



UvA-DARE (Digital Academic Repository)

Deciphering authenticity in the age of AI

how AI-generated disinformation images and AI detection tools influence judgements of authenticity

Farooq, Aqsa; de Vreese, Claes

DOI

[10.1007/s00146-025-02416-5](https://doi.org/10.1007/s00146-025-02416-5)

Publication date

2026

Document Version

Final published version

Published in

AI and Society

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Farooq, A., & de Vreese, C. (2026). Deciphering authenticity in the age of AI: how AI-generated disinformation images and AI detection tools influence judgements of authenticity. *AI and Society*, 41(1), 493-504. <https://doi.org/10.1007/s00146-025-02416-5>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Deciphering authenticity in the age of AI: how AI-generated disinformation images and AI detection tools influence judgements of authenticity

Aqsa Farooq¹ · Claes de Vreese¹

Received: 7 April 2025 / Accepted: 29 May 2025 / Published online: 29 June 2025
© The Author(s) 2025

Abstract

An ongoing surge of Artificial Intelligence (AI)-enabled false content has been spreading its way through the information ecosystem, including AI-generated images, which have been used as part of political disinformation campaigns. Thus, there remains a pressing need to understand which factors individuals rely upon when determining whether images are AI-generated, particularly when they can be used to spread disinformation. AI-generated images have been characterised by their aesthetic realism, which can be leveraged to deceive users, and those who use generative AI to create deceptive content also tend to exploit its ability to convey and elicit emotion. This experimental study explored how aesthetic realism and emotional salience, as key features of both AI-generated content and disinformation, may influence authenticity judgements of AI-generated disinformation images. In this study, 292 UK-based participants were presented with both AI-generated and non-AI-generated disinformation images which varied in aesthetic realism and emotional salience. Results showed that participants were more likely to judge realistic-looking AI-generated images as being authentic compared with less realistic-looking AI-generated images, but did so with less confidence in their decision. Emotional salience was not a significant predictor of judgements. When participants were presented with the correct verdict of an AI detection tool, their reliance on the tool to update their own judgements was predicted by the aesthetic realism of the image and their confidence levels. These findings may assist with the development of disinformation detection tools, as well as strategies that mitigate the spread of deceptive, synthesised visual content in the digital age.

Keywords Generative Artificial Intelligence · Authenticity · Heuristic-systematic model · Disinformation

1 Introduction

The proliferation of Large Language Models (LLMs) has revolutionised the way we interact with and consume information on the internet. These sophisticated AI systems, such as OpenAI's GPT, possess an unparalleled ability to generate coherent and contextually relevant text, thereby raising concerns about the misuse and impact of their output. One pressing concern is the emergence of false content generated by these models and its potential influence on individuals, with an alarming surge of AI-enabled false content

already spreading its way through the information ecosystem (Newsguard 2023; The New York Times, 2023) demonstrating a deceptive power that surpasses even human-generated disinformation (Spitale et al. 2023). AI-generated disinformation in the form of visual information, such as images, presents a further challenge due to the faster speed at which visual information is processed and its longer-lasting cognitive influence, compared with textual information (Graber 1990; Jakus 2018). Already, AI-generated imagery has been used to propagate political disinformation campaigns (Carnevale et al. 2023; Klepper & Swenson 2023) and incite hostile narratives about vulnerable groups (Raymond 2023) leading to concerns that AI-generated imagery could play a destabilising role in democratic processes (Wirtschafter 2024). Amidst this landscape, individuals are expected to navigate the images they come across and decipher their authenticity to protect themselves from false and even harmful narratives or agendas. Thus, there remains a pressing

✉ Aqsa Farooq
a.farooq@uva.nl

Claes de Vreese
c.h.devreese@uva.nl

¹ University of Amsterdam, Amsterdam, Netherlands

need to understand which factors individuals rely upon when determining the authenticity of images generated by AI, particularly when such images can be used to spread disinformation.

The aim of the present study is to investigate how key visual features of AI-generated disinformation, such as aesthetic realism and emotional salience, influence individuals' authenticity judgements about AI-generated, image-based disinformation. With the ongoing development of AI-powered tools to assist with the detection of AI-generated images, it is also important to understand whether authenticity judgements may be influenced by the verdict of an AI detector, and whether the key visual features of the AI-generated disinformation (aesthetic realism and emotional salience) play a role in individuals' judgement changes. This investigation may assist with the development of disinformation detection tools, as well as strategies that mitigate the spread of deceptive, synthesised visual content in the digital age. For the present study, the focus is on disinformation, as opposed to misinformation, where the former refers to false communication spread intentionally, and the latter refers to false communication which can be spread accidentally (Wardle 2018). This is because the generation of images using AI necessitates a certain level of deliberate intentionality.

Past research has proposed that when users make judgements about the credibility of content in an environment of information overload, they are less likely to systematically investigate the credibility of the information they come across—instead, they use mental shortcuts, otherwise known as *heuristics* (Metzger et al. 2010). This is in line with the heuristic-systematic model (Chen & Chaiken 1999) which argues that heuristics provide us with quick judgement rules without needing to do the cognitively effortful task of systematically analysing whether the information is as credible or trustworthy as it claims to be (e.g. thinking a statement must be true because it cites a source). This model can be applied to the modern experiences of users navigating online spaces, where they are inundated with content of varying credibility. Sometimes, users may be able to apply the additional cognitive effort needed to make deliberate, systematic judgements about whether the content they encounter is credible or not. However, it is a reality that due to the sheer amount of content users are regularly exposed to and required to process, heuristic processing by relying on mental shortcuts offers a way of processing information that is substantially less demanding of cognitive effort, particularly when determining the credibility of online content (Sundar 2008). Determining the credibility of content is particularly important when it comes to online misinformation and disinformation. Being able to systematically process information, as opposed to simply processing it on a heuristic level, can be a crucial way of avoiding belief in misleading information by being able to assess credibility beyond surface-level cues

(Buchanan 2020; Vishwanath et al. 2011) and can also lead to identification of misinformation generated by AI (Shin et al. 2024). According to Appelman and Sundar's (2016) conceptualisation of credibility, users may consider the accuracy, believability and authenticity of content to establish credibility. In the context of AI-generated disinformation, the authenticity of the content (whether it is AI-generated or not) is central to determining whether it is credible, which is why the present study focuses on capturing participants' judgements of authenticity (*authenticity judgement*).

As such, users may rely on external cues to inform their authenticity judgements of the content they come across, which can act as heuristics and facilitate quick, intuitive thinking. In relation to AI-generated disinformation, malicious actors may utilise such cues to manipulate reactions towards, and engagement with, misleading content. For instance, the aesthetic realism of content can act as a heuristic, such as by appearing to look, on the surface, just like authentic content. Engagement with disinformation is often facilitated by its ability to imitate realistic information (Sundar et al. 2021) and most AI-generated content is characterised by its hyper-realistic quality (Asperti et al. 2025). This feature of AI-generated disinformation can be leveraged to spread propaganda and panic that has been circulated by even authority figures (Shoab et al. 2023). Consequently, the realistic appearance of AI-generated images may act as a heuristic and inform authenticity judgements.

Furthermore, another heuristic misleading content tends to employ is strong emotional appeals. This is a trait of misinformation and disinformation that aims to evoke arousal and promote engagement (Paschen 2019). The subsequent priming of emotions can lead to an impairment in an individual's ability to distinguish between fake and real content (Martel et al. 2020) and undermine their ability to reason logically, and thus makes them less likely to judge credibility systematically and analytically (Olanipekun 2025). For example, emotional language has been used in health misinformation (Peng et al. 2023) and the negative feelings elicited from such material, such as feelings of anger, can subsequently lead to belief in and sharing of misinformation (Han et al. 2020). Thus, the link between emotional salience in misleading content and a faltering in individuals' ability to systematically process the credibility of the content has been well-established by the literature. Using affective cues to prime emotions is also a tactic used by disinformation actors with the assistance of generative AI, as witnessed in a number of distressing cases documented by mainstream media (BBC News 2023; Evans & Novak 2023; Upton-Clark 2023). Analyses of AI technologies have highlighted how audiences' emotions can be exploited by visual AI-generated disinformation both to mislead the general public about political campaigns, and to manipulate specific vulnerable groups, such as senior citizens (Olanipekun

2025). The emotional appeals that often characterise AI-generated visual disinformation, including images, have indeed been identified as one of its key methods of targeting audiences via social engineering (Davis 2025). Thus, it is increasingly becoming evident that to capture audiences and incite engagement, disinformation actors utilising AI tools to create visuals are deliberately using emotional appeals as a tactic. Investigating disinformation through the tactics employed by the disinformation actors themselves can arguably lead to a better understanding of how to adapt preventative strategies against visual disinformation and its spread (Lecheler & Egelhofer 2022). Therefore, it is important to explore how aesthetic realism and emotional salience, as key features of *both* disinformation and AI-generated content, may influence individuals' authenticity judgements of AI-generated disinformation images. In the present study, this will be done by presenting participants with both AI-generated and non-AI-generated disinformation images which will vary in how aesthetically realistic they are, and the level of emotionality they convey, before participants will be asked to assess whether they judge those images to be authentic (not AI-generated) or not (AI-generated). It is expected that—in line with the literature on heuristics—AI-generated images high in aesthetic realism and AI-generated images high in emotional salience will be more likely to be judged as being authentic compared with AI-generated images low in aesthetic realism and AI-generated images low in emotional salience.

H1a: *AI-generated images high in aesthetic realism will be more likely to be judged as being authentic (not AI-generated) compared with AI-generated images low in aesthetic realism.*

H1b: *AI-generated images high in emotional salience will be more likely to be judged as being authentic (not AI-generated) compared with AI-generated images low in emotional salience.*

Given the variation of aesthetic realism and emotional salience, it is possible that the confidence with which participants assess the authenticity of the images will also vary. To capture the strength of their conviction in the assessments they have made, we also asked participants to provide a measurement of how confident they were in each of their authenticity assessments (*confidence level*). This is an important measure that has been used in previous misinformation research (e.g. Basol et al. 2020), but is also relevant to the interests of the current study as certainty levels can be influenced by the variables of interest. We predicted that both the aesthetic realism of the AI-generated images and their emotional salience will influence confidence levels.

H2a: *Confidence levels will be influenced by the aesthetic realism of the AI-generated images.*

H2b: *Confidence levels will be influenced by the emotional salience of the AI-generated images.*

With the surge of AI-generated content making it harder to decipher authenticity, the development of AI-powered tools to assist with the detection of AI-generated images has been underway (Sensity 2023), prompting the possibility that an additional heuristic may play a pivotal role in how users determine the authenticity of AI-generated images—the machine heuristic. This heuristic can be represented by the perceived logical and unbiased nature of machines in relation to humans, that can afford them a sense of legitimacy and signal credibility even when there is none (Sundar 2008). The spread of targeted disinformation can be facilitated by the influence of the machine heuristic (Bradshaw 2019; Bradshaw & Howard 2019). However, the machine heuristic can also contribute to mitigation strategies against the spread of misinformation and disinformation via the same mechanism. For instance, amongst individuals who believed that machines could be objective, greater trust was demonstrated in an AI agent's fact-checking verdict when it assessed a claim as being true (Banas et al. 2022). In contrast, even when individuals *do not* trust an algorithmically driven machine learning warning, it can still be effective at improving their awareness of misinformation (Seo et al. 2019). This suggests that the machine heuristic can have an influence on user's judgements of credibility in the wake of exposure to misinformation and disinformation. As AI agents begin to play a more significant role in the information ecosystem, particularly as potential counter measures to disinformation, the machine heuristic is, therefore, of particular importance. Whilst the development of tools to detect synthetic images has been raised as a possible solution to the problem of AI-generated disinformation (Wirtschatter 2024), these AI detectors have performed poorly at recognising AI-generated misinformation (Zhou et al. 2023) and run the risk of being used to legitimise actual disinformation that has not been correctly identified (e.g. by claiming an image is authentic because an AI classifier failed to classify it as AI). However, it is currently unknown whether individual users' judgements would involve relying on the machine heuristic (e.g. an AI detector's verdict) when making decisions about the authenticity of AI-generated or even non-AI-generated images. This is an important research gap to fill, particularly with the ongoing launch of AI detection tools as well as their increased reliance on the machine heuristic to enable trust in their mechanisms. In the current study, we aim to understand the influence of an AI detector by presenting participants with its verdict in relation to the disinformation images, and giving participants the option to change their authenticity judgements in light of the verdict (*change in authenticity judgement*). Understanding whether the aesthetic realism of disinformation images, or their level

of emotional salience, influences individuals' decisions to align with a verdict or not can be useful for the development and adoption of such detection tools. We expected participants' change in authenticity judgement to be influenced by the aesthetic realism and the emotional salience of the images, and that participants with low confidence levels will be more likely to change their authenticity judgement.

H3a: Change in authenticity judgements will be influenced by the aesthetic realism of the disinformation images.

H3b: Change in authenticity judgements will be influenced by the emotional salience of the disinformation images.

H4: Low confidence levels will predict change in authenticity judgement.

2 Method

2.1 Participants

Participants were recruited from the survey company Kantar, and were all based in the United Kingdom. The company was instructed to collect a varied sample of UK-based participants representative of the current demographic statistics on age, gender and education level. A power analysis for the required analyses revealed that 288 participants would be needed to achieve sufficient power. Initially, 315 participants were recruited, and based on the pre-registered exclusion criteria, we eliminated from the final sample participants who were outside of the average range of time majority of participants took to complete the survey. Also excluded were those who failed the attention check and those who admitted to looking up the images online. The final sample consisted of 292 participants, of which 52.4% were female, 11.1% were aged 18–25, 37.5% were aged 26–40, 26.4% were aged 41–60 and 25% were aged 60 and above. The hypotheses, design, data collection procedures, sampling and sample size were pre-registered at the Open Science Framework.¹ With the exception of H2, all of the aforementioned hypotheses were pre-registered, and some wording was changed for clarity purposes. Ethical approval was received from the first author's university's ethics committee.

Design and procedure.

Participants were told that their help was needed for the development of a tool, which can help with the detection of AI-generated images. To develop this tool, they were told that we need as many people as possible to train our algorithm by indicating whether a certain image is likely to be AI

generated or not, and provide additional information about the image. This information can help the algorithm become better at identifying AI-generated images from authentic images. We also told them that to understand how our tool compares to existing AI detecting tools, they will also be shown the verdict of a currently developed AI detector. At the end, they were told that they would also be asked some questions about themselves to see if certain people are betting at predicting AI-generated images from others.

Next, participants were randomly assigned to one of four conditions, based on the 2 (aesthetic realism: high vs low) × 2 (emotional salience: high vs low) between-subjects factorial design of the study. All images presented to a participant were either high or low on aesthetic realism and also either high or low on emotional salience. Before they were shown the images, they were told that some of the images are AI generated and some of them are authentic (not AI-generated, e.g. photographs), and that the image will only be shown for 10 s before it will disappear from the screen, after which they will be asked to make their judgement of authenticity and indicate their confidence level. Finally, they were told that they would be shown the verdict of an existing AI detector, indicating the percentage likelihood that the image is AI generated, and that they will be given the option to change their answer should they wish to. Here, the verdict of AI detection tool Hive Moderation,² which was chosen for its accuracy at predicting the authenticity of all of the images used in the study, was shown to participants. At the end, participants were asked questions about themselves, such as demographics (age, gender) and we also measured their trust in AI using the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway 2020). When the survey was complete, participants were debriefed about the nature of the study and were given the chance to see which images were AI generated and which were authentic.

2.2 Materials

We recruited participants ($N=75$) from the UK to pre-test the stimuli that would be used in the present study. Stimuli for the pre-test involved gathering both AI-generated and non-AI-generated images which had been circulating online as disinformation, and had been debunked by fact-checking organisations such as FullFact or Snopes, and images that were similar to these pre-existing disinformation images. The decision to use images that were already circulating online, and images similar to them, was based on the guidance from Pennycook et al. (2021) on doing misinformation research using materials that have successfully deceived audiences. The participants recruited for the pre-test were

¹ <https://osf.io/5sem9>

² <https://hivemoderation.com/ai-generated-content-detection>

Table 1 Participants' average responses to the study measures by experimental condition

Experimental condition	Measures					
	Authenticity judgements		Confidence levels		Change in authenticity judgements	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Aesthetic realism: low	.67	.47	5.40	1.48	.34	.47
Aesthetic realism: high	.54	.50	5.06	1.59	.41	.49
Emotional salience: low	.58	.49	5.15	1.52	.40	.49
Emotional salience: high	.61	.49	5.26	1.58	.37	.48

M mean, *SD* standard deviation

asked to indicate their familiarity with the images (we eliminated images that most participants stated that they were familiar with), their aesthetic realism (we asked: *To what extent does this image look realistic (like it is an actual photograph taken by a camera)?* And *To what extent does this image appear to show a realistic event (like it depicts an event that could actually occur)?* Both of which they answered on a 7-point Likert scale; 7 = looks completely realistic. The answers to these were combined to create a mean score of aesthetic realism and their emotional salience (we asked: *To what extent does this image portray something negative or positive?* Which they answered on a 9-point Likert scale; 9 = extremely positive). The 20 AI-generated and 20 non-AI-generated images that were rated to be lowest and highest on aesthetic realism and emotional salience were then chosen for the present study. See Appendix A for examples of images according to the experimental condition.

2.3 Measures

2.3.1 Authenticity judgement

“Is this image not-AI-generated or AI-generated?” (0 = Not AI-generated; 1 = AI-generated).

2.3.2 Confidence level

“Overall, how confident do you feel about your previous judgment?” (1 = not confident at all; 7 = completely confident).

2.3.3 Change in authenticity judgement

“Next, you will see the verdict of an existing AI detector, which will use its own algorithm to determine whether the image you just saw is AI-generated or not. Remember that these tools are not 100% accurate. This is the verdict made by an existing AI image detector: [AI detector verdict]. You said this image is not AI-generated/AI-generated. Would you like to change your answer?” (0 = No, keep as

not AI-generated/AI-generated; 1 = yes, change to AI generated/not AI generated).

2.4 Data analysis plan

To explore H1–H2, analyses were conducted only on the AI-generated images that participants were shown, and were conducted at the observation level ($N_{\text{observations}} = 1460$). To explore H3–H4, analyses were conducted on all of the images that participants were shown, also at the observation level ($N_{\text{observations}} = 2920$). This was to examine reliance on the AI detection tool for both AI-generated and non-AI-generated images, in line with the purpose and likely use of such tools. To test all hypotheses, regression analyses were run with the variables of interest added as predictors. For all analyses, attitudes towards AI (measured using the GAAIS) was controlled for to separate participants' attitudes towards AI from their responses to the measures. The positive GAAIS score was higher for participants with more positive attitudes towards AI, whereas the negative GAAIS score was higher for participants who were more negative about AI. Further control variables included participants' age and gender. The control variables were added first to the models, followed by the predictors.

3 Results

3.1 Authenticity judgements

H1a: *AI-generated images high in aesthetic realism will be more likely to be judged as being authentic compared with AI-generated images low in aesthetic realism.*

H1b: *AI-generated images high in emotional salience will be more likely to be judged as being authentic compared with AI-generated images low in emotional salience.*

For an overview of participants’ average responses to the study variables by the experimental condition, see Table 1.

The results of a binary logistic regression analysis showed that the model with the variables of interest added (aesthetic realism, emotional salience) was significant at predicting authenticity judgements, $\chi^2(2, 1460) = 23.35$, Nagelkerke $R^2 = 0.03$, $p < 0.001$. Aesthetic realism was a significant predictor of authenticity judgements (see Table 2). This meant that the higher the aesthetic realism of the image, the more likely that the image was classified as being not AI-generated. This supported H1a, but did not support H1b.

3.2 Confidence levels

H2a: *Confidence levels will be influenced by the aesthetic realism of the AI-generated images.*

H2b: *Confidence levels will be influenced by the emotional salience of the AI-generated images.*

The results of a OLS regression showed that the model with predictors added (aesthetic realism, emotional salience) was significant at predicting confidence levels, $F(2, 1434) = 12.82$, $R^2 = 0.06$, $p < 0.001$. Again, only aesthetic realism was a significant predictor, this time of confidence levels (see Table 3). Images low in aesthetic realism were judged with higher levels of confidence ($M = 5.40$, $SD = 1.48$) than images high in aesthetic realism ($M = 5.06$, $SD = 1.59$), $t = 4.19$, $p < 0.001$, $d = 1.54$). However, levels of confidence did not differ as a function of the emotional

salience of the image, $t = -1.42$, $p = 0.077$, $d = 1.55$ (see Fig. 1). This supports H2a, but does not support H2b.

3.3 Change in authenticity judgements

H3a: *Change in authenticity judgements will be influenced by the aesthetic realism of the disinformation images.*

H3b: *Change in authenticity judgements will be influenced by the emotional salience of the disinformation images.*

H4: *Low confidence levels will predict change in authenticity judgements.*

To explore the predictors of participants’ likelihood to change their authenticity judgements, we looked at their judgements of all of the disinformation images (AI-generated and non-AI-generated). The results of the binary logistic regression showed that the model with the variables of interest added (aesthetic realism, emotional salience, confidence levels) was significant at predicting participants’ change in authenticity judgements, $\chi^2(3, 2920) = 61.02$, Nagelkerke $R^2 = 0.06$, $p < 0.001$. Aesthetic realism and confidence levels were significant predictors of change in authenticity judgements; however, emotional salience was not (see Table 4). These findings support H3a and H4. The higher the aesthetic realism of the disinformation image, the more likely participants were to change their authenticity judgements. Regarding participants’

Table 2 Binary logistic regression with predictors of authenticity judgement

	B	SE	Wald	Sig	OR	95% CI	
						Lower	Upper
Constant	.97	.22	.01	<.001	2.65		
Participant age	-.01	.01	.758	.384	1.00	.99	1.00
Participant gender (0= male)	.09	.11	.609	.435	1.09	.87	1.37
Positive GAAIS score	.09	.07	1.463	.226	1.09	.95	1.27
Negative GAAIS score	.14	.07	3.456	.063	1.15	.99	1.32
Aesthetic realism (0=low)	-.91	.16	21.35	<.001	.40	.29	.56
Emotional salience (0=low)	-.28	.17	1.82	.097	.75	.54	1.05

Table 3 Linear regression with predictors of confidence levels

	B	SE	B	t	Sig	95% CI	
						Lower	Upper
Constant	6.16	0.15		41.56	<.001	4.01	5.38
Participant age	-.01	0.01	-.14	-5.25	<.001	-.02	-.01
Participant gender (0= male)	-.16	0.08	-.05	-1.93	0.054	-.33	0.01
Positive GAAIS score	0.14	0.05	0.08	2.59	0.01	0.03	0.25
Negative GAAIS score	0.3	0.05	0.15	5.57	<.001	0.19	0.4
Aesthetic realism (0=low)	-.31	0.08	-.10	-3.9	<.001	-.47	-.16
Emotional salience (0=low)	0.1	0.08	0.03	1.24	0.215	-.06	0.26

Fig. 1 Confidence levels (7 = completely confident) as a function of emotional salience (low vs high) and aesthetic realism (low vs high) (with standard error bars $\pm .1$)

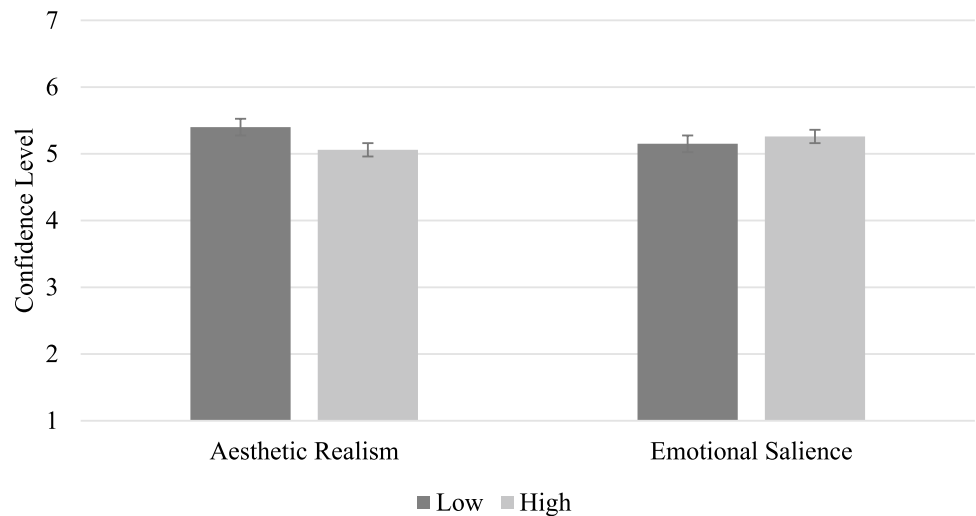


Table 4 Binary logistic regression with predictors of change in authenticity judgements

	B	SE	Wald	Sig	OR	95% CI	
						Lower	Upper
Constant	-.49	.44	1.24	.266	.61		
Participant age	-.01	.00	21.89	<.001	.99	.98	.99
Participant gender (0 = male)	.18	.10	3.11	.078	1.20	.98	1.47
Positive GAAIS score	.24	.07	13.63	<.001	1.28	1.12	1.45
Negative GAAIS score	-.09	.07	1.88	.170	.92	.81	1.04
Aesthetic realism (0 = low)	.38	.10	14.19	<.001	1.46	1.20	1.78
Emotional salience (0 = low)	-.08	.10	.61	.433	.93	.77	1.12
Confidence level	-.21	.03	40.51	<.001	.82	.77	.87

confidence levels, low confidence levels in authenticity judgements did indeed predict the likelihood to change authenticity judgements in line with the AI detection tool’s verdict.

Finally, differences were observed according to participants’ demographics and their attitudes towards AI. Age was a negative predictor of both confidence levels and change in authenticity judgements. With age, the confidence with which images were assigned authenticity decreased, as did the likelihood to change authenticity judgements in line with the AI detection tool. Gender was not associated with predicting judgements or confidence levels. In terms of the GAAIS score, which measured participants’ attitudes towards AI, both positive and negative GAAIS scores were predictive of confidence levels, whereas positive GAAIS score was a predictor of change in authenticity judgements. This meant that the more positive participants’ attitudes were about AI, the more confidence with which they judged the AI-generated images, and the more likely they were to change their authenticity judgement to match the AI detection tool’s verdict. Those with more negative attitudes towards AI were also more likely to judge the AI-generated images with more confidence.

4 Discussion

In this study, the aim was to understand how common features of AI-generated disinformation images may influence users’ judgements of what is authentic and what is not. Identified as characteristics of *both* synthesised content and disinformation, aesthetic realism and emotional salience were explored as the features that may be influential to authenticity judgements, as well as the confidence levels behind those judgements. With the continuous rollout of AI detection tools, we also set out to explore whether these key features would have an impact on individuals’ likelihood to adjust their previous judgement of authenticity in light of an AI detection tool’s verdict, and whether confidence levels would play a role in this. The results of this study suggest a relationship between characteristics of the image and both participants’ authenticity judgements and their confidence levels, which in turn may be linked to their likelihood to change authenticity judgements.

First, we found that the level of aesthetic realism in AI-generated disinformation images was a key indicator of how likely participants were to judge them as being

authentic. It is likely that the inconsistencies or unnatural features in the less realistic-looking images may have acted as the visual cues suggestive of their authenticity. Despite the small effect size, this finding was congruent with predictions as well as past research on heuristics (Chen & Chaiken 1999; Metzger et al. 2010). Furthermore, aesthetic realism also modestly predicted participants' confidence levels: when images were less realistic in appearance, participants judged their authenticity with more confidence than when the images were highly realistic looking. This suggests that AI-generated images that lack surface-level traits of realism are less likely to convince users of authenticity, as individuals from our sample were more confident in their judgements of authenticity for such images relative to images lower on surface-level traits of realism. With the rapid pace at which LLMs are becoming capable of producing highly aesthetically realistic-looking images, it is likely that in the near future, such cues will become less visible. This raises questions about the risks associated with generative AI models, given the discourse surrounding the false, stereotypical and biased representations often depicted by AI-driven image generators (Bendel 2023). Furthermore, disinformation actors may also capitalise on the influence of surface-level features to propagate a narrative that is facilitated by AI-generated images, or to instil doubt in the authenticity of *non*-AI-generated images as a way of devaluing the truth—something already being seen in relation to political narratives worldwide (Stockwell et al. 2024). Developers, regulators and policy makers may need to consider ways of increasing AI image detection literacy amongst the public that go beyond the surface-level characteristics of the images, lest the public become too reliant on such cues.

Furthermore, where AI-powered detection tools may be able to provide guidance for those who are uncertain about the authenticity of an AI-generated *or* a non-AI-generated image, the aesthetic realism of an image can also predict whether or not individuals will rely on such tools to update their decisions—albeit modestly. When a disinformation image was perceived as highly aesthetically realistic, participants were more likely to rely on the AI detection tool's verdict for those images. We know from past research on people's trust in artificial agents, that in conditions of uncertainty, there is usually an element of having to take a *leap of faith* when trusting agents that may be able to help them achieve their goal (Papagni et al. 2023). In the context of the study, facing highly realistic-looking images may have positioned participants in a comparable situation of uncertainty, where they may have felt compelled to rely on the AI detection tool—whereas the varying levels of emotional salience in the images did not prompt participants to use the tool any differently. Overtime, as LLMs become more advanced, and become capable of producing more realistic-looking content,

there is a likelihood that reliance on these tools will also increase to facilitate human decision-making. Past research has found that the machine heuristic can even lead individuals to be more likely to trust an AI tool's judgement for morally salient judicial decision-making in cases where objectivity is needed (Kim & Peng 2024). Scholarly discussions of individuals' interactions with generative AI and AI-related content, which tend to consider factors such as individuals' perceived competence of AI and their proximity to the impacts of AI (Kirkpatrick et al. 2024), as well as different accounts of trust in AI (Sheir et al. 2024) may also extend to include dependence on AI technology as a direct result of the increasing sophistication and perceived objectivity of AI-generated outputs, as well as the uncertainty in which individuals find themselves when making decisions.

Furthermore, low confidence levels predicted reliance on the verdict of the AI detection tool, suggesting that a perceived lack of trust in one's *own* ability for a given task may also contribute to dependence on AI-driven tools. In situations where individuals lack confidence in their own ability to accurately determine whether an image is AI generated or not, they may, therefore, be more inclined to defer to an external source perceived as more accurate or competent—in this case, the AI detection tool. The small but statistically significant relationship between confidence and reliance on the AI detection tool may reflect broader trends in human–technology interaction, where technological tools are increasingly positioned as authoritative sources of knowledge and decision-making. This finding, therefore, also has implications for the literature on the effectiveness of the machine heuristic (Banas et al. 2022; Seo et al. 2019; Sundar 2008), as it suggests that when users have low confidence in themselves, the AI system may provide an effective alternative. British systems continue to be integrated into various domains—including journalism, social media moderation, and content verification—understanding how user confidence influences reliance on such systems is crucial. This is because deference to such systems can be problematic in situations where the AI tool provides misleading or inaccurate feedback, or when disinformation actors adapt to create content that can evade systems (Rahman-Jones & Gerken 2024). Journalists and photographers in particular have voiced concerns about tools such as those implemented by Meta falsely labelling authentic images as being “Made with AI” (Mehta 2024). Those responsible for the development of AI-generated image detection tools should heed the fact that those relying on their tools may feel less confident about the image in question. Companies developing such systems may focus on better informing users about the tell-tale signs of the image's AI origins, in addition to the verdict, or provide disclaimers about the error rate of their tools. This way, the public can more cautiously integrate the verdict of AI detection tools into their own decision-making





about the authenticity of AI-generated images, particularly when they are less confident. Recent research has identified techniques for quantifying uncertainties that can be applied to AI-driven tools and AI-generated outputs, and ultimately guide users' decision-making (Chakraborti et al. 2025). Explanatory interventions that help users understand how algorithms work have also been successful in experimental research (Rader et al. 2018; Shin et al. 2024). Some scholars argue, however, that using AI systems to detect AI-generated disinformation is not a viable solution, due to the problems of non-transparency and bias inherent in algorithmic forms of content moderation (Gorwa et al. 2020). Furthermore, informing the public about the inner workings or fallibilities of their AI detection tools may not be in the interest of the companies and their business models. Thus, for individuals with low confidence levels, it is perhaps more paramount that efforts are centred around tackling the misuse of AI systems to create misleading content in the first place, as well as independent media literacy initiatives that are tailored to providing users with the skills necessary to better understand their information environment (Bontridder & Pouillet 2021). In the meantime, regulation that focuses on making AI systems safer and more accountable, such as the European Union's AI Act, may be necessary to legislate the use of generative AI tools and tools that claim to detect their output. In the present study, the AI detection tool's verdict was always correct, which is not always the case for all accessible detection tools, or for the various types of images created using LLMs. Even though participants were duly informed about the erroneous nature of the existing AI detection tool in the study, it is clear that confidence levels are a potential factor that can influence authenticity judgements. Exploring alternative conditions under which individuals would change their authenticity judgements when the tool is not always correct may be an avenue for future research in this area.

Having positive attitudes towards AI, as well as being younger in age, also predicted the likelihood of participants turning to the AI detector's verdict, an observation supported by previous research (Araujo et al. 2020; Hoff & Bashir 2015; Kim & Peng 2024). As AI-driven technologies become further integrated into everyday use, it is increasingly possible that reliance will continually increase, and these individual-level factors may give us insight into not just *whether* certain individuals are more likely to rely on AI tools than others, but eventually *how* certain individuals interact with AI tools compared with others. Interestingly, confidence levels when making authenticity judgements were influenced by age and attitudes towards AI. The relationship between age and confidence levels, as per previous trends in age, could be attributed to younger individuals feeling more comfortable judging AI-generated content compared to older individuals. However, we also found that individuals with strong attitudes towards AI—in either

direction—were also more likely to be confident whilst judging the authenticity of AI-generated images. It is possible that people with positive attitudes towards AI may feel more trusting of AI systems and their own ability to interact with such systems effectively, hence why they were also more likely to change their authenticity judgements in light of the AI detector's verdict. As for those with negative attitudes towards AI, they may also feel confident, but for different reasons—perhaps they are confident because they distrust AI and rely on their own judgement rather than on AI predictions, as demonstrated by their unlikelihood to rely on the AI detector's verdict. Overall, this finding highlights the complexity of how attitudes towards technology can influence the confidence behind judgement processes. The present study was conducted only amongst individuals based in the UK, a country that is unique for being very nearly split down the middle with regards to the frequent use and perceptions of the trustworthiness of AI systems, when compared globally (Gillespie et al. 2025). However, the example of the UK, and the present study's findings, cannot be applied cross-culturally, and so must be interpreted as being representative of the British context only. As such, there is scope for future research to further explore the intersection between strong attitudes towards AI and interactions with AI-generated content and AI-powered systems—but also to do so in alternative national contexts so that this phenomenon can be understood cross-culturally.

Finally, it is also important to note the limitations within this study. We found that, contrary to predictions, the level of emotional salience portrayed by the images presented to participants was not influential to participants' authenticity judgements, confidence levels, or their likelihood to change their authenticity judgements. This suggests that whilst previous research has established a link between emotional content and disinformation tactics (Olanipekun 2025; Paschen 2019; Peng et al. 2023), it is possible that in the current context of the study, this feature of the images was not an important indicator of authenticity. Usually, emotional content encountered online is embedded within a salient context. As the present study took emotionally salient images out of context, there is a likelihood that this feature of the images was diminished. Future research should explore whether AI-generated disinformation images that vary in emotionality may influence authenticity judgements when they are encountered in context. Another key limitation of this study is the use of a between-subjects factorial design, which, whilst methodologically advantageous in minimising carry-over effects, restricts the ability to control for individual differences in traits such as perceptual sensitivity or media literacy. These factors could influence how participants interpret and respond to AI-generated images, potentially affecting both authenticity judgements and confidence levels. Future research should consider employing mixed or

Table 5 Example stimuli by experimental condition

	AI-Generated Images	Emotional Salience: High	Emotional Salience: Low
Aesthetic Realism: High			
Aesthetic Realism: Low			

within-subjects designs, or including individual difference measures as covariates, to more directly account for such variability and provide a clearer picture of the mechanisms driving responses to AI-generated disinformation. Furthermore, our study only involved AI-generated images. AI-generated text, video and audio can all also share features of disinformation, and certainly can capitalise on appearing realistic and consisting of emotional appeals (Davis 2025). Another avenue for future research could involve exploring alternative modalities of AI-generated disinformation, and the extent to which their key traits influence authenticity judgements.

In conclusion, this study sheds light on the factors influencing judgements of authenticity in the context of AI-generated disinformation images. Our findings highlight the critical role of aesthetic realism in shaping authenticity judgements, with hyper-realistic images being more likely to be perceived as authentic, regardless of their emotionality. Furthermore, confidence levels of the participants were significantly influenced by the perceived realism of the images, indicating the importance of considering ways to make users better capable of identifying AI-generated images beyond their visual cues—particularly with the rapid development of LLMs. Finally, although our results were supported by small effect sizes, they raise important considerations for the development and deployment of AI detection tools, as our findings suggest that reliance on these tools may be influenced by the aesthetic realism of images as well as users' confidence in their own judgements. This underscores the need for AI detection tools to develop transparent and informative communication

strategies which empower users in navigating the complexities of AI-generated disinformation.

5 Funding

The authors are grateful to Amsterdam School of Communication Research (ASCoR) and the European Digital Media Observatory (EDMO) for the funding for this research.

Appendix

See Table 5

Author Contribution Conceptualization: AF and CdV, Methodology: AF and CdV, Formal analysis and investigation: AF, Writing—original draft preparation: AF; Writing—review and editing: CdV.

Data availability Anonymised data is available at this link: https://osf.io/kw7pb/?view_only=fc2f02d23f29499da62f05728e086bc6.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Appelman A, Sundar SS (2016) Measuring message credibility: construction and validation of an exclusive scale. *Journal Mass Comm Quart* 93(1):59–79. <https://doi.org/10.1177/1077699015606057>
- Araujo T, Helberger N, Kruikemeier S, de Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Soc* 35(3):611–623. <https://doi.org/10.1007/s00146-019-00931-w>
- Asperti A, George F, Marras T, Stricescu RC, & Zanotti F. (2025). A critical assessment of modern generative models' ability to replicate artistic styles (arXiv:2502.15856). arXiv. <https://doi.org/10.48550/arXiv.2502.15856>
- Banas JA, Palomares NA, Richards AS, Keating DM, Joyce N, Rains SA (2022) When machine and bandwagon heuristics compete: Understanding users' response to conflicting AI and crowdsourced fact-checking. *Hum Commun Res* 48(3):430–461
- Basol M, Roozenbeek J, van der Linden S (2020) Good news about bad news: gamified inoculation boosts confidence and cognitive immunity against fake news. *J Cogn*. <https://doi.org/10.5334/joc.91>
- Bendel O (2023) Image synthesis from an ethical perspective. *AI & Soc*. <https://doi.org/10.1007/s00146-023-01780-4>
- Bontridder N, Pouillet Y (2021) The role of artificial intelligence in disinformation. *Data & Policy* 3:e32
- Bradshaw S (2019) Disinformation optimised: gaming search engine algorithms to amplify junk news. *Internet Policy Review* 8(4):1–24
- Bradshaw S, & Howard PN. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*
- Buchanan T (2020) Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0239666>
- Carnevale A, Falche Del Gado C, Bisconti Lucidi P (2023) Hybrid ethics for generative AI: Some philosophical inquiries on GANs. *Humana Mente* 16(44):33–56
- Chakraborti T, Banerji CRS, Marandon A, Hellon V, Mitra R, Lehmann B, Bräuninger L, McGough S, Turkay C, Frangi AF, Bianconi G, Li W, Rackham O, Parashar D, Harbron C, MacArthur B (2025) Personalized uncertainty quantification in artificial intelligence. *Nature Mach Intellig* 7(4):522–530. <https://doi.org/10.1038/s42256-025-01024-8>
- Chen S, & Chaiken S. (1999). *The heuristic-systematic model in its broader context*. In *Dual-process theories in social psychology*. The Guilford Press, New York
- Davis J (2025) Disinformation in the era of generative ai: challenges, detection strategies, and countermeasures. In *Public Relations and the Rise of AI*, Routledge
- Evans C, & Novak A. (2023). *Scammers use AI to mimic voices of loved ones in distress—CBS News*. <https://www.cbsnews.com/news/scammers-ai-mimic-voices-loved-ones-in-distress/>
- Gillespie N., Lockey S, Ward T, Macdade A, & Hassed G. (2025). Trust, attitudes and use of Artificial Intelligence: A global study 2025. The University of Melbourne and KPMG. <https://mbs.edu/faculty-and-research/trust-and-ai>
- Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data Soc*. <https://doi.org/10.1177/2053951719897945>
- Grabner DA (1990) Seeing is remembering: how visuals contribute to learning from television News. *J Commun* 40(3):134–156. <https://doi.org/10.1111/j.1460-2466.1990.tb02275.x>
- Han J, Cha M, & Lee, W. (2020). Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, 1(3). <https://misinfoforeview.hks.harvard.edu/article/anger-contributes-to-the-spread-of-covid-19-misinformation/>
- Hoff KA, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Factors* 57(3):407–434. <https://doi.org/10.1177/0018720814547570>
- Jakus D (2018) Visual communication in public relations campaigns. *Market Scient Res Organiz* 27(1):25–36
- Kim T, Peng W (2024) Do we want AI judges? The Acceptance of AI Judges' Judicial Decision-Making on Moral Foundations. *AI & Soc*. <https://doi.org/10.1007/s00146-024-02121-9>
- Kirkpatrick AW, Boyd AD, Hmielowski JD (2024) Who shares about AI? Media exposure, psychological proximity, performance expectancy, and information sharing about artificial intelligence online. *AI & Soc*. <https://doi.org/10.1007/s00146-024-01997-x>
- Klepper D, & Swenson A. (2023). AI-generated disinformation poses threat of misleading voters in 2024 election. *PBS NewsHour*. <https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election>
- Lecheler S, Egelhofer JL (2022) Disinformation, misinformation, and fake news: Understanding the supply side. *Knowl Resist High Choice Inform Environ*. <https://doi.org/10.4324/9781003111474-4>
- Martel C, Pennycook G, Rand DG (2020) Reliance on emotion promotes belief in fake news. *Cogn Res* 5(1):47–47
- Mehta I. (2024). Meta is tagging real photos as “Made with AI,” say photographers. *TechCrunch*. <https://techcrunch.com/2024/06/21/meta-tagging-real-photos-made-with-ai/>
- Metzger MJ, Flanagin AJ, Medders RB (2010) Social and heuristic approaches to credibility evaluation online. *J Commun* 60(3):413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- BBC News. (2023, July 7). Martin Lewis felt “sick” seeing deepfake scam ad on Facebook. *BBC News*. <https://www.bbc.com/news/uk-66130785>
- Newsguard. (2023). Could ChatGPT Become A Monster Misinformation Superspreader? *NewsGuard*. <https://www.newsguardtech.com/misinformation-monitor/jan-2023>
- Olanipekun SO (2025) Computational propaganda and misinformation: AI technologies as tools of media manipulation. *World J Adv Res Rev*. <https://doi.org/10.30574/wjarr.2025.25.1.0131>
- Papagni G, de Pagter J, Zafari S, Filzmoser M, Koeszegi ST (2023) Artificial agents' explainability to support trust: Considerations on timing and context. *AI & Soc* 38(2):947–960. <https://doi.org/10.1007/s00146-022-01462-7>
- Paschen J (2019) Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *J Product Brand Manage* 29(2):223–233. <https://doi.org/10.1108/JPBM-12-2018-2179>
- Peng W, Lim S, Meng J (2023) Persuasive strategies in online health misinformation: a systematic review. *Inf Commun Soc* 26(11):2131–2148. <https://doi.org/10.1080/1369118x.2022.2085615>
- Pennycook G, Binnendyk J, Newton C, Rand DG (2021) A practical guide to doing behavioral research on fake news and misinformation. *Collabra*. <https://doi.org/10.1525/collabra.25293>
- Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. *Proceed Confer Human Factors Comp Syst*. <https://doi.org/10.1145/3173574.3173677>

- Rahman-Jones I, & Gerken T. (2024). Facebook and Instagram to label all fake AI images. BBC News. <https://www.bbc.com/news/technology-68215619>
- Raymond S. (2023). Debunked: Picture of Irish homeless woman and children is an AI-generated image. TheJournal.Ie. <https://www.thejournal.ie/homeless-image-dublin-ai-generated-imagery-artificial-6248519-Dec2023/>
- Schepman A, Rodway P (2020) Initial validation of the general attitudes towards artificial intelligence scale. *Comp Human Behavior Report*. <https://doi.org/10.1016/j.chbr.2020.100014>
- Sensity. (2023). How to detect AI generated images with Sensity in 2023. <https://sensity.ai/blog/deepfake-detection/how-to-detect-ai-generated-im/>
- Seo H, Xiong A, Lee D (2019) Trust it or not: effects of machine-learning warnings in helping individuals mitigate misinformation. *Proceed Confer Web Sci*. <https://doi.org/10.1145/3292522.3326012>
- Sheir S, Manzini A, Smith H, Ives J (2024) Adaptable robots, ethics, and trust: A qualitative and philosophical exploration of the individual experience of trustworthy AI. *AI & Soc*. <https://doi.org/10.1007/s00146-024-01938-8>
- Shin D, Koerber A, Lim JS (2024) Impact of misinformation from generative AI on user information processing: how people understand misinformation from generative AI. *New Media Soc*. <https://doi.org/10.1177/14614448241234040>
- Shoaib MR, Wang Z, Ahvanooy MT, Zhao J (2023) Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *Intern Confer Comp Applicat* 2023:1–7
- Spitale G, Biller-Andorno N, Germani F (2023) AI model GPT-3 (dis) informs us better than humans. *Sci Adv*. <https://doi.org/10.1126/sciadv.adh1850>
- Stockwell S, Hughes M, Swatton P, & Bishop K. (2024). AI-Enabled Influence Operations: The Threat to the UK General Election (CETaS Briefing Papers). <https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election>
- Sundar SS (2008) The MAIN model: A heuristic approach to understanding technology effects on credibility. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA
- Sundar SS, Molina MD, Cho E (2021) Seeing is believing: is video modality more powerful in spreading fake news via online messaging apps? *J Comput-Mediat Commun* 26(6):301–319. <https://doi.org/10.1093/jcmc/zmab010>
- The New York Times. (2023). A.I.-generated content discovered on news sites, content farms and product reviews. the New York times. <https://www.nytimes.com/2023/05/19/technology/ai-generated-content-discovered-on-news-sites-content-farms-and-product-reviews.html>
- Upton-Clark E. (2023). The terrifying new AI scam. Business Insider. <https://www.businessinsider.com/ai-voice-generator-phone-scam-imposter-crime-money-cash-2023-6>
- Vishwanath A, Herath T, Chen R, Wang J, Rao HR (2011) Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decis Support Syst* 51(3):576–586. <https://doi.org/10.1016/j.dss.2011.03.002>
- Wardle C (2018) Information disorder: The essential glossary. Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School, Harvard, MA
- Wirtschafter V. (2024). The impact of generative AI in a global election year. Brookings. <https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/>
- Zhou J, Zhang Y, Luo Q, Parker AG, & De Choudhury M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.