most users are not aware of manipulated fake videos, let alone deepfake videos, and expect fake videos to be of low quality. Furthermore, most users lack the skills to recognize a fake video.

Tahir et al. (2021) designed a training program for humans to improve deepfake detection skills, which was found to be effective. The training involved exposure to deepfakes, detailed examples of common inconsistencies in deepfake videos, and example strategies to identify these inconsistencies. This raises the question of how effective these strategies will be in the long term, as deepfake technology is expected to improve together with human detection abilities.

In another study investigating deepfake detection interventions, Somoray and Miller (2023) showed people 20 videos, of which half were deepfakes, and reported a mean accuracy of 60.7%. Half of all participants were given a detection strategy to improve deepfake detection, but this intervention did not have an effect on accuracy.

Korshunov and Marcel (2020) compared the human performance at detecting deepfake videos with the performance of algorithms. This study made use of 120 videos (60 authentic and 60 deepfake videos). The deepfake videos were defined in five difficulty categories. Participants were shown the videos and had to classify the videos as "deepfake", "real", or "I am not sure". In the category "very difficult" participants were confused in 75.5% of the cases. It was also found that algorithms differ greatly from humans at detecting deepfake videos, having trouble with identifying some deepfakes that look obviously fake to humans, but also easily detecting videos that are difficult to human participants.

Khodabakhsh et al. (2019) studied the performance of 30 participants in distinguishing deepfake videos from authentic videos. It was found that being familiar with a person portrayed in a video is associated with a higher accuracy at distinguishing authentic from deepfake. No significant effect was found between age and accuracy, possibly due to the sample size. Next to this, it was found that humans rely on a small number of cues when assessing videos, mostly looking at the face area of the video.

Groh et al. (2022) studied deepfake detection by humans, machines, and humans informed by machines. Key findings were that specialized visual processing of faces helps humans distinguish authentic from fake visual media. The authors also found that participants were significantly better at detecting deepfake videos of well-known political leaders than videos of unknown persons.

Across studies, human deepfake detection accuracy ranges from 57.6% to 88.9% (Somoray and Miller, 2023), although methodological differences make studies difficult to compare.

Finally, although not investigating classification performance, Ahmed (2021) held a survey on the likelihood of individuals sharing deepfakes, shedding light on the relationship between social media use and deepfake distribution. Key findings were a negative correlation between cognitive ability and the sharing of deepfake videos. No correlation was found between social network size and the sharing of deepfake videos. However, individuals with an extensive social network are more likely to be exposed to disinformation.

2. Present study

Most research on deepfakes so far has focused on algorithmic detection, and the relatively little research described above has not shed light on the *predictors* of deepfake detection performance. In the current study we investigated the relationship between several personal characteristics and performance at recognizing deepfake videos, helping us better understand how humans recognize deepfake videos.

As previous research has shown that people with a higher conspiracy mentality are more likely to share and believe fake news (Halpern et al., 2019), it could be expected that people with a higher conspiracy mentality are likely to perform worse at recognizing deepfake videos.

As mentioned above, age is negatively correlated with digital competence (Krumsvik et al., 2016). Additionally, older people are more likely to share fake news (Guess et al., 2019). Assuming that persons who are less digital competent are less familiar with deepfake videos and how to recognize them, we expect a negative correlation between age and deepfake detection performance.

Time spent on social media is negatively correlated to believing fake news (Halpern et al., 2019). This might be explained by the idea that avid social media users have developed more awareness of fake news. A similar correlation might be present between social media use and believing deepfake videos. Therefore, it is expected that people who spend more time on social media are likely to perform better at recognizing deepfake videos.

Recent studies on the human recognition of deepfake videos found that being familiar with a person portrayed on a deepfake video is associated with a higher deepfake detection accuracy (Khodabakhsh et al., 2019; Groh et al., 2022). We expected to replicate this finding.

In summary, the current study had the following hypotheses:

1. Conspiracy mentality is inversely associated with deepfake classification performance.
2. Age is inversely associated with deepfake classification performance.
3. Time spent on social media is positively associated with deepfake classification performance.
4. Familiarity with the depicted person is positively associated with deepfake classification performance.

3. Methods

3.1. Participants

Initially we recruited 101 participants (14 males, 86 females, 1 participant preferred not to say; age $M = 20.0$ years old, $SD = 2.2$) through the SONA Leiden University research participant platform and social networks. We used a Bayesian optional stopping criterion of $BF_{10} = 5$ to stop data collection after reaching the required number of participants. In a second phase, we recruited 29 more through the SONA Leiden University research participant platform due to some analyses being underpowered (i.e. $\frac{1}{3} < BF < 3$), ending up with a sample of 130 participants (16 males, 110 females, 4 other; age $M = 20.0$ years old, $SD = 2.9$). The second phase of recruitment used the same advertisement on SONA with an identical description. In both phases of recruitment, inclusion criteria were age between 18–65 years old. There were no exclusion criteria. Participants were rewarded with 2 course credits. The participant with the highest accuracy score at detecting deepfake videos was rewarded with €10. The study was approved by the Leiden University Psychology Research Ethics Committee.

3.2. Materials

Several questionnaires and a deepfake detection task were administered on Qualtrics, after participants completed a demographics questionnaire.

3.2.1. Conspiracy mentality

Conspiracy mentality was measured using the Conspiracy Mentality Questionnaire (Bruder et al., 2013). This validated questionnaire contains five questions with 11-point scales from 0% to 100% agreement and is designed to measure the generic tendency to engage in conspiracy mentality within and across cultures. The internal consistency of the English 5-item CMQ is good ($\alpha = 0.84$). An example item from the CMQ includes "I think that there are secret organizations that greatly influence political decisions." A higher score on the CMQ indicated higher conspiracy mentality. The score used in our analyses is the sum of all five agreement measures, normalized from 0 to 10. (e.g. answering all five items with 50% leads to a score of 5.0. The full CMQ is included in Appendix A.

3.2.2. Social media

Social media usage was assessed via a set of questions about participants' social media platform use, weekly time spent on social media, number of social media contacts and weekly number of posts created. An example item includes "How often do you post something on social media per week?" The complete list of questions is included in Appendix B.

3.2.3. Deepfake detection performance task

We designed a task to assess the participants' performance at detecting deepfake videos, making use of the Celeb-DF dataset (Li et al., 2020). This dataset was originally created to train algorithms to detect deepfake videos and to assess the performance of deepfake detection algorithms. The dataset consists of videos of 59 celebrities, with a diverse distribution in gender, age and ethnicity. The videos provided in this dataset do not contain audio. The dataset contains a total of 6229 videos, of which 590 are authentic and 5639 are deepfakes based on the authentic videos, with an average length of 13 s. The deepfake videos were generated by swapping faces of the 59 subjects. Fig. 1 shows an example frame of an authentic video and a corresponding deepfake video.

For each of the 59 celebrities in the dataset, we randomly selected one authentic video. We then randomly selected, for each authentic video, two corresponding deepfakes. For three of the authentic videos only one corresponding deepfake was present, leading to a total of 174 videos, of which 59 were authentic videos of different celebrities, and 115 were deepfake videos corresponding to the authentic videos. Participants were shown all videos in randomized order, and authentic videos were not paired with their corresponding deepfakes. For each of the videos, participants were asked to assess whether it was an authentic or a deepfake video, and whether they were familiar with the celebrity portrayed or not as a binary question.

As a measure of classification performance we used the Matthews correlation coefficient (MCC), defined as

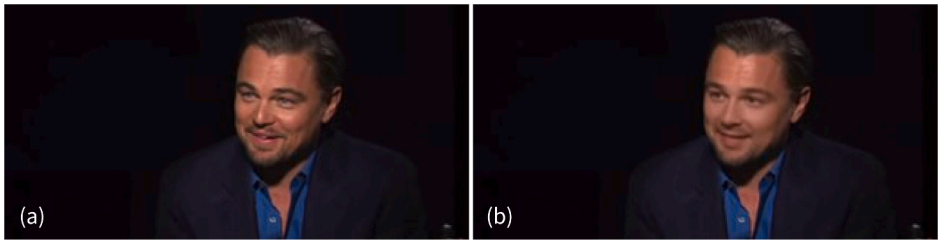


Fig. 1. Example frames of the videos used in the study taken from the dataset of Li et al. (2020). Frame (a) is an authentic video, frame (b) is a deepfake video corresponding to. the left video.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

with TP, FP, TN, FN being true and false positives and negatives.

The MCC has several advantages over accuracy and F1 score, two other widely used measures of binary classification. Accuracy is defined as the ratio between the number of correctly classified items and the total number of items, overestimating classification performance on the majority class in unbalanced datasets such as the current one (Chicco and Jurman, 2020). The F1 score is better suitable for unbalanced datasets, although it has been criticized for being misleading as it ignores true negatives, and is sensitive to class swapping (swapping positive and negative class labels; Sitarz (2022)). The MCC uses all four confusion matrix categories, and takes a value between −1 and +1, with +1 indicating perfect classification, 0 indicating chance-level performance, and −1 indicating incorrect classification for all items, and is insensitive to class swapping.

3.3. Design and procedure

The participants took part in this study from their computers at home. All tasks were performed on Qualtrics. First, participants were asked for their informed consent. Following this, the demographics questionnaire, conspiracy mentality questionnaire and the social media questionnaire were administered. Finally, the deepfake detection task was performed. Trials were presented in random order. In the end, the participants were debriefed.

3.4. Statistical analysis

Following the data collection, all statistical analyses were performed in R 4.2.2 (R Core Team, 2022). Bayesian analyses were performed using the RStan, bayestestR, and BayesFactor packages. For all Bayesian analyses for correlation, we used the standard beta (3, 3) prior for ρ. For Bayesian *t*-tests we used the standard prior scale of $\sqrt{2}/2$. Unless otherwise indicated, interpretation of effect sizes and Bayes factors correspond to the interpretations of Cohen (1988) and Jeffreys (1999), respectively. Following the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework (Makowski et al., 2019), for Bayesian analyses we report the median of the posterior distribution and its 95% CI (Highest Density Interval), along the probability of direction (pd), the probability of significance (i.e. being non-negligible) and the probability of being large. The thresholds beyond which the effects are considered as non-negligible and large are |0.05| and |0.30|, respectively.

4. Results

4.1. General classification performance

For each participant, we first computed their MCC score. A confusion matrix was then calculated for each participant. Mean accuracy was 0.80 and mean MCC was 0.59 ($SD = 0.22$). The mean false positive (classifying an authentic as a deepfake) rate was 0.18 ($SD = 0.13$). The mean false negative (classifying a deepfake as an authentic) rate was 0.21 ($SD = 0.15$). The mean true positive (classifying a deepfake as a deepfake) rate was 0.79 ($SD = 0.15$). The mean true negative (classifying an authentic as an authentic) rate was 0.82 ($SD = 0.13$). The confusion matrix over all responses given is listed under Table 1. Total MCC over all responses was 0.58, with a 0.15 false positive rate and a 0.24 false negative rate.

Over all responses given, mean accuracy for authentic videos was 0.82 ($SD = 0.11$, range 0.39–0.96) and mean accuracy for deepfakes was 0.79 ($SD = 0.13$, range 0.43–0.98), suggesting a wide range in difficulty. To look at the difference between authentic and deepfake classification performance, we fitted a logistic mixed model to predict accuracy with stimulus type (authentic or deepfake). As a fixed effect, we entered stimulus type into the model. As random effects, we had intercepts for participants, as well as by-participant random slopes for the effect of stimulus type. P-values were obtained by likelihood ratio tests of the full model with stimulus type as a fixed effect against the model without this fixed effect. The point estimate for the effect of stimulus type is 0.11, 95% confidence interval [−0.08, 0.30]. We conclude that there is no evidence for an effect of stimulus type on classification performance, $\chi^2 = 1.28$, $p = 0.258$.

Table 1
Confusion matrix over all responses given. NPV stands for negative predicted value.

		Predicted class		
		Fake	Real	
Actual class	Fake	11,782	1,413	Sensitivity 89%
	Real	3,168	6,257	Specificity 66%
		Precision 79%	NPV 82%	Accuracy 80%

4.2. Age and gender

The relationship between age and MCC is shown in Fig. 2. The mean age of the participants was 20.0 ($SD = 2.9$). Pearson's correlation test revealed that there was no correlation between participants' age and classification performance, $r(128) = 0.102$, $p = 0.249$. A Bayesian analysis shows that ρ (median = 0.09, 95% HDI [-0.07, 0.26]) has a 87.0% probability of being positive, 69.9% of being non-negligible, and 0.80% of being large. The observed data are 2.61 times more likely under the null hypothesis than under the alternative hypothesis. We conclude that there is anecdotal evidence against an effect of age on classification performance.

There was no difference between male ($M = 0.57$, $SD = 0.18$) and female ($M = 0.60$, $SD = 0.23$) participants on MCC, $t(124) = 0.497$, $p = 0.620$. A Bayesian analysis shows that the difference in means (median = 0.02, 95% HDI [−0.08, 0.13]) has a 31.7% probability of being non-negligible, and 0.01% of being large. The observed data are 3.34 times more likely under the null hypothesis than under the alternative hypothesis. We conclude that there is substantial evidence against an effect of gender on classification performance.

4.3. Conspiracy mentality

The relationship between conspiracy mentality and classification performance is shown in Fig. 3. The mean score on the Conspiracy Mentality Questionnaire was 5.65 ($SD = 1.58$). Pearson's correlation test revealed that the score on the Conspiracy Mentality Questionnaire was positively correlated with classification performance as measured by MCC, $r(128) = 0.212$, $p = 0.015$, corresponding to a small-to-medium effect size.

A Bayesian analysis shows that ρ (median = 0.20, 95% HDI [0.04, 0.36]) has a 99.0% probability of being positive, 96.6% of being non-negligible, and 10.7% of being large. The observed data are 3.39 times more likely under the alternative hypothesis than under the null hypothesis which is substantial evidence for the existence of an effect. We conclude that there is an effect of conspiracy mentality on classification performance.

4.4. Social media

The relationship between time spent on social media per week and classification performance is shown in Fig. 4. Mean time spent on social media per week was 16.4 h ($SD = 11.5$). Pearson's test for correlation revealed that time spent on social media was positively correlated with classification performance, $r(128) = 0.287$, $p < 0.001$, corresponding to a small-to-medium effect size.

A Bayesian analysis shows that ρ (median = 0.27, 95% HDI [0.12, 0.43]) has a 99.98% probability of being positive, 99.7% of being non-negligible, and 36.3% of being large. The observed data are 38.8 times more likely under the alternative hypothesis than under the null hypothesis, which is considered very strong evidence for the existence of an effect. To verify that this effect was not dependent on one extreme observation with 70 h per week spend on social media, we also ran the analysis with this outlier removed. This did not change the main outcome of the analysis, $r(127) = 0.263$, $p = 0.003$, $BF_{10} = 16.2$. We conclude that there is an effect of time spent on social media on classification performance.

Three other measure of social media activity, i.e. number of different social media platforms used, number of published social media posts per week, and number of social media contacts did not show a relationship with MCC, with $ps > 0.035$ and $BF_{01}s > 0.59$. There is moderate evidence for the absence of an effect on these measures, except for the number of different social media platforms used, where there is insufficient evidence to draw conclusions. More detailed statistics are shown in Table 2.

Finally, we tested a linear model including both conspiracy thinking as well as time spent on social media as predictors to see if they both independently predict classification performance. This resulted in a significant model, $F(2, 127) = 9.46$, $p < .001$, $R^2_{adj} = 0.116$. Both conspiracy thinking ($t = 2.63$, $p < .01$) and time spent on social media ($t = 3.52$, $p < .001$) were significant predictors of

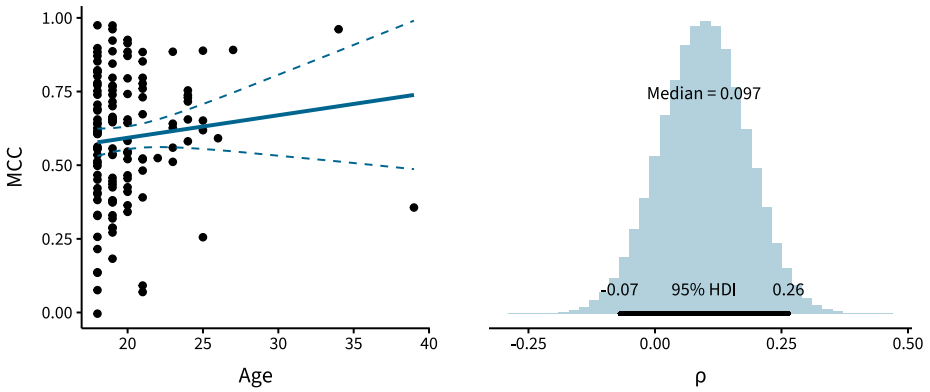


Fig. 2. Left: Participants' MCC as a function of age. Dotted lines indicate 95% CI of the regression line. Right: posterior distribution of ρ with the range of the 95% highest density interval and median value for ρ indicated.

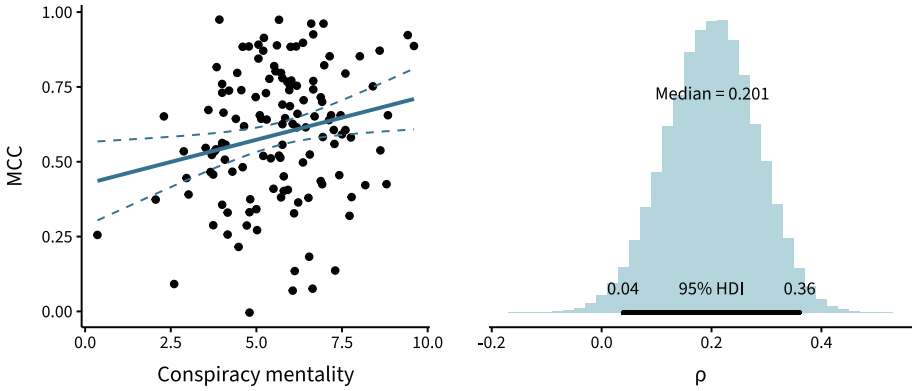


Fig. 3. Left: Participants' MCC as a function of conspiracy mentality as measured by the Conspiracy Mentality Questionnaire. Dotted lines indicate 95% CI of the regression line. Right: posterior distribution of ρ with the range of the 95% highest density interval and median value for ρ indicated.

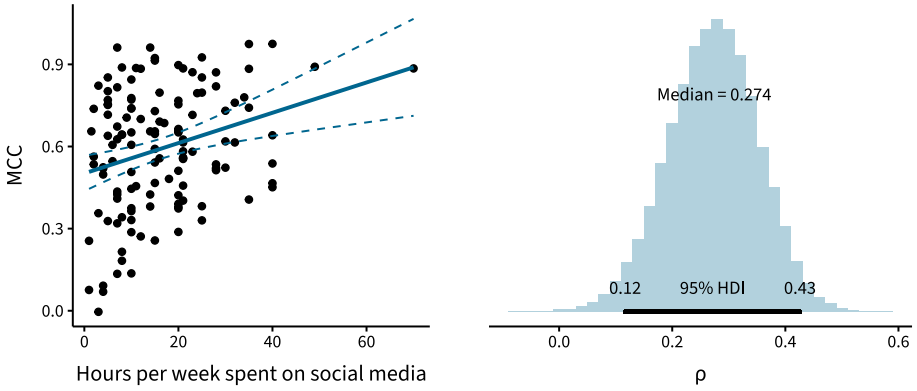


Fig. 4. Left: Participants' MCC as a function of weekly time spent on social media. Dotted lines indicate 95% CI of the regression line. Right: posterior distribution of ρ with the range of the 95% highest density interval and median value for ρ indicated.

Table 2
Relationship between different measures of social media use and MCC, expressed as *p*-values for Pearson's test for correlation, 95% highest density interval for ρ and BF_{01}.

Measure	*p*-value	95% HDI	BF_{01}
Number of different social media platforms used	0.035	[0.01, 0.34]	0.59
Number of published social media posts per week	0.249	[−0.07, 0.26]	2.61
Number of social media contacts	0.479	[−0.11, 0.23]	3.88

classification performance. There was no correlation between the two predictors, $r(128) = -0.02$, $p = 0.826$, $BF_{01} = 4.82$.

4.5. Familiarity

To investigate the effect of stimulus familiarity on classification performance, we used binary accuracy scores coding misses and false alarms as 0 and hits and correct rejections as 1. We used item-wise binary accuracy to perform this analysis, as MCC is defined as a summary statistic and does not apply to individual trials.

We fitted a crossed-level logistic mixed model to predict accuracy with familiarity. As a fixed effect, we entered familiarity into the model. As random effects, we had intercepts for participant and stimulus, as well as by-participant and by-stimulus random slopes for the effect of familiarity. P-values were obtained by likelihood ratio tests of the full model with familiarity as a fixed effect against the model without this fixed effect. The point estimate for the effect of familiarity is 0.81, 95% confidence interval [0.62, 0.99]. We conclude that familiarity affects classification performance ($\chi^2 = 68.84$, $p < .001$) with a 124% increase in the odds of correctly classifying a video as authentic or deepfake for familiar persons compared to unfamiliar persons.

4.6. Summary of results

We conclude that there is substantial evidence for a relationship between conspiracy mentality and deepfake classification performance, and strong evidence for a relationship between the number of hours per week spent on social media and deepfake classification performance. There is no evidence for a relationship between the other social media measures, age, and deepfake classification performance.

5. Discussion

The present study investigated human performance at recognizing deepfake videos and several predictors. What is the current human performance at detecting deepfake videos? Our results show a total accuracy score of 0.80 and a total MCC of 0.58, with a 0.15 false positive rate and a 0.24 false negative rate. This is higher than the 0.73 accuracy found by Groh et al. (2022) and an MCC of 0.46 (calculated by us using their confusion matrix), although it should be noted that Groh et al. (2022) used a different measure of accuracy, incorporating a certainty estimate. In addition, their design was not identical to ours, with a different dataset which could either reflect more realistic deepfakes or a different performance by participants. Regardless, this is a cause for concern, considering that deepfake videos will become increasingly realistic over time even if humans will become more familiar with deepfakes over time, enhancing their abilities to detect them.

We hypothesized a negative correlation between conspiracy beliefs and deepfake detection performance. Our results provide substantial evidence for a *positive* correlation between these two variables, suggesting that humans who believe in conspiracy theories perform better at detecting deepfake videos. Our hypothesis was based on the study of Halpern et al. (2019), which suggested that humans who believe in conspiracy theories are more likely to believe in fake news. A possible explanation for the contrast in findings is the difference between fake news and deepfake videos. Halpern et al. (2019) presented participants with real and fake news stories, whereas the current study used videos that did not contain audio. This means that the samples in the current study did not contain any text or contextual cues. One possible explanation for the effect itself is that people who believe in conspiracy theories tend to be more skeptical and suspicious of information they encounter. This suspiciousness may lead to more critical and alert viewing behavior when watching videos, including deepfakes, whereas news stories provide fewer cues for distinguishing authentic from fake.

Second, we hypothesized a negative relationship between age and deepfake detection performance based on several theories, one of which was that older persons tend to have a lower digital competence (Krumsvik et al., 2016). Our results do not indicate a significant correlation and show anecdotal evidence ($BF_{01} = 2.61$) against a relationship between age and deepfake detection performance.

Our third hypothesis was a positive relationship between social media activity and deepfake detection performance. We have used several measures to test this hypothesis, i.e. time spent on social media, the number of different social media platforms used, the number of published social media posts per week, and the number of social media contacts. In line with the hypothesis, the results of the present study provide supporting evidence for a positive correlation between time spent on social media and deepfake detection performance. Halpern et al. (2019) found a similar relationship between social media activity and ability to recognize fake news. A possible explanation for this effect is that individuals using social media frequently have more awareness of deepfake videos and fake news (Halpern et al., 2019). No significant correlations were found between deepfake detection performance, and the number of social media platforms used, the number of social media contacts and the number of social media posts made. These results are in line with the study of Ahmed (2021), which found no direct impact of social network size on sharing deepfake videos. Though, it is important to note that social network size might result in a higher exposure to deepfake videos.

Our last hypothesis stated a positive relationship between familiarity with the person portrayed on a deepfake video, and classification performance. The results show decisive evidence for this hypothesis, suggesting that it is easier for individuals to distinguish between authentic and deepfake videos of familiar persons than of unfamiliar persons, similar to what was found in Groh et al. (2022). A possible explanation for this finding is the idea that humans remember familiar faces and their characteristics, making it easier to spot discrepancies between real and fake. As mentioned in the introduction, famous persons are the most likely victims of deepfake technology (Westerlund, 2019). It is important to note that while famous persons are the likeliest targets of deepfake technology, humans perform better at detecting deepfakes of famous persons, making overall deepfake detection performance higher.

We can identify three potential limitations to the current study. First, only visual content was used in the samples. It is plausible that audio plays a significant role in recognizing deepfake videos, although much progress has been made in AI-generated voice. Second, it is important to consider that participants of the current study were given the assignment to distinguish real videos from deepfake videos. This made the participants aware of deepfake videos being present. In a real-life setting, individuals may not be aware at all that they might be looking at a deepfake. Third, as with any study investigating novel technologies, the results presented in this paper are based on a dataset that was recent at the time of the study design phase, and not necessarily state-of-the-art at the time of manuscript publication. While this does not detract from the main messages of the current study, which are focused on predicting classification performance instead of classification performance proper, it should be considered an inherent limitation that is, alas, unavoidable.

The results of the current study strongly imply the importance of an intervention to prevent humans from believing deepfake videos. Potential intervention strategies are either focused on humans, computers or a combination of both. Keeping in mind that a large part of the human population is not aware of the existence of deepfake videos (Shahid et al., 2022), it may be beneficial to set up campaigns to create awareness of deepfake videos. This could make individuals more alert to deepfake videos, and with this improve detection performance. Another potential human intervention is creating a deepfake detection training, improving detection skills of

humans (Tahir et al., 2021; Somoray and Miller, 2023). A dedicated training may not be feasible for the average person. However, it may be useful for people in specific occupations where more careful judgment of videos is required such as in the legal profession. Finally, algorithms can be used to detect deepfake videos. For example, social media could implement deepfake detection algorithms, to automatically filter out deepfake videos, or to alert users that a video is potentially fake. A combination of the interventions mentioned above may reduce the threat of deepfake technology being used for fake news.

In terms of future research, it would be useful to extend the current findings by studying participants from different age groups to draw better conclusions on the relationship between age and deepfake detection performance. Second, it would be useful to examine the relationship between digital skills and deepfake detection performance, the role of specific cognitive abilities in detecting deepfake videos, and to what extent deepfakes can influence a person's attitude towards a person portrayed on the video. Third, future research could include deepfakes that have actually been shared on social media, although these videos are then often already known to be deepfakes which could moderate classification performance.

6. Conclusion

The purpose of this study was to gain a better understanding of human deepfake recognition. There are three key findings of the present research. First, humans perform better at distinguishing deepfake from authentic videos of familiar people opposed to unfamiliar people. Second, individuals who spend more time on social media are better at distinguishing deepfake from authentic videos. Third, individuals who score higher on conspiracy thinking are better at distinguishing deepfake from authentic videos. With deepfake videos from the past five years or so, humans only have around 60–90% accuracy at distinguishing real from fake. Given the fast development of deepfake technology and its quality, it will likely become increasingly difficult to distinguish real videos from fake videos. Awareness campaigns about deepfake videos, training programs to improve deepfake recognition skills, as well as algorithmic deepfake detection might be a solution for this problem.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Conspiracy Mentality Questionnaire (CMQ)

Questions:
I think that…

1. many very important things happen in the world, which the public is never informed about.
2. politicians usually do not tell us the true motives for their decisions.
3. government agencies closely monitor all citizens.
4. events which superficially seem to lack a connection are often the result of secret activities.
5. there are secret organizations that greatly influence political decisions.

Answer options:

- 0% – certainly not
- 10% – extremely unlikely
- 20% – very unlikely
- 30% – unlikely
- 40% – somewhat unlikely
- 50% – undecided
- 60% – somewhat likely
- 70% – likely
- 80% – very likely
- 90% – extremely likely
- 100% – certain

Appendix B. Social media use questions

1. Which social media do you use? (checkboxes)
   (a) Facebook
   (b) Instagram
   (c) Pinterest
   (d) Reddit
   (e) Telegram
   (f) TikTok
   (g) YouTube
   (h) Twitter
2. How many hours do you on average spend on social media per week? If you don't know, try to make an estimate.
3. How often do you post something on social media per week?
4. How many contacts do you have on social media in total? Do not count double contacts.

References

Ahmed, S., 2021. Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics Inform. 57, 101508.

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. J. Econ. Perspect. 31 (2), 211–236. https://doi.org/10.1257/jep.31.2.211.

Brashier, N., Schacter, D., 2020. Aging in an era of fake news. Curr. Directions Psychol. Sci. 29 (3), 316–323. https://doi.org/10.1177/0963721420915872.

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., Imhoff, R., 2013. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. Front. Psychol. 4 https://doi.org/10.3389/fpsyg.2013.00225.

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.

Fletcher, J., 2018. Deepfakes, artificial intelligence, and some kind of dystopia: the new faces of online post-fact performance. Theatre J. 70 (4), 455–471. https://doi.org/10.1353/tj.2018.0097.

Goertzel, T., 1994. Belief in conspiracy theories, Vol. 15. Political psychology.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. p. 27.

Goreis, A., Voracek, M., 2019. A systematic review and meta-analysis of psychological research on conspiracy beliefs: Field characteristics, measurement instruments, and associations with personality traits. Front. Psychol. 10, 205. https://doi.org/10.3389/fpsyg.2019.00205/pdf.

Groh, M., Epstein, Z., Firestone, C., Picard, R., 1 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. Proc. Natl. Acad. Sci. 119 (1), 2110013119. doi: 10.1073/pnas.2110013119.

Guess, A., Nagler, J., Tucker, J., 2019. Less than you think: prevalence and predictors of fake news dissemination on Facebook. Sci. Adv. 5 (1), 4586. https://doi.org/10.1126/sciadv.aau4586.

Halpern, D., Valenzuela, S., Katz, J., Miranda, J., 2019. From belief in conspiracy theories to trust in others: Which factors influence exposure, believing and sharing fake news. In: Meiselwitz, G. (Ed.), Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019. Vol. Proceedings, Part I 21. Springer International Publishing, Orlando, FL, USA, p. 217–232.

Jeffreys, H., 1999. Theory of probability, Vol. 94. Oxford University Press, 3rd ed.

Keersmaecker, J., Roets, A., 2017. Fake news: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. Intelligence 65, 107–110. https://doi.org/10.1016/j.intell.2017.10.005.

Khodabakhsh, A., Ramachandra, R., Busch, C., 6 2019. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX. IEEE, p. 1–6.

Korshunov, P., Marcel, S., 2018. Deepfakes: A new threat to face recognition? assessment and detectionArXiv:1812.08685. URL: http://arxiv.org/abs/1812.08685.

Korshunov, P., Marcel, S.,????a. Deepfake detection: Humans vs machines. ArXiv:2009.03155. URL: http://arxiv.org/abs/2009.03155.

Krumsvik, R., Jones, L., Øfstegaard, M., Eikeland, O., 2016. Upper secondary school teachers' digital competence: Analysed by demographic, personal and professional characteristics. Nordic J. Digital Literacy 11 (3), 143–164. https://doi.org/10.18261/issn.1891-943x-2016-03-02.

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., Zittrain, J., 3 2018.. The science of fake news. Science 359 (6380), 1094–1096. https://doi.org/10.1126/science.aao2998.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 6 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR. IEEE, p. 3204–3213. URL: doi: 10.1109/CVPR42600.2020.00327.

Makowski, D., Ben-Shachar, M.S., Lüdecke, D., 2019. bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J. Open Source Software 4 (40), 1541. https://doi.org/10.21105/joss.01541.

Maras, M.-H., Alexandrou, A., 7 2019.. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. Int. J. Evidence Proof 23 (3), 255–262. https://doi.org/10.1177/1365712718807226.

Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M., 2019. Fake News Detection on Social Media using Geometric Deep LearningArXiv:1902.06673). arXiv. URL: http://arxiv.org/abs/1902.06673.

Nightingale, S., Wade, K., Watson, D., 2017. Can people identify original and manipulated photos of real-world scenes? Cogn. Res.: Principles Implications 2 (1), 30. https://doi.org/10.1186/s41235-017-0067-2.

Ostrovsky, Y., Cavanagh, P., Sinha, P., 2005. Perceiving illumination inconsistencies in scenes. Perception 34 (11), 1301–1314. https://doi.org/10.1068/p5418.

Poulin, M., Haase, C., 2015. Growing to trust: evidence that trust increases and sustains well-being across the life span. Soc. Psychol. Personality Sci. 6 (6), 614–621. https://doi.org/10.1177/1948550615574301.

Salthouse, T., 2009. When does age-related cognitive decline begin? Neurobiol. Aging 30 (4), 507–514. https://doi.org/10.1016/j.neurobiolaging.2008.09.023.

Shahid, F., Kamath, S., Sidotam, A., Jiang, V., Batino, A., Vashistha, A., 2022. It matches my worldview: Examining perceptions and attitudes around fake videos. In: Barbosa, S.D.J., Lampe, C., Appert, C., Shamma, D.A., Drucker, S.M., Williamson, J.R., Yatani, K. (Eds.), CHI Conference on Human Factors in Computing Systems. ACM, p. 1–15. doi: 10.1145/3491102.3517646.

Sitarz, M., 2022. Extending F1 metric, probabilistic approachArXiv:2210.11997. URL: http://arxiv.org/abs/2210.11997.

Somoray, K., Miller, D.J., 2023. Providing detection strategies to improve human detection of deepfakes: An experimental study. Comput. Hum. Behav. 149, 107917.

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M., Zaffar, M., 5 2021. Seeing is believing: Exploring perceptual differences in deepfake videos. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S.M. (Eds.), Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–16. doi: 10.1145/3411764.3445699.

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

Westerlund, M., 2019. The emergence of deepfake technology: a review. Technol. Innov. Manage. Rev. 9 (11), 40–53. https://doi.org/10.22215/timreview/1282.

Yadlin-Segal, A., Oppenheim, Y., 2021. Whose dystopia is it anyway? Deepfakes and social media regulation. Convergence 27 (1), 36–51. https://doi.org/10.1177/1354856520923963.