GENAI MIRAGE: THE IMPOSTOR BIAS AND THE DEEPFAKE DETECTION CHALLENGE IN THE ERA OF ARTIFICIAL ILLUSIONS

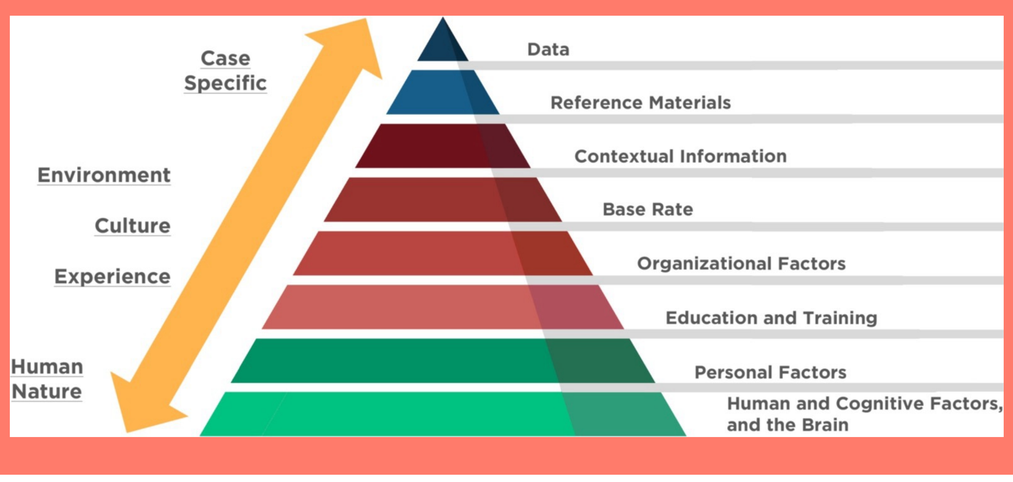
Casu M.[1], Guarnera L.[1], Caponnetto P.[1], Battiato S.[1]
University of Catania [1]
mirko.casu@phd.unict.it, {luca.guarnera, p.caponnetto}@unict.it, battiato@dmi.unict.it

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. Cognitive bias in forensic sciences

Sources of bias fall into three categories: case-related, analyst-related, and human nature. These biases can distort processes like sampling, observation, test strategies, analysis, and conclusions, even when performed by experts.

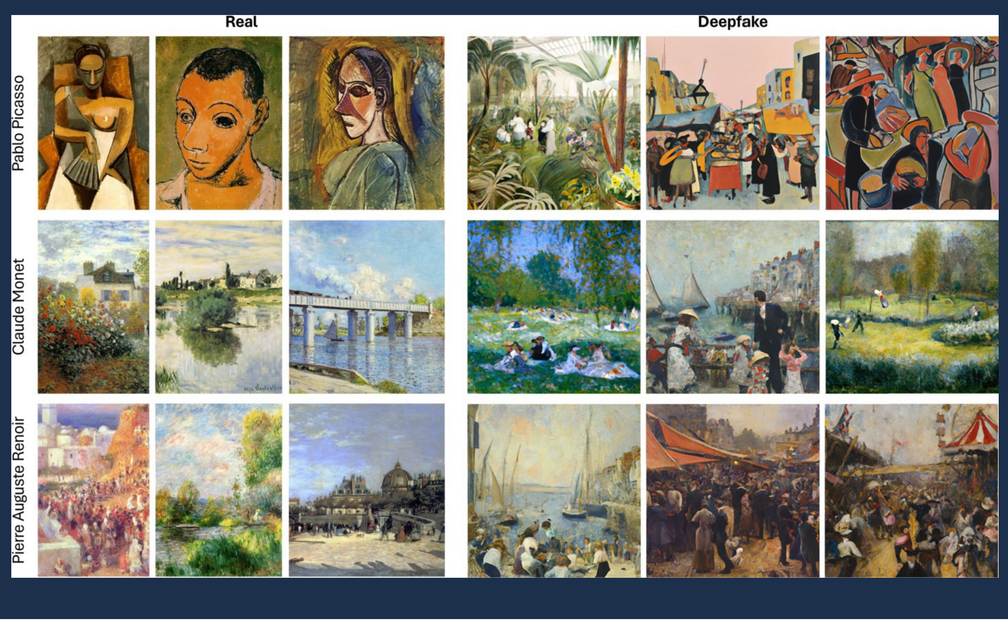


2. Case study: confirmation bias

Surveillance footage was analyzed to determine the presence of a vehicle passenger in a murder case, but findings suggested the absence of a passenger. An experiment with students evaluating ambiguous images revealed perceptual biases, with only two certain identifications of a human face.

3. The «Impostor Bias»

The *Impostor Bias* arises from the growing use of AI-generated media, leading people to doubt the authenticity of multimedia content due to knowledge of AI's ability to create realistic fakes. This bias creates systematic distrust, even when AI-generated content is indistinguishable from the real thing, posing challenges for fields like forensics and copyright protection.



4. 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)

5. Test yourself: REAL or FAKE?

Do you think you can correctly distinguish between real and fake images? Scan these QR codes and *test yourself*!

6. Key takeaways

- Cognitive biases impact decision-making in forensics and digital forensics.
- Limited understanding of cognitive bias among forensic examiners.
- Effective bias mitigation strategies include game-based interventions and LSU-E approach.
- Generative Adversarial Networks and Diffusion Models deepfake detection methodologies and their performances.
- The *Impostor Bias* is introduced as a concern that arises from the advances of Generative AI.