Forensic Science International: Digital Investigation 50 (2024) 301795

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions

Mirko Casu [a,b,*], Luca Guarnera [a], Pasquale Caponnetto [b], Sebastiano Battiato [a]

[a] Department of Mathematics and Computer Science, University of Catania, Viale Andrea Doria 6, Catania, 95126, CT, Italy
[b] Department of Educational Sciences, Section of Psychology, University of Catania, Via Teatro Greco 84, Catania, 95124, CT, Italy

ARTICLE INFO

MSC:
0000
1111

Keywords:
Forensic sciences
Cognitive biases
Cognitive psychology
Digital forensics
Synthetic data
Impostor bias
Generative AI
GAN
Diffusion models
Deepfake detection

ABSTRACT

This paper examines the impact of cognitive biases on decision-making in forensics and digital forensics, exploring biases such as confirmation bias, anchoring bias, and hindsight bias. It assesses existing methods to mitigate biases and improve decision-making, introducing the novel “Impostor Bias”, which arises as a systematic tendency to question the authenticity of multimedia content, such as audio, images, and videos, often assuming they are generated by AI tools. This bias goes beyond evaluators’ knowledge levels, as it can lead to erroneous judgments and false accusations, undermining the reliability and credibility of forensic evidence. Impostor Bias stems from an a priori assumption rather than an objective content assessment, and its impact is expected to grow with the increasing realism of AI-generated multimedia products. The paper discusses the potential causes and consequences of Impostor Bias, suggesting strategies for prevention and counteraction. By addressing these topics, this paper aims to provide valuable insights, enhance the objectivity and validity of forensic investigations, and offer recommendations for future research and practical applications to ensure the integrity and reliability of forensic practices.

1. Introduction

In forensic sciences, the objectivity of judgment in analyzing data for justice purposes is paramount. Beyond technical expertise, awareness of cognitive biases is crucial. These biases, rather than being deficits, are systematic preferences that influence the way we process, select, and retain information (Lester et al., 2011; Grisham et al., 2014). They can have both positive and negative effects, depending on the context and situation, facilitating swift decision-making when time is critical but also leading to poor decisions and adverse outcomes (Meterko and Cooper, 2022; Berthet, 2022).

Some of these biases have been identified as a significant factor impacting the objectivity and accuracy of forensic science (Bhadra, 2021). For example, law enforcement professionals showed vulnerable to *confirmation bias* (Meterko and Cooper, 2022), which is a tendency to search for, interpret, and remember information in a way that confirms one’s preexisting beliefs or hypotheses (Nickerson, 1998). Another recurring bias in forensics is the *anchoring bias* or *effect* (Edmond et al., 2015), which occurs when an individual relies too heavily on an initial piece of information (the “anchor”) when making decisions (Chapman and Johnson, 1994, 1999). One more relevant bias related to forensic science is the *hindsight bias* (Giroux et al., 2016), which occurs when people believe that an event is more predictable after it becomes known, involving memory distortion, beliefs about objective likelihoods, and subjective beliefs about one’s own prediction abilities (Roese and Vohs, 2012). The discipline of digital forensics is not exempt from these challenges (Sunde and Dror, 2019). Consequently, an increasing number of scholarly investigations are focusing on this particular field (Sunde and Dror, 2021).

This paper explores the impact of cognitive biases in forensics and digital forensics, with a focus on deepfakes and Artificial Intelligence (AI)-generated multimedia content, which pose threats such as manipulating public opinion and impersonating individuals. We introduce the *Impostor Bias*, an inherent distrust of multimedia authenticity due to the prevalence of AI-generated content. Effective detection of deepfakes is crucial to prevent confusion between real and fake multimedia, which can lead to erroneous decisions. We analyze advanced deepfake detection systems to aid operators in distinguishing authentic content,

* Corresponding author at: Department of Mathematics and Computer Science, University of Catania, Viale Andrea Doria 6, Catania, 95126, CT, Italy.
E-mail address: mirko.casu@phd.unict.it (M. Casu).

https://doi.org/10.1016/j.fsidi.2024.301795
Received 24 December 2023; Received in revised form 16 May 2024; Accepted 11 June 2024
Available online 19 June 2024

addressing the challenges posed by deepfakes, and ensuring accurate multimedia evaluation in digital forensics.

Particularly, the following points encapsulate the salient findings of this article:

- Cognitive biases in digital forensics: we discuss how cognitive biases can affect the perception and judgment of digital forensic investigators, especially in the face of complex and large-scale data.
- Deepfake detection methods: the state-of-the-art methods for detecting deepfakes, which are synthetic media created by advanced AI technologies, such as GANs and DMs, are reviewed.
- The Impostor Bias: we unveil the new concept of the Impostor Bias, which is the tendency to doubt the authenticity of real media due to the proliferation of deepfakes and the difficulty of distinguishing them from reality.
- Biases mitigating strategies: some strategies to reduce the impact of cognitive biases in digital forensics, such as using objective and standardized procedures, are proposed to enhance the training and education of forensic experts, and to adopt ethical and legal guidelines.

Finally, the paper is structured as follows: Section 1 introduced the concept of cognitive biases. Section 2 analyses their impact on forensic sciences. Section 3 explores some examples of cognitive biases in digital forensics, such as confirmation bias and pareidolia bias. Section 4 explores various bias mitigation strategies in both forensics and digital forensics. Section 5 presents the deepfakes and how they can be generated and managed. Section 6 introduces the Impostor Bias, a new type of bias triggered by AI media that affects the perception of reality. Section 7 reviews some of the most recent and relevant methods for deepfake detection, which is crucial to counter the Impostor Bias, as well as the problem of model attribution. Section 8 discusses the potential impact of Impostor Bias in digital forensics and everyday life. Section 9 concludes the paper and provides some directions for future research.

2. Cognitive biases impact on forensic sciences

Cognitive biases have been a concern in forensic science since 1984, when Larry Miller published a work discussing the presence of bias in forensic document examiners (Stoel et al., 2014). He suggested the introduction of procedural modifications to reduce cognitive bias, which could potentially result in incorrect outcomes.

These biases can significantly impact expert judgments and the criminal justice process (Stoel et al., 2014; Dror and Rosenthal, 2008; Neal and Grisso, 2014), are influenced by various factors, and can lead to errors and misinterpretation of evidence (Kassin et al., 2013; Cooper and Meterko, 2019; Bhadra, 2021). Both forensic experts and law enforcement professionals are susceptible to these biases, and one of the most common of them is the *confirmation bias* (Moser, 2013; Thompson and Newman, 2015; van den Eeden et al., 2019; Cooper and Meterko, 2019; Meterko and Cooper, 2022): if a forensic analyst believes a suspect is guilty, they might unconsciously interpret ambiguous evidence as incriminating (Kassin et al., 2013; van den Eeden et al., 2019). Gardner et al. (2019) found that task-irrelevant information can bias forensic analysts' decisions, suggesting that extraneous details can inadvertently sway the interpretation of evidence. This is echoed by Nakhaeizadeh et al. (2014), who discovered that irrelevant information can lead to confirmation bias in forensic anthropology, potentially leading to skewed conclusions based on pre-existing beliefs rather than objective evidence. Dror et al. (2021) further explored this issue, finding that base-rate neglect could bias forensic pathologists' decisions in child death cases. This suggests that statistical information about the prevalence of certain causes of death may be overlooked, leading to potential misinterpretations. In the realm of DNA forensics, Jeanguenat et al. (2017) found that suspect-driven bias and the presence of rare alleles can influence interpretation. This highlights the potential for preconceived notions about a suspect, as well as the rarity of certain genetic markers, to affect the analysis of DNA evidence. In a study by Douglass et al. (2023), evaluators successfully differentiated accurate from inaccurate witnesses based on videos of identification procedures alone, but their ability to discern accuracy was disrupted when extraneous incriminating evidence was also provided. This aligns with confirmation bias, where evaluators tend to favor information that confirms their existing beliefs or expectations, even when it contradicts objective evidence. Regarding the *anchoring bias*, in a forensic context, an analyst might give undue weight to the first piece of evidence they examine, which could skew their interpretation of subsequent evidence (Edmond et al., 2015; Meterko and Cooper, 2022). For the *hindsight bias*, this could lead to overconfidence in the accuracy of a forensic analysis after a suspect has been identified (Giroux et al., 2016; Beltrani et al., 2018; Meterko and Cooper, 2022). Neal et al. (2022) conducted a systematic review on cognitive biases and debiasing techniques in forensic mental health, finding significant bias effects specifically for confirmation and hindsight bias. Lastly, Stevenage and Bennett (2017) found that irrelevant DNA test outcomes could bias fingerprint matching tasks, confirming the presence of contextual bias. This suggests that unrelated test results can influence the interpretation of fingerprint evidence.

The sources of bias can be categorized into three groups related to the case, the analyst, and human nature. Inspired by the work of Dror (2020), we define a taxonomy of potential sources of bias. These biases can introduce cognitive distortions into the processes of sampling, observing, strategizing tests, analyzing, and drawing conclusions, even when conducted by experts. This taxonomy is synthetically sketched in Fig. 1.

3. Exploring cognitive biases in digital forensics

The definition of Digital Forensic Science often referred to is the one from the Digital Forensic Research Workshop (DFRWS) in Palmer et al. (2001): "The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations". A key area within this field involves the proper acquisition of digital content such as images, videos, and audio to produce evidence for forensic investigations. Multimedia forensics focuses on verifying the authenticity of data and reconstructing the history of an image since its acquisition (Battiato et al., 2016; Arceri et al., 2023; Giudice et al., 2017; Piva, 2013).

In the realm of digital forensics, cognitive bias emerges as a subtle yet potent force that can shape the outcomes of investigations. Sunde and Dror (2021) explored the susceptibility of digital forensics examiners to bias, revealing how preconceived notions and contextual information can skew their observations and interpretations. Despite the digital evidence's facade of objectivity, the human factor introduces variability, leading to inconsistent conclusions among experts analyzing identical datasets (Sunde and Dror, 2021).

3.1. Confirmation bias in text and face recognition

Fontani's blog post on Amped Software depicted a hypothetical situation featuring two characters, John and Lucy (Fig. 2) (Fontani, 2021). In this scenario, John inadvertently influenced Lucy's interpretation of a license plate by prematurely sharing his own interpretation. The cognitive bias present in this example could be the *confirmation bias*, since this latter is a type of cognitive bias where individuals are more likely to seek out, interpret, and remember information that confirms their pre-existing belief (Nickerson, 1998; Cooper and Meterko, 2019;

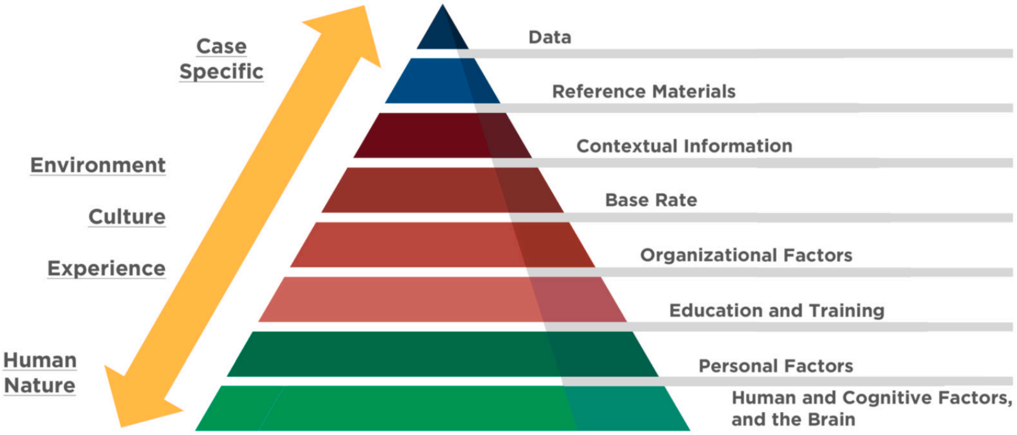
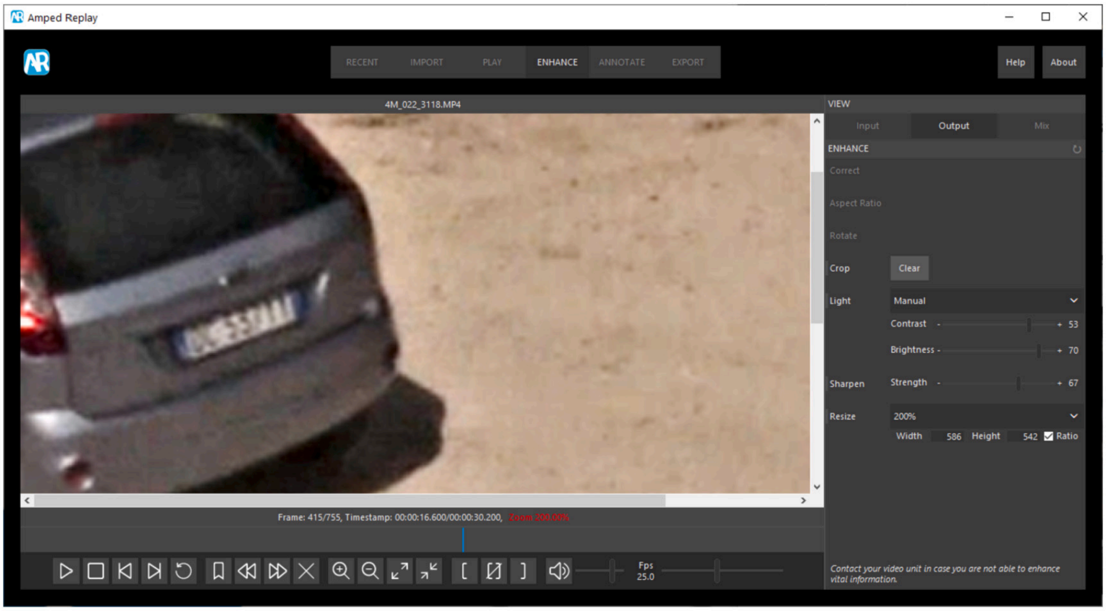


Fig. 1. Eight potential sources of bias that can influence forensic decision-making, as detailed by Dror (2020). These sources range from data and reference materials to human and cognitive factors.



4M_022_3118.MP4

Frame: 415/755, Timestamp: 00:00:16.600/00:00:30.200, Zoom 200.00%

Fps 25.0

Light: Manual
Contrast 53
Brightness 70
Sharpen: Strength 67
Resize: 200%
Width 586 Height 542 ☑ Ratio

Contact your video unit in case you are not able to enhance vital information.

Fig. 2. John asked Lucy for help with a license plate, reading “BC 537”, but unsure about the last two characters and the first one. Credit: Fontani (2021).

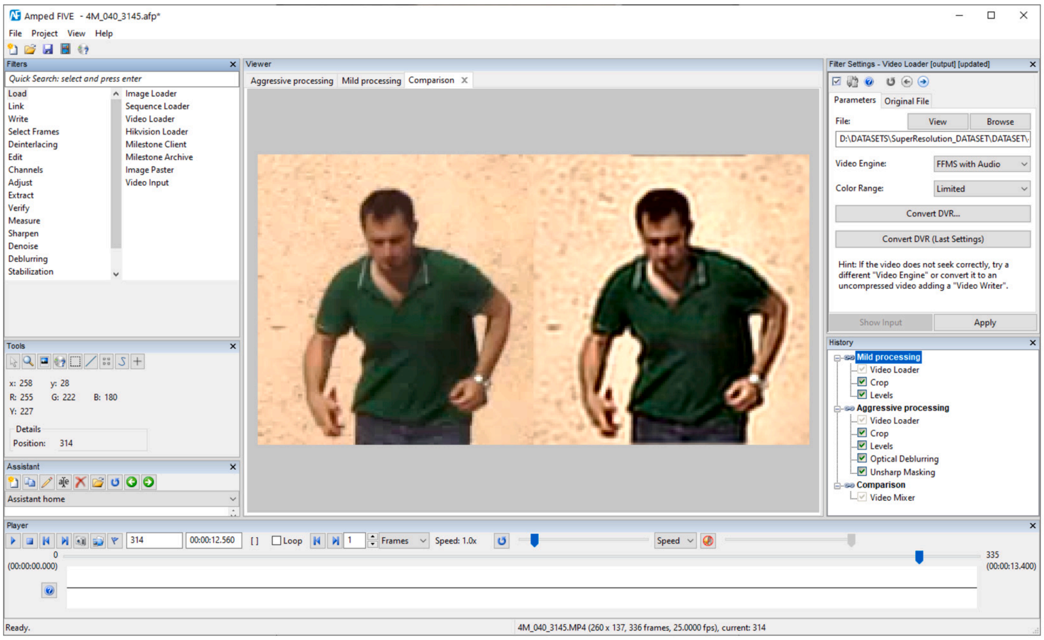
Meterko and Cooper, 2022). John’s premature sharing of his interpretation could lead Lucy to interpret the license plate in the same way, confirming John’s interpretation rather than considering other possible interpretations. It may lead Lucy to unconsciously process pixels and select frames that align with John’s interpretation. The post further exploited into the intricacies of face comparison, noting that varying processing techniques can result in significantly different facial appearances. It underscored the necessity of withholding the suspect’s face from the examiner prior to the enhancement process to prevent unconscious bias towards a match. Fontani also emphasized that different processing techniques can significantly alter facial appearances during enhancement (Fig. 3). Therefore, it’s crucial that the examiner doesn’t see the suspect’s face before the enhancement process to avoid unconsciously adjusting the enhancement to create a match.

Furthermore, Sunde and Dror (2019) underscored the importance of digital forensics as a rapidly growing field within forensic science. The authors analyzed seven specific sources of cognitive and human error within the digital forensics process and propose relevant countermeasures, concluding that while some cognitive and bias issues are common across forensic domains, others are unique and dependent on the specific characteristics of the domain, such as digital forensics.

3.2. Image processing could lead to pareidolia

A study by Di Lazzaro et al. (2013) focused on the potentially misleading effects of software techniques used for elaborating low-contrast images. The researchers used the Shroud of Turin, one of the most studied archeological objects in history, as an example (Fig. 4). They demonstrated that image processing of both old and recent photographs of the Shroud could lead researchers to perceive inscriptions and patterns that do not actually exist. The study further emphasized that the limited static contrast of our eyes can make the perception of low-contrast images problematic. The brain’s ability to retrieve incomplete information can interpret false image pixels after image processing. This phenomenon, named “pareidolia”, can lead to the perception of patterns in Shroud photographs that do not exist in reality (Fig. 5).

The enhancement of images extracted from video cameras can indeed lead to a degradation of the overall quality of information. This degradation can be attributed to factors such as excessive compression (Maity et al., 2023), distance from the recording plane (Wang et al., 2023), and limited overall resolution (Maity et al., 2023; Wang et al., 2023). International best practices suggest verifying on a case-by-case basis whether the level of information is sufficient to extract useful data for investigations (Tenopir et al., 2020; Soltani and Nikou, 2020).



x: 258 y: 28
R: 255 G: 222 B: 180
Y: 227
Details
Position: 314

File: D:\DATASETS\SuperResolution_DATASET\DATASET\
Video Engine: FFMS with Audio
Color Range: Limited

Hint: If the video does not seek correctly, try a different "Video Engine" or convert it to an uncompressed video adding a "Video Writer".

314 00:00:12.560 [ ] □ Loop 1 Frames Speed: 1.0x

Ready. 4M_040_3145.MP4 (260 x 137, 336 frames, 25.0000 fps), current: 314

Fig. 3. Image processing could lead to produce a noticeable different image, with different face characteristics. Credit: Fontani (2021).

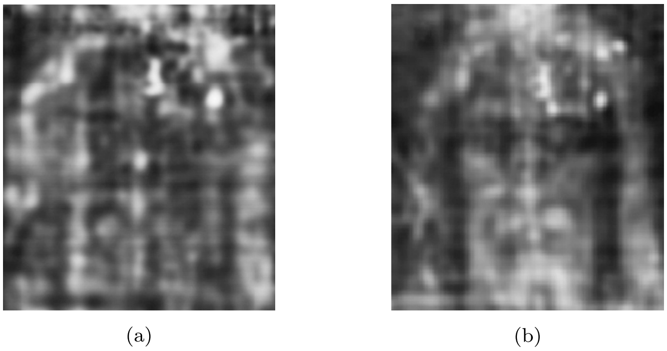

(a) (b)

Fig. 4. A supposed concealed image of a face on the back side of the Shroud is revealed through advanced image processing of a photograph published in a book. The image is flipped from right to left (b). A negative image of the face that can be seen on the front side of the Shroud, processed in the same way as (a). Credit: Fanti and Maggiolo (2004).

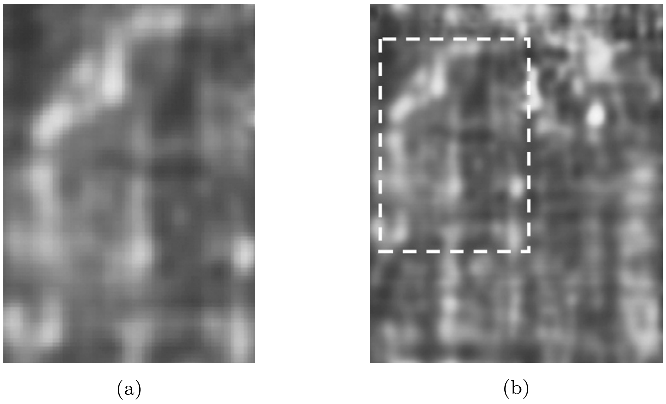

(a) (b)

Fig. 5. A magnified version of Fig. 4b (a). A face resembling the Shroud that we discern in the top-left section of Fig. 4b (as depicted on the right) (b). We can also discern another face in the bottom left section of Fig. 4b. Pareidolia leads to false positives, enabling us to see faces in Fig. 4b that aren't actually there. Credit: Di Lazzaro et al. (2013).

However, when examining low-contrast images that present pseudo-random visual patterns after an initial enhancement process, it is crucial to mitigate the risk of pareidolia. This bias is particularly potent when the object of interest refers to "human faces" or more generally to "letters/numbers" or known human structures (Wang and Yang, 2018; Zhou and Meng, 2020). Pareidolia is a subconscious illusion that tends to associate random shapes with known forms, especially human figures and faces. Classic examples include seeing animals or human faces in clouds, or a human face on the moon. In forensic investigations, we also suggest to entrust the analysis and interpretation to automatic methods (Solanke and Biasiotti, 2022) or experts who can follow a "blind testing" approach, i.e., an interpretation detached from the knowledge of details and the reference context (Cowan and Koppl, 2011; Servick, 2015). This helps to ensure the search for "certain" evidence is as rigorous and unbiased as possible.

Zhou and Meng (2020) discussed how some individuals exhibit a looser decision criterion for detecting faces, making them more prone to perceive faces where none exist. This relates to a concept in signal detection theory known as response bias (Nguyen and Beins, 2013). In digital forensics, examiners may fall prey to response biases when analyzing ambiguous digital evidence, predisposing them to validate or dismiss forensic hypotheses based on non-diagnostic features. Just as some are more likely to see faces in random patterns due to biases in how they set thresholds for face judgments, forensic analysts could have biases influencing how strictly they apply standards of evidence to digital artifacts. Understanding individual differences in cognitive biases like threshold placement could help address potential sources of error and increase objectivity in digital forensic examinations (Berthet, 2021; Horsman, 2024).

3.3. Case study: confirmation bias and pareidolia in surveillance camera footage

We present a case study with the objective of determining the presence of a passenger in a vehicle involved in a murder criminal case, through the analysis of surveillance camera footage. The methodology included the scrutiny of various cameras and a detailed analysis of the vehicle's passages from different surveillance cameras, considering image overlaps and real passage times, including sunset. The data for this study is primarily derived from the surveillance camera footage. Fig. 6

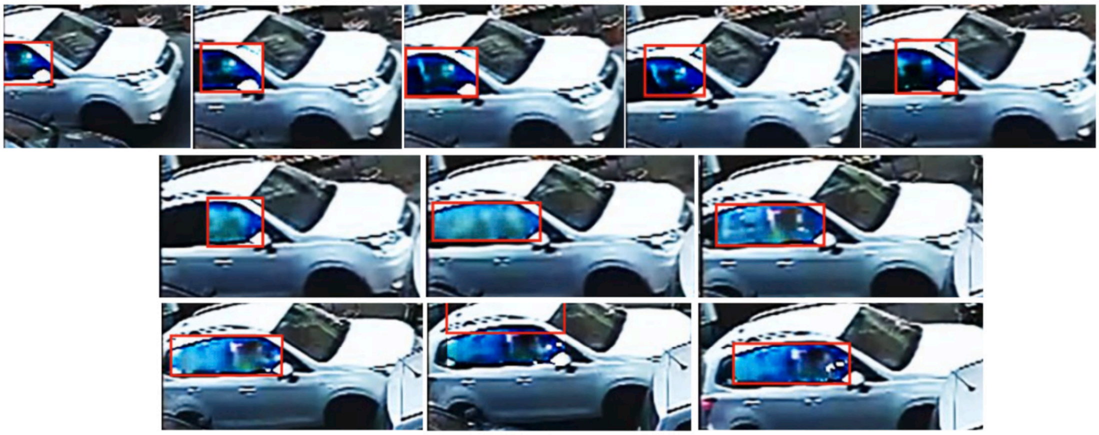
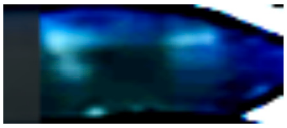
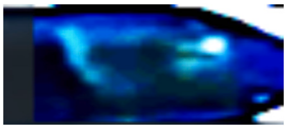
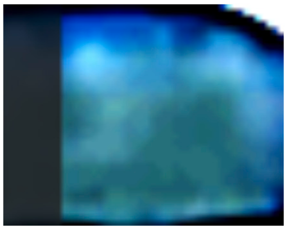


Fig. 6. Sequence of frames that have been processed by the technical consultant.

shows the only sequence of frames (among all the collected surveillance footage) in which the passenger's presence is doubtful, and that has been processed by a technical consultant. These frames demonstrate a clear reflection effect that disappears as the vehicle moves, and exhibit strong chromatic variability due to video compression. The technical advisor claimed to discern a human face in these images, a perception *potentially* induced by confirmation bias, as his intervention was required precisely to identify the possible presence of the passenger. Specifically, on the passage of the car identified by the technical advisor he identifies with certainty, in the first frames of the passage, a face as a "clear" silhouette traceable to a passenger placed in the right front seat. A clear spot can be seen that he traces back to a silhouette. This spot disappears completely in the second part of the passage and then reappears but in inverted colors, that is, it becomes dark. In reality, the images contain pseudo-random blobs, and their temporal persistence can be attributed to a simpler and more evident reflection. The study highlights the lack of scientific rigor in these approaches and proposes new analyses that suggest the absence of a passenger in the vehicle.

An experiment was conducted involving college students. The students were anonymously and without context presented with carefully selected frames from the surveillance camera footage, focusing on the most controversial and doubtful portions. The images shown to the students were a representative subset of the entire set of images extracted and processed by the technical advisor (Fig. 6). To ensure impartiality in their observations, no preliminary information was provided to the students. As part of the experiment, a questionnaire was administered for each image, consisting of the question: "What do you observe in this image?", with response options such as "Nothing" and "Other," along with an open text field for further clarification. This design encouraged a spectrum of responses and allowed students to give concise yet descriptive explanations. A total of 15 subjects participated in evaluating the selected images, resulting in 165 responses, as summarized in Table 1. Notably, only two of the responses conveyed a "certain" identification of a human face. This underscores the necessity of meticulous and unbiased analysis when interpreting surveillance footage. The primary objective of the experiment was to assess the students' ability to detect the presence of a passenger in the vehicle. By presenting them with ambiguous visuals, the experiment aimed to understand their perceptual limitations and biases. The absence of explicit certainty measurement in the questionnaire was addressed by considering the students' chosen responses as an indication of their certainty levels. This highlights the importance of rigorous and unbiased analysis in interpreting surveillance footage. The study emphasizes the need to maintain an objective perspective in observational tasks (Altmann, 1974) and underscores the value of scientific rigor in such analyses.

Table 1
Table representing students' evaluation of the various images extracted from Fig. 6.

Image	Students' Evaluation
	Nothing: 8 Other: 7 Those who answered "other" identified: • Automobile elements
	Nothing: 9 Other: 6 Those who answered "other" identified: • Automobile elements
	Nothing: 10 Other: 5 Those who answered "other" identified: • Silhouette with sun reflection • Automobile elements
	Nothing: 9 Other: 6 Those who answered "other" identified: • A car seat and a person • Reflected human silhouette • A cow
	Nothing: 13 Other: 2 Those who answered "other" identified: • Indistinct silhouettes behind glass
	Nothing: 9 Other: 6 Those who answered "other" identified: • There are two human silhouettes • Silhouettes of hands

4. Strategies for mitigating cognitive bias

Among other fields, software engineering is currently experiencing a significant gap in the area of cognitive bias mitigation techniques, with a notable lack of both practical strategies and theoretical foundations (Mohanani et al., 2017). However, other fields have seen success in this area. For instance, in social work, the use of a nomogram tool and an online training course has been shown to effectively mitigate cognitive bias, leading to improvements in the accuracy of clinical reasoning (Featherston et al., 2019). Despite these advancements, it's important to note that there is currently insufficient evidence to suggest that cognitive bias mitigation interventions significantly improve decision-making in real-life situations (Korteling et al., 2021).

Nevertheless, the potential benefits of countering cognitive biases are clear. In healthcare, for example, addressing cognitive biases in decision-making can help reduce low-value care and enhance the impact of campaigns aimed at reducing such care (Scott et al., 2017). In the realm of gaming, the MACBETH serious game has been found to effectively mitigate cognitive biases such as the fundamental attribution error (Miller and Lawson, 1989) and confirmation bias. The game's effectiveness is further enhanced through explicit instruction and repetitive play, which serve to reinforce learning (Dunbar et al., 2014).

4.1. Strategies for mitigating cognitive bias in forensic science

In the field of forensic science, cognitive biases can significantly impact the accuracy and impartiality of examiner decisions. Specifically, Dror (2013) focused on strategies to mitigate confirmation bias, contextual influences (Nakhaeizadeh et al., 2014), and base-rate regularities (Thakur et al., 2021). Dror proposed that recognizing the spectrum of biases, not only those that can arise from knowing irrelevant case information, but also biases that emerge from base rate regularities, working "backwards" from the suspect to the evidence, and from the working environment itself, can strengthen forensic science.

To mitigate these effects, several strategies have been proposed. First, cognitive training programs raise awareness among examiners about potential biases. Second, blind verification procedures, where the second examiner is unaware of the first examiner's decision, help minimize bias. Third, linear examination processes, starting with evidence analysis before considering the suspect, reduce contextual contamination. Fourth, a triage approach tailors procedures based on case complexity, ensuring that resources are allocated effectively. Lastly, cognitive profiles aid in selecting the best-suited individuals for forensic work, enhancing overall objectivity and performance.

In the pursuit of objectivity and accuracy in forensic analysis, the Forensic Science Regulator of the United Kingdom Government has proposed and implemented several strategies designed to ensure that the analysis is not influenced by any form of bias (Science Regulator, 2020).

1. Blinding Precautions:
   (a) Analysts should be shielded from information that is not directly relevant to the analysis.
   (b) Sequential unmasking can be used, where decisions on suitability are made before comparison with reference samples.
   (c) Careful records should be kept to ensure the order of disclosure and analysis is transparent.
2. Structured Approach (ACE-V and CAI):
   (a) ACE-V (Analysis, Comparison, Evaluation, and Verification) provides a structured process for fingerprint comparison (Reznicek et al., 2010).
   (b) CAI (Case Assessment and Interpretation) uses Bayesian thinking and balances prosecution and defense hypotheses (Jackson, 2011).
   (c) Both approaches emphasize transparency and avoid post hoc rationalization.
3. Awareness, Training, and Competence Assessment:
   (a) Practitioners need training on cognitive bias risks and mitigation strategies.
   (b) Proficiency testing and regular assessment help maintain competence.
4. Avoidance of Reconstructive Effects:
   (a) Contemporaneous notes or technical records prevent reconstructive bias.
   (b) Analysts should rely on memory as little as possible.
5. Avoidance of Role Effects:
   (a) Organizational structures should insulate scientists from potential biasing pressures.
   (b) Scientists must prioritize their duty to the court over any other obligations.

Controlling the flow of information to analysts is crucial to prevent unnecessary influences on their judgment. Dror and Kukucka (2021) introduced Linear Sequential Unmasking–Expanded (LSU-E), a methodology that reduces noise and bias in forensic decision-making. LSU-E involves initial analysis of raw data without reference material, followed by a sequential consideration of relevant information based on objectivity and relevance. This approach optimizes information presentation to enhance utility and minimize cognitive biases. LSU-E also offers guidelines for documenting the influence of information on the decision-making process, ensuring transparency and accountability. Furthermore, Camilleri et al. (2019) emphasized the need for a systematic assessment of cognitive bias risks in forensic laboratories, proposing a risk management framework. Key mitigation strategies include raising awareness through training, developing guidance documents, and limiting access to task-irrelevant information. Redacting irrelevant details, implementing blind known tests, and conducting independent casefile reviews enhance objectivity and reduce expectation bias. These measures may ensure the integrity of forensic interpretations and minimize the impact of cognitive biases.

5. The intersection of generative AI and the craftsmanship of deepfakes

Generative Artificial Intelligence (GenAI) is an increasingly popular technology that has significant implications across various fields (Bockting et al., 2023; Stokel-Walker and Van Noorden, 2023). It refers to AI systems that can generate new content, such as text, images, and audio, in response to human prompts. These systems, including deepfakes and AI chatbots like Generative Pre-trained Transformers (e.g., GPT-4), use complex algorithms to produce outputs that are often indistinguishable from content created by humans (Stokel-Walker and Van Noorden, 2023). The technology is advancing rapidly, with each new version adding capabilities that increasingly encroach on human skills; however, the use of these "black box" AI tools can introduce biases and inaccuracies, potentially distorting scientific facts while still sounding authoritative.

The emergence of these sophisticated AI technologies has brought about fresh challenges in this domain. One such challenge is the detection of deepfakes, the research area of which is constantly expanding (as shown in Fig. 7). Deepfakes are synthetic media created through generative models based mainly on Generative Adversarial Networks (GANs) and Diffusion Models (DMs) (Goodfellow et al., 2014). GANs are composed of a Generator (G) and a Discriminator (D) trained simultaneously through a competitive process. The Generator is trained to capture the data distribution of the training set Ts. The Discriminator is trained to distinguish the images created by G from the set Ts. When G creates images with the same data distribution as Ts, D will no longer be able to solve its task and the training phase can be considered completed. Currently, researchers demonstrated that synthetic images created by DMs are better than those generated by GAN engines in terms of photorealism, as the creation process follows a more accurate and "controlled" flow. The basic idea of DMs is to iteratively add noise to an input random noise vector for synthetic data generation in order to model complex data distributions. Fig. 8 and Fig. 9 show generic GAN and DM schemes related to the creation of synthetic people's faces.

Deepfakes can pose significant challenges in distinguishing real images from manipulated ones, thereby complicating the task of digital forensic investigators: in fact, the problem of deepfake detection has been addressed extensively by the scientific community (Masood et al., 2023; Verdoliva, 2020; Lin et al., 2024). In this context, preventing

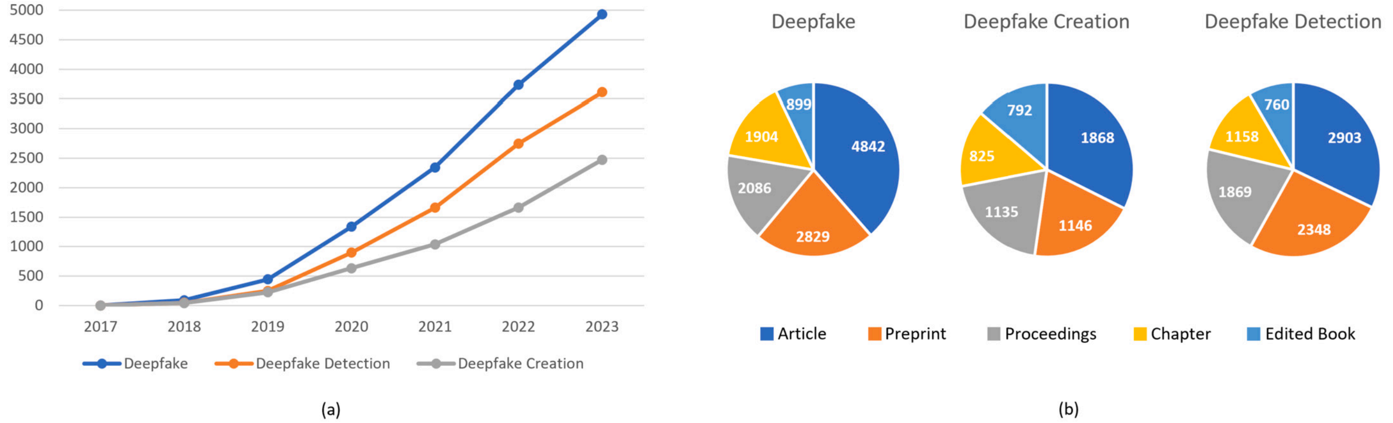


Fig. 7. Statistics of papers published in the deepfake field. (a) Papers published from 2017 to 2023 with the keywords deepfake, deepfake creation, deepfake detection. (b) Numbers of papers published in: Article, Preprint, Proceedings, Chapter and Edited Book.

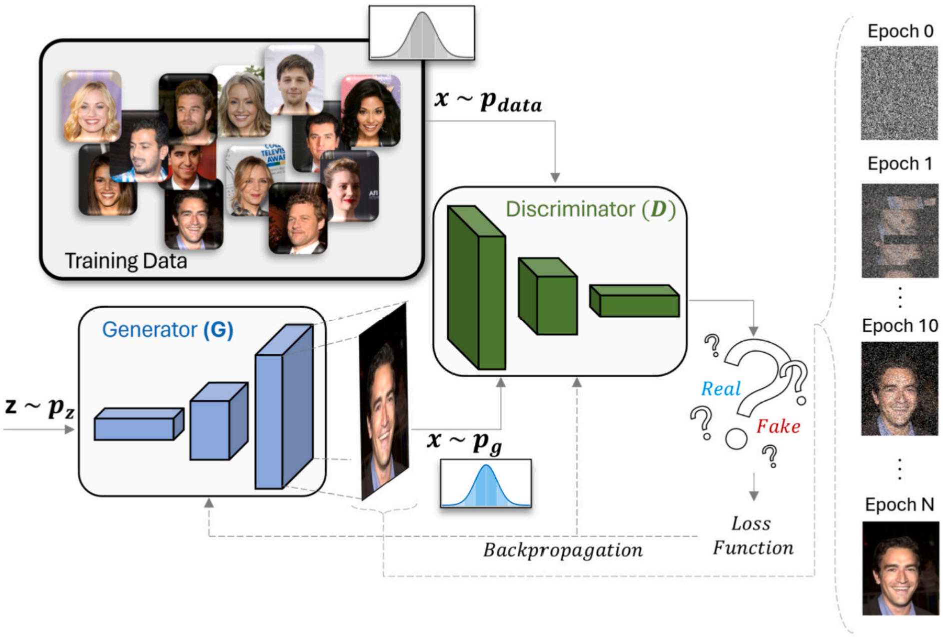


Fig. 8. A standard GAN framework. A Generator (G) creates data samples from noise, aiming to mimic the training set. A Discriminator (D) differentiates between real and G-generated data. Training ends when D can't distinguish G's images from training samples.

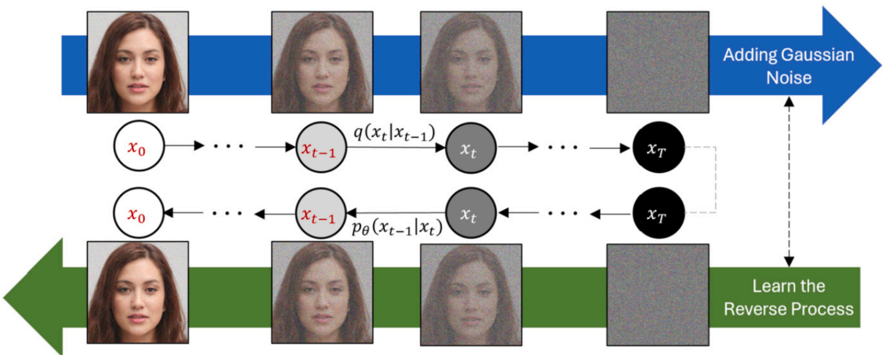


Fig. 9. A Diffusion Model architecture. Training data is corrupted with added Gaussian noise. From this data (x_T step), a reverse process is constructed to generate new samples resembling the original ones.

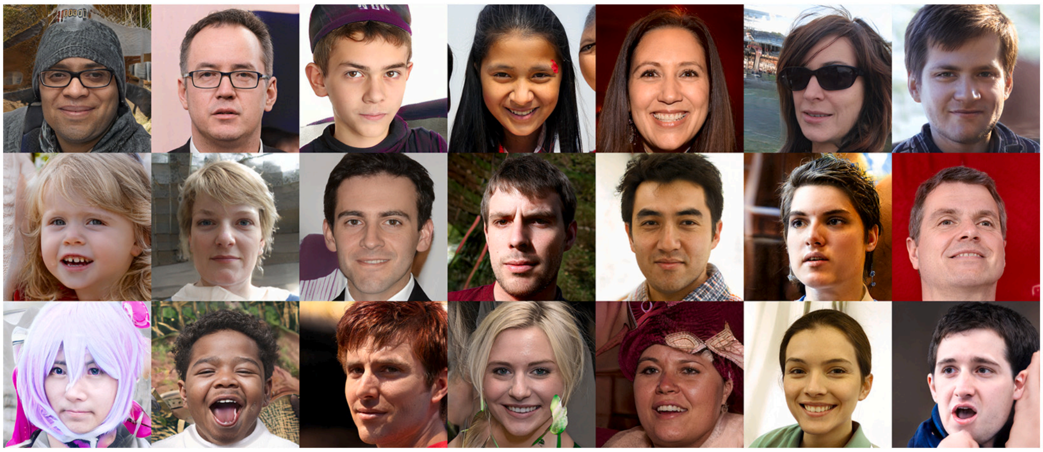


Fig. 10. Figs. 8 and 9 illustrate the architectures of generative models that produce these images. The verisimilitude of these faces could potentially lead observers to question the existence of the depicted individuals, thereby giving rise to the Impostor Bias. Credit: Guarnera et al. (2022).

cognitive biases in digital forensics becomes crucial to ensure the objectivity and neutrality of judgments.

6. The impostor bias: how AI media triggers bias and doubt in perception

The emergence of GenAI has brought about a sea change in the multimedia landscape, opening up novel avenues for crafting and modifying content. It is largely attributed to the development of a class of machine learning models known as foundation models (FMs) (Rabowsky, 2023). These models, which include the likes of ChatGPT released in November 2022, marked the beginning of a new era in Artificial Intelligence. Foundation models are distinguished by their powerful applications to GenAI, which involves the use of models to generate new content and transform existing content. GenAI models can produce high-quality artistic media for visual arts, concept art, music, fiction, literature, video, and animation; distinguishing the real from the fake is becoming increasingly complex, as in the case of human face recognition and its veracity (Fig. 10).

The existence of GenAI and its knowledge can lead to the development of a new type of cognitive bias, which we have identified as the “Impostor Bias”. This is a hypothetical bias with no empirical basis yet, and the term “Impostor” is derived from the “Impostor Syndrome”, a psychological phenomenon in which people doubt their competence and fear being exposed as fraudulent. This bias, however, refers to a different context: distrust of multimedia content generated by Artificial Intelligence. In the context of AI, “Impostor Bias” refers to the tendency to doubt the veracity of multimedia elements such as videos, photos and audios, due to the knowledge that these can be realistically generated by AI models. This bias manifests itself as an a priori distrust, regardless of the quality or context of the multimedia content. The name “Impostor Bias” is appropriate because, just as in Impostor Syndrome, there is a persistent doubt about the veracity of something - in this case, AI-generated media content. Although they may appear authentic and realistic, awareness of the possibility that they are AI-generated “impostors” can lead to doubt and distrust.

The term “bias” is used to describe Impostor Bias because it refers to a systematic and predictable tendency in the way people perceive media content, regardless of the evidence presented. For instance, this is not simply a variation in evaluations due to different levels of knowledge of the evaluator. In other words, even when presented with AI-generated media content that is indistinguishable from the real thing, people with Impostor Bias may still doubt its authenticity due to the knowledge that such content may be AI-generated. This is a bias because it is based on an a priori assumption rather than an objective assessment of the content itself. Furthermore, the term “bias” implies that this tendency can lead to distortions in perception and judgment. For example, “Impostor Bias” may lead people to discard or devalue authentic media content because they perceive it as potentially false or misleading.

Judicial practitioners and investigators may struggle with determining the authenticity of multimedia content, especially as AI-generated media becomes more realistic and descriptive. The risk of falling into cognitive traps or suspecting original content as AI-generated underscores the complexity of this issue. Additionally, GenAI raises concerns about artwork counterfeiting, copyright infringement, and the potential for forged masterpieces to be sold as genuine. Indeed, such technologies are now capable of simulating the technique and artistic style of the most famous artists, thereby compromising the ability to correctly discern between real and simulated works (Epstein et al., 2023; Leotta et al., 2023) (Fig. 11). Forensic examiners from 21 countries showed limited understanding and appreciation of cognitive bias, with fewer than half supporting blind testing, highlighting the need for procedural reforms to blind them to potentially biasing information (Kukucka et al., 2017).

7. Generative AI and deepfake detection methods

The mitigation of the potential impacts of the hypothetical “Impostor Bias” could significantly rely on Generative AI and deepfake detection methods. As AI-generated media products become increasingly realistic (Verdoliva, 2020; de Lima-Santos and Ceron, 2021), the prevalence of “Impostor Bias” is expected to rise, posing significant challenges in various sectors, including the justice system. In this context, technologies for deepfake detection become increasingly important. These technologies are evolving to not only recognize manipulated or altered multimedia artifacts but also those generated from scratch using descriptive textual inputs.

Detecting content generated by Generative AI, such as ChatGPT, is a rapidly evolving field of study. Recent research has focused on developing machine learning tools capable of distinguishing between human-written text and machine-generated content (Prillaman, 2023; Nature, 2023; Weber-Wulff et al., 2023; Perkins et al., 2024; Liu et al., 2024). These tools, such as the one described in a study by Perkins et al. (2024), analyze various features of writing style, including variation in sentence lengths, and the frequency of certain words and punctuation marks. However, the efficacy of these detection tools can be significantly reduced when they are confronted with machine-generated content that has been modified using techniques designed to evade detection (Perkins et al., 2024). For instance, the study found that the detectors' already low accuracy rates (39.5%) showed major reductions in accuracy (17.4%) when faced with manipulated content. Despite these challenges, the development of more accurate and robust GenAI

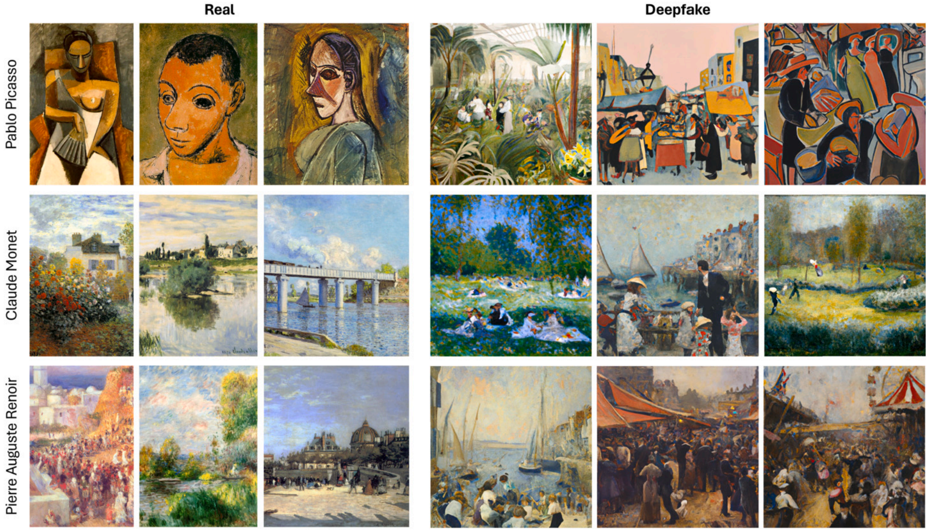


Fig. 11. Real vs deepfake images of famous artists.

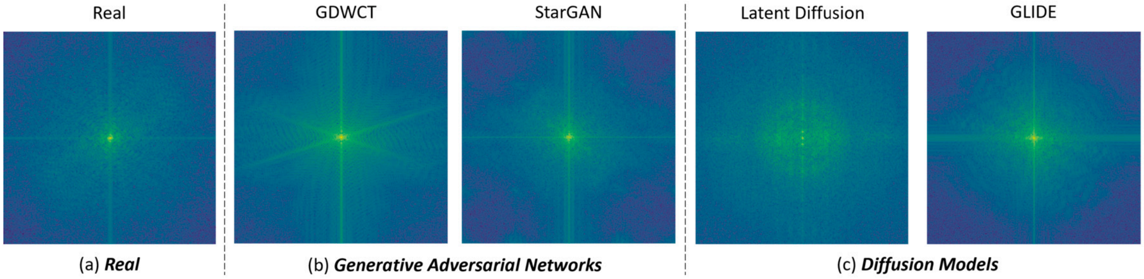


Fig. 12. Fourier spectrum of different categories of data: (a) real images; (b) images generated by two GAN architectures (GDWCT (Cho et al., 2019) and StarGAN (Choi et al., 2018)); (c) images generated by two DM architectures (Latent Diffusion (Ramesh et al., 2022) and GLIDE (Nichol et al., 2022)). The abnormal frequencies (light peaks) are mainly visible in the images generated by artificial intelligence engines.

detection methods continues to be a critical area of research, given the increasing prevalence of AI-generated content in various domains (Perkins et al., 2024; Liu et al., 2024). For example, a tool called Check-GPT has been developed, which examines 20 features of writing style to determine whether an academic scientist or ChatGPT wrote a piece of text (Liu et al., 2024). The tool was found to be highly accurate, achieving an average classification accuracy of 98% to 99% for task-specific discipline-specific detectors and the unified detectors.

Furthermore, researchers have demonstrated that generative engines leave traces on synthetic content that can be identified and detected in the frequency domain (Guarnera et al., 2020c; Zhang et al., 2019; Marra et al., 2019; Giudice et al., 2021; Dzanic et al., 2020; Durall et al., 2020) (Fig. 12). These traces are characterized by both the network architecture (number and type of layers) and its specific parameters (Yu et al., 2019). In order to distinguish real data from deepfake, Guarnera et al. (2020a,b) proposed methods based on the Expectation-Maximization (Moon, 1996) algorithm capable of capturing traces defined as the correlation of pixels left by convolutional layers.

Wang et al. (2020) used ResNET-50 to distinguish real and ProGAN-generated images, showing generalization across different GANs. FakeSpotter, proposed by Wang et al. (2021), detects GAN-generated faces by monitoring CNN neuron behaviors. Vision Transformer-based solutions for deepfake detection have also been proposed (Wodajo and Atnafu, 2021; Coccomini et al., 2022; Heo et al., 2023; Wang et al., 2022). Wodajo and Atnafu (2021) combined transformers with a convolutional network to extract patches from detected faces in videos. Several studies (Yu et al., 2019, 2021; Girish et al., 2021; Asnani et al., 2023; Yu et al., 2020; Guarnera et al., 2022, 2024) have explored identifying specific GAN models used in creation (Model Attribution Task). Guarnera et al. (2022) distinguished 100 StyleGAN2 instances using ResNET-18 (He et al., 2016) and metric learning (Liu et al., 2012), demonstrating the method's effectiveness in deepfake model recognition.

The scientific community is also working extensively on the creation of advanced techniques for the detection of synthetic images created by diffusion models. Corvi et al. (2023) have been trying to understand how difficult it is to distinguish synthetic images generated from diffusion models from real ones, and whether current state-of-the-art detectors are suitable for this task. Sha et al. (2023) proposed DE-FAKE, a machine-learning classifier-based method for diffusion model detection on four popular text-image architectures. Guarnera et al. (2023) proposed a hierarchical approach based on different architectures in order to define: whether the image is real or manipulated via any generative architecture (AI-generated); the specific framework used among GAN or DM; defines the specific generative architecture used among a predefined set. Furthermore, another practical digital forensic tool, Transfer learning-based Autoencoder with Residuals (TAR), was proposed (Lee et al., 2021). The ultimate goal of TAR was to develop a unified model to detect various types of deepfake videos with high accuracy, with only a small number of training samples that can work well in real-world settings. A short summary of these methods can be found in Table 2.

Experimental results of all these methods show that, in general, all generative models leave unique traces that can solve all previously listed tasks with high accuracy. Therefore, these methods can be used in order to help the general user in countering what we called the *Impostor Bias* phenomenon. However, we want to highlight an important

Table 2
A summary of the discussed deepfake detection methods.

Reference	Generation Models	Database(s) Used	Precision (avg)
He et al. (2016)	StyleGAN, StyleGAN2-ADA	FFHQ (Flickr-Faces-HQ)	96.2%
Wang et al. (2020)	ProGAN	CelebA	99.1%
Guarnera et al. (2023)	GANs: (AttGAN, CycleGAN, GDWCT, IMLE, ProGAN, StarGAN, StarGAN-v2, StyleGAN, StyleGAN2) DMs: (DALL·E 2, GLIDE, Latent Diffusion, Stable Diffusion)	CelebA, FFHQ, ImageNet	97,6% (Level 1) 98,0% (Level 2) 97,8% (Level 3, GANs) 98,0% (Level 3, DMs)
Wang et al. (2021)	StyleGAN, StyleGAN2, BigGAN, ProGAN	FaceForensics++	90.6%
Wodajo and Atnafu (2021)	FaceSwap, Face2Face, FaceShifter, NeuralTextures, DeepFakeDetection	FaceForensics++, UADFV	91.5%
Lee et al. (2021)	FaceSwap, Face2Face, DeepFake, NeuralTextures	FaceForensics++	98.0% deepfake type detection 89.5% on DW videos
Sha et al. (2023)	GLIDE, Latent Diffusion, Stable Diffusion, DALL·E 2	MSCOCO (a), Flickr30k (b)	90.2%(a), 84.6% (b)

element of the previously listed methods and not just deepfake detection. All methods in the literature achieve extremely high results in "constrained" contexts, that is, with architectures known a priori. In practice, current deepfake detectors fail to generalize with synthetic images generated by novel architectures (different from those used during the training procedure), resulting in a drastic drop in classification performance. Some recent new methods (Dong et al., 2023; Coccomini et al., 2023) published by the scientific community seem to be good starting points in order to achieve generalization.

8. Discussion

The concept of Impostor Bias, though not yet widely recognized, is increasingly relevant in the era of deepfakes and digital forensics. The recent Ukraine war, marked by propaganda and misinformation on social media (Ciuriak, 2022; Suciu, 2022), has heightened this bias. The constant exposure to videos, photos, and statements fosters an inherent suspicion of media manipulation or AI generation (Linehan et al., 2023). This "digital warfare" has led to immediate doubt and scrutiny of all published media, sometimes uncovering deepfakes (Bond, 2023). The prevalence of Impostor Bias will persist, especially with sensitive content like wartime communications (Linehan et al., 2023). Thus, novel techniques and strategies are essential for effective mitigation.

The unique nature of the Impostor Bias necessitates specialized approaches beyond the strategies discussed in Section 4. While those strategies offer valuable insights, addressing Impostor Bias effectively requires a combination of targeted measures. This includes shielding analysts from irrelevant information and providing specific training on cognitive bias risks. Additionally, proficiency testing and regular assessments maintain practitioner competence. However, to truly counter Impostor Bias, advanced technical tools for synthetic content detection are essential. Practical deep learning algorithms can detect generative models' traces, aiding investigators in distinguishing authentic evidence. Given the success of deepfake detection algorithms (Pei et al., 2024; Gong and Li, 2024), these tools can assist in counteracting Impostor Bias and other digital biases. Similar detection methods for forgery images (Baumy et al., 2022; Zanardelli et al., 2023; Singh and Kumar, 2024) and multimedia data manipulation (Galante et al., 2023; Dunsin et al., 2024) further reinforce the value of algorithmic assistance in avoiding bias. The increasing sophistication of digital data demands the use of advanced deepfake detection techniques to maintain trust and accuracy in forensic examinations.

Detection methods play a crucial role in identifying deepfakes, but interpreting and explaining the results is equally important. While some methods use classic machine learning techniques, others employ deep learning techniques, each with its advantages and limitations. Classic machine learning approaches are more interpretable but lack robustness, while deep learning methods offer higher performance and robustness but are more complex to explain. The question of trusting the predictions of detection algorithms is indeed a separate but related study, as explored by Mathews et al. (2023), Lim et al. (2022), and others (Cantero-Arjona and Sánchez-Macián, 2024; Pinhasov et al., 2024; Pandey et al., 2024; Pontorno et al., 2024). These studies enhance the "trust" in detection methods and underscore the need for further exploration of cognitive biases in digital forensic science. Creating specialized training programs for practitioners can enhance awareness, reduce the influence of biases, and preserve the objectivity of judgments by disregarding irrelevant elements and separating personal experiences.

9. Conclusions and future works

In this article, we have explored the significant impact of cognitive biases in the fields of forensics and digital forensics, specifically focusing on confirmation bias, anchoring bias, and hindsight bias. To reduce the influence of these biases, we have discussed strategies such as game-based interventions and the Linear Sequential Unmasking-Expanded approach. With the growing prevalence of deepfakes, synthetic media that can manipulate or impersonate individuals, the integrity of digital evidence is at risk. We have surveyed current deepfake detection methods and their limitations, introducing the novel concept of Impostor Bias. This bias, influenced by the widespread use of deepfakes, may lead to false negatives and reduced confidence in digital forensic findings. To address these challenges, we propose future research directions, including the development of advanced deepfake detection methods and a deeper understanding of the factors contributing to Impostor Bias. Furthermore, we emphasize the importance of interventions to mitigate this bias and the need to explore the ethical, legal, and social implications of deepfakes. By doing so, we aim to enhance the reliability and integrity of digital forensic practices, ensuring the accuracy and objectivity of forensic investigations.

CRediT authorship contribution statement

Mirko Casu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luca Guarnera:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pasquale Caponnetto:** Visualization, Validation, Supervision, Project administration, Investigation, Formal analysis, Conceptualization. **Sebastiano Battiato:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgements

This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

Altmann, J., 1974. Observational study of behavior: sampling methods. Behaviour 49 (3), 227–267. https://doi.org/10.1163/156853974X00534.

Arceri, N.F., Giudice, O., Battiato, S., 2023. An innovative tool for uploading/scraping large image datasets on social networks. In: 2023 IEEE International Conference on Metrology for EXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 549–554.

Asnani, V., Yin, X., Hassner, T., Liu, X., 2023. Reverse engineering of generative models: inferring model hyperparameters from generated images. IEEE Trans. Pattern Anal. Mach. Intell.

Battiato, S., Giudice, O., Paratore, A., 2016. Multimedia forensics: discovering the history of multimedia contents. In: Proceedings of the 17th International Conference on Computer Systems and Technologies 2016. Association for Computing Machinery, New York, NY, USA, pp. 5–16.

Baumy, A., Algarni, A.D., Abdalla, M., El-Shafai, W., El-Samie, A., Fathi, E., Soliman, N.F., 2022. Efficient forgery detection approaches for digital color images. Comput. Mater. Continua 71. https://doi.org/10.32604/cmc.2022.021047.

Beltrani, A., Reed, A.L., Zapf, P., Otto, R., 2018. Is hindsight really 20/20?: the impact of outcome information on the decision-making process. Int. J. Forensic Ment. Health 17, 285–296. https://doi.org/10.1080/14999013.2018.1505790.

Berthet, V., 2021. The measurement of individual differences in cognitive biases: a review and improvement. Front. Psychol. 12. https://doi.org/10.3389/fpsyg.2021.630177.

Berthet, V., 2022. The impact of cognitive biases on professionals' decision-making: a review of four occupational areas. Front. Psychol. 12. https://doi.org/10.3389/fpsyg.2021.802439.

Bhadra, P., 2021. Is Forensic Evidence Impartial? Cognitive Biases in Forensic Analysis. Springer, Singapore, Singapore, pp. 215–227.

Bockting, C.L., van Dis, E.A.M., van Rooij, R., Zuidema, W., Bollen, J., 2023. Living guidelines for generative AI — why scientists must oversee its use. Nature 622, 693–696. https://doi.org/10.1038/d41586-023-03266-1.

Bond, S., 2023. How Russia is losing — and winning — the information war in Ukraine. NPR. (Accessed 13 December 2023).

Camilleri, A., Abarno, D., Bird, C., Coxon, A., Mitchell, N., Redman, K.E., Sly, N., Wills, S., Silenieks, E., Simpson, E., Lindsay, H., 2019. A risk-based approach to cognitive bias in forensic science. Science & Justice: Journal of the Forensic Science Society 59 (5), 533–543. https://doi.org/10.1016/J.SCIJUS.2019.04.003.

Cantero-Arjona, P., Sánchez-Macián, A., 2024. Deepfake detection and the impact of limited computing capabilities. arXiv:2402.14825.

Chapman, G., Johnson, E.J., 1994. The limits of anchoring. J. Behav. Decis. Mak. 7, 223–242. https://doi.org/10.1002/BDM.3960070402.

Chapman, G., Johnson, E.J., 1999. Anchoring, activation, and the construction of values. Organ. Behav. Hum. Decis. Process. 79 (2), 115–153. https://doi.org/10.1006/OBHD.1999.2841.

Cho, W., Choi, S., Park, D.K., Shin, I., Choo, J., 2019. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10639–10647.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797.

Ciuriak, D., 2022. Social media warfare is being invented in Ukraine. https://www.cigionline.org/articles/social-media-warfare-is-being-invented-in-ukraine/. (Accessed 13 December 2023).

Coccomini, D.A., Caldelli, R., Falchi, F., Gennaro, C., 2023. On the generalization of deep learning models in video deepfake detection. Journal of Imaging 9, 89.

Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F., 2022. Combining efficientnet and vision transformers for video deepfake detection. In: International Conference on Image Analysis and Processing. Springer, pp. 219–229.

Cooper, G.S., Meterko, V., 2019. Cognitive bias research in forensic science: a systematic review. Forensic Sci. Int. 297, 35–46. https://doi.org/10.1016/j.forsciint.2019.01.016.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L., 2023. On the detection of synthetic images generated by diffusion models. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1–5.

Cowan, E.J., Koppl, R., 2011. An experimental study of blind proficiency tests in forensic science. Rev. Austrian Econ. 24, 251–271. https://doi.org/10.1007/s11138-010-0130-4.

Di Lazzaro, P., Murra, D., Schwortz, B., 2013. Pattern recognition after image processing of low-contrast images, the case of the shroud of turin. Pattern Recognit. 46, 1964–1970.

Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z., 2023. Implicit identity leakage: the stumbling block to improving deepfake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3994–4004.

Douglass, A.B., Charman, S.D., Matuku, K.P., Shambaugh, L.J., Lapar, M.P., Lamere, E., 2023. Case information biases evaluations of video-recorded eyewitness identification evidence. Journal of Applied Research in Memory and Cognition.

Dror, I., 2013. Practical solutions to cognitive and human factor challenges in forensic science. Forensic Science Policy & Management: An International Journal 4, 105–113. https://doi.org/10.1080/19409044.2014.901437.

Dror, I., Melinek, J., Arden, J.L., Kukucka, J., Hawkins, S., Carter, J., Atherton, D.S., 2021. Cognitive bias in forensic pathology decisions. J. Forensic Sci. 66, 1751–1757.

Dror, I., Rosenthal, R., 2008. Meta-analytically quantifying the reliability and biasability of forensic experts. J. Forensic Sci. 53, 900–903. https://doi.org/10.1111/j.1556-4029.2008.00762.x.

Dror, I.E., 2020. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. Anal. Chem. 92 (12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704.

Dror, I.E., Kukucka, J., 2021. Linear sequential unmasking–expanded (lsu-e): a general approach for improving decision making as well as minimizing noise and bias. Forensic Science International: Synergy 3, 100161.

Dunbar, N.E., Miller, C.H., Adame, B.J., Elizondo, J., Wilson, S.N., Lane, B.L., Kauffman, A.A., Bessarabova, E., Jensen, M.L., Straub, S.K., Lee, Y.H., Burgoon, J.K., Valacich, J.J., Jenkins, J., Zhang, J., 2014. Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game. Comput. Hum. Behav. 37, 307–318. https://doi.org/10.1016/j.chb.2014.04.053.

Dunsin, D., Ghanem, M.C., Ouazzane, K., Vassilev, V., 2024. A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. Forensic Science International: Digital Investigation 48, 301675. https://doi.org/10.1016/j.fsidi.2023.301675.

Durall, R., Keuper, M., Keuper, J., 2020. Watch your up-convolution: cnn based generative deep neural networks are failing to reproduce spectral distributions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7887–7896.

Dzanic, T., Shah, K., Witherden, F., 2020. Fourier spectrum discrepancies in deep network generated images. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc, pp. 3022–3032.

Edmond, G., Tangen, J.M., Searston, R.A., Dror, I.E., 2015. Contextual bias and cross-contamination in the forensic sciences: the corrosive implications for investigations, plea bargains, trials and appeals. Law Probab. Risk 14, 1–25. https://doi.org/10.1093/lpr/mgu018.

van den Eeden, C.A.J., de Poot, C.J., van Koppen, P.J., 2019. The forensic confirmation bias: a comparison between experts and novices. J. Forensic Sci. 64, 120–126. https://doi.org/10.1111/1556-4029.13817.

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M.R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., et al., 2023. Art and the science of generative AI: a deeper dive. Preprint. arXiv:2306.04141.

Fanti, G., Maggiolo, R., 2004. The double superficiality of the frontal image of the turin shroud. J. Opt. A, Pure Appl. Opt. 6, 491.

Featherston, R.J., Shlonsky, A., Lewis, C., Luong, M.L., Downie, L.E., Vogel, A.P., Granger, C., Hamilton, B., Galvin, K., 2019. Interventions to mitigate bias in social work decision-making: a systematic review. Res. Soc. Work Pract. 29, 741–752. https://doi.org/10.1177/1049731518819160.

Fontani, M., 2021. Cognitive bias: steering conclusions irrationally. https://blog.ampedsoftware.com/2021/04/20/cognitive-bias-steering-conclusions-irrationally. (Accessed 7 November 2023).

Forensic Science Regulator, 2020. Cognitive bias effects relevant to forensic science examinations. https://www.gov.uk/government/publications/cognitive-bias-effects-relevant-to-forensic-science-examinations. (Accessed 7 November 2023).

Galante, N., Cotroneo, R., Furci, D., Lodetti, G., Casali, M.B., 2023. Applications of artificial intelligence in forensic sciences: current potential benefits, limitations and perspectives. Int. J. Leg. Med. 137, 445–458. https://doi.org/10.1007/s00414-022-02928-5.

Gardner, B.O., Kelley, S., Murrie, D.C., Blaisdell, K.N., 2019. Do evidence submission forms expose latent print examiners to task-irrelevant information? Forensic Sci. Int. 297, 236–242.

Girish, S., Suri, S., Rambhatla, S.S., Shrivastava, A., 2021. Towards discovery and attribution of open-world gan generated images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14094–14103.

Giroux, M.E., Coburn, P.I., Harley, E.M., Connolly, D.A., Bernstein, D.M., 2016. Hindsight bias and law. Z. Psychol. 224, 190–203. https://doi.org/10.1027/2151-2604/a000253.

Giudice, O., Guarnera, L., Battiato, S., 2021. Fighting deepfakes by detecting GAN DCT anomalies. Journal of Imaging 7, 128. https://doi.org/10.3390/jimaging7080128.

Giudice, O., Paratore, A., Moltisanti, M., Battiato, S., 2017. A classification engine for image ballistics of social data. In: Image Analysis and Processing-ICIAP 2017: 19th

International Conference. Catania, Italy, September 11–15, 2017, Proceedings, Part II 19. Springer, pp. 625–636.

Gong, L.Y., Li, X.J., 2024. A contemporary survey on deepfake detection: datasets, algorithms, and challenges. Electronics 13. https://doi.org/10.3390/electronics13030585.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680.

Grisham, J.R., Becker, L., Williams, A.D., Whitton, A.E., Makkar, S.R., 2014. Using cognitive bias modification to deflate responsibility in compulsive checkers. Cogn. Ther. Res. 38, 505–517.

Guarnera, L., Giudice, O., Battiato, S., 2020a. Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 666–667.

Guarnera, L., Giudice, O., Battiato, S., 2020b. Fighting deepfake by exposing the convolutional traces on images. IEEE Access 8, 165085–165098.

Guarnera, L., Giudice, O., Battiato, S., 2023. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. Preprint. arXiv:2303.00608.

Guarnera, L., Giudice, O., Battiato, S., 2024. Mastering deepfake detection: a cutting-edge approach to distinguish gan and diffusion-model images. ACM Trans. Multimed. Comput. Commun. Appl.

Guarnera, L., Giudice, O., Nastasi, C., Battiato, S., 2020c. Preliminary forensics analysis of deepfake images. In: 2020 AEIT International Annual Conference (AEIT), IEEE, pp. 1–6.

Guarnera, L., Giudice, O., Nießner, M., Battiato, S., 2022. On the exploitation of deepfake model recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 61–70.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heo, Y.J., Yeo, W.H., Kim, B.G., 2023. Deepfake detection algorithm based on improved vision transformer. Appl. Intell. 53, 7512–7527.

Horsman, G., 2024. Sources of error in digital forensics. Forensic Science International: Digital Investigation 48, 301693. https://doi.org/10.1016/j.fsidi.2024.301693.

Jackson, G., 2011. The Development of Case Assessment and Interpretation (CAI) in Forensic Science. Ph.D. thesis. University of Abertay Dundee.

Jeanguenat, A.M., Budowle, B., Dror, I., 2017. Strengthening forensic DNA decision making through a better understanding of the influence of cognitive bias. Science & Justice: Journal of the Forensic Science Society 57 (6), 415–420. https://doi.org/10.1016/j.scijus.2017.07.005.

Kassin, S.M., Dror, I.E., Kukucka, J., 2013. The forensic confirmation bias: problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition 2, 42–52.

Korteling, J., Gerritsma, J.Y., Toet, A., 2021. Retention and transfer of cognitive bias mitigation interventions: a systematic literature study. Front. Psychol. 12, 629354.

Kukucka, J., Kassin, S., Zapf, P.A., Dror, I., 2017. Cognitive bias and blindness: a global survey of forensic science examiners. Journal of Applied Research in Memory and Cognition 6, 452–459. https://doi.org/10.1016/J.JARMAC.2017.09.001.

Lee, S., Tariq, S., Kim, J., Woo, S.S., 2021. Tar: generalized forensic framework to detect deepfakes using weakly supervised learning. In: IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, pp. 351–366.

Leotta, R., Giudice, O., Guarnera, L., Battiato, S., 2023. Not with my name! Inferring artists' names of input strings employed by diffusion models. In: International Conference on Image Analysis and Processing. Springer, pp. 364–375.

Lester, K.J., Mathews, A., Davison, P.S., Burgess, J.L., Yiend, J., 2011. Modifying cognitive errors promotes cognitive well being: a new approach to bias modification. J. Behav. Ther. Exp. Psychiatry 42, 298–308.

Lim, S.Y., Chae, D.K., Lee, S.C., 2022. Detecting deepfake voice using explainable deep learning techniques. Appl. Sci. 12. https://doi.org/10.3390/app12083926.

de Lima-Santos, M.F., Ceron, W., 2021. Artificial intelligence in news media: current perceptions and future outlook. Journalism and Media. https://doi.org/10.20944/preprints202110.0020.v1.

Lin, L., Gupta, N., Zhang, Y., Ren, H., Liu, C.H., Ding, F., Wang, X., Li, X., Verdoliva, L., Hu, S., 2024. Detecting multimedia generated by large ai models: a survey. Preprint. arXiv:2402.00045.

Linehan, C., Murphy, G., Twomey, J.J., 2023. Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine. http://theconversation.com/deepfakes-in-warfare-new-concerns-emerge-from-their-use-around-the-russian-invasion-of-ukraine-216393. (Accessed 13 December 2023).

Liu, E.Y., Guo, Z., Zhang, X., Jojic, V., Wang, W., 2012. Metric learning from relative comparisons by minimizing squared residual. In: 2012 IEEE 12th International Conference on Data Mining, IEEE, pp. 978–983.

Liu, Z., Yao, Z., Li, F., Luo, B., 2024. On the detectability of chatgpt content: benchmarking, methodology, and evaluation through the lens of academic writing. arXiv:2306.05524.

Maity, A., Pious, R., Lenka, S.K., Choudhary, V., Lokhande, P.S., 2023. A survey on super resolution for video enhancement using gan. arXiv:2312.16471.

Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G., 2019. Do GANs leave artificial fingerprints? In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, pp. 506–511.

Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., Malik, H., 2023. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl. Intell. 53, 3974–4026.

Mathews, S., Trivedi, S., House, A., Povolny, S., Fralick, C., 2023. An explainable deepfake detection framework on a novel unconstrained dataset. Complex Intell. Syst. 9, 4425–4437. https://doi.org/10.1007/s40747-022-00956-7.

Meterko, V., Cooper, G., 2022. Cognitive biases in criminal case evaluation: a review of the research. J. Police Crim. Psychol. 37, 101–122.

Miller, A., Lawson, T., 1989. The effect of an informational option on the fundamental attribution error. Pers. Soc. Psychol. Bull. 15, 194–204. https://doi.org/10.1177/0146167289152006.

Mohanani, R., Salman, I., Turhan, B., Rodríguez Marín, P., Ralph, P., 2017. Cognitive biases in software engineering: a systematic mapping study. IEEE Trans. Softw. Eng. 46, 1318–1339. https://doi.org/10.1109/TSE.2018.2877759.

Moon, T.K., 1996. The expectation-maximization algorithm. IEEE Signal Process. Mag. 13, 47–60.

Moser, S., 2013. Confirmation bias: the pitfall of forensic science. Themis: Research Journal of Justice Studies and Forensic Science 1. https://doi.org/10.31979/THEMIS.2013.0107. https://scholarworks.sjsu.edu/themis/vol1/iss1/7.

Nakhaeizadeh, S., Dror, I.E., Morgan, R.M., 2014. Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. Sci. Justice 54, 208–214.

Nature, 2023. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 613, 612. https://doi.org/10.1038/d41586-023-00191-1.

Neal, T., Grisso, T., 2014. The cognitive underpinnings of bias in forensic mental health evaluations. Psychol. Public Policy Law 20, 200–211. https://doi.org/10.1037/A0035824.

Neal, T., Lienert, P., Denne, E., Singh, J., 2022. A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. Law Hum. Behav. https://doi.org/10.1037/lhb0000482.

Nguyen, A.M.D., Beins, B.C., 2013. Response bias (response style). The Encyclopedia of Cross-Cultural Psychology, 1098–1103.

Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M., 2022. Glide: towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning, PMLR, pp. 16784–16804.

Nickerson, R.S., 1998. Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psychol. 2, 175–220.

Palmer, G., et al., 2001. A road map for digital forensic research. In: First Digital Forensic Research Workshop, pp. 27–30.

Pandey, M., Singh, S., Malik, A., Kumar, R., 2024. Detecting low-resolution deepfakes: an exploration of machine learning techniques. Multimed. Tools Appl. https://doi.org/10.1007/s11042-024-18235-7.

Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., Tao, D., 2024. Deepfake generation and detection: a benchmark and survey. arXiv:2403.17881.

Perkins, M., Roe, J., Vu, B.H., Postma, D., Hickerson, D., McGaughran, J., Khuat, H.Q., 2024. Genai detection tools, adversarial techniques and implications for inclusivity in higher education. arXiv:2403.19148.

Pinhasov, B., Lapid, R., Ohayon, R., Sipper, M., Aperstein, Y., 2024. Xai-based detection of adversarial attacks on deepfake detectors. arXiv:2403.02955.

Piva, A., 2013. An overview on image forensics. In: ISRN Signal Processing 2013, p. 496701.

Pontorno, O., Guarnera, L., Battiato, S., 2024. On the exploitation of DCT-traces in the generative-AI domain. Preprint. arXiv:2402.02209.

Prillaman, M., 2023. 'ChatGPT detector' catches AI-generated papers with unprecedented accuracy. Nature. https://doi.org/10.1038/d41586-023-03479-4.

Rabowsky, B., 2023. Applications of generative AI to media. SMPTE Motion Imaging Journal 132, 53–57. https://doi.org/10.5594/JMI.2023.3297238.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. Preprint. arXiv:2204.06125.

Reznicek, M., Ruth, R.M., Schilens, D.M., 2010. Ace-v and the scientific method. J. Forensic Identif. 60, 87.

Roese, N., Vohs, K., 2012. Hindsight bias. Perspect. Psychol. Sci. 7, 411–426. https://doi.org/10.1177/1745691612454303.

Scott, I.A., Soon, J., Elshaug, A.G., Lindner, R., 2017. Countering cognitive biases in minimising low value care. Med. J. Aust. 206, 407–411. https://doi.org/10.5694/mja16.00999.

Servick, K., 2015. Forensic labs explore blind testing to prevent errors: evidence examiners get practical about fighting cognitive bias. Science 349, 462–463. https://doi.org/10.1126/science.349.6247.462.

Sha, Z., Li, Z., Yu, N., Zhang, Y., 2023. De-fake: detection and attribution of fake images generated by text-to-image generation models. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, pp. 3418–3432.

Singh, S., Kumar, R., 2024. Image forgery detection: comprehensive review of digital forensics approaches. J. Comput. Soc. Sci. https://doi.org/10.1007/s42001-024-00265-8.

Solanke, A.A., Biasiotti, M.A., 2022. Digital forensics ai: evaluating, standardizing and optimizing digital evidence mining techniques. Künstl. Intell. 36, 143–161. https://doi.org/10.1007/s13218-022-00763-9.

Soltani, S., Nikou, S., 2020. An assessment of academic library services: international and domestic students perspectives. Libr. Manage. 41, 631–653. https://doi.org/10.1108/LM-04-2020-0071.

Stevenage, S.V., Bennett, A., 2017. A biased opinion: demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results. Forensic Sci. Int. 276, 93–106.

Stoel, R., Dror, I., Miller, L., 2014. Bias among forensic document examiners: still a need for procedural changes. Aust. J. Forensic Sci. 46, 91–97. https://doi.org/10.1080/00450618.2013.797026.

Stokel-Walker, C., Van Noorden, R., 2023. What ChatGPT and generative AI mean for science. Nature 614, 214–216. https://doi.org/10.1038/d41586-023-00340-6.

Suciu, P., 2022. Is Russia's invasion of Ukraine the first social media war? https://www.forbes.com/sites/petersuciu/2022/03/01/is-russias-invasion-of-ukraine-the-first-social-media-war/. (Accessed 13 December 2023).

Sunde, N., Dror, I.E., 2019. Cognitive and human factors in digital forensics: problems, challenges, and the way forward. Digit. Investig. 29, 101–108.

Sunde, N., Dror, I.E., 2021. A hierarchy of expert performance (hep) applied to digital forensics: reliability and biasability in digital forensics decision making. Forensic Science International: Digital Investigation 37, 301175. https://doi.org/10.1016/j.fsidi.2021.301175.

Tenopir, C., Rice, N.M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., Sandusky, R.J., 2020. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS ONE 15, 1–26. https://doi.org/10.1371/journal.pone.0229003.

Thakur, V.N., Basso, M.A., Ditterich, J., Knowlton, B.J., 2021. Implicit and explicit learning of Bayesian priors differently impacts bias during perceptual decision-making. Sci. Rep. 11, 16932. https://doi.org/10.1038/s41598-021-95833-7.

Thompson, W., Newman, E.J., 2015. Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. Law Hum. Behav. 39 (4), 332–349. https://doi.org/10.1037/lhb0000134.

Verdoliva, L., 2020. Media forensics and deepfakes: an overview. IEEE J. Sel. Top. Signal Process. 14, 910–932.

Wang, H., Yang, Z., 2018. Face pareidolia and its neural mechanism. Advances in Psychological Science 26, 1952.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.G., Li, S.N., 2022. M2tr: multimodal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval, pp. 615–623.

Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y., 2021. Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 3444–3451.

Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A., 2020. Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695–8704.

Wang, Y., Ming, J., Jia, X., Elder, J.H., Lu, H., 2023. Blind image super-resolution with degradation-aware adaptation. In: Computer Vision – ACCV 2022. Springer Nature, Switzerland, Cham, pp. 69–85.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L., 2023. Testing of detection tools for ai-generated text. International Journal for Educational Integrity 19. https://doi.org/10.1007/s40979-023-00146-z.

Wodajo, D., Atnafu, S., 2021. Deepfake video detection using convolutional vision transformer. CoRR. arXiv:2102.11126 [abs].

Yu, N., Davis, L.S., Fritz, M., 2019. Attributing fake images to GANs: learning and analyzing gan fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7556–7566.

Yu, N., Skripniuk, V., Abdelnabi, S., Fritz, M., 2021. Artificial fingerprinting for generative models: rooting deepfake attribution in training data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14448–14457.

Yu, N., Skripniuk, V., Chen, D., Davis, L., Fritz, M., 2020. Responsible disclosure of generative models using scalable fingerprinting. Preprint. arXiv:2012.08726.

Zanardelli, M., Guerrini, F., Leonardi, R., Adami, N., 2023. Image forgery detection: a survey of recent deep-learning approaches. Multimed. Tools Appl. 82, 17521–17566. https://doi.org/10.1007/s11042-022-13797-w.

Zhang, X., Karaman, S., Chang, S.F., 2019. Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, pp. 1–6.

Zhou, L.F., Meng, M., 2020. Do you see the "face"? Individual differences in face pareidolia. Journal of Pacific Rim Psychology 14. https://doi.org/10.1017/prp.2019.27.