

# AI Mirage: The Impostor Bias and the Deepfake Detection Challenge in the Era of Artificial Illusions

Mirko Casu<sup>a,b,\*</sup>, Luca Guarnera<sup>a</sup>, Pasquale Caponnetto<sup>b</sup>, Sebastiano Battiato<sup>a</sup>

<sup>a</sup>*Department of Mathematics and Computer Science, University of Catania, Viale Andrea Doria 6, Catania, 95126, CT, Italy*

<sup>b</sup>*Department of Educational Sciences, Section of Psychology, University of Catania, Via Teatro Greco 84, Catania, 95124, CT, Italy*

---

## Abstract

This paper provides a comprehensive analysis of cognitive biases in forensics and digital forensics, examining their implications for decision-making processes in these fields. It explores the various types of cognitive biases that may arise during forensic investigations and digital forensic analyses, such as confirmation bias, expectation bias, overconfidence in errors, contextual bias, and attributional biases. It also evaluates existing methods and techniques used to mitigate cognitive biases in these contexts, assessing the effectiveness of interventions aimed at reducing biases and improving decision-making outcomes. Additionally, this paper introduces a new cognitive bias, called “impostor bias”, that may affect the use of generative Artificial Intelligence (AI) tools in forensics and digital forensics. The impostor bias is the tendency to doubt the authenticity or validity of the output generated by AI tools, such as deepfakes, in the form of audio, images, and videos. This bias may lead to erroneous judgments or false accusations, undermining the reliability and credibility of forensic evidence. The paper discusses the potential causes and consequences of the impostor bias, and suggests some strategies to prevent or counteract it. By addressing these topics, this paper seeks to offer valuable insights into understanding cognitive biases in forensic practices and provide recommendations for future research and practical applications to enhance the objectivity and validity of forensic investigations.

---

\*mirko.casu@phd.unict.it; Department of Mathematics and Computer Science, University of Catania, Viale Andrea Doria 6, Catania, 95126, CT, Italy

*Keywords:* Forensic Sciences, Cognitive Biases, Cognitive Psychology,  
Digital Forensics, Synthetic Data, Impostor Bias, Generative AI, GAN,  
Diffusion Models, Deepfake Detection  
*PACS:* 0000, 1111  
*2000 MSC:* 0000, 1111

---

## 1. Introduction

In the realm of forensic sciences, perhaps more than in other disciplines, the so-called objectivity of judgment in analyzing data and interpreting them for justice purposes is one of the crucial aspects. Beyond the specific technical aspects of individual disciplines, the authoritativeness of studies, and the competence of experts, there is a need to be informed about the so-called cognitive biases. Cognitive biases, rather than being deficits, are systematic preferences that influence the way we process, select, and retain information (Grisham et al., 2014; Lester et al., 2011). These biases can significantly impact our judgment and decision-making processes.

The tendency to *jump to conclusions* involves making hasty judgments or drawing premature conclusions without adequate evidence (Warman et al., 2013). *Confirmation bias* is another common cognitive bias where individuals selectively pay attention to information that aligns with their pre-existing beliefs, often disregarding contradictory evidence (Peters, 2020). *Expectation bias*, also known as experimenter’s bias, occurs when an individual’s anticipated outcome influences the actual outcome observed (Forensic Science Regulator, 2020). *Overconfidence in errors* is a related bias where individuals exhibit unwarranted certainty in their judgments and predictions, even when proven incorrect (Han, 2020; Sauvé et al., 2020). *Contextual bias* refers to situations where an individual’s judgment is influenced, either consciously or subconsciously, by additional information beyond what is being directly evaluated (Forensic Science Regulator, 2020). *Attributional biases* involve the unfair attribution of external events or experiences to oneself or others, thereby influencing our perception of causality (Sauvé et al., 2020). Cognitive biases can have both positive and negative effects, depending on the context and situation. They can facilitate swift decision-making when time is critical but can also lead to poor decisions and adverse outcomes (Berthet, 2022). Cognitive biases have been found to significantly influence decision-making processes in various professional fields such as management, finance,

medicine, law, and psychology (Acciarini et al., 2021; Berthet, 2022; Neal and Grisso, 2014). Overconfidence bias, in particular, is a recurring bias that impacts decisions in these areas (Grežo, 2021; Wu et al., 2012). A summary of these biases can be found in Table 1.

Bias	Description
Jump to Conclusions	Making quick judgments or drawing premature conclusions without sufficient evidence.
Confirmation Bias	Selectively paying attention to information that confirms our pre-existing beliefs while ignoring contradictory evidence.
Expectation Bias	An individual’s anticipated outcome influences the actual outcome that is observed.
Overconfidence in Errors	Individuals display an unwarranted certainty in their judgments and predictions, even when they are proven wrong.
Contextual Bias	An individual’s judgment is influenced (either consciously or subconsciously) by additional information beyond what is being directly evaluated.
Attributional Biases	Unfairly attributing external events or experiences to oneself or others, influencing our perception of causality.

Table 1: Summary of Cognitive Biases.

This paper delves into the influence of cognitive biases in the realm of forensics and digital forensics, with a particular focus on the detection of deepfakes. Deepfakes, which are Artificial Intelligence (AI)-generated images, videos, and audio, have a substantial impact on these fields. They can shape attitudes even when viewers are cognizant of the fact that they are viewing deepfakes, a phenomenon attributable to cognitive biases. Deepfakes pose threats such as the manipulation of public opinion and the impersonation of individuals, which could potentially lead to financial and reputational damage. In this context, we will introduce the concept of the *impostor bias*, which we define as an inherent distrust concerning the authenticity of multimedia elements like videos, photos, and audio. This distrust stems from the awareness that these elements can be realistically generated by AI mod-

els. Effective detection of deepfake products is crucial to prevent falling prey to the impostor bias and to avoid the potential confusion between real and fake multimedia products. Such confusion can lead to erroneous decisions and perceptions across all layers of multimedia evaluation in digital forensics. The challenges presented by deepfakes necessitate ongoing research and the development of advanced solutions. Consequently, we will analyze the most recent and efficient deepfake detection systems to aid operators and investigators in distinguishing between real and fake multimedia.

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

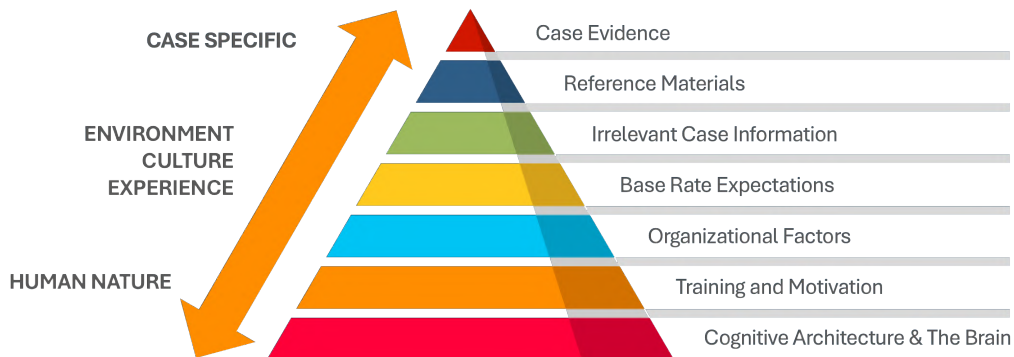


Figure 1: Seven potential sources of bias that could impact forensic decision-making.

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 put forth a paper 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, including confirmation bias (Thompson and Newman, 2015; Meterko and Cooper, 2022). The sources of bias can be categorized into three groups related to the case, the analyst, and human nature (Dror, 2020). Seven primary factors can introduce bias within the laboratory setting (Jeanguenat et al., 2017) (As synthetically sketched in Figure 1).

Gardner et al. (2019) found that task-irrelevant information can bias forensic analysts' decisions. Similarly, Nakhaeizadeh et al. (2014) discovered that irrelevant information can lead to confirmation bias in forensic anthro-

pology. Dror et al. (2021) found that base-rate neglect could bias forensic pathologists’ decisions in child death cases. In DNA forensics, Jeanguenat et al. (2017) found that suspect-driven bias and rare alleles can influence interpretation. Neal et al. (2022) conducted a systematic review on cognitive biases and debiasing techniques in forensic mental health, finding significant bias effects. Lastly, Stevenage and Bennett (2017) found that irrelevant DNA test outcomes could bias fingerprint matching tasks, confirming the presence of contextual bias.

### **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; Fabio Arceri et al., 2023; Giudice et al., 2017; Piva, 2013). In the realm of digital forensics, cognitive biases can significantly impact the interpretation of data.

#### *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 (Figure 2) (Fontani, 2021). In this scenario, John inadvertently influenced Lucy’s interpretation of a license plate by prematurely sharing his own interpretation. This phenomenon, known as confirmation bias, may lead Lucy to unconsciously process pixels and select frames that align with John’s interpretation. The post further delved 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

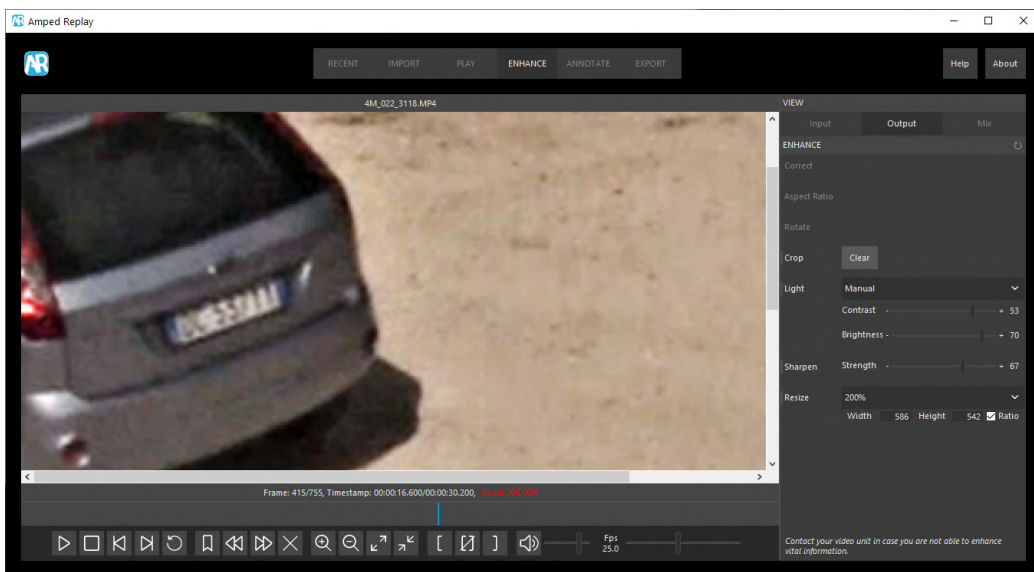


Figure 2: John asked Lucy for help with a license plate, reading “BC 537”, but unsure about the last two characters and the first one. Credit: Fontani (2021).

alter facial appearances during enhancement (Figure 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. This unconscious adjustment is a form of cognitive bias.

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 (Figure 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.

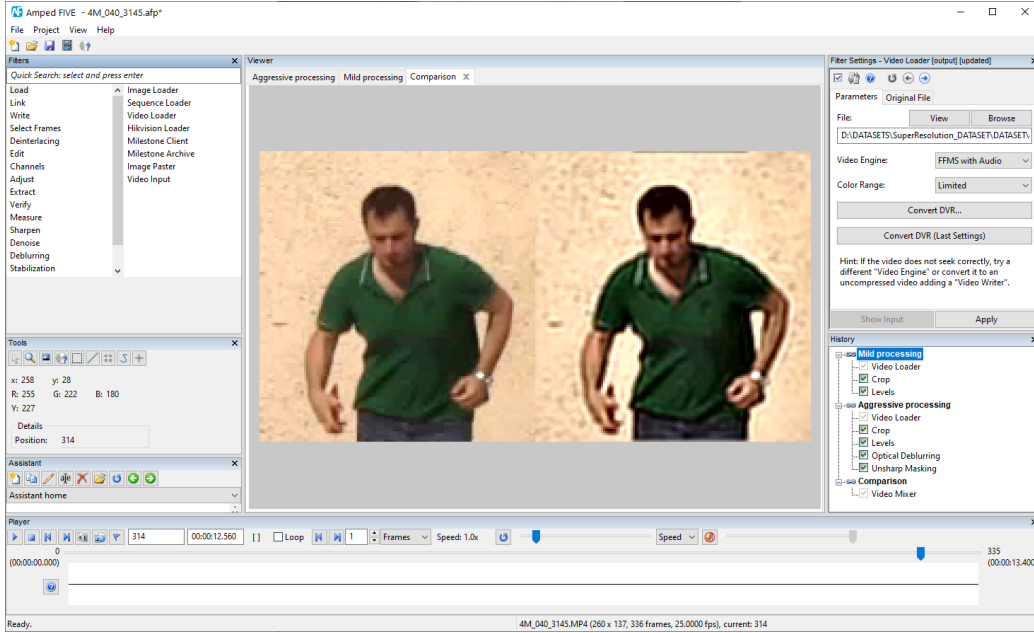


Figure 3: Image processing could lead to produce a noticeable different image, with different face characteristics. Credit: Fontani (2021).

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 (Figure 5).

The enhancement of images extracted from video cameras can lead to a degradation of the overall quality of information. This is due to factors such as excessive compression, distance from the recording plane, and limited overall resolution. International best practices suggest verifying on a case-by-case basis whether the level of information is sufficient to extract useful data for investigations (Khin et al., 2020). 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, 2020a). Pareidolia is a subconscious illusion



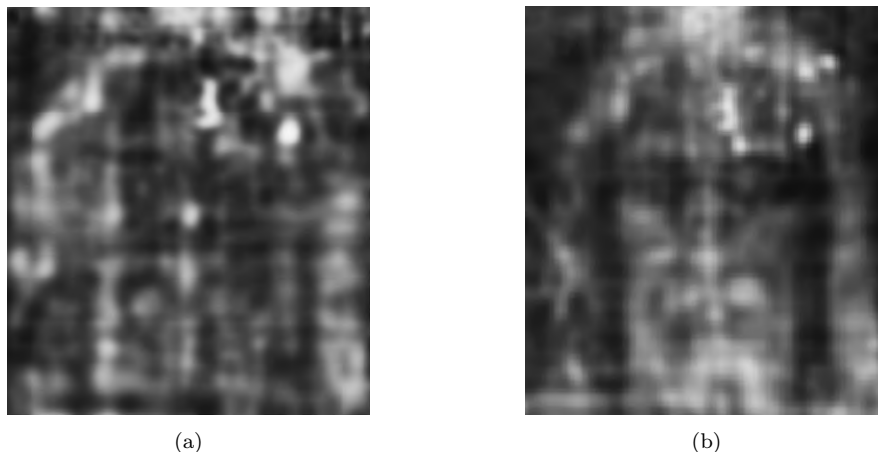


Figure 4: A supposed concealed image of a face on the back side of the Shroud is revealed through advanced image processing of a photograph published in a book. The image is flipped from right to left (b). A negative image of the face that can be seen on the front side of the Shroud, processed in the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

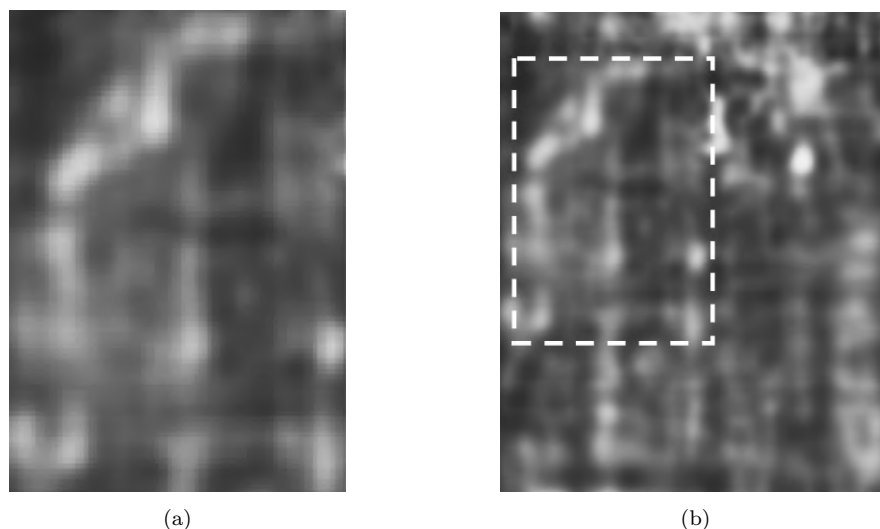


Figure 5: A magnified version of Figure 4b (a). A face resembling the Shroud that we discern in the top-left section of Figure 4b (as depicted on the right) (b). We can also discern another face in the bottom left section of Figure 4b. Pareidolia leads to false positives, enabling us to see faces in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

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 or experts who can follow a “blind testing” approach, i.e., an interpretation detached from the knowledge of details and the reference context. This helps to ensure the search for “certain” evidence is as rigorous and unbiased as possible.

Zhou and Meng (2020b) 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.

Furthermore, we present a case study which involved the analysis of surveillance camera footage to determine the presence of a passenger in a vehicle involved in a legal case. The analysis included scrutiny of various cameras and raised concerns about the validity of conclusions drawn by experts. The frames are part of a detailed analysis of the vehicle’s passages from different surveillance cameras, considering image overlaps and real passage times, including sunset. The aim was to evaluate the methodologies and conclusions of technical consultants and expert witnesses regarding the analysis of vehicle images, highlighting the lack of scientific rigor and proposing new analyses suggesting the absence of a passenger in the vehicle. Specifically, Figure 6 shows a sequence of frames that have been processed by a technical consultant. The frames demonstrate a clear reflection effect that disappears as the vehicle moves. The effect is attributed to the ease with which contour lines can introduce aberrations. The frames also exhibit strong chromatic variability due to video compression, which is known to heavily penalize the color component.

The technical advisor claimed to see the image of human faces in such images, a factor induced by confirmation bias: in reality, what is present in



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

the images are pseudo-random blobs and their temporal hold is instead to be traced to a simpler and more obvious reflection. To demonstrate this, we conducted an experiment with college students. In the experiment, students were presented with frames from surveillance camera footage and tasked with identifying and interpreting the images. Crucially, no prior information was provided to the students to prevent any preconceived notions or confirmation biases from influencing their observations. The primary focus was on the students’ ability to discern whether a vehicle contained a passenger or not. The experiment was designed to rigorously assess the students’ cognitive perception, comprehension of the images, and their capacity to make accurate identifications solely based on the visual data at hand. This approach underscores the importance of maintaining an unbiased perspective in observational tasks. Table 2 summarizes the outcome of this analysis process, which reports a number of responses of 165 (15 subjects multiplied by the number of images viewed) in which in no case was the ”certain” presence of a human face evidenced.

#### 4. Bias Mitigating Strategies

Mitigating these cognitive biases could significantly enhance decision-making across society, promoting long-term human well-being. Extensive research has been conducted on whether and how these biases can be mitigated (Acciarini et al., 2021).

Korteling et al. (2021) conducted a systematic review of the literature on



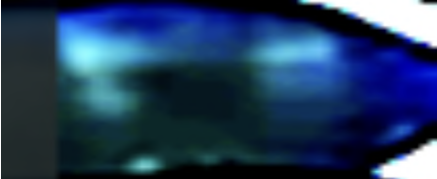



Image	Students' Evaluation
	Nothing: 8 Other: 7 Those who answered "other" identified: • Automobile elements
	Nothing: 9 Other: 6 Those who answered "other" identified: • Automobile elements
	Nothing: 10 Other: 5 Those who answered "other" identified: • Silhouette with sun reflection • Automobile elements
	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

Table 2: Table representing students' evaluation of the various images cutted from Figure 6.

the retention and transfer of bias mitigation interventions, which yielded 12 peer-reviewed studies that adequately studied retention over a period of at least 14 days or transfer to different tasks and contexts. Most of these studies indicated that game-based interventions were effective after the retention interval and were more effective than video interventions. One study found indications of transfer of bias mitigation training across contexts.

#### *4.1. Biases Mitigating Strategies in Forensics*

Dror (2013) focused on strategies to mitigate these biases. He 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. He suggested measures such as a triage approach and case managers to enhance the quality and effectiveness of the work carried out by forensic examiners.

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 (Forensic Science Regulator, 2020). One of these strategies is the adoption of a structured and validated approach for the analysis and evaluation of evidence. Two models that exemplify this approach are the ACE-V model and the CAI model. The ACE-V model, which stands for Analysis, Comparison, Evaluation, and Verification, is a widely accepted method for comparing fingerprints (Reznicek et al., 2010). On the other hand, the CAI model, which stands for Case Assessment and Interpretation, is a framework based on Bayesian thinking (Jackson, 2011). It provides clarity on the role of forensic scientists within the criminal justice process and allows for the design of effective, efficient, and robust examination strategies. Documentation, independent verification, and maintaining ethical standards are other key strategies in the forensic process. Documentation involves taking contemporary notes and recording observations, decisions, and changes, ensuring transparency and enabling a thorough review. Independent verification of critical results by a competent peer, unaware of the original examiner’s judgment, adds an extra layer of scrutiny for accuracy. Furthermore, forensic operators must uphold ethical standards such as honesty, integrity, objectivity, and impartiality, and avoid conflicts of interest and external pressures to ensure unbiased and reliable analysis (Forensic Science Regulator, 2020). Lastly, controlling the flow of information to the

analyst is paramount. This is achieved using methods such as “sequential unmasking” or “linear unmasking”, which help to prevent unnecessary information from influencing the analyst’s judgment. In this regard, Dror and Kukucka (2021) introduced Linear Sequential Unmasking–Expanded (LSU-E), a methodology aimed at improving decision-making in forensic science by reducing noise and bias. The approach advocates for an initial analysis of raw data or evidence without any reference material or context, followed by a sequential consideration of other relevant information based on its objectivity, relevance, and potential for bias. LSU-E is applicable to all types of decisions within forensic science, not just comparative ones. Its main objective is to optimize the sequence of information exposure to maximize information utility and minimize cognitive and psychological influences. In addition, LSU-E provides guidelines for determining the optimal sequence for presenting task-relevant information and emphasizes the need for transparent documentation of the influence and role of various pieces of information on the decision-making process.

Camilleri et al. (2019) proposed that while some laboratories have implemented bias mitigating strategies with varying impact on operational efficiency, there has been no systematic assessment of the risk posed by cognitive bias. They suggested using a risk management framework to assess the potential impact of bias on forensic interpretations across multiple disciplines. This approach proved useful in assessing the effectiveness of existing bias mitigating strategies and identifying the latent level of risk posed. Furthermore, the authors affirm that all forensic organizations should seek to implement bias limiting measures that are simple, cost-effective, and do not adversely impact efficiency. Recently, Douglass et al. (2023) provided an examination of cognitive biases in digital forensics, particularly focusing on the potential of video-recorded eyewitness identification procedures for assisting evaluators in assessing eyewitness accuracy. The authors conducted three studies involving a total of 1,630 participants. The results revealed that evaluators, on average, successfully differentiated accurate from inaccurate witnesses based on videos of identification procedures alone. However, this ability was compromised when extraneous incriminating evidence was also provided. The authors found that instructions highlighting the limitations of forensic evidence did not preserve evaluators’ ability to discern accuracy when extraneous incriminating case evidence was provided. Moreover, case information affected other judgments, such as perceptions of the witness’s view.

#### *4.2. Biases Mitigating Strategies in Digital Forensics*

Grubl and Lallie (2022) addressed one of the most significant challenges in digital forensics: the precise age estimation of victims in child sexual abuse and exploitation cases. Investigators often need to determine the age of victims by examining images and interpreting the sexual development stages and other human characteristics. However, this process can be negatively impacted by a large forensic backlog, cognitive bias, and the immense psychological stress that this work can entail. In response to these challenges, the authors evaluated existing facial image datasets and proposed a new dataset tailored to the needs of similar digital forensic research contributions. This dataset, which included 327 images of individuals aged 0 to 20, is tested on the Deep EXpectation algorithm pre-trained on the IMDB-WIKI dataset (Grubl and Lallie, 2022). The results for young adolescents aged 10 to 15 and older adolescents/adults aged 16 to 20 are very encouraging, achieving Mean Absolute Errors as low as 1.79.

Karie et al. (2019) also delved into the role of cognitive biases in digital forensics, particularly in the context of combating cybercrimes. The authors acknowledged that the increasing complexity of cyber-attacks and the vast amounts of data, also known as Big Data, that investigators need to sift through to unveil Potential Digital Evidence have made the process of forensic investigation more challenging. In response to these challenges, the authors proposed a generic framework for diverging DL cognitive computing techniques into Cyber Forensics (CF), referred to as the DLCF Framework. Such solutions can range from reducing bias in forensic investigations to challenging what evidence is considered admissible in a court of law or any civil hearing.

### **5. The Art and Science of Deepfakes**

The advent of advanced AI technologies has introduced new challenges in this field. One such challenge is the detection of deepfakes, the research area of which is constantly expanding (as shown in Figure 7). Deepfakes are synthetic media created through generative models based mainly on Generative Adversarial Networks (GANs) and Diffusion Models (DMs) (Goodfellow et al., 2014; Sohl-Dickstein et al., 2015). 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

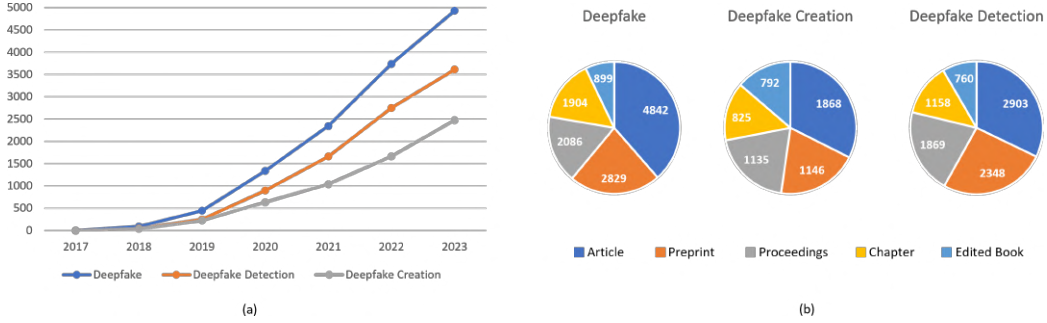


Figure 7: Statistics of papers published in the deepfake field. (a) Papers published from 2017 to 2023 with the keywords deepfake, deepfake creation, deepfake detection. (b) Numbers of papers published in: Article, Preprint, Proceedings, Chapter and Edited Book.

training set  $T_s$ . The Discriminator is trained to distinguish the images created by  $G$  from the set  $T_s$ . When  $G$  creates images with the same data distribution as  $T_s$ ,  $D$  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 (Sohl-Dickstein et al., 2015) 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. Figure 8 and Figure 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). In this context, preventing 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 Generative AI (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).



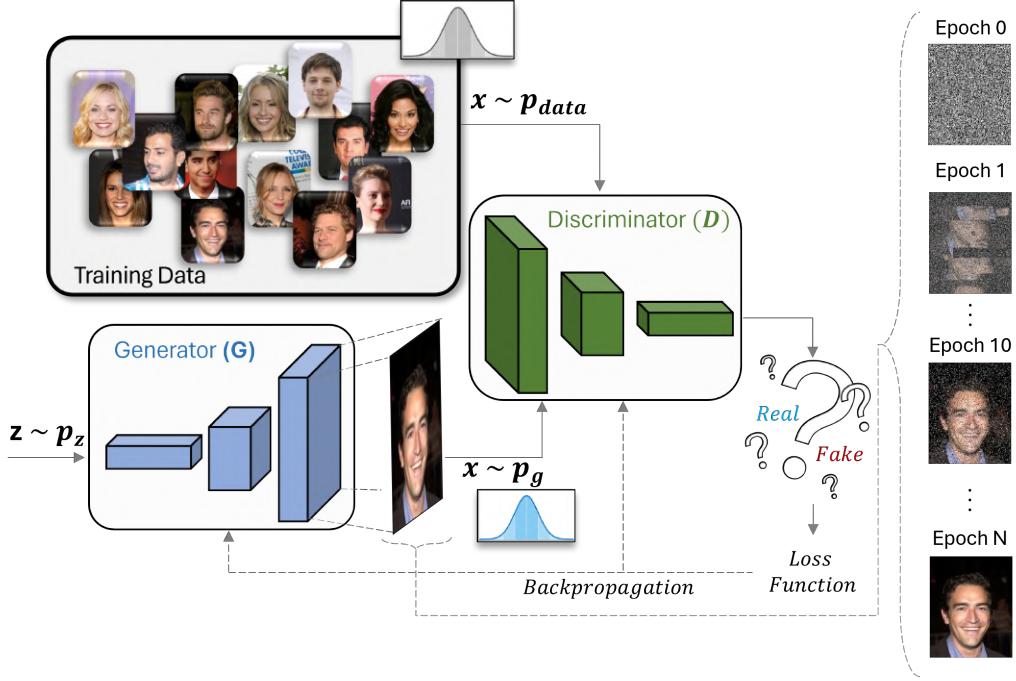


Figure 8: A standard GAN framework including a Generator ( $G$ ) that creates data samples from random noise, aiming to mimic the training set's data distribution; a Discriminator ( $D$ ) differentiates between real and  $G$ -generated data. Both are trained competitively.

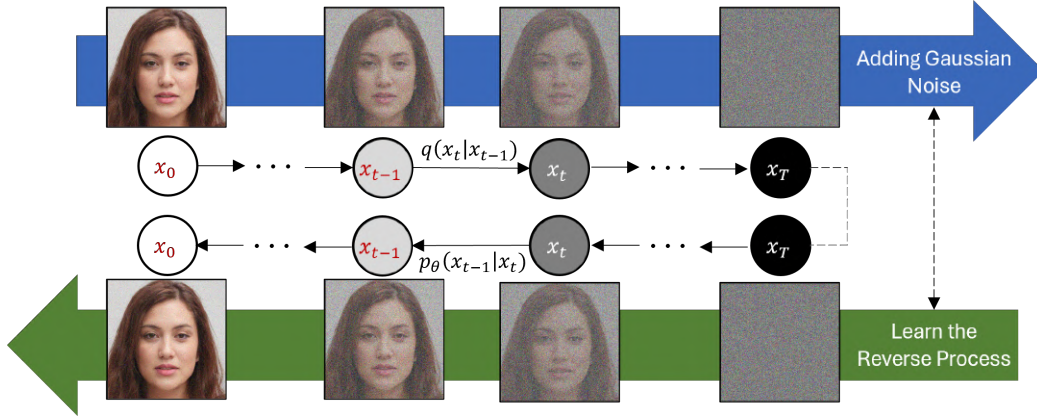


Figure 9: A generic Diffusion Model learning a latent variable model that maps the latent space with a fixed Markov chain. The data is corrupted by adding Gaussian noise gradually to get the approximate posterior  $q(x_t|x_{t-1})$ . The aim of training is to learn the inverse process ( $p_\theta(x_{t-1}|x_t)$ ) so that new data can be generated by running the chain backward. The latent variables  $x_1, \dots, x_t$  have the same dimension as the data  $x_0$ .



Figure 10: Do these people exist? Credit: Guarnera et al. (2022)

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 (Rabowsky, 2023). 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 (Figure 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 refers to an a priori distrust regarding the veracity of multimedia elements such as videos, photos, and audio, due to the awareness that these can be realistically generated by Artificial Intelligence models. Being subject to this bias, a condition towards which we believe we are increasingly heading in parallel with the increase in realism of GenAI multimedia products, entails and will entail a constant doubt that will also be influential when the victims are, for example, court operators, investigators, or magistrates themselves. In this regard, the technologies previously analyzed regarding deepfake detection will increasingly have to orient themselves not



Figure 11: Real vs deepfake images of famous artists.

only towards the recognition of manipulated or altered multimedia artifacts, but also those generated from scratch starting from descriptive textual inputs. This is a great challenge on which we will have to work to avoid falling into cognitive traps dictated by prejudices of technological suspicion, risking declaring or considering original multimedia products instead generated by AI, and vice versa. Another risk stemming from GenAI pertains to the counterfeiting of artworks, the sale of forged masterpieces, and the infringement of copyright laws. 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) (Figure 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. Deepfake Detection Methods

Researchers have demonstrated that generative engines leave traces on synthetic content that can be identified and detected in the frequency do-



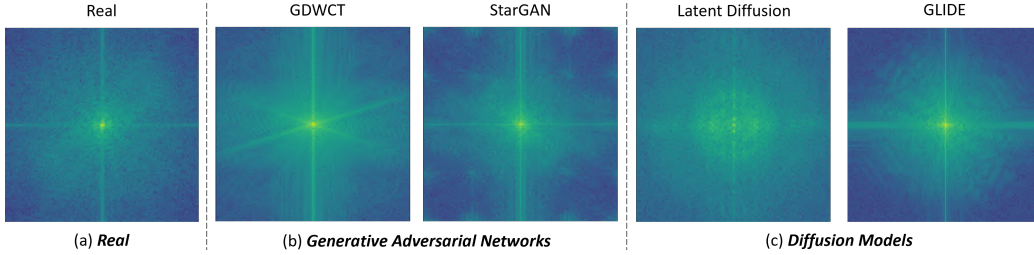


Figure 12: Fourier spectrum of different categories of data: (a) real images; (b) images generated by two GAN architectures (GDWCT (Cho et al., 2019) and StarGAN (Choi et al., 2018)); (c) images generated by two DM architectures (Latent Diffusion (Ramesh et al., 2022) and GLIDE (Nichol et al., 2022)).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

main (Guarnera et al., 2020c; Zhang et al., 2019; Marra et al., 2019; Giudice et al., 2021; Dzanic et al., 2020; Durall et al., 2020) (Figure 12). These traces are characterized by both the network architecture (number and type of layers) and its specific parameters (Yu et al., 2019a). 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., 2019b, 2021; Girish et al., 2021; Asnani et al., 2023; Yu et al., 2020; Guarnera et al., 2022) 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

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)

Table 3: A summary of the discussed deepfake detection methods. Legend: AUC = Area Under the Curve; TAR = Transfer learning-based Autoencoder with Residuals; ViT = Vision Transformer.

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 3.

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 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

We highlighted the importance of deepfakes in the field of digital forensics, which we consider to be a topic of great relevance and timeliness: in particular, the concept of impostor bias that we introduced is increasingly prevalent, although it had not yet been named and defined precisely. A concrete example of impostor bias was seen with the recent war in Ukraine, the first war characterized by constant propaganda from both sides carried out on all media, and especially on social networks (Ciuriak, 2022; Suciu, 2022). Being flooded with videos, photos and statements inevitably leads to believe a priori that everything that is seen can be modified, or generated by AI models through a simple prompt (Linehan et al., 2023). The habit of suspicion developed during this “digital warfare” with propaganda and false information led to immediate doubt of all published media, sometimes leading to the discovery that some of this media was deepfake (Bond, 2023).

This condition of prior doubt on the veracity of the media will be more and more invasive and constant, especially when it comes to delicate material, such as a communication from a head of state in wartime (Linehan et al., 2023): for this reason, we believe that it is necessary to use the best tools for detecting deepfakes and AI-generated multimedia to avoid being deceived, especially to avoid considering fake multimedia that, paradoxically, are instead real.

Therefore, the data presented herein, along with previous findings, underscore the imperative need for further exploration of cognitive biases in digital forensic science. The focus should not be solely on the biases themselves, but rather on the strategies that can be employed to counteract and disengage from them. One potential approach could involve the creation and implementation of specialized training programs for forensic practitioners. These programs would aim to address cognitive biases in various scenarios, such as those examined in this review. By enhancing awareness of cognitive biases, we could potentially reduce the likelihood of practitioners being influenced or overwhelmed by these biases. This would enable them to disregard irrelevant elements of the investigation, separate personal experiences and other contextual factors that may “contaminate” decision-making processes, thereby preserving the objectivity of their judgments.

## 9. Conclusions and Future Works

This article explores cognitive biases in forensics and digital forensics, including confirmation bias, expectation bias, and overconfidence in errors. It discusses methods to reduce these biases, such as game-based interventions and the Linear Sequential Unmasking-Expanded approach. The article highlights the challenge of deepfakes, synthetic media that can manipulate or impersonate individuals, threatening the integrity of digital evidence. It surveys current deepfake detection methods and their limitations. The article introduces the impostor bias, a new cognitive bias that may lead to false negatives and reduced confidence in digital forensic findings due to the prevalence of deepfakes. We conclude by proposing new research directions in digital forensics, suggesting the creation of advanced deepfake detection methods, exploring the factors contributing to impostor bias, and recommending interventions to mitigate this bias. We also underscore the need to examine the ethical, legal, and social implications of deepfakes.

## 10. Acknowledgments

This research is supported by Azione IV.4 - “Dottorati e contratti di ricerca su tematiche dell’innovazione” del nuovo Asse IV del PON Ricerca e Innovazione 2014-2020 “Istruzione e ricerca per il recupero - REACT-EU”-CUP: E65F21002580005.

## References

- Acciarini, C., Brunetta, F., Boccardelli, P., 2021. Cognitive biases and decision-making strategies in times of change: A systematic literature review. *Management Decision* 59, 638–652.
- Asnani, V., Yin, X., Hassner, T., Liu, X., 2023. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .
- 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*, pp. 5–16.

- Berthet, V., 2022. The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology* 12, 802439.
- Bhadra, P., 2021. Is forensic evidence impartial? cognitive biases in forensic analysis. *Criminal Psychology and the Criminal Justice System in India and Beyond* , 215–227.
- Bond, S., 2023. How Russia is losing — and winning — the information war in Ukraine. NPR Accessed: 2023-12-13.
- 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. doi:10.1016/J.SCIJUS.2019.04.003.
- 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: 2023-12-13.
- 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 Science International* 297, 35–46.
- 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.
- Di Lazzaro, P., Murra, D., Schwartz, B., 2013. Pattern recognition after image processing of low-contrast images, the case of the shroud of turin. *Pattern Recognition* 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. doi: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. *Journal of Forensic Sciences* 66, 1751–1757.
- Dror, I., Rosenthal, R., 2008. Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences* 53, 900–903.
- Dror, I.E., 2020. Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry* 92, 7998–8004.
- 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.

- 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. doi:10.1109/CVPR42600.2020.00791.
- 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.
- 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. arXiv preprint arXiv:2306.04141 .
- Fabio Arceri, N., Giudice, O., Battiato, S., 2023. An innovative tool for uploading/scraping large image datasets on social networks. arXiv e-prints , arXiv–2311.
- Fanti, G., Maggiolo, R., 2004. The double superficiality of the frontal image of the turin shroud. Journal of Optics A: Pure and Applied Optics 6, 491.
- Fontani, M., 2021. Cognitive bias: Steering conclusions irrationally. <https://blog.ampedsoftware.com/2021/04/20/cognitive-bias-steering-conclusions-irrationally>. Accessed: 2023-11-07.
- 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: 2023-11-07.
- 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 Science International 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.

- Giudice, O., Guarnera, L., Battiato, S., 2021. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging* 7, 128. doi: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.
- 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.
- Grežo, M., 2021. Overconfidence and financial decision-making: A meta-analysis. *Review of Behavioral Finance* doi:10.1108/rbf-01-2020-0020.
- 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. *Cognitive Therapy and Research* 38, 505–517.
- Grubl, T., Lallie, H.S., 2022. Applying artificial intelligence for age estimation in digital forensic investigations. *arXiv:2201.03045*.
- 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. *arXiv preprint arXiv:2303.00608*.
- 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. doi:10.23919/AEIT50178.2020.9241108.
- 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. doi:10.1109/CVPRW56347.2022.00016.

- Han, W.H., 2020. Consumers' overconfidence biases in relation to social exclusion. *The Journal of Asian finance, economics, and business* 7, 303–308.
- 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. *Applied Intelligence* 53, 7512–7527.
- 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. doi:10.1016/j.scijus.2017.07.005.
- Karie, N.M., Kebande, V.R., Venter, H., 2019. Diverging deep learning cognitive computing techniques into cyber forensics. *Forensic Science International: Synergy* 1, 61–67.
- 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.
- Khin, N.A., Francis, G., Mulinde, J., Grandinetti, C., Skeete, R., Yu, B., Ayalew, K., Cho, S., Fisher, A., Kleppinger, C., Ayala, R., Bonapace, C., Dasgupta, A., Kronstein, P.D., Vinter, S., 2020. Data integrity in global clinical trials: Discussions from joint us fda and mhra uk good clinical practice workshop. *Clinical Pharmacology and Therapeutics* doi:10.1002/cpt.1794.
- Korteling, J., Gerritsma, J.Y., Toet, A., 2021. Retention and transfer of cognitive bias mitigation interventions: A systematic literature study. *Frontiers in Psychology* 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. doi: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. *Journal of Behavior Therapy and Experimental Psychiatry* 42, 298–308.
- 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: 2023-12-13.
- 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.
- 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. *Applied Intelligence* 53, 3974–4026.
- Meterko, V., Cooper, G., 2022. Cognitive biases in criminal case evaluation: A review of the research. *Journal of Police and Criminal Psychology* 37, 101–122.

- Moon, T.K., 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* 13, 47–60.
- Nakhaeizadeh, S., Dror, I.E., Morgan, R.M., 2014. Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. *Science & Justice* 54, 208–214.
- Neal, T., Grisso, T., 2014. The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy and Law* 20, 200–211. doi: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 and Human Behavior* doi: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.
- Palmer, G., et al., 2001. A road map for digital forensic research, in: *First digital forensic research workshop, utica, new york*, pp. 27–30.
- Peters, U., 2020. What is the function of confirmation bias? *Erkenntnis* 87, 1351–1376. doi:10.1007/s10670-020-00252-1.
- Piva, A., 2013. An overview on image forensics. *International Scholarly Research Notices* 2013.
- Rabowsky, B., 2023. Applications of generative ai to media. *SMPTE Motion Imaging Journal* 132, 53–57. doi:10.5594/JMI.2023.3297238.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical Text-Conditional Image Generation with Clip Latents. *arXiv preprint arXiv:2204.06125*.
- Reznicek, M., Ruth, R.M., Schilens, D.M., 2010. Ace-v and the scientific method. *Journal of Forensic Identification* 60, 87.

- Sauvé, G., Lavigne, K.M., Pochiet, G., Brodeur, M.B., Lepage, M., 2020. Efficacy of psychological interventions targeting cognitive biases in schizophrenia: A systematic review and meta-analysis. *Clinical Psychology Review* 78, 101854. doi:<https://doi.org/10.1016/j.cpr.2020.101854>.
- 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.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics, in: *International Conference on Machine Learning*, PMLR. pp. 2256–2265.
- 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 Science International* 276, 93–106.
- Stoel, R., Dror, I., Miller, L., 2014. Bias among forensic document examiners: Still a need for procedural changes. *Australian Journal of Forensic Sciences* 46, 91 – 97. doi:10.1080/00450618.2013.797026.
- 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: 2023-12-13.
- Sunde, N., Dror, I.E., 2019. Cognitive and human factors in digital forensics: Problems, challenges, and the way forward. *Digital Investigation* 29, 101–108.
- Thompson, W., Newman, E.J., 2015. Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and human behavior* 39 4, 332–49. doi:10.1037/lhb0000134.
- Verdoliva, L., 2020. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing* 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: Multi-modal 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.
- Warman, D., Martin, J.M., Lysaker, P., 2013. Jumping to conclusions and delusions: The impact of discussion of the bias on the bias. *Schizophrenia Research* 150, 575–579. doi:10.1016/j.schres.2013.09.003.
- Wodajo, D., Atnafu, S., 2021. Deepfake video detection using convolutional vision transformer. CoRR abs/2102.11126. URL: <https://arxiv.org/abs/2102.11126>, arXiv:2102.11126.
- Wu, S.W., Johnson, J.E.V., Sung, M., 2012. Overconfidence in judgements: the evidence, the implications and the limitations. *The Journal of Prediction Markets* 2, 73–90. doi:10.5750/JPM.V2I1.436.
- Yu, N., Davis, L.S., Fritz, M., 2019a. 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., Davis, L.S., Fritz, M., 2019b. 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. arXiv preprint arXiv:2012.08726 .
- 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., 2020a. Do you see the “face”? individual differences in face pareidolia. *Journal of Pacific Rim Psychology* 14. doi:10.1017/prp.2019.27.
- Zhou, L.F., Meng, M., 2020b. Do you see the “face”? individual differences in face pareidolia. *Journal of Pacific Rim Psychology* 14, e2.