


Human Versus Machine: A Comparative Analysis in Detecting Artificial Intelligence-Generated Images

Luca Maiano  | Sapienza University of Rome

Alexandra Benova  | Osnabrück University

Lorenzo Papa  , **Mara Stockner**  , **Michela Marchetti**, **Gianmarco Convertino**  ,
Giuliana Mazzoni, and **Irene Amerini**  | Sapienza University of Rome

This article delves into the intricate process of artificial intelligence-generated content detection, shedding light on automated detectors' challenges and revealing human detection biases, strengths, and weaknesses.

Generative artificial intelligence (AI) dominated the main emerging technologies of 2023. The enormous advances in this field have gained the general public's interest with mainstream tools such as ChatGPT (<https://openai.com/chatgpt>) or Midjourney (<https://www.midjourney.com/>). The results of these tools are simply astonishing, opening the door to the adoption of AI in numerous new sectors. In parallel with all this, however, the interest of a large part of society is growing in relation to the ethical and social impact that these technologies will have on our lives. In addition to creative and industrial uses, the problem arises of understanding how to distinguish real from generated content. In this regard, despite generally shared concerns, we still know little about whether and how humans perceive the difference between real and

artificial content. Knowing this boundary and monitoring it is essential to understand how far we still are from generating content that is indistinguishable from real.

The main aim of this article is to investigate the limits of humans and AI in the detection of fake images. In fact, we propose, on the one hand, an analysis of the human detection of artificially generated static images created with some of the most recent generative techniques. On the other hand, we compare these results with some automatic detection models. Our results demonstrate that the interaction between photo likeability and the type of photo can lead one to believe a photo is real. Moreover, human confidence seems to predict detection accuracy, especially for deepfake faces. From a machine perspective, we propose an analysis of some automatic detectors. In particular, we introduce an architecture called ResFormer that combines the benefits of convolutional networks with transformers and propose a comparison between this network

Digital Object Identifier 10.1109/MSEC.2024.3390555
Date of current version: 10 May 2024

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see <https://creativecommons.org/licenses/by/4.0/>

and other commonly used baselines and show that it achieves state-of-the-art performance. Our experiments confirm results reported in previous studies. Although the performance of AI systems is superior to human performance when a model is trained and tested on images generated with the same family of techniques, the generalization of these models is a problem yet to be solved.

Despite the provocative title of this work, our intent with this article is not to pit humans against machines to decide a winner but, rather, to study both limits to understand how to create more robust and human-friendly deepfake detectors. As we show, both have limitations that can give us some suggestions on how to design automatic detectors that are more robust and more applicable in real contexts.

The remainder of this article is organized as follows. The following section, “[Related Works](#),” provides an overview of the state of the art.

We then present this article’s methodology for assessing human and AI performance. Next, we introduce a new dataset and describe the design guidelines we adopted to create it. We then show the results of experiments and discuss the difference between human and AI performance. Finally, in the last section, we draw conclusions and discuss possible future directions.

Related Works

This section provides an overview of the state of the art and the positioning of this article. In particular, we begin discussing recent literature on static deepfake photo detection by humans, also highlighting factors already shown to impact detection accuracy. We provide an overview of studies in which human performance was compared with that of AI. Finally, we describe the challenges still to be solved in automatic detection.

Human Detection

Face processing is one of the most investigated fields in cognitive sciences (e.g., Fysh et al.⁴). However, since AI is becoming increasingly pervasive in people’s lives, this new environment is challenging what we know about face processing, and the detection of deepfake versus real faces has gained notable attention. Humans’ ability to detect deepfake faces has been reported to range from chance level to up to 75% accuracy (e.g., Papa et al.¹¹). Interestingly, recent literature has also highlighted that deepfake

detection performance might not be related to other face processing abilities (Ramon et al.¹²). The authors found that deepfake detection abilities of dynamic stimuli (videos) did not differ between a control group and super-recognizers (i.e., individuals with an innate superiority in face identity processing), nor did they find individual differences in face identity processing to be related to the deepfake detection performance in a large sample of police officers. This suggests the potential specificity of deepfake detection processes in humans. Still, little is known about which cognitive factors impact deepfake detection ability. In this field, seminal studies have shown that attractiveness, for instance, influences performance

in face detection, leading to judging faces as deepfake (e.g., Miller et al.¹⁰ and Tucciarelli et al.¹³). However, classical studies have shown that likeability is one of the first impressions that people have when they see a face (Willis and Todorov¹⁵). Here, we propose to replicate

and extend results on human deepfake detection ability, assessing the impact of likeability on detection accuracy. Knowing whether likable (deepfake) faces are more prone to detection errors will prove useful to deeply understand how humans interact with and make decisions about AI.

Automatic Detection

The rapid progress of new generative techniques has attracted growing interest from the scientific community regarding the importance of detecting these contents. As with content generation, a large part of automatic deepfake detection techniques is based on deep learning.¹ Most detectors focus primarily on at least one of the semantic features and frequency domain features. The first category includes all those models that try to take advantage of semantic, geometrical, or physical errors produced by generative models, such as light or perspective. Despite many promising semantic-based approaches, it is clear that these errors will dramatically reduce in the future with the advancement of generative techniques. As a result, it is important to understand whether alternatives exist. Alternatively, differences in the frequency domain can often give us essential information about the model used to generate an image. These techniques,³ although very promising because they can extract a fingerprint of the generative models, are more challenging to interpret and require a sufficient number



Knowing this boundary and monitoring it is essential to understand how far we still are from generating content that is indistinguishable from real.



of images to estimate a quality fingerprint. Regardless of the type of feature taken into consideration, generalization remains the biggest challenge to overcome. The literature proposes specific solutions to address this issue, like few-shot learning.⁸ However, a simple yet impactful approach to enhance generalization is augmentation⁵ based on JPEG compression, Gaussian noise blurring, geometric transformations, cutout, brightness, or contrast changes. Although a human easily perceives these augmented variations of the same image as identical to the unaugmented image, these examples are perceived by an automatic model as completely new examples. Therefore, by adding these augmented examples at training time, the detector can learn from



This new environment is challenging what we know about face processing, and the detection of deepfake versus real faces has gained notable attention.



a more significant number of examples. Finally, utilizing ensemble methods significantly aids in enhancing performance and producing more robust detectors.²

Human Versus Machine

To date, very few studies have focused on studying the human detection of deepfakes compared to automatic algorithms. A primary study by Korshunov and Marcel⁷ proved that people are confused by good quality deepfakes in 75.5% of cases. In contrast, algorithms detect deepfakes differently than human observers. They frequently fail to recognize fake videos that people can readily tell are false, yet specific algorithms, depending on their training data and selected threshold, can efficiently recognize videos that are difficult for humans to see. Subsequently, based on the first generation of deepfake generative models [generative adversarial network (GAN) based, from the Deepfake Detection Challenge (<https://ai.meta.com/datasets/dfdc/>)], Groh et al.⁶ show that ordinary individuals can notice artifacts in deepfake videos. Notably, between 13% and 37% of regular individuals outperform leading deepfake detection models, particularly in challenging scenarios, while the models perform slightly better in low-quality videos labeled as grainy, blurry, or dark. Despite these exciting results, generative techniques have recently made giant strides. In this sense, GANs have left room for the most recent diffusion models, allowing even more impressive results based on a textual description of what we want to generate. In this direction, Papa et al.¹¹ demonstrated how a careful engineering procedure of prompts for image generation could guide diffusion models toward creating realistic human faces, outperforming

previous generation methodologies. Starting from this latest study, we propose an analysis of the human and automatic detection of these contents.

Method

Our goal for this article is to compare the performance of an automatic deep learning-based detector with human performance. The first part introduces the methodology we followed to measure human detection

of fake images. Next, we present a hybrid architecture based on convolutional layers and transformers for deepfake detection.

Human Detection

To evaluate human performance, we recruited 120 online participants [63 females, age $M = 26.03$; standard deviation

(SD) = 8.13]. For the experiments, we used 24 deepfake photos generated through the method from Papa et al.¹¹ and 24 real photos from the FFHQ (<https://github.com/NVLabs/ffhq-dataset>) dataset. The photos were matched for the gender, age, and ethnicity of the faces. The dimension of both the deepfake and real photos was 512×512 pixels, with PsychoPy (<https://www.psychopy.org/>) height units of 0.5×0.5 . After signing an informed consent and giving sociodemographic information, the participants were randomly presented, one at a time, with 48 photos, with a request to indicate whether the photos were real or fake (the presentation lasted until a response was given). After pressing the corresponding key on the keyboard, the participants were asked to rate their confidence as well as the likeability of the face in the photo, with both ratings on a scale from zero (not at all) to six (very

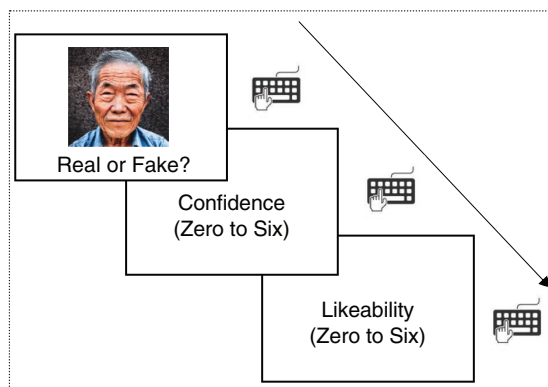


Figure 1. The procedure for human performance evaluation.

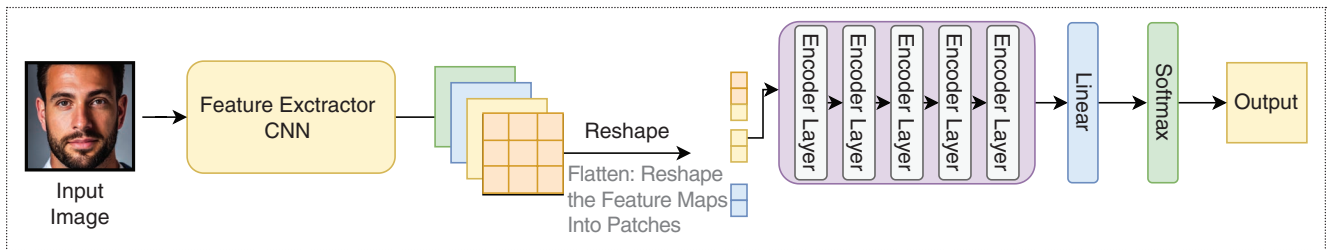


Figure 2. The proposed ResFormer architecture.

much), as shown in Figure 1. The participants performed the task on their personal computers. The whole procedure was constructed on PsychoPy and then carried out on Pavlovía (<https://pavlovía.org/>).

The accuracy of each response was computed by assigning a score to each correct (one) or incorrect answer (zero). Our statistical analyses aimed at verifying whether our predictors (the type of photo, photo likeability, and confidence), i.e., fixed effects, predicted accuracy (the dependent variable). To address this problem, we chose to apply generalized linear mixed models (i.e., an extension of logistic regressions), which allow us to control for the variability of both participants and stimuli (by including them as random effects). All statistical analyses were performed on R-Studio (version 4.13), using the lme4 and emmeans packages.

Automatic Detection

Convolutional neural networks (CNNs) have proved effective in detecting AI-generated content; however, these architectures have inherent disadvantages. CNNs operate on fixed-sized local receptive fields, which restricts their ability to understand global contexts effectively. Differently, transformers introduced the self-attention mechanism, allowing

models to weigh the importance of different parts of the input sequence when making predictions. Similar to how the human brain can focus attention on specific aspects of our environment selectively, the self-attention mechanism allows transformers to capture the global context of the input sequence and long-range dependencies.

Despite the superiority of transformers over CNNs in several tasks, their effectiveness is not as good in forensic studies, where the limited availability of data leads the models to overfit. To overcome this problem, we propose a hybrid architecture called ResFormer. The hybrid model (depicted in Figure 2) consists of two main parts. The first is the convolutional part for extracting spatial relationships. This component extracts feature maps that effectively capture all the essential parts in the images. After that, we turn these feature maps into patches that we feed into a transformer model, which is expected to spot connections and context across the entire feature map. The transformer structure is based on a multihead self-attention (MSA) mechanism, which is composed of several single self-attention layers running in parallel. Formally, given an input feature map, the transformer layer first computes three matrices: the query Q , the key K , and the value V , of sizes $d_q = d_k = d_v$. Then, we use the softmax dot product self-attention operation introduced by Vaswani et al.,¹⁴ which is defined as follows:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

The multihead attention layer allows the model to attend to information from different representation subspaces at different positions and operates by concatenating several attention heads. Formally,

$$\text{MSA}(Q, K, V) = \text{Concat}(\text{Att}_1, \dots, \text{Att}_h)W^o \quad (2)$$

where $\text{Att}_i = \text{Att}(QW_i^q, KW_i^k, VW_i^v)$, $W_i^q \in \mathcal{R}^{d_{\text{model}} \times d_k}$, $W_i^k \in \mathcal{R}^{d_{\text{model}} \times d_k}$, and $W_i^v \in \mathcal{R}^{d_{\text{model}} \times d_v}$.

The output of the network classifies the images as real or fake.



Figure 3. The improved generated images: (a) the new dataset and (b) Papa et al.¹⁴

Prompt Analysis

Despite the astonishing results obtained from the most recent text-to-image generative models, the choice of one textual prompt over another still makes a difference concerning images' quality. Our goal is to generate images that look like photos taken with a smartphone in everyday life. This enables us to assess human capabilities in detecting AI-generated images that look as realistic as the ones we commonly observe on social media. To do this, we rely on the prompt engineering strategy proposed in Papa et al.¹¹ but propose further improvement. Figure 3 gives an example of the newly generated images and compares them with the ones from Papa et al.¹¹ Unlike their method, which used Stable Diffusion v1.5, our solution is based on the Attend-and-Excite (https://huggingface.co/docs/diffusers/api/pipelines/attend_and_excite) pipeline. We chose this model over Midjourney or DALL-E because it is open source, enabling us to produce any number of images for free.

The advantage of the Attend-and-Excite model over Stable Diffusion comes from introducing negative prompts that allow us to guide image generation with greater flexibility. In fact, we noticed that the model often tends to generate faces with deformed eyes and teeth, as depicted in Figure 4. Often, these areas are the most obvious indicator of artificiality. To solve this problem, we realized that the specific words we used in negative suggestions had a big impact. Experimentally, we have noticed that applying the following negative prompt to all the generated images can obtain very good quality results, as in Figure 3: "disfigured, ugly, bad, immature, cartoon, anime, 3d, painting, b&w."

These negative prompts were carefully curated to discourage the generation of unrealistic and disfigured images. Following the prompt generation procedure proposed by Papa et al. and the negative prompts explained above, we generated a dataset of 10,000 images. For all the images, we set the guidance scale parameter to seven, which encourages the model to generate images closely linked to the text prompt.

Compared to Papa et al., our images are more realistic in the details of the eyes (see the first row in Figure 3); the mouth, and especially the teeth (see the last row in Figure 3); wrinkles, which in Papa et al. were excessively marked; and in the backgrounds, which in our case are very realistic (see the first two rows of Figure 3).

Human Versus Machine

To compare AI and human performance, we followed the methodology of Groh et al.⁶ For the AI model, we considered the accuracy value for each photo, whereas

for the human sample, we computed the mean accuracy for each stimulus (i.e., the participants' responses averaged per photo). The human confidence rates were transformed by dividing the raw Likert scale scores by six (i.e., the number of intervals), achieving values in line with the AI model's confidence range, i.e., zero to one. Moreover, we applied receiver operating characteristic (ROC) analysis. ROC analysis uses the area under the curve (AUC) as a metric to express accuracy in the discrimination of two categories, in our case, real and deepfake photos.

Experiments

In this section, we report the results we obtained with humans and AI detectors.

Human Detection

Table 1 reports the descriptive statistics of the human participants' accuracy and confidence levels. Our statistical model explained 27% of the variance in accuracy [Pseudo - $R^2(\text{total}) = 0.27$; Pseudo - $R^2(\text{fixed effects}) = 0.08$]. For all the results, χ^2 represents the chi-square distribution with one degree of freedom for the tested model, while p indicates the probability to obtain the hypothesized results. In regard to individual predictors, both photo likeability [$\chi^2(1) = 5.71, p < 0.05$] and confidence [$\chi^2(1) = 87.23, p < 0.0001$] were shown to

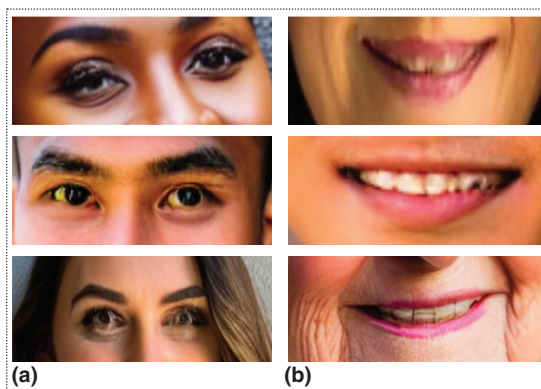


Figure 4. Common semantic errors produced by the model without negative prompts.

Table 1. Descriptive statistics of the human participants.

Class	Accuracy	Confidence
Fake	0.694 (0.175)	4.42 (0.929)
Real	0.718 (0.158)	4.3 (0.934)
Total	0.708 (0.456)	4.36 (1.43)

For each column, we report the mean and SD.

be significantly associated with overall detection accuracy, whereas the photo type ($p = 0.42$) was not related with the general ability of the participants to correctly detect the photos. The latter result indicates that there was no statistical difference in accuracy between real and deepfake photos. Moreover, photo likeability and confidence presented significant interactions with the photo type ($p < 0.001$), indicating that the variables were associated with the accuracy of real and deepfake photos in different ways (see Table 2).

We then ran post hoc analyses on the regression model, where the estimate indicates the estimated coefficient of the model, SE indicates the standard error, and z represents the index of the ratio of the estimated coefficient to its standard error. As illustrated in Figure 5(a), photo likeability was significantly associated with the accuracy of both real and deepfake photos, but in opposite directions: whereas higher likeability was significantly related to higher accuracy for real photos (estimate = 0.246, SE = 0.037, z - ratio = 6.75, $p < 0.0001$), it was also associated with lower

accuracy for deepfake photos (estimate = -0.397 , SE = 0.038, z - ratio = -10.48 , $p < 0.0001$). In fact, the contrast between the two slopes is significantly different (estimate = 0.634, SE = 0.049, z - ratio = 12.872, $p < 0.0001$). Finally, higher confidence was shown to significantly relate to higher accuracy for both real (estimate = 0.126, SE = 0.036, z - ratio = 3.504, $p < 0.01$) and deepfake photos (estimate = 0.377, SE = 0.036, z - ratio = 10.44, $p < 0.0001$) [see Figure 5(b)]. However, this effect appears to be significantly stronger for deepfake stimuli (estimate = -0.251 , SE = 0.048, z - ratio = -5.239 , $p < 0.0001$).

AI Detection

We compare the performance of the AI model proposed in the previous section against state-of-the-art models (i.e., Resnet18, Resnet50, ViT-B/16, and Grag2021⁵) on two different datasets: our proposed new dataset, which we call Diffusion Model Human Detection Dataset (DMHD), and a modified version of the CDDB⁹ dataset, which we refer to as CDDB-s. We identify these two datasets based on particular aspects: the DMHD was purposely generated to include highly realistic images that were difficult for humans to detect as fake. The second, i.e., the CDDB-s, is more challenging for an automatic detector since it was created using different generative techniques, thus allowing us to analyze the generalization capabilities of the model. The DMHD dataset is composed of 10,000 fake images generated as explained in the previous section, 10,000 fake images from Papa et al., and 20,000 real images taken from the FFHQ.¹¹ We use an 80/10/10 split for training, validation, and testing. For the CDDB-s dataset, we select a subset of fake classes containing human

Table 2. The regression model predicting detection accuracy.

Predictor	χ^2	p Value
Photo likeability	5.71	0.017*
Confidence	87.23	0***
Photo type (real versus fake)	.65	0.421
Photo likeability \times photo type	165.68	0***
Confidence \times photo type	27.44	0***

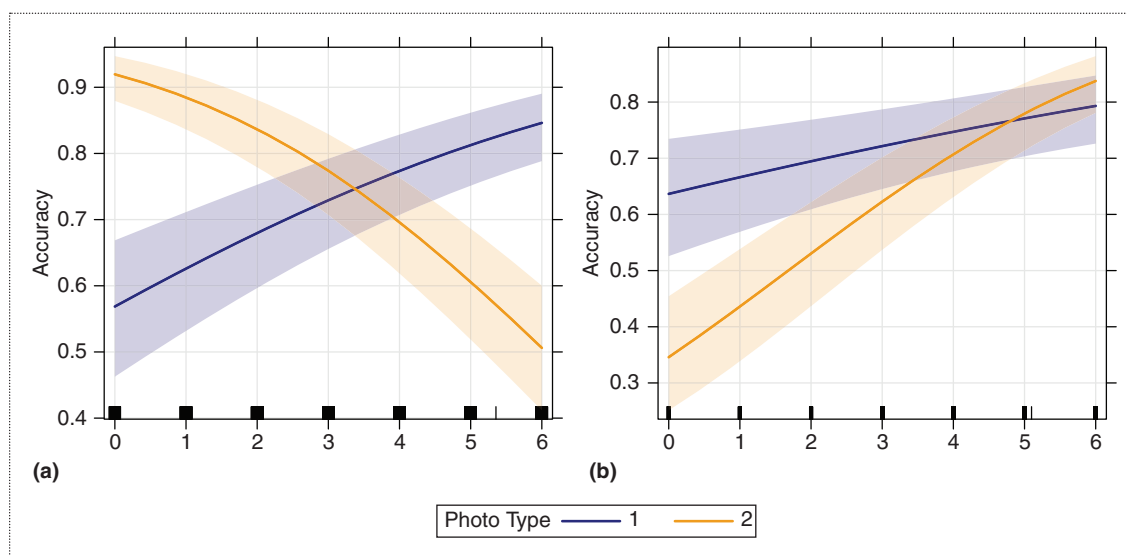


Figure 5. The detection accuracy for real (1) and deepfake (2) photos, impacted by (a) photo likeability and (b) confidence.

faces. Specifically, we use ProGAN, StyleGAN, BigGAN, and CycleGAN from the CDDb, adding stable diffusion from Papa et al. for training. For testing, we use NewFake, Glow, and StarGAN from the CDDb and the Attend-and-Excite-generated images proposed in this article. Due to the way it is constructed, this dataset is much more complex than the first concerning the models used for generation, and it is essential to emphasize that the generative models used in testing are different from those used in training. All manual annotations and generated images are available on our GitHub repository (<https://github.com/lucamaiano/humanvsmachines>; this will become available on the acceptance of the paper).

Table 3 reports the performance of all the models on both datasets. Our proposed model achieves the best performance on the CDDb-s dataset and is the runner-up for the DMHD dataset. In general, we can see that on our dataset, which is very difficult for humans, the models' performances are, on average, high. This suggests that, if properly trained on very realistic data, models can recognize features that are less visible to us as humans. At the same time, such high performance by all the models is a sign of overfitting. In fact, in a more complex and heterogeneous scenario, such as that of the CDDb-s, the models' performances are much lower, suggesting that the information learned during training is probably too specific and not very generalizable compared to new generative techniques not seen at training time. To confirm this assumption, in Table 4, we present the results in the more complex scenario. In the first column, we measure the performance of the models trained on the CDDb-s and tested on the DMHD, while the second column describes the opposite scenario. We can see that performance on the DMHD is generally lower than the previous experiment. In particular, the deeper models (Resnet50 and ViT-B/16) record a more marked drop in accuracy, while ResFormer, Grag2021, and Resnet18 appear to be more robust. On the CDDb-s, the results are slightly improved compared to the previous scenario. This result is due to the fact that the model trained on the DMHD is able, in most cases, to correctly recognize the majority of the images generated with the diffusion models. However, the performance on images generated by GANs is still very low.

Generally speaking, the results seem to align with what has been seen in the state of the art. Tables 5 and 6 show the performance of ResFormer when trained on the DMHD and CDDb-s and tested on the other dataset, respectively. As we can see from the results, generalization to new distributions (i.e., new generative models) remains an open issue. The generalization problem arises when training and testing an AI model on different distributions. In fact, images generated with

different models will belong to different distributions. It has been shown in Corvi et al.³ that these models will introduce different traces in the generated images and therefore make detection harder for an AI model. In

Table 3. Performance in terms of the accuracy of state-of-the-art neural networks.

Model	DMHD	CDDb-s
ViT-B/16	0.97	0.58
Resnet18	0.98	0.62
Resnet50	0.97	0.55
Grag2021 ⁵	1	<u>0.64</u>
ResFormer (ours)	<u>0.99</u>	0.65

Table 4. The generalization performance.

Model	DMHD	CDDb-s
ViT-B/16	0.83	0.62
Resnet18	<u>0.94</u>	<u>0.66</u>
Resnet50	0.81	0.58
Grag2021 ⁵	0.97	0.68
ResFormer (ours)	<u>0.94</u>	<u>0.66</u>

The first column shows training on the CDDb-s and testing on the DMHD. The second column shows training on the DMHD and testing on the CDDb-s.

Table 5. The descriptive statistics of ResFormer trained on the DMHD and tested over images employed in the human detection experiment.

Class	Accuracy	Confidence
Fake	1	4.955 (0.165)
Real	0.958	4.998 (0.005)
Total	0.979	4.976 (0.117)

The second column reports the mean and SD.

Table 6. Descriptive statistics of ResFormer trained on GAN images and tested over images employed in the human detection experiment.

Class	Accuracy	Confidence
Fake	0.875	4.886 (0.294)
Real	1	4.82 (0.454)
Total	0.937	4.853 (0.379)

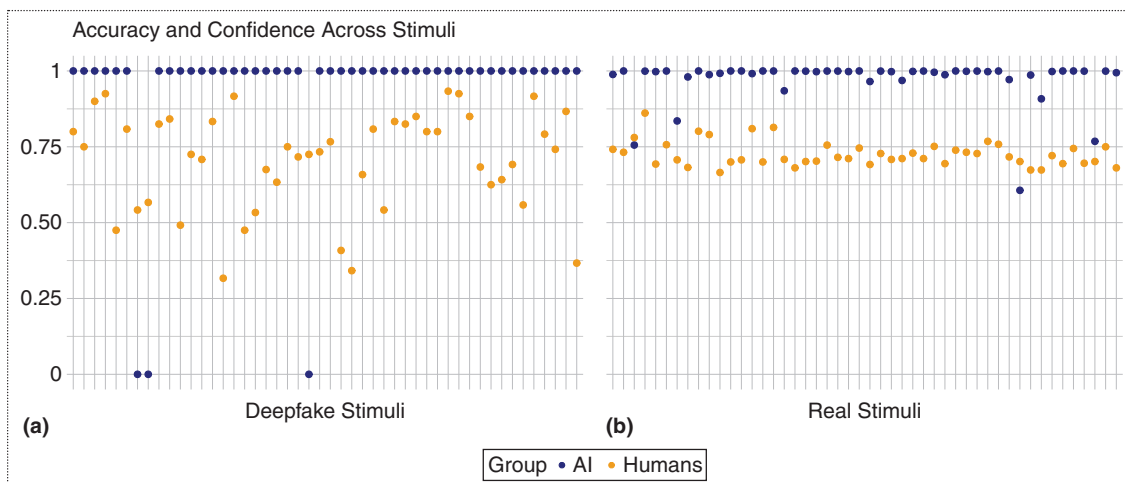


Figure 6. The detection (a) accuracy and (b) confidence of ResFormer (single observations for each stimulus) and humans (the participants' means).

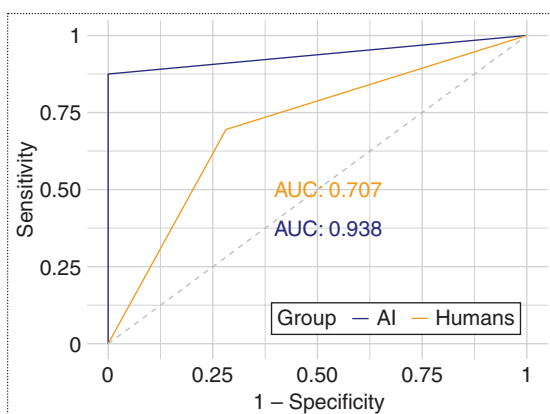


Figure 7. The ROC curves of human and AI (ResFormer) performance. Sensitivity = the percentage of deepfake photos correctly categorized as deepfake, and 1 - specificity = the percentage of real photos erroneously classified as deepfake. The dashed gray line illustrates random performance (AUC: 0.5).

the next section, we compare human performance with automatic performance.

Humans Versus AI

As reported above, the overall accuracy of human detection was 71.61% (SD = 0.16), whereas the corresponding accuracy of the AI model was 93.7% (SD = 0.24). As detailed in Figure 6, the AI model erroneously classified three deepfake photos that, on the other hand, were correctly identified by 54.17%, 56.67%, and 72.5% of our participants, respectively. Overall, no human outperformed the AI model, even though one participant reached the same accuracy as the AI model (93.75%, with three erroneous classifications of real photos). To test the difference in performance between AI and humans, we ran paired t-tests on the above-described

values. The accuracy of the AI model was significantly higher compared to humans [$t(47) = 5.86, p < 0.001$], as were confidence rates [$t(47) = 20.47, p < 0.001$]. Humans showed a mean confidence in photo classification of 72.63% (SD = 0.04), whereas the AI model showed a mean confidence of 97.08% (SD = 0.08). ROC analyses showed the human AUC to be 0.707 [confidence interval (CI): 0.695-0.0718], whereas the AI AUC was found to be 0.938 (CI: 0.87-1), confirming the AI model's higher accuracy in discriminating between the two categories (see Figure 7).

Discussion

Our analysis shows that AI models can outperform humans in detection accuracy. However, from the results presented in the previous section, it must be highlighted that AI has significant limitations. Its generalization is still far behind that of humans. In fact, it must be considered that none of the human subjects involved in this study had any specific training for detecting false images. The recorded performances depended solely on the subjects' experience. This allows us to understand the impact of fake content on the average population. In this regard, it will be interesting in the future to understand whether a change in performance can be observed when people are trained to detect fakes.

Our findings also reveal interesting facts regarding human detection of fake content. First, we showed that photo likeability created a general sense of realness for our stimuli, leading to high accuracy for real photos and low accuracy rates for deepfake photos. Interestingly, this is in contrast to above-cited studies on attractiveness, which is a related construct found to be associated with the judgment of faces as fake (e.g., Miller et al.¹⁰ and Tucciarelli et al.¹³). However, while likeability represents a global affective response to

an image (Willis and Todorov¹⁵), attractiveness refers to the specific evaluation of facial elements. Tucciarelli et al.¹³ argue that less attractive deepfake photos are seen as more real because they might be more similar to mental templates of faces and their characteristics, created on the basis of everyday experiences. Likeability, on the other hand, might be sustained by broader mechanisms influenced by additional sociocognitive variables (e.g., motivations, future interactions, believability, and so on). Thus, we argue that these underlying variables led to higher likeability being associated with the categorization of images as real. This finding may be highly relevant for future studies, which should directly investigate the differential impact of the underlying mechanisms of both likeability and attractiveness on (deepfake) face detection. Moreover, measuring the effect of specific training in face detection (based on face attractiveness and/or likeability) would help to understand whether specific knowledge about deepfake stimuli can improve detection performance. Our findings not only show an association between accuracy and confidence in human deepfake detection but also highlight that for deepfake photos, participants in low-confidence conditions reach accuracy levels below chance level. Participants seem to be aware of their potentially erroneous decisions, as low confidence levels usually indicate awareness of difficulty in selecting adequate detection criteria. Thus, future research should also consider investigating which criteria humans base their (erroneous) detection decisions on as well as which elements make them aware of their incorrect decisions. Since one limitation of the present study is the lack of a standardized (or strictly controlled) screen size to present faces, our preliminary results should also be tested in a strictly controlled laboratory setting, for instance, using different screen resolutions. In this line, further studies should also assess the reliability of this pattern of performance, for instance, by adopting a test-retest approach, for both humans and AI detection algorithms. Finally, future directions should also investigate individual and cultural differences in (deepfake) face detection abilities, i.e., Fysh et al.⁴ and Ramon et al.¹²

In this article, we presented a comparative study between human and AI performance in the context of artificially generated image detection. The results show that AI can surpass human performance when appropriately trained and tested on images generated with a specific type of generative technique. Still, it is far behind humans in terms of its ability to generalize to new manipulations.

Furthermore, we believe the limitations highlighted in this work can guide the design of new detectors supporting a human operator. Interpretability, therefore, plays a fundamental role in obtaining detectors that can be applied in a real context. Our next studies will focus

on these aspects as well as on the study of generalization, which remains the most critical challenge to solve for automatic detectors. ■

Acknowledgment

This article has been partially supported by the Security and Rights in Cyberspace Foundation (Grant PE00000014), under the Ministry of Universities and Research National Recovery and Resilience Plan, funded by the European Union's NextGenerationEU plan and Sapienza University of Rome project "EV2" (Grant 003_009_22). This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the ethical committee of the Department of Dynamic and Clinical Psychology, and Health Studies of "Sapienza," University of Rome (Prot. n. 0001705 - [UOR: SI000092 - Classif. VII/15]) and performed in line with the principles of the declaration of Helsinki.

References

1. I. Amerini et al., "Deep learning for multimedia forensics," *Found. Trends Comput. Graph. Vis.*, vol. 12, no. 4, pp. 309–457, 2021, doi: 10.1561/06000000096.
2. N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Piscataway, NJ, USA: IEEE, 2021, pp. 5012–5019, doi: 10.1109/ICPR48806.2021.9412711.
3. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095167.
4. M. C. Fysh, L. Stacchi, and M. Ramon, "Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection," *Roy. Soc. Open Sci.*, vol. 7, no. 9, 2020, Art. no. 200233, doi: 10.1098/rsos.200233.
5. D. Gagnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428429.
6. M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proc. Nat. Acad. Sci.*, vol. 119, no. 1, 2022, Art. no. e2110013119, doi: 10.1073/pnas.2110013119.
7. P. Korshunov and S. Marcel, "Subjective and objective evaluation of deepfake videos," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE, 2021, pp. 2510–2514, doi: 10.1109/ICASSP39728.2021.9414258.

8. P. Korshunov and S. Marcel, "Improving generalization of deepfake detection with data farming and few-shot learning," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 386–397, Jul. 2022, doi: 10.1109/TBIOM.2022.3143404.
9. C. Li et al., "A continual deepfake detection benchmark: Dataset, methods, and essentials," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1339–1349, doi: 10.1109/WACV56688.2023.00139.
10. E. J. Miller, B. A. Steward, Z. Witkower, C. A. Sutherland, E. G. Krumhuber, and A. Dawel, "Ai hyperrealism: Why ai faces are perceived as more real than human ones," *Psychol. Sci.*, vol. 34, no. 12, pp. 1390–1403, 2023, doi: 10.1177/09567976231207095.
11. L. Papa, L. Faiella, L. Corvitto, L. Maiano, and I. Amerini, "On the use of stable diffusion for creating realistic faces: From generation to detection," in *Proc. 11th Int. Workshop Biometrics Forensics (IWBF)*, 2023, pp. 1–6, doi: 10.1109/IWBF57495.2023.10156981.
12. M. Ramon, M. J. Vowels, and M. Groh, "Deepfake detection in super-recognizers and police officers," *IEEE Security Privacy*, early access, 2024, doi: 10.1109/MSEC.2024.3371030.
13. R. Tucciarelli, N. Vehar, S. Chandaria, and M. Tsakiris, "On the realness of people who do not exist: The social processing of artificial faces," *Iscience*, vol. 25, no. 12, 2022, Art. no. 105441, doi: 10.1016/j.isci.2022.105441.
14. A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010, doi: 10.5555/3295222.3295349.
15. J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychol. Sci.*, vol. 17, no. 7, pp. 592–598, 2006, doi: 10.1111/j.1467-9280.2006.01750.x.

Luca Maiano is a research fellow in the Department of Computer, Control, and Management Engineering "Antonio Ruberti," Sapienza University of Rome, 00185 Rome, Italy. His research interests include computer vision and multimedia forensics. Maiano received a Ph.D. in data science from Sapienza University of Rome. Contact him at luca.maiano@uniroma1.it.

Alexandra Benova is an undergraduate student in cognitive science at the Institute of Cognitive Science, Osnabrück University, 49073 Osnabrück, Germany, and a visiting student in the Department of Computer, Control, and Management Engineering "Antonio Ruberti," Sapienza University of Rome, Rome, Italy. Contact her at abenova@uni-osnabrueck.de.

Lorenzo Papa is a Ph.D. student in the Department of Computer, Control, and Management Engineering "Antonio Ruberti," Sapienza University of Rome, 00185

Rome, Italy. His research interests include efficient deep learning methodologies in computer vision applications. Papa received a Master of Science in artificial intelligence and robotics from Sapienza University of Rome. Contact him at lorenzo.papa@uniroma1.it.

Mara Stockner is a Ph.D. candidate with the Faculty of Medicine and Psychology, Sapienza University of Rome, 00185 Rome, Italy. Her research interests include perception, mental simulation, and embodied cognition. Stockner received a Master of Science in psychology from Sapienza University of Rome. Stockner received a Master of Science in psychology from Sapienza University of Rome. Contact her at mara.stockner@uniroma1.it.

Michela Marchetti is a Ph.D. candidate with the Faculty of Medicine and Psychology, Sapienza University of Rome, 00185 Rome, Italy. Her research interests include lineup identification and face recognition cognitive biases. Marchetti received a Master of Science in psychology from Sapienza University of Rome. Contact her at michela.marchetti@uniroma1.it.

Gianmarco Convertino is a Ph.D. candidate with the Faculty of Medicine and Psychology, Sapienza University of Rome, 00185 Rome, Italy. His research interests include deception detection and cognitive forensic assessment. Convertino received a Master of Science in psychology from the University of Bari Aldo Moro. Contact him at gianmarco.convertino@uniroma1.it.

Giuliana Mazzoni is a full professor in psychology at Sapienza University of Rome, 00185 Rome, Italy, and an emeritus professor at the University of Hull, Hull, U.K. Her research interests include eyewitness testimony and false memories. Contact her at mazzoni@uniroma1.it.

Irene Amerini is an associate professor in computer science at Sapienza University of Rome, 00185 Rome, Italy. Her research interests include digital image processing, computer vision, and multimedia forensics. She is a Member of IEEE; a member of the IEEE Information Forensics and Security Technical Committee; a member of the European Association for Signal Processing Technical Area Committee on Biometrics, Data Forensics, and Security; and a member of International Association for Pattern Recognition Technical Committee 6—Computational Forensics Committee. She is a Member of IEEE. Contact her at irene.amerini@uniroma1.it.