Wiley
Human Behavior and Emerging Technologies
Volume 2025, Article ID 9977058, 22 pages
https://doi.org/10.1155/hbe2/9977058

WILEY

*Research Article*

# A (Mid)journey Through Reality: Assessing Accuracy, Impostor Bias, and Automation Bias in Human Detection of AI-Generated Images

**Mirko Casu**,[1] **Luca Guarnera**,[1] **Ignazio Zangara**,[1] **Pasquale Caponnetto**,[2] **and Sebastiano Battiato**[1]

[1]*Department of Mathematics and Computer Science, University of Catania, Catania, Italy*
[2]*Department of Educational Sciences, University of Catania, Catania, Italy*

Correspondence should be addressed to Mirko Casu; mirko.casu@phd.unict.it

Received 5 March 2025; Revised 30 July 2025; Accepted 9 August 2025

Academic Editor: Mona Alhasani



While the challenge of distinguishing AI-generated from real images is widely acknowledged, the specific cognitive biases that systematically shape human judgment in this domain remain poorly understood. It is particularly unclear how a general awareness of AI capabilities fosters novel biases, like a pervasive skepticism ("impostor bias"), and how this interacts with established phenomena like "automation bias". This study addresses this gap by providing the first quantitative analysis of how these two biases operate across five distinct experimental variants designed to test the context-dependency of human perception. Through a mixed-methods study with 746 participants, we demonstrate that human authentication accuracy hovered around chance levels (ranging from 47.0% to 55.5%). However, our analysis provides robust evidence for the systematic operation of cognitive biases. We validate the presence of "impostor bias" through a consistent pattern of higher doubt for AI-generated images and confirm "automation bias" through significant opinion changes following algorithmic suggestions. Our findings reveal that these biases are not uniform across populations: gender was a consistent predictor of automation bias, with males in all five variants showing a significantly stronger and more consistent tendency (Cohen's $d$ = 0.254–0.683) to be influenced by algorithmic suggestions. In contrast, age and academic background had minimal and highly localized effects. Furthermore, we identified a significant interaction between experimental stimuli and performance over time, isolating a pronounced fatigue effect to a single questionnaire variant where accuracy progressively declined (by approximately 1.7% per trial). By integrating human feedback with Grad-CAM visualizations, we confirm a divergence between human holistic evaluation and the localized focus of machine learning models. These findings carry direct implications for policy, as discussed within the context of the European AI Act, and inform the design of human–AI systems and media literacy programs aimed at mitigating these critical cognitive vulnerabilities.

**Keywords:** cognitive biases; cognitive psychology; deepfake detection; generative AI; impostor bias

## 1. Introduction

The challenge of establishing the authenticity of digital media is a long-standing concern in forensic science. For years, research focused on identifying traces left by editing software or artifacts introduced by the distribution platforms themselves, such as the unique compression and resizing signatures left by social networks [1]. However, recent advancements in deep learning have revolutionized generative artificial intelligence (GenAI), particularly through foundation models capable of processing multiple data modalities. These models demonstrate remarkable capabilities across text, images, audio, and video domains, transforming fields from entertainment to medicine [2, 3].

Large language models (LLMs) exemplify this transformation, with models like the GPT series showcasing unprecedented natural language processing abilities through transformer architectures [4, 5]. The field has expanded

beyond text with the emergence of text-to-image models like Imagen and DALL·E, which combine language understanding with diffusion processes to generate high-fidelity images [6–8]. This technological progression has been accompanied by growing research into image sentiment analysis, exploring how digital images evoke and communicate emotional responses [9, 10]. Unified multimodal architectures, such as UnIVAL or LLaVA-Mini, have further advanced the field by integrating multiple data types within a single framework, achieving competitive performance through efficient multimodal learning [11, 12].

Furthermore, the relevance of explainable AI in decision-making processes has been increasingly emphasized across disciplines. For example, Toumaj et al. [13] highlight the role of deep learning and explainability methods such as gradient-weighted class activation mapping (Grad-CAM) in supporting diagnostic reliability in Alzheimer's disease. Similarly, Wang et al. [14] present a holistic analysis of explainable AI techniques in the context of neurodegenerative disorders, underscoring their impact on user trust and system transparency. Moreover, Heidari et al. [15] provide a wide overview of the architecture and capabilities of ChatGPT, one of the foundational generative models that inform contemporary approaches to synthetic content creation. These studies reinforce the need to address human–AI interaction not only from a technical but also from a cognitive and perceptual standpoint—perspectives that our study brings into focus within the forensic evaluation of AI-generated media.

A critical development—and concern—in this domain is the deepfake technology, which uses generative engines (e.g., GANs and diffusion models) to create increasingly convincing synthetic media [16, 17]. While offering creative opportunities, deepfakes pose significant risks for misinformation and manipulation [18], potentially undermining public trust in digital information [19, 20]. Recent empirical research reveals significant challenges in human deepfake detection. Somoray and Miller [21] and Soto-Sanfiel et al. [22] found that participants' ability to identify synthetic media was moderately successful, with average detection accuracy around 60%. Providing explicit detection strategies did not significantly improve identification, suggesting that human perception of synthetic media is complex and influenced by psychological factors like narrative transportation and familiarity. Although researchers have developed various detection methods [23–26], generative technologies often outpace these countermeasures [4]. The emerging body of research highlights the need for sophisticated approaches that combine technological detection with an understanding of human cognitive processes in interpreting synthetic media.

Navigating this new landscape requires an understanding of novel cognitive phenomena that emerge from our interactions with synthetic media. In this study, we define and describe two specific types of cognitive bias that influence how people judge whether images are real or AI-generated.

- The *impostor bias*, a hypothetical cognitive bias where individuals systematically doubt digital content (audio, images, and videos) authenticity due to the awareness of AI's generative capabilities [27], can be distinguished from healthy skepticism. While skepticism is a reasoned and cautious approach to evaluation, impostor bias becomes a systemic cognitive error when this doubt is applied indiscriminately, leading to a higher rate of specific misclassifications (i.e., labeling real content as fake). Theoretically, this bias is rooted in a cognitive dissonance between the long-held expectation of human-centric creativity and the new reality of hyperrealistic AI outputs [28, 29]. It can also be seen as an overapplication of the "availability heuristic," where the vivid and widely discussed examples of deepfakes make the possibility of synthetic origin feel more common than it actually is in any given context. This bias then interacts with other established processes, such as confirmation bias [30], where users might actively search for flaws that support their initial suspicion of inauthenticity.

- The *automation bias*, defined as the tendency to overrely on automated systems in decision-making [31–33], leads to both omission and commission errors in content authentication [34]. In our case, we will analyze the tendency to align one's judgment with algorithmic suggestions, even when those suggestions are incorrect. It is driven by overreliance on perceived machine authority. We will assess automation bias by comparing participants' initial responses to their final decisions after being exposed to the classifier's output, quantifying shifts that occur toward algorithm-driven choices.

Our study empirically investigates these theoretical constructs through a systematic experimental design, examining participants' ability to distinguish AI-generated from authentic images, their inherent skepticism regarding image authenticity, and their susceptibility to algorithmic suggestion. By analyzing discrimination accuracy, confidence levels, and response patterns, we provide the first quantitative assessment of how generative AI awareness influences digital media trust and judgment, contributing to our understanding of emerging cognitive biases in the AI era.

Finally, the paper is structured as follows: Section 1 introduced the challenges of human–AI interaction in media authentication, contextualizing cognitive biases such as impostor and automation bias. Section 2 details the mixed-methods experimental design, including dataset construction, participant recruitment, and algorithmic tools. Section 3 presents the results, analyzing demographic patterns, accuracy rates, and susceptibility to algorithmic influence. Section 4 compares human perceptual feedback with Grad-CAM visualizations, highlighting divergences between human and machine decision-making. Section 5 discusses content-specific authentication patterns, demographic influences, and implications of the European AI Act on deepfake regulation. It also addresses study limitations and proposes future research directions. Section 6 concludes by synthesizing key findings on human detection capabilities, bias dynamics,

and strategies to address evolving generative AI challenges. Acknowledgments and references follow thereafter.

## 2. Materials and Methods

*2.1. Study Design.* This study employed a mixed-methods approach to investigate participants' ability to distinguish between real and AI-generated images, their susceptibility to the impostor bias, and their trust in algorithmic assessments. The experimental design involved five variants ($\alpha$, $\beta$, $\gamma$, $\delta$, and $\varepsilon$) of a standardized questionnaire, each containing 15 images with associated questions.

Each questionnaire variant was balanced with three categories of images: five real photographs, five images generated using Midjourney [35], and five images created with Midjourney and subsequently enhanced using Magnific AI [36]. As illustrated in Figure 1, the evaluation process for each image was structured in two steps. In Step 1, participants were asked three initial questions: to determine the image's authenticity (real vs. AI-generated), to indicate whether they had any doubts about their choice (yes/no), and to provide a brief written explanation for their decision. In Step 2, participants were presented with a hypothetical algorithmic assessment (e.g., "an algorithm suggests to you that, at 86%, the previous image was generated by AI."). They were then asked to either confirm their initial choice or change it, allowing us to measure their trust in the algorithm's suggestion.

*2.2. Participant Recruitment and Sample.* Recruitment was conducted through multiple channels, including social media advertisements and word of mouth dissemination. Additional recruitment efforts included promoting the study during university courses, conferences, and lectures. In these settings, the questionnaire was initially introduced as a generic tool for assessing the visual perception of images, in order to avoid conditional biases. Subsequently, after obtaining participants' consent, a QR code was projected that provided access to the different versions of the questionnaire, which they could complete independently.

*2.3. Materials*

*2.3.1. Dataset Characteristics.* The dataset consists of five questionnaire variants ($\alpha$–$\varepsilon$), each containing 15 images spanning various subjects and scenarios (see Table 1). The collection encompasses portraits and group photos, action shots of people engaged in activities, environmental and landscape photography, cultural and social situations, and urban and natural settings.

Each variant maintains consistent technical specifications and a uniform distribution of image sources. The first category comprises five real photographs featuring authentic human interactions, natural lighting, and spontaneous moments. The second category includes five Midjourney v6 generated images that demonstrate advanced AI capabilities in creating portraits and scenes. The third category consists of five Midjourney v6 images enhanced with Magnific AI, combining AI generation with improved detail clarity and texturization.

*2.3.2. AI Image Generation and Enhancement Tools.* Our study employed a multifaceted approach to AI-generated imagery, utilizing different sets of models for the human-facing stimuli and the training of our backend classifier.

For the images presented directly to participants, we exclusively selected Midjourney v6 as the primary generation tool, with a subset of these images enhanced by Magnific AI. This decision was deliberate: we aimed to test human perception against a state-of-the-art pipeline known for its high photorealism and coherence, thus creating a challenging and ecologically valid detection task. Midjourney v6 builds upon previous versions with significant improvements in aesthetics and prompt adherence [37–39], while Magnific AI was employed to further increase sharpness and realism. Göring et al. [40] demonstrated that the quality and realism of generated images vary significantly based on prompt engineering and methodological approaches.

In contrast, the ResNet-101 classifier used to provide algorithmic suggestions was trained on a much broader and more diverse dataset of synthetic images. As detailed in Section 2.3.4, this training set included outputs from numerous models, such as DALL·E 2, Latent Diffusion (the basis for Stable Diffusion), and StyleGAN2. This design allowed us to evaluate human performance on a novel, high-quality generator (Midjourney) that the classifier had not been explicitly trained on, simulating a more realistic "zero-shot" detection scenario.

*2.3.3. Image Generation and Selection Protocol.* The prompts were crafted using a strategy that combined detailed scene descriptions with technical and stylistic specifications. This included camera types (e.g., "Canon R6 Mark II" and "iPhone 14"), lens information ("35 mm" and "85 mm"), film types ("Kodak Portra 400" and "Fujicolor Pro 400H"), and lighting conditions ("golden hour" and "neon lights"). To control the output, command flags such as -style raw and -stylize 0 were used to favor photorealism over artistic stylization. Prompts also incorporated various aspect ratios (e.g., 9:16, 3:2, and 16:9) to match common photography formats. The complete list of all 72 prompts used for generation is available from the corresponding author upon reasonable request.

From the large pool of generated images, a final set was selected based on several criteria designed to ensure ecological validity and challenge. The primary criterion was high photorealism, meaning that each selected image had to convincingly resemble a real photograph. Consequently, images with clear giveaways of their synthetic origin, such as distorted hands, unnatural facial features, or nonsensical text, were systematically excluded. In addition to technical quality, the selection process aimed for thematic and stylistic diversity, covering a wide range of subjects including individual portraits, group shots, and various cultural settings to reflect the breadth of the prompts. Finally, the overall plausibility of the depicted scene was a key factor in the selection.

The selection and categorization process was conducted by two researchers. An image was included in the final dataset only after both researchers reached a consensus that it met

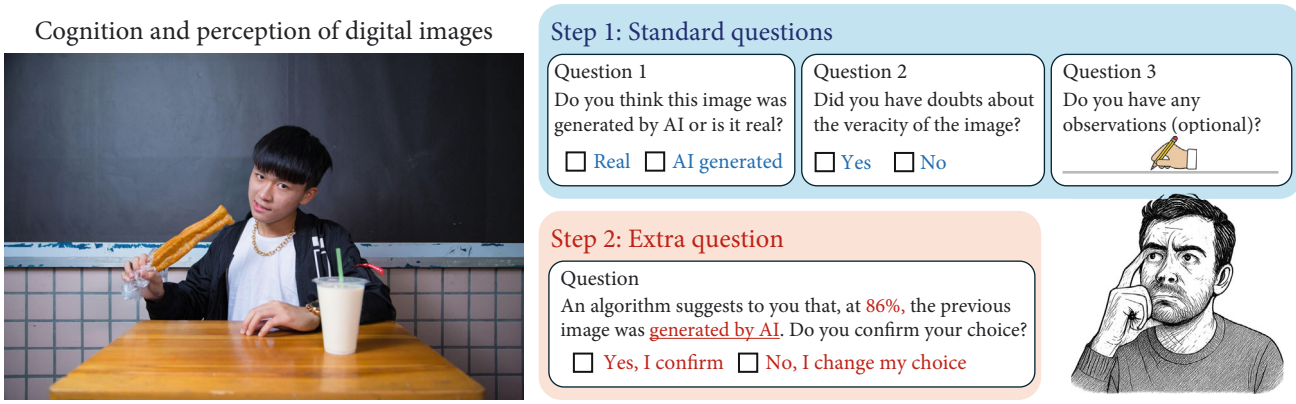



Figure 1: An illustration of the two-step experimental procedure presented to participants for each image. Step 1 (standard questions) prompted users to assess the image's authenticity and express their doubt. Step 2 (extra question) measured their trust in algorithmic suggestions by asking them to either confirm or change their initial choice after being shown a hypothetical AI-generated confidence score.

all the established criteria. This consensus-based approach was implemented to ensure consistency and intersubjective agreement on the final set of images, thereby addressing the need for reliability in the categorization process. A subset of these selected Midjourney images was then processed with Magnific AI to create the third category of stimuli (MJM), as illustrated in Figure 2. To ensure full transparency and reproducibility, the complete dataset of all images used in the study, along with the full list of generation prompts and the anonymized raw participant data, will be made publicly available in a dedicated repository upon publication.

*2.3.4. Algorithmic Probability Assessment.* Detecting deepfake images remains an open challenge in Computer Vision, with numerous methods proposed in the literature. Recent approaches include convolutional neural networks (CNNs) trained specifically for deepfake detection [41–44] and hybrid architectures combining spatial and frequency domain analysis [45–47]. The field of digital forensics has also explored more advanced questions, such as "image ballistics," which aims to determine not only the origin of a synthetic image but also its processing history, for instance, by detecting how many times it has been manipulated by a generative model [48]. However, despite these advancements, deepfake detection models still face limitations, particularly in their generalizability across different datasets and manipulation techniques.

We employed a ResNet-101 architecture [49] as the core algorithm for generating probabilistic assessments of image authenticity, primarily because its performance characteristics matched the requirements of the experimental design. Although ResNet-101 is not the most recent architecture and presents some known limitations in terms of robustness and generalizability, its selection was deliberate and aligned with the goals of our study. Extremely accurate and state-of-the-art models (such as ViT-B16 or ensemble-based detectors) could have introduced strong anchoring effects, leading participants to trust algorithmic outputs uncritically. This would have reduced our ability to observe genuine manifestations of automation bias. ResNet-101, by contrast, provides a balance between solid classification performance and probabilistic outputs that remain open to participant interpretation. As shown in our internal benchmark (see Table 2), it outperformed EfficientNet-b4/b7 and achieved results comparable to ViT-B16 while being significantly more efficient and easier to integrate within a real-time behavioral experiment. Its stability, interpretability, and moderate confidence calibration made it a pragmatic and methodologically consistent choice for our task.

Participants were first asked to classify each image as either real or fake and to indicate any doubts regarding their judgment; subsequently, they were presented with a probability score—computed by the algorithm—that suggested the likelihood that the image was fake (e.g., "XX% chance that the image is fake").

In detail, we used the same dataset and data splitting (train, validation, and test set) described in Guarnera et al. [50]:

1. Real images: CelebA [51], FFHQ[1], and ImageNet [52].
2. Deepfakes: AttGAN [53], CycleGAN [54], GDWCT [55], IMLE [56], ProGAN [57], StarGAN [58], StarGAN-v2 [59], StyleGAN [60], StyleGAN2 [61]), DALL·E 2 [62], GLIDE [63], and Latent Diffusion [64][2].

The PyTorch implementations were used of the ResNet-101 architecture, pretrained with ImageNet[3]. A fully connected layer with an output size equal to 2 followed by a SoftMax was added to the last layer in order to solve the binary classification task (real vs. deepfake).

Notably, the training dataset for the ResNet-101 did not contain any images generated by Midjourney or processed by Magnific AI. This inherent limitation was considered advantageous, as an overly precise classifier (yielding uniformly high or confident probability values) might unduly bias the participants' subsequent responses.

## 2.4. Data Collection Methods

*2.4.1. Questionnaire Design.* The questionnaire was administered via Google Forms, designed to be easily accessible on both desktop and mobile devices. Although each version featured different image contents, the overall structure and

**TABLE 1**: A detailed summary of image descriptions categorized by variant and generation type (R, real photo; MJ, Midjourney v6; MJM, Midjourney v6 enhanced with Magnific AI).

| Var. | Img # | Type | Image description |
|---|---|---|---|
| | 1 | R | Young man seated holding churro with milkshake in restaurant |
| | 2 | MJM | Three women standing before sunset, lens flare obscuring faces |
| | 3 | R | Young man on cliff photographing with tripod during sunny day |
| | 4 | MJ | Two men kissing in dimly lit nightclub with colorful lights |
| | 5 | MJM | Two women in crop tops and sunglasses posing on street |
| | 6 | MJ | Young people dancing in dimly lit nightclub, illuminated by spotlights |
| | 7 | R | Man in suit speaking into microphone, seated with masked woman |
| $\alpha$ | 8 | MJ | Two women taking selfie on beach during sunset |
| | 9 | R | Man in black jacket holding red object before lit doorway |
| | 10 | MJM | Group in traditional attire enjoying food at lively market |
| | 11 | R | Young girl eating heart-shaped lollipop by lake with trees |
| | 12 | MJ | Young woman in glasses, long brown hair posing by water |
| | 13 | MJM | Woman with bun, dressed in black, facing left against gray |
| | 14 | MJ | Older man with beard, tattoos, holding fishing net in workshop |
| | 15 | MJM | Three women in casual attire having coffee at cafe |
| | 1 | MJM | Two young girls in floral dresses sharing a moment in a BW photo |
| | 2 | MJ | Woman in green hijab with neutral expression indoors |
| | 3 | MJM | Elderly man repairing fishing net by the water |
| | 4 | MJM | Three young people enjoying drinks at an outdoor cafe |
| | 5 | MJ | Young men performing with microphones and guitars on stage |
| | 6 | MJM | Woman checking phone against graffiti walls, others seated behind her |
| | 7 | R | Woman in black shirt addressing seated young people |
| $\beta$ | 8 | R | Young woman standing arms outstretched in lively bowling alley |
| | 9 | MJM | Young girl in Hogwarts robe holding staff at themed event |
| | 10 | R | Close-up of hand holding Bible and pen with "FAITH" wristband |
| | 11 | R | Brown dog by food bowls on patio with garden background |
| | 12 | MJ | Three women in dresses taking mirror selfie in pink-lit room |
| | 13 | R | Young boy drinking from bottle in park with bouquet foreground |
| | 14 | MJ | Woman holding blanket to face, crying indoors with teal walls |
| | 15 | MJ | Two children lying in yellow tent on cloudy day in field |
| | 1 | R | Woman swimming in pool, performing breaststroke with goggles |
| | 2 | MJ | Young woman with brown hair in yellow top against teal wall |
| | 3 | MJM | Man in dimly lit room playing video games on a large screen |
| | 4 | R | Young boy with two friends in room with blurred gray wall |
| | 5 | MJM | Woman picking tea leaves in lush tea garden on sunny day |
| | 6 | MJ | Woman relaxing in bed with light blue floral top |
| | 7 | MJM | Woman crouched down interacting with cat on street |
| $\gamma$ | 8 | R | Man wearing life jacket by water, with blurred river in background |
| | 9 | R | Rusty 1936 Ford V8 displayed in city square with people admiring |
| | 10 | R | Group of young girls in gymnastics leotards seated |
| | 11 | MJM | Sepia-toned photo of young man in vintage clothing |
| | 12 | MJM | Couple silhouetted in nightclub, vibrant lights and intimate moment |
| | 13 | MJ | Man standing behind table with cured meats in professional setting |
| | 14 | MJ | Young man with sunglasses in front of cityscape with obelisk |
| | 15 | MJ | Young man in vintage outfit outdoors with blurred background |

**TABLE 1**: Continued.

| Var. | Img # | Type | Image description |
|---|---|---|---|
| | 1 | MJM | Young woman in green hijab, gray and white shirt in indoor setting |
| | 2 | R | Young boy sitting on floor with basketball, gray floor background |
| | 3 | MJ | Black and white close-up of tattoo artist's face with glasses |
| | 4 | MJM | Two clowns in city street at night with people and neon lights |
| | 5 | R | Pregnant woman in bedroom, hands on belly, calm atmosphere |
| | 6 | MJ | Black cat and golden retriever sleeping on bed |
| | 7 | R | Diverse protest group advocating women's rights, peaceful atmosphere |
| $\delta$ | 8 | MJM | Young couple standing on sidewalk, casual clothes, smartphone in hand |
| | 9 | MJ | Young woman in red hijab, black shirt, gazing right by wall |
| | 10 | MJM | Man in dark suit and tie standing in rain on city street |
| | 11 | R | Glasses with brown-tinted lenses on granite countertop, casual setting |
| | 12 | MJM | Three cadets walking in park, outdoor activity |
| | 13 | MJ | Black and white photo of tattooed man leaning on table |
| | 14 | R | Industrial furnace with worker in protective gear in factory setting |
| | 15 | R | Group of four people in airport with "ready for action" sign |
| | 1 | MJ | Elderly man with leaf crown sitting cross-legged in forest |
| | 2 | R | Man in canoe taking photo in tranquil setting |
| | 3 | MJ | Two women in urban attire standing on street |
| | 4 | MJM | Two children lying on blanket at campsite |
| | 5 | MJM | Man taking selfie on narrow European street |
| | 6 | MJM | Woman lying on bed, gazing at camera |
| | 7 | R | Man and woman standing in park or garden |
| $\varepsilon$ | 8 | MJ | Three women posing at nightclub under neon lights |
| | 9 | MJM | Couple walking in front of Colosseum on cobblestone street |
| | 10 | R | Man in bucket hat performing on stage |
| | 11 | R | Tree-lined park pathway with trimmed trees |
| | 12 | MJ | Conductor leading orchestra, gesturing to musicians |
| | 13 | R | Group of eight people posing in hallway |
| | 14 | MJM | Man and girl having breakfast at table |
| | 15 | R | Black cat with white chest patch on surface |

question formats were consistent throughout. Participants first encountered a question asking whether they believed the image was real or AI-generated, setting the stage for the study. This was followed by an inquiry into whether they had any doubts about the image's authenticity, which aimed to capture their inherent skepticism. An optional open-ended field then allowed respondents to detail any specific visual cues that influenced their judgment. Finally, participants were shown an algorithmic confidence level and asked to confirm their choice, thereby testing the impact of numerical cues on trust. In addition to these core components, the questionnaire gathered demographic information, including age, biological sex, and university affiliation and provided an option to submit an email address to receive the true authentication status of the images.

*2.5. Ethical Considerations.* The study adhered to the guidelines set forth by the Declaration of Helsinki and followed the principles of the General Data Protection Regulation (GDPR), as well as the guidelines established by the Italian Association of Psychology (AIP). The study received approval from the Internal Ethic Review Board of Psychology Research (Prot. no. Ierb-Edunict-2024.03.07/04). All participants provided written informed consent, and the study ensured both confidentiality and anonymity throughout the research process.

## 3. Results and Data Analysis

*3.1. General Demographics.* The final sample consisted of 746 Italian university students, distributed unevenly across the five questionnaire variants: $\alpha$ (n = 256, 34.3%), $\beta$ (n = 150, 20.1%), $\gamma$ (n = 142, 19.0%), $\delta$ (n = 106, 14.2%), and $\varepsilon$ (n = 92, 12.3%).

The sample was composed predominantly of males (56.2%; $n = 419$) over females (43.7%; $n = 326$), with a mean age of 22.1 years (SD = 1.1). The majority of participants were aged 23 or older (58.6%). Regarding academic background, the sample was heavily skewed toward computer science (46.0%), followed by healthcare (25.7%) and

(a) Midjourney (b) Midjourney + Magnific AI

(c) Midjourney (d) Midjourney + Magnific AI

FIGURE 2: Examples of images generated with Midjourney and enhanced with Magnific AI. Each pair shows the original version (left) and enhanced version (right).

education (11.8%). The detailed distribution of these demographic characteristics across all five questionnaire variants is illustrated in Figure 3.

*3.2. Accuracy, Changes of Opinion, and Doubts.* To establish a baseline for the algorithmic suggestion shown to participants, we first evaluated the performance of the ResNet-

**TABLE 2**: Evaluation of various architectures for the classification task conducted to select a suitable model for the study. The results indicate that ResNet-101 offered performance ($F1 = 0.96$) comparable to the state-of-the-art ViT-B16 model, providing an effective balance between accuracy and the avoidance of anchoring effects critical for studying automation bias.

| | Level 1 | | | |
|---|---|---|---|---|
| | **Prec** | **Rec** | ***F*1** | **Acc** |
| ResNet-18 | 0.84 | 0.98 | 0.89 | 0.965 |
| ResNet-34 | 0.88 | 0.98 | 0.92 | 0.9761 |
| ResNet-50 | 0.9 | 0.98 | 0.94 | 0.9818 |
| ResNet-101 | 0.94 | **0.99** | **0.96** | 0.9893 |
| ResNet-101 | 0.86 | 0.98 | 0.91 | 0.9729 |
| DenseNet-121 | 0.88 | **0.99** | 0.92 | 0.9768 |
| EfficientNet-b4 | 0.7 | 0.93 | 0.76 | 0.8996 |
| EfficientNet-b7 | 0.84 | 0.98 | 0.89 | 0.9661 |
| ViT-B16 | **0.95** | 0.98 | **0.96** | **0.9904** |

*Note:* The highest performing metrics are highlighted in bold.

101 classifier on our image set. The algorithm demonstrated varied performance, achieving an accuracy of 25% on real images ($p \approx 0.003$) while correctly identifying AI-generated images at rates of 73.2% (Midjourney-only, $p < 0.001$) and 70.4% (Midjourney-enhanced, $p < 0.001$).

Against this backdrop of a moderately reliable but imperfect algorithmic aid, the analysis of participant responses revealed several notable patterns in their own authentication accuracy, susceptibility to influence, and decision uncertainty. Overall mean accuracy rates varied across variants: 53.4% ($\alpha$), 55.5% ($\beta$), 54.1% ($\gamma$), 47.9% ($\delta$), and 47.0% ($\varepsilon$). These figures are accompanied by high variability, with standard deviations ranging from 18.7% ($\delta$) to 24.4% ($\varepsilon$). A deeper statistical analysis reveals that none of these accuracy rates were statistically significant from a 50% chance baseline (all $p$-values > 0.33), and a conclusion reinforced after applying a Bonferroni correction for multiple comparisons. The 95% confidence intervals for all variants were wide and overlapped with the 50% chance mark, indicating a lack of statistical certainty in performance. Real photographs consistently elicited higher accuracy rates, particularly evident in the outer edges of the blue accuracy lines for images labeled "R" across all variants (Figure 3).

The changes of opinion, visualized by the orange markers in Figure 3, demonstrated varying levels of automation bias. While Variant $\alpha$ showed the lowest average changes (6.2%), followed by Variant $\beta$ (7.6%), Variants $\gamma$, $\delta$, and $\varepsilon$ exhibited notably higher rates (9.9%, 10.7%, and 10.4%, respectively). The radar plots reveal particular hotspots of algorithmic influence, such as Positions 10 and 11 in Variants $\gamma$ and $\varepsilon$, where changes of opinion peaked at 24.6% and 23.9%, respectively.

The distribution of participant uncertainty showed intriguing patterns across image types, as illustrated in Figure 4. While average doubt rates remained relatively consistent across variants (ranging from 44.3% in $\alpha$ to 49.0% in $\gamma$), the box plots reveal a systematic difference in uncertainty between real and AI-generated images. Notably, Variant $\alpha$ shows the widest disparity, with real images (blue box) generating a broader range of doubt levels compared to AI-generated images (orange box). This pattern shifts in Variants $\gamma$ and $\varepsilon$, where AI-generated images prompted more concentrated levels of uncertainty, particularly evident in the tighter interquartile ranges of the orange boxes.

The radar plots (Figure 5) further illuminate this uncertainty pattern, with the green doubt lines showing pronounced peaks around specific image positions. For instance, the substantial overlap between areas of high doubt and low accuracy in Variants $\alpha$ and $\gamma$ suggests a correlation between participant uncertainty and classification errors. This relationship is particularly evident in Magnific AI-enhanced images (MJM), where the green doubt lines consistently expand outward, indicating heightened uncertainty in evaluating these enhanced images.

A comparative analysis of Variants $\beta$ and $\gamma$ reveals an interesting trade-off. While Variant $\beta$ achieved the highest mean accuracy (55.5%), Cohen's $d$ effect size analysis shows its performance advantage over chance was small ($d = 0.255$), and its advantage over Variant $\gamma$ (54.1%) was negligible ($d = 0.064$). In contrast, the performance difference between the higher performing variants (e.g., $\beta$) and the lower performing ones (e.g., $\delta$) was more substantial, with a small effect size ($d = 0.378$). Variant $\gamma$ maintained a similar accuracy to $\beta$ but exhibited both higher changes of opinion and more pronounced differences in doubt patterns. This suggests that the specific combination of images in Variant $\gamma$ might have created conditions more conducive to algorithmic influence while simultaneously polarizing participant certainty based on image type.

A comprehensive statistical evaluation was performed for each variant to assess participant performance in greater detail. The analysis included descriptive statistics (mean and standard deviation), 95% confidence intervals for mean accuracy, Cohen's $d$ effect sizes, and single-sample $t$-tests to compare accuracy against a 50% chance baseline. These key findings are summarized in Table 3.

While the statistical analysis indicates that the observed differences in mean accuracy across variants are not statistically significant (all $p > 0.05$, see Table 2), the results provide a more nuanced view of participant behavior. The notable variance in other key metrics, such as changes of opinion (ranging from 6.2% to 10.7%) and average doubt (44.3%–49.0%), suggests that participants' responses were not uniform. Therefore, even though the overall accuracy did not reliably exceed chance levels, the specific composition of images and the presence of algorithmic suggestions in each variant had a measurable, albeit complex, influence on the decision-making process and user confidence.

*3.3. Demographic Patterns in Variant Performance.* Our analysis of demographic factors revealed that gender was a consistent predictor of susceptibility to automation bias, whereas age effects were highly localized to a specific experimental context.

Gender differences emerged as a significant and recurring factor influencing changes of opinion. In all five

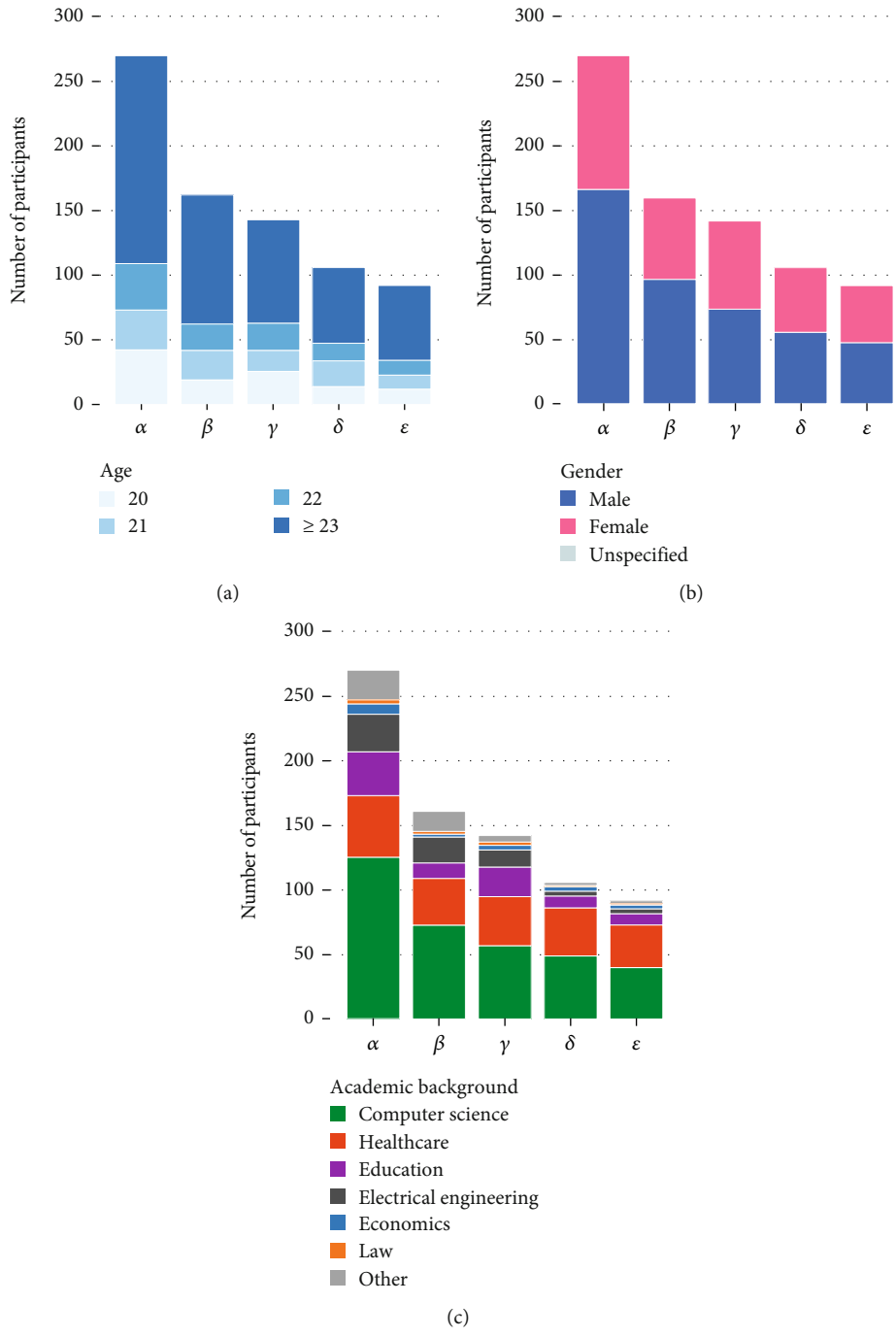



FIGURE 3: Demographic distribution of participants across questionnaire variants ($\alpha$, $\beta$, $\gamma$, $\delta$, or $\varepsilon$): (a) age distribution showing the proportion of participants aged 20, 21, 22, and ≥ 23 years; (b) gender distribution showing male, female, and unspecified responses; and (c) distribution of participants' affiliation areas including computer science, healthcare, education, electrical engineering, economics, law, and other.

questionnaire variants, males demonstrated a significantly higher propensity to change their initial judgment after receiving an algorithmic suggestion. The magnitude of this effect ranged from small in Variant $\alpha$ ($p = 0.047$, $d = 0.254$) to moderate and large in Variants $\beta$, $\gamma$, $\delta$, and $\varepsilon$ ($d$ values from 0.457 to 0.683), indicating a robust pattern of males being more susceptible to automation bias. In contrast, gender's impact on accuracy was limited to a single instance in Variant $\alpha$, where males performed slightly better ($p = 0.039$, $d = 0.265$). No significant gender-based differences were found for doubt levels in any of the variants.

Conversely, age appeared to have a much more limited influence on performance, with ANOVAs revealing no statistically significant differences across age groups for nearly all metrics and variants. The sole, notable exception was a strong and significant effect on change of opinion in Variant

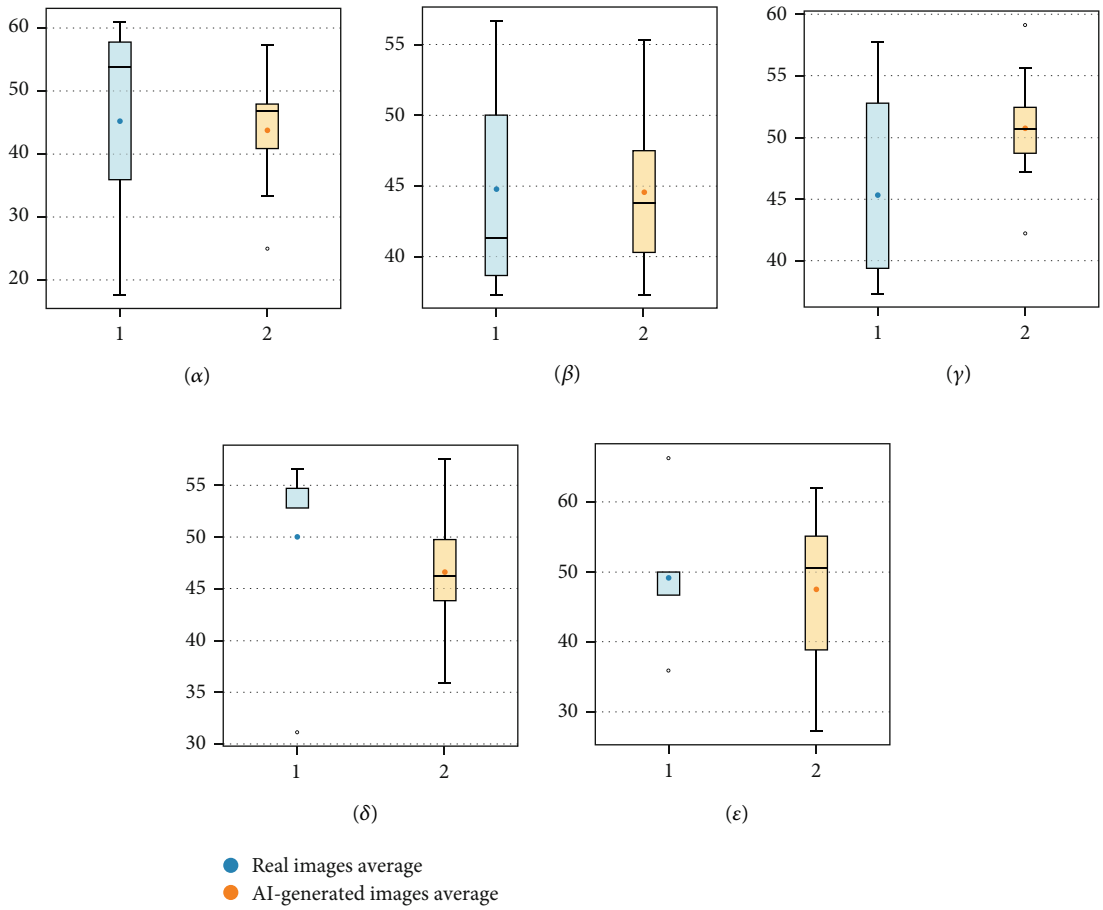



FIGURE 4: Distribution of doubts by type of image (percent) across questionnaire variants (α–ε). For each variant, blue boxes represent the distribution of doubts for real images, while orange boxes show the distribution for AI-generated images (including both Midjourney v6 and Magnific AI–enhanced images). The boxes display the interquartile range, with the middle line representing the median. Whiskers extend to the minimum and maximum values, excluding outliers (shown as individual points).

ε ($F(3, 88) = 6.65$, $p < 0.001$, $\eta p^2 = 0.185$). In this context, younger participants (specifically 21 year olds, $M = 24.2\%$) were substantially more likely to alter their judgment compared to other age groups, suggesting that the specific stimuli in this variant interacted uniquely with age-related decision-making strategies. A comprehensive summary of all significant demographic findings is presented in Table 4.

*3.4. Interdepartmental Comparison of Results.* The analysis of performance based on academic background revealed that departmental affiliation was not a pervasive factor in determining outcomes. One-way ANOVAs conducted for each variant showed no statistically significant differences in accuracy or doubt levels across departments in any of the five experimental conditions.

Significant interdepartmental differences were, however, isolated to the change of opinion metric and only emerged in two specific contexts: Variant β ($F(6, 143) = 2.65$, $p = 0.018$, $\eta p^2 = 0.100$) and Variant δ ($F(6, 99) = 2.52$, $p = 0.026$, $\eta p^2 = 0.133$). This indicates that the specific combination of images in these two variants triggered different levels of susceptibility to algorithmic influence depending on a participant's field of study. For instance, in both of these variants, participants from electrical engineering were among those most likely to change their opinion after seeing the algorithmic suggestion.

While most other differences were not statistically significant, some descriptive trends are notable. The computer science department, for example, consistently demonstrated moderate to high accuracy across all variants. In contrast, the law and economics departments often exhibited more extreme patterns in doubt and opinion change, though these trends did not consistently reach statistical significance. The detailed performance metrics for each department and the results of all ANOVA tests are provided in Tables 5, 6, 7, 8, and 9.

*3.5. Accuracy Sequential Effect Analysis.* To investigate whether participants exhibited learning or fatigue effects as they progressed through the experimental session and whether these sequential effects varied across different questionnaire variants, we conducted a comprehensive mixed-effects analysis examining accuracy trends across the 15 trials. We first implemented a linear mixed-effects model with trial number (centered at the midpoint) and questionnaire variant as fixed effects and participant ID as a random effect

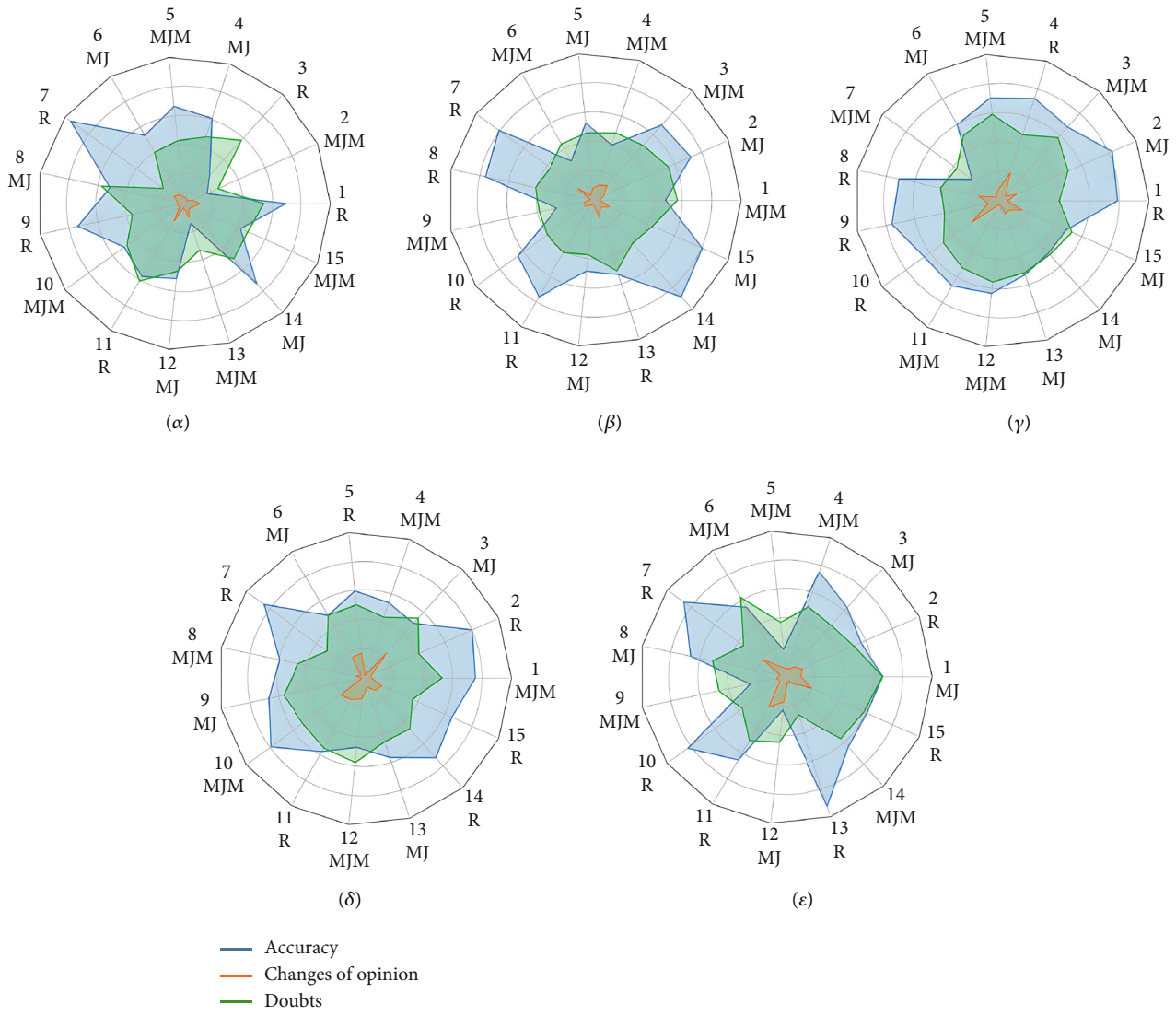



FIGURE 5: Radar plots displaying participant responses across questionnaire variants (α–ε). Each plot represents three key metrics: accuracy in image classification (blue), changes in initial assessment after algorithmic suggestion (orange), and reported uncertainty (green). Numbers 1–15 indicate individual images, labeled according to their source: real photographs (R), Midjourney v6-generated images (MJ), and Midjourney v6 images enhanced with Magnific AI (MJM). Each concentric circle represents a 20% increment from the center. $n = 746$ total participants (α: $n = 256$, β: $n = 150$, γ: $n = 142$, δ: $n = 106$, and ε: $n = 92$).

to account for individual differences in baseline performance. The analysis revealed no significant main effect of trial number ($\beta = 0.0036$, $p = 0.100$), indicating no overall sequential effect across all variants. However, we observed a significant *trial* × *variant* interaction ($\beta = -0.0016$, $p = 0.038$), suggesting that sequential effects were not uniform across experimental conditions but depended on the specific combination and ordering of images presented.

To further investigate this interaction, we performed individual linear regression analyses for each questionnaire variant, examining the relationship between trial number and mean accuracy (Figure 6).

These analyses revealed striking heterogeneity in sequential patterns across variants. Only Variant γ demonstrated a statistically significant sequential effect, characterized by a negative slope ($\beta = -0.017$, $p = 0.048$) indicating progressive fatigue, with participants losing approximately 1.72% accuracy per trial, culminating in a total decline of 24.1% from the first to the final trial. In contrast, Variants α ($\beta = 0.001$, $p = 0.956$), β ($\beta = 0.013$, $p = 0.271$), δ ($\beta = -0.003$, $p = 0.730$), and ε ($\beta = 0.003$, $p = 0.836$) showed no significant sequential trends, suggesting stable performance throughout the experimental session.

*3.6. Logistic Regression Analysis of Impostor Bias and Automation Bias Interaction.* To investigate whether individuals with high impostor bias (measured through initial doubt levels) respond differently to algorithmic suggestions compared to those with low impostor bias, we conducted a logistic regression analysis using data from the 746

**TABLE 3**: This table details the key performance metrics for each questionnaire variant, including the number of participants ($N$), mean accuracy with standard deviation (SD), and the 95% confidence interval (CI) for accuracy. It also reports the mean doubt rate and the rate of automation bias (opinion changes). Inferential statistics are provided to compare accuracy against a 50% chance baseline, including Cohen's $d$ effect size and the $p$ value from a single-sample $t$-test. Notably, the results show that none of the variants achieved an accuracy level statistically significantly different from chance (all $p > 0.05$).

| Variant | $N$ (participants) | Mean accuracy (%) | Accuracy SD (%) | 95% CI (accuracy) | Mean doubt (%) | Automation bias (%) | Cohen's $d$ (vs. 50%) | $p$ value (vs. 50%) |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 256 | 53.4 | 22.1 | [41.2%, 65.7%] | 44.3 | 6.2 | 0.155 | 0.5578 |
| $\beta$ | 150 | 55.5 | 21.4 | [43.6%, 67.3%] | 44.6 | 7.6 | 0.255 | 0.3398 |
| $\gamma$ | 142 | 54.1 | 19.7 | [43.2%, 65.1%] | 49.0 | 9.9 | 0.209 | 0.4312 |
| $\delta$ | 106 | 47.9 | 18.7 | [37.5%, 58.2%] | 47.7 | 10.7 | −0.114 | 0.6646 |
| $\varepsilon$ | 92 | 47.0 | 24.4 | [33.5%, 60.5%] | 48.0 | 10.4 | −0.122 | 0.6440 |

**TABLE 4**: Results of all statistically significant independent sample $t$-tests (for gender) and one-way ANOVAs (for age) across the five variants. For brevity, nonsignificant results are omitted from this table but are described in the main text. Effect sizes are reported as Cohen's $d$ for $t$-tests and partial eta-squared ($\eta p^2$) for ANOVA.

| Analysis | Variant | Metric | Test statistic | $p$ value | Effect size |
|---|---|---|---|---|---|
| Gender | $\alpha$ | Accuracy | $t(254) = 2.08$ | 0.039 | $d = 0.265$ |
| | $\alpha$ | Change of opinion | $t(254) = 1.99$ | 0.047 | $d = 0.254$ |
| | $\beta$ | Change of opinion | $t(147) = 2.98$ | 0.003 | $d = 0.496$ |
| | $\gamma$ | Change of opinion | $t(140) = 2.72$ | 0.007 | $d = 0.457$ |
| | $\delta$ | Change of opinion | $t(104) = 3.51$ | < 0.001 | $d = 0.683$ |
| | $\varepsilon$ | Change of opinion | $t(90) = 2.70$ | 0.008 | $d = 0.563$ |
| Age | $\varepsilon$ | Change of opinion | $F(3, 88) = 6.65$ | < 0.001 | $\eta p^2 = 0.185$ |

participants across 11,170 observations. The dependent variable was the binary change of opinion (0 = *no change*, 1 = *opinion change*) following the ResNet-101 recommendation, while the independent variables included initial doubt level (continuous proxy for impostor bias), AI agreement status (binary: 1 if AI disagreed with participant's initial assessment, 0 if agreed), and their interaction term.

The logistic regression model revealed significant main effects for both initial doubt ($\beta = 0.5063$, $p = 0.006$) and AI disagreement ($\beta = 1.8259$, $p < 0.001$), indicating that participants with higher initial doubt were more likely to change their opinion, and AI disagreement substantially increased the probability of opinion change across all participants. However, the critical interaction term between initial doubt and AI disagreement was not statistically significant ($\beta = 0.2261$, $p = 0.258$), suggesting that the effect of AI disagreement on opinion change does not significantly vary based on participants' initial doubt levels. As illustrated in Figure 7, the logistic curves for AI agreement and disagreement conditions run nearly parallel across the full range of initial doubt levels, visually confirming the absence of a meaningful interaction effect.

The model achieved a Pseudo $R^2$ of 0.098, and the parallel nature of the regression lines demonstrates that automation bias (susceptibility to changing opinion based on AI feedback) operates consistently regardless of an individual's impostor bias level.

## 4. Analysis and Comparison of Participant Comments and Grad-CAM Activations

In this section, we analyze and compare the qualitative feedback obtained from human participants with the regions highlighted by Grad-CAM in our ResNet-101 classification model. Grad-CAM is a technique that generates visual explanations for decisions from CNNs by using the gradients of any target concept flowing into the final convolutional layer to produce a localization map highlighting important regions in the image for predicting the concept [65].

*4.1. Data Extraction Methodology.* Observations from 75 digital images were systematically collected via an optional free-text field provided with each image in the survey. The corresponding columns, labeled *Hai qualche osservazione?*

**TABLE 5**: Mean accuracy, doubt, and opinion change percentages by department for Variant $\alpha$. A one-way ANOVA revealed no statistically significant differences among departments for accuracy ($F(6, 249) = 2.09$, $p = 0.055$, $\eta p^2 = 0.048$), doubts ($F(6, 249) = 1.30$, $p = 0.257$, $\eta p^2 = 0.030$), or changes of opinion ($F(6, 249) = 1.68$, $p = 0.127$, $\eta p^2 = 0.039$).

| Department | Accuracy (%) | Doubts (%) | Changes of opinion (%) |
|---|---|---|---|
| Comp. science | 55.68 | 43.52 | 7.68 |
| Electr. eng. | 53.33 | 50.72 | 10.14 |
| Education | 48.43 | 42.16 | 6.08 |
| Other | 60.42 | 54.17 | 5.00 |
| Law | 57.78 | 55.56 | 0.00 |
| Healthcare | 50.56 | 39.72 | 3.61 |
| Economics | 48.57 | 44.76 | 7.62 |

**TABLE 6**: Mean accuracy, doubt, and opinion change percentages by department for Variant $\beta$. A one-way ANOVA found a significant effect on changes of opinion ($F(6, 143) = 2.65$, $p = 0.018$, $\eta p^2 = 0.100$) but not on accuracy ($F(6, 143) = 1.35$, $p = 0.238$, $\eta p^2 = 0.054$) or doubts ($F(6, 143) = 0.48$, $p = 0.820$, $\eta p^2 = 0.020$).

| Department | Accuracy (%) | Doubts (%) | Changes of opinion (%) |
|---|---|---|---|
| Comp. science | 57.13 | 44.44 | 9.54 |
| Electr. eng. | 50.98 | 36.86 | 16.47 |
| Education | 60.56 | 46.67 | 3.89 |
| Other | 62.22 | 42.96 | 11.85 |
| Healthcare | 53.52 | 47.59 | 5.74 |
| Economics | 46.67* | 63.33* | 3.33* |
| Law | 46.67* | 40.00* | 0.00* |

*Note:* Groups with fewer than three participants (economics and law) were excluded from the ANOVA due to insufficient sample size and are reported here for descriptive purposes only. Asterisks (*) indicate these groups.

*(facoltativo)* [eng. *Do you have any observations? (optional)*] with sequential numeric suffixes, were preprocessed by removing nontextual entries, null or empty strings, and extraneous whitespace. Duplicate responses were then filtered out, preserving each unique entry in its original form. This procedure was implemented in Python using the `pandas` library—leveraging functions such as `dropna()` and `unique()`—to ensure a comprehensive, clean, and semantically faithful dataset.

*4.2. Mapping Participant Insights to Grad-CAM Visualizations.* The five images included in this analysis were selected because they represent some of the most challenging cases in our dataset: they elicited the greatest number of comments from participants and yielded the least accurate probability estimates from the algorithm. Figure 8 provides a qualitative comparison between the classifier's Grad-CAM activation maps and the visual reasoning expressed by participants in their open-ended responses. Each selected

**TABLE 7**: Mean accuracy, doubt, and opinion change percentages by department for Variant $\gamma$. A one-way ANOVA found no statistically significant effects on any metric: accuracy ($F(6, 135) = 1.41$, $p = 0.214$, $\eta p^2 = 0.059$), doubts ($F(6, 135) = 1.21$, $p = 0.307$, $\eta p^2 = 0.051$), or changes of opinion ($F(6, 135) = 1.35$, $p = 0.238$, $\eta p^2 = 0.057$).

| Department | Accuracy (%) | Doubts (%) | Changes of opinion (%) |
|---|---|---|---|
| Comp. science | 61.99 | 46.08 | 11.46 |
| Electr. eng. | 53.85 | 39.49 | 12.82 |
| Education | 62.03 | 49.28 | 7.25 |
| Other | 66.67 | 42.67 | 10.67 |
| Healthcare | 57.19 | 55.09 | 6.84 |
| Economics | 51.67 | 50.00 | 15.00 |
| Law | 50.00* | 86.67* | 20.00* |

*Note:* The law group ($n = 2$) was excluded from the ANOVA due to insufficient sample size and is reported here for descriptive purposes only. Asterisks (*) indicate this group.

image includes comments in which users explicitly referred to perceived visual anomalies (e.g., "hands look distorted," "light reflections are unnatural," and "skin texture is too smooth"), which allow us to assess whether the classifier and participant focused on the same visual regions.

In the Grad-CAM heatmaps, red areas indicate the regions that contributed most to the classifier's decision, while blue or cooler areas contributed less or were completely ignored. This visual coding helps reveal where the model "looks" when assigning a label. It is interesting to observe a systematic divergence between classifier attention and human perceptual strategies. Grad-CAM often highlights low-level visual regions such as backgrounds, abrupt texture changes, or color transitions, elements rarely mentioned by participants. In contrast, users focused on high-level semantic inconsistencies and contextual cues, such as anatomical distortions, lighting consistency, or facial symmetry, which were not emphasized by the model.

This discrepancy suggests that the classifier pays attention to statistically informative but perceptually opaque regions, while humans rely on high-level intuitive semantic cues. The lack of overlap between model and human attention raises important questions about the interpretability of algorithmic decisions in real-world contexts. In particular, it shows that explanations based on attention maps may not bridge the gap between human and automatic reasoning. Rather than reinforcing trust, these visualizations may highlight a deeper disconnect, which should be carefully considered when employing AI systems in user-facing applications, where transparency and alignment with the user are critical.

*4.2.1. Variant α—Image 1 (Young Man Seated Holding Churro With Milkshake in Restaurant, R).* Participants offered a rich set of observations. Several comments focused on specific local details (e.g., the convincing gaze, unusual artifacts on the board, and atypical rendering of accessories such as the necklace and hand positioning). Notably, comments regarding the readability of text and the lighting

**TABLE 8**: Mean accuracy, doubt, and opinion change percentages by department for Variant $\delta$. A one-way ANOVA found a significant effect on changes of opinion ($F(6, 99) = 2.52$, $p = 0.026$, $\eta p^2 = 0.133$) but not on accuracy ($F(6, 99) = 0.76$, $p = 0.601$, $\eta p^2 = 0.044$) or doubts ($F(6, 99) = 0.43$, $p = 0.856$, $\eta p^2 = 0.026$).

| Department | Accuracy (%) | Doubts (%) | Changes of opinion (%) |
|---|---|---|---|
| Comp. science | 59.46 | 47.07 | 14.69 |
| Education | 58.00 | 38.00 | 8.00 |
| Healthcare | 60.54 | 51.35 | 5.41 |
| Electr. eng. | 48.33 | 46.67 | 18.33 |
| Economics | 48.89 | 55.56 | 6.67 |
| Law | 26.67* | 53.33* | 0.00* |
| Other | 43.33* | 33.33* | 20.00* |

*Note:* The law and other groups ($n < 3$) were excluded from the ANOVA due to insufficient sample size and are reported here for descriptive purposes only. Asterisks (*) indicate these groups.

**TABLE 9**: Mean accuracy, doubt, and opinion change percentages by department for Variant $\epsilon$. A one-way ANOVA found no statistically significant effects on any metric: accuracy ($F(6, 85) = 1.27$, $p = 0.279$, $\eta p^2 = 0.082$), doubts ($F(6, 85) = 1.04$, $p = 0.407$, $\eta p^2 = 0.068$), or changes of opinion ($F(6, 85) = 1.70$, $p = 0.132$, $\eta p2 = 0.107$).

| Department | Accuracy (%) | Doubts (%) | Changes of opinion (%) |
|---|---|---|---|
| Comp. science | 60.83 | 46.50 | 12.17 |
| Education | 59.26 | 40.74 | 7.41 |
| Other | 64.44 | 40.00 | 20.00 |
| Healthcare | 57.37 | 52.73 | 6.67 |
| Electr. eng. | 45.00 | 43.33 | 23.33 |
| Economics | 50.00* | 76.67* | 6.67* |
| Law | 46.67* | 6.67* | 13.33* |

*Note:* The economics ($n = 2$) and law ($n = 1$) groups were excluded from the ANOVA due to insufficient sample size and are reported here for descriptive purposes only. Asterisks (*) indicate these groups.

effects suggest sensitivity to artifacts typically associated with AI generation. In contrast, the Grad-CAM visualization concentrated predominantly on the left side of the image, specifically targeting the background region around the board. This disparity is significant because the classifier's focus on the board's boundary details seems to have been a key factor in predicting an 86% probability of AI generation, despite the diversity of human observations.

*4.2.2. Variant α—Image 4 (Two Men Kissing in Dimly Lit Nightclub, MJ).* The participant feedback for this image centered on anomalous lighting, color saturation, and specific anatomical distortions (e.g., misaligned fingers and unrealistic reflections). These comments point to subtle artifacts in the depiction of figures and ambient lighting. The Grad-CAM output, however, predominantly highlighted the hand of the left individual (positioned in the bottom-right area) and a portion of the right individual's face. The limited overlap between human observations and the algorithm's attention is reflected in the 74% AI-generation prediction, underlining a divergence between human scrutiny and machine-learned features.

*4.2.3. Variant β—Image 3 (Elderly Man Repairing Fishing Net by the Water, MJM).* For this image, feedback was more restrained, with participants noting specific aspects such as the unusual color tone of the hands and hair and an overall smoothness that appeared less natural. In contrast, the Grad-CAM analysis revealed a primary focus on the lower right segment of the image—particularly on the fishing net and parts of the subject's hands—with a secondary focus on the face. The relatively low 24% AI-generation probability assigned by the classifier suggests that the localized features it deemed important (i.e., inconsistencies in the net and hand details) were less pronounced than the broader textural anomalies flagged by participants.

*4.2.4. Variant β—Image 5 (Young Men Performing on Stage, MJ).* Participant observations for this dynamic image focused on distorted textual elements (e.g., nonsensical logos and incoherent slogans on clothing) and a general mismatch between visual cues and expected physical properties. In contrast, Grad-CAM attention was concentrated on the foreground microphone and parts of a secondary subject, specifically in the right and lower right regions. With the classifier assigning only a 20% probability of AI generation, the divergence between participant emphasis on textual artifacts and the algorithm's focus on performance-related elements illustrates the different strategies used by human evaluators and the model.

*4.2.5. Variant γ—Image 2 (Young Woman With Brown Hair in Yellow Top Against Teal Wall, MJ).* In this case, participants raised concerns regarding overall image smoothness, such as pixel homogeneity in the subject, and doubts about the background details (e.g., the incompletely rendered painting and inconsistent earring). Grad-CAM activation was mainly concentrated on the face—especially the lower facial region, neck, and part of the hair—indicating that the classifier's attention was focused on the subject's defining features. This focus corresponds to a modest 29% probability of AI generation, suggesting that the classifier relied on more consistent, localized cues even when participants noted broader background anomalies.

## 5. Discussion

*5.1. Content-Specific Patterns in Image Authentication.* The analysis uncovered systematic relationships between image content categories and authentication performance, with distinct accuracy–uncertainty patterns tied to scene characteristics across experimental variants. Structured environments, such as Variant $\delta$'s industrial setting (R, Image 14) and Variant $\beta$'s classroom (R, Image 7), facilitated high accuracy rates (exceeding 82%) likely due to discernible spatial layouts and object relationships that aided technical evaluation. In contrast, emotionally charged or intimate scenes

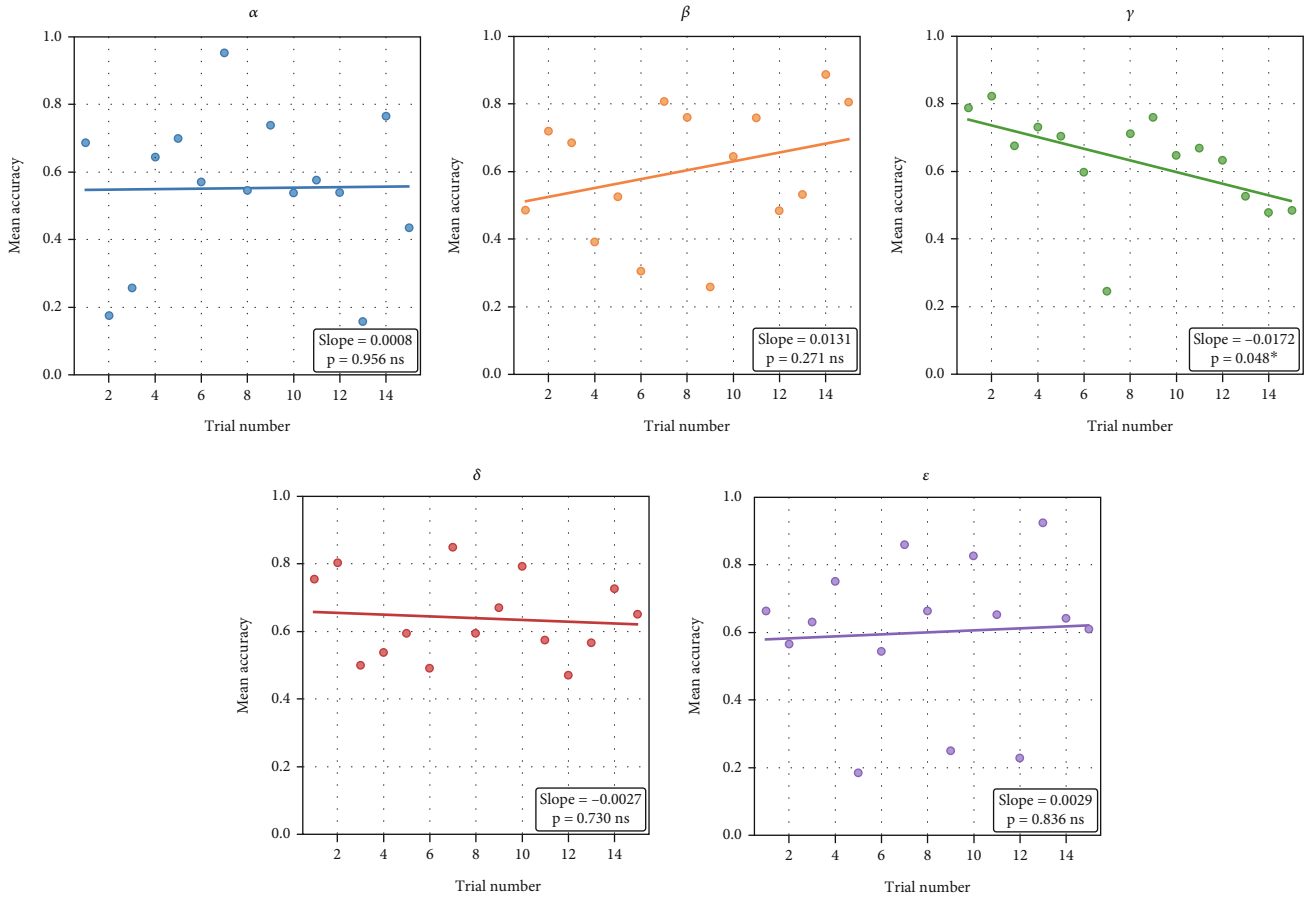



FIGURE 6: Trial accuracy trends across five questionnaire variants (α–ε). Per variant: *points = mean accuracy per trial*, *lines = trend*. Only Variant γ shows significant fatigue (negative slope, $p < 0.001$); all others stable (ns). Slopes: γ ($\beta < 0$, ***), α/β/δ/ε (ns). Demonstrates fatigue occurs only with specific image sequences. $N = 746$ participants (~149/variant), 15 trials each.

in low-light conditions, such as Variant α's nightclub kiss (MJ, Image 4) and Variant γ's silhouetted couple (MJM, Image 12), elicited significantly higher uncertainty, suggesting that affective salience may compete with analytical scrutiny during authentication tasks.

Scenes depicting human interaction revealed a paradoxical pattern. Authentic outdoor landscapes, such as Variant ε's natural setting (Image 11), were identified reliably (89% accuracy), whereas AI-generated social scenes, including Variant γ's human-populated environment (Image 5), triggered disproportionately high uncertainty. This contrast peaked in Variant α's social interaction images, which exhibited a 33% accuracy gap between real and synthetic content, underscoring the challenge of evaluating plausibility in socially complex scenes.

AI-enhanced MJM-processed images introduced unique perceptual conflicts, particularly in portraits and group scenes. For example, Variant β's children's portrait (MJM, Image 1) and Variant δ's street scene (MJM, Image 8) maintained moderate accuracy despite elevated participant uncertainty. This dissociation implies that subtle inconsistencies in AI-enhanced details—such as unnatural textures or lighting gradients—may raise suspicion without providing definitive classification cues.

Technical and artistic content appeared to amplify susceptibility to algorithmic influence, with participants revising judgments more frequently after receiving automated feedback for specialized scenes like Variant γ's gaming setup Image 3) and Variant ε's orchestral performance (Image 12). This aligns with theoretical models of automation bias, where users defer to algorithmic guidance when evaluating domains perceived as requiring expertise. Collectively, these patterns demonstrate that image authentication operates as a context-dependent perceptual process, modulated by content-driven cognitive strategies and fluctuating trust in automated systems.

*5.2. Demographic Influences on Variant Performance.* Our analysis revealed that demographic factors influenced performance in specific and consistent ways, with gender emerging as a significant modulator of automation bias, while age effects were highly localized.

A striking and robust pattern was observed in relation to gender and changes of opinion. Across all five experimental variants, male participants demonstrated a significantly higher propensity to change their initial judgment after receiving an algorithmic suggestion. The magnitude of this effect ranged from small ($d = 0.254$) to large ($d = 0.683$),

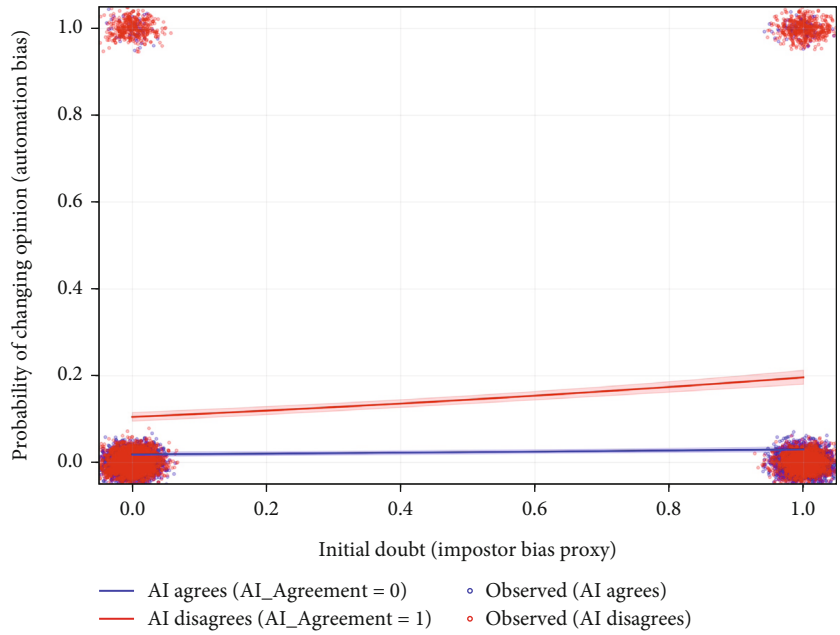



FIGURE 7: Predicted opinion change probability (automation bias) by initial doubt (impostor bias proxy) and AI agreement. Blue line: AI agrees (AI_Agreement = 0); red line: AI disagrees (AI_Agreement = 1). Shaded areas: 95% CIs; points: observed data (jittered). Parallel curves indicate no significant interaction ($\beta = 0.226$, $p = 0.258$) showing AI disagreement consistently increases opinion change probability regardless of doubt level. Model: Pseudo $R^2 = 0.098$; $N = 11{,}170$ (746 participants).

suggesting that males were consistently more susceptible to automation bias than females in this study. This finding may reflect underlying differences in decision-making strategies, risk aversion, or confidence in initial judgments when confronted with algorithmic authority. In contrast, gender's impact on overall accuracy was minimal, reaching statistical significance only in a single instance (Variant $\alpha$).

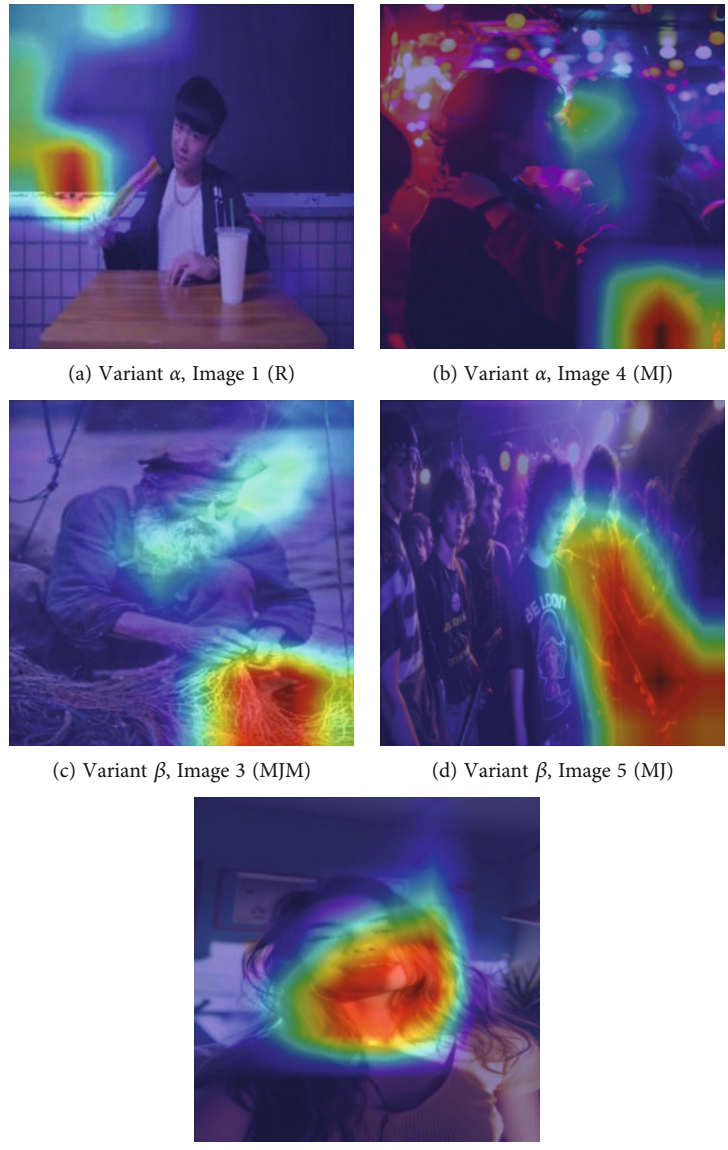
Conversely, age appeared to have a very limited influence on participant behavior. Across nearly all variants and metrics, ANOVAs found no statistically significant differences between age groups. The sole exception was a strong and notable effect on changes of opinion in Variant $\varepsilon$ ($\eta p^2 = 0.185$), where younger participants (specifically 21-year-olds) were substantially more likely to alter their judgment. This isolated finding suggests that the specific stimuli in that variant may have interacted uniquely with generational differences in trust or engagement with technology rather than indicating a general age-related trend. The differential impact of these demographic variables highlights that susceptibility to cognitive biases is not uniform and can be consistently influenced by factors like gender.

*5.3. Interdepartmental Variability.* The analysis of interdepartmental performance indicates that academic background was not a pervasive factor in determining outcomes. Contrary to what might be expected, our one-way ANOVAs revealed no statistically significant differences in accuracy or doubt levels across departments in any of the five experimental conditions. This suggests that the core ability to authenticate images or the general level of skepticism was not significantly shaped by a participant's field of study in this sample.

However, departmental affiliation did have a localized effect on susceptibility to automation bias. We found significant interdepartmental differences in the rate of opinion change in two specific contexts: Variant $\beta$ ($p = 0.018$, $\eta p^2 = 0.100$) and Variant $\delta$ ($p = 0.026$, $\eta p^2 = 0.133$). This finding suggests that while the baseline ability to detect fakes is consistent, the tendency to defer to an algorithmic suggestion can be influenced by academic training, but only when triggered by specific combinations of stimuli. For instance, in both variants, participants from electrical engineering were among the most likely to change their opinion.

While not statistically significant, some descriptive trends are worth noting. The computer science department, for example, consistently achieved descriptively moderate to high accuracy, perhaps reflecting a baseline familiarity with the task. The law and economics departments occasionally showed more extreme patterns, such as the high doubt rate for law students in Variant $\gamma$, though these did not translate into significant overall effects. These results suggest that while disciplinary culture may influence certain decision-making tendencies, it does not appear to confer a general advantage or disadvantage in the fundamental task of image authentication.

*5.4. Comparing Human Perception and ML Model Interpretations.* The comparison of human feedback and Grad-CAM activations in these challenging images reveals a noteworthy divergence in feature prioritization. Human evaluators tend to consider a wide range of perceptual cues—from global compositional inconsistencies to local details such as textual artifacts and anatomical peculiarities—whereas the ResNet-101 classifier appears to rely on



(a) Variant $\alpha$, Image 1 (R)

(b) Variant $\alpha$, Image 4 (MJ)

(c) Variant $\beta$, Image 3 (MJM)

(d) Variant $\beta$, Image 5 (MJ)

(e) Variant $\gamma$, Image 2 (MJ)

FIGURE 8: The five challenging images analyzed in this study along with their corresponding Grad-CAM heatmaps. These images were selected due to their high volume of participant comments and lower prediction accuracy by the algorithm.

more localized features. For example, in Variant $\alpha$, Image 1, the classifier's focus on the board contrasts with the broader human scrutiny, and in Variant $\beta$, Image 5, the textual inconsistencies noted by participants were not as influential for the classifier. These observations underscore the value of integrating human perceptual feedback with machine-based feature attribution to gain a more comprehensive understanding of the artifacts present in AI-generated imagery.

*5.5. Aims and Empirical Validation.* This study was designed to address three primary aims: (i) to quantify human discrimination ability for state-of-the-art synthetic media, (ii) to empirically validate the existence and influence of impostor bias, and (iii) to characterize automation bias as evidenced by trust in algorithmic suggestions. Our empirical findings, supported by the rigorous statistical analyses presented in Section 3, provide a nuanced validation of these aims.

Regarding the first aim, our analysis of accuracy rates across five questionnaire variants ($n = 746$) revealed that participant performance hovered around chance levels, with mean accuracy ranging from 47.0% (Variant $\varepsilon$) to 55.5% (Variant $\beta$). A key finding from our deeper statistical analysis is that none of these accuracy rates were statistically significant from a 50% baseline (all $p$ values $> 0.33$), a conclusion reinforced by wide 95% confidence intervals that overlapped the 50% mark and the application of a Bonferroni correction for multiple comparisons (see Table 3). This result aligns with and extends recent research [21, 22] by demonstrating the profound challenge that modern generative models, like Midjourney v6, pose to human perception. While the results indicate that real photographs consistently

elicited descriptively higher accuracy rates (as seen in Figure 3), the overall ability to reliably discriminate between authentic and synthetic content was not statistically robust.

The second aim, the validation of impostor bias, was supported by our descriptive data. As illustrated by the box plots in Figure 5, the analysis of participant uncertainty revealed a systematic difference in doubt levels between real and AI-generated images. For instance, Variant $\alpha$ showed a wider disparity in doubt levels for real images compared to AI-generated ones, while this pattern shifted in Variants $\gamma$ and $\varepsilon$, where AI-generated images prompted more concentrated levels of uncertainty. This finding suggests that a general awareness of generative AI leads to a pervasive skepticism that disproportionately affects the evaluation of synthetic content, complicating the landscape of digital trust.

The third aim, characterizing automation bias, was also clearly validated. The rate of opinion change following an algorithmic suggestion—our operationalization of automation bias—ranged from 6.2% to 10.7% across variants (see Table 3). These shifts were consistent and impactful. The demographic analysis further reinforced the reality of this bias, showing, for example, that male participants were consistently and significantly more susceptible to it across all five experimental conditions, with effect sizes ranging from small ($d = 0.254$) to large ($d = 0.683$) (see Table 4).

Finally, we explored the interplay between these two biases. Our logistic regression analysis (Section 3.6) confirmed that both initial doubt (a proxy for impostor bias) and a contradictory AI suggestion (AI disagreement) were significant predictors of opinion change. Specifically, higher initial doubt ($\beta = 0.5063$, $p = 0.006$) and the AI disagreeing with the user's initial assessment ($\beta = 1.8259$, $p < 0.001$) both increased the likelihood of a participant changing their opinion. However, we found no significant interaction effect between them ($p = 0.258$). This suggests that impostor bias and automation bias operate as two largely independent forces. As illustrated by the parallel logistic curves in Figure 7 (Section 3.6), a high level of initial skepticism does not appear to insulate a user from the powerful influence of a contradictory algorithmic suggestion; rather, both biases contribute to the complexity of the final judgment.

*5.6. Broader Implications: Ethics, Policy, and Practical Applications.* The European Regulation 2024/1689 (AI Act) seeks to ethically steer the development of artificial intelligence systems, ensuring that both deontological principles and societal norms guide human behavior and computational processes to safeguard life and dignity. While ethics is interpreted differently across various contexts, in continental Europe, it is primarily seen as a means to protect human beings and their fundamental rights. Yet, geographic differences and emotional influences render a unified ethical framework for artificial intelligence impracticable, especially given the amplified deviations—such as bias and hallucinations—observed in GenAI systems, which pose significant risks to ethical values [66, 67].

Deepfake technology, which generates high-quality synthetic content, further complicates trust in legal and media interactions by blending misinformation with factual content. Furthermore, beyond the risks of misinformation, these generative technologies raise complex legal and ethical issues related to copyright, as it is possible to instruct models to mimic an artist's style simply by using their name in the input prompt. This has led to research focused on forensically inferring whether an artist's name was improperly used in the generation process [68]. Given the pervasiveness of deepfakes and their ability to deceive individuals or the public, in order to prevent misunderstandings and misinformation, the AI Act has explicitly addressed this phenomenon and established very stringent regulations for producers and distributors. In the spirit of transparency, they are required to mark (with watermarking technologies) all synthetic content intended for online sharing. It also provides, however, that the disclosure requirement may be relaxed when it could hinder the display or enjoyment of the work, but must still be adequate to disclose the existence of artificial content in cases where it is part of an evidently artistic, creative, satirical, or fictitious program. The European legislator's intervention, therefore, is not aimed at demonizing the phenomenon of deepfakes in the various ways they are implemented, as described in this paper. Rather, it has decided to regulate their use, as with other communication technologies, in order to bring all its manifestations within an ethical dimension. Our findings directly inform this ethical debate by highlighting the specific cognitive vulnerabilities that make such regulations necessary.

For example, our results on automation bias show that people may rely too heavily on algorithmic suggestions. In situations where malicious actors gain control over automated detection systems or content-ranking algorithms, even small manipulations in the outputs could sway public opinion or distort trust in authentic media. Similarly, impostor bias could be exploited to cast doubt on genuine images or videos, which is particularly concerning in the context of journalism, justice, or public health communication. Bad actors could exploit this bias to discredit truthful evidence simply by creating an environment of general suspicion toward visual media.

Our findings suggest that these cognitive vulnerabilities deserve greater attention in the design of AI-based tools, especially those involving automated decision support or misinformation detection. Interfaces should be carefully designed not to encourage blind trust but also not to induce unnecessary skepticism. At a broader level, these insights highlight the importance of media literacy and user education. Raising awareness of these biases is not just a psychological concern, it is a practical step toward making digital platforms safer and more trustworthy.

Translating this awareness into action requires concrete interventions in several key domains. Our findings reveal that even domain experts struggle to consistently identify content generated by artificial intelligence, especially when algorithmic suggestions are involved. This highlights several concrete directions for real-world applications. In the context of education, our experimental design could be adapted into interactive learning tools aimed at training students and professionals to critically evaluate visual media, recognize

cognitive biases such as automation or impostor bias, and develop the ability to resist misleading cues.

Regarding media literacy, our results suggest that traditional training focused on spotting technical artifacts may not be sufficient. Educational programs should incorporate uncertainty management strategies, helping individuals become aware not only of manipulated content but also of their own confidence levels and judgment calibration when faced with ambiguity. Finally, in terms of policy, the presence of measurable behavioral biases requires normative attention. Although content labeling (e.g., synthetic media watermarking) is essential, our study suggests that the "presentation" of algorithmic suggestions and the "context" in which they are provided can significantly influence users' decisions. Policymakers should therefore consider guidelines that account for the psychological dimensions of human–AI interaction, not just technical transparency.

*5.7. Practical Recommendations for Improving Human Detection Abilities.* Our findings point to several practical recommendations for enhancing human resilience to synthetic media, moving beyond simple technical training to focus on metacognitive skills and improved human–AI interaction. Instead of merely teaching users to spot artifacts, which can inadvertently increase impostor bias, a more effective approach is metacognitive training focused on bias awareness. The goal should be to foster calibrated trust, teaching users when to be skeptical and when to trust their own judgment. This can be supported by encouraging context-aware evaluation strategies, where users are prompted to be more vigilant with content types, like emotionally charged scenes, that our data show can impair analytical scrutiny. Furthermore, to mitigate automation bias, the design of AI detection aids is critical. Rather than presenting a single, authoritative probability score, interfaces should encourage critical engagement by visualizing model uncertainty (e.g., via confidence intervals) and providing explanations (e.g., heatmaps) that empower users to verify an algorithmic suggestion rather than blindly accepting it. Ultimately, effective training should be interactive and adaptive, helping users identify their specific biases and calibrate their judgment over time.

*5.8. Limitations and Future Perspectives.* Despite the valuable insights generated by this study, several limitations must be acknowledged, particularly concerning the generalizability of our findings. We structure these limitations around three key areas: the participant sample, the experimental design, and the stimulus dataset.

*5.8.1. Sample Generalizability.* Our recruitment strategy resulted in a convenience sample with a specific demographic profile. The participant pool consisted largely of young university students, with a notable overrepresentation of individuals from computer science and related technical fields. While this allowed us to explore cognitive biases in a population with high exposure to synthetic media, it also limits how directly these results can be generalized to more diverse demographic groups, including older adults or those with lower digital literacy. Future work should aim to replicate these findings across a broader demographic spectrum to evaluate whether similar patterns hold in the general population.

*5.8.2. Design Generalizability and Methodological Considerations.* Our experimental setup adopted a within-subject design, which, while powerful for capturing individual decision shifts, may introduce order effects. As discussed in our sequential effects analysis, the specific ordering of images can induce fatigue, and future studies employing a between-subject design could better isolate the causal impact of algorithmic suggestions. Furthermore, our study is limited by the absence of response time data, which prevented us from analyzing cognitive effort. Other methodological considerations include the potential for demand characteristics from the presentation of the algorithmic assessment and the inherent interpretative biases in our qualitative analysis.

*5.8.3. Stimulus Generalizability.* Although our image dataset was diverse, it represents only a limited subset of real-world scenarios and generative models. We focused on outputs from Midjourney v6 and Magnific AI for the human-facing task to test perception against a state-of-the-art pipeline. While this provides a strong basis for analysis, future work should incorporate a wider range of generators (such as DALL·E and Stable Diffusion) as direct stimuli to test the generalizability of our findings across a greater variety of visual artifacts and styles. Moreover, the ResNet-101 classifier was not trained on Midjourney outputs, a deliberate choice to create a challenging detection scenario but also a factor that limits the direct applicability of its performance metrics to real-world, fine-tuned detectors.

Future research should address these limitations by employing more balanced and representative sampling strategies, expanding the diversity of datasets and generative models, and utilizing mixed-design methodologies to validate our findings.

## 6. Conclusions

Our findings reveal a critical vulnerability in human perception: participants could not distinguish state-of-the-art AI-generated images from real photographs at rates statistically better than chance. This fallibility is compounded by two distinct psychological drivers: impostor bias, which manifests as a systematic difference in doubt applied to real versus synthetic content, and automation bias, a consistent propensity to defer to imperfect algorithmic suggestions. Crucially, demographic factors like gender and academic background did not reliably predict accuracy but were strong indicators of susceptibility to automation bias, suggesting that vulnerability to AI influence is linked to specific user characteristics. Our analysis further shows that these biases operate independently; high initial skepticism does not shield individuals from the influence of a contradictory AI, a finding with urgent implications for real-world detection systems pairing human moderators with AI tools. These results underscore the inadequacy of relying on innate human perception for

image authentication. They highlight an urgent need for public awareness strategies and human-in-the-loop systems designed not just to flag content but to counteract the predictable, yet complex, cognitive biases that define trust and vulnerability in our current digital ecosystem.

## Data Availability Statement

The data generated and analyzed in this study are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Funding

The work of Luca Guarnera and Sebastiano Battiato has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union, NextGenerationEU.

## Endnotes

[1]https://github.com/NVlabs/ffhq-dataset.
[2]a.k.a. stable diffusion: https://github.com/CompVis/stable-diffusion.
[3]https://pytorch.org/vision/stable/models.html.

## References

[1] M. Moltisanti, A. Paratore, S. Battiato, and L. Saravo, "Image Manipulation on Facebook for Forensics Evidence," in *International Conference on Image Analysis and Processing* (Springer International Publishing, 2015), 506–517, https://doi.org/10.1007/978-3-319-23234-8_47.

[2] Z. Epstein, A. Hertzmann, L. Herman, et al., "Art and the Science of Generative AI," *Science* 380, no. 6650 (2023): 1110–1111, https://doi.org/10.1126/science.adh4451.

[3] M. Schubert, M. Lasotta, F. Sahm, W. Wick, and V. Venkataramani, *Evaluating the Multimodal Capabilities of Generative AI in Complex Clinical Diagnostics* (Medrxiv, 2023), https://doi.org/10.1101/2023.11.01.23297938.

[4] M. R. Douglas, "Large Language Models," *Communications of the ACM* 66, no. 8 (2023): 7, https://doi.org/10.1145/3606337.

[5] R. Zhang, D. Zou, G. Cheng, and H. Xie, "Flow in ChatGPT-Based Logic Learning and Its Influences on Logic and Self-Efficacy in English Argumentative Writing," *Computers in Human Behavior* 162 (2025): 108457, https://doi.org/10.1016/j.chb.2024.108457.

[6] M. Balasubramanian and T. Periyaswamy, "Rapid Design Prototyping Using Generative Artificial Intelligence: A Case Study Comparing DALL-E, Midjourney and Firefly," in *International Textile and Apparel Association Annual Conference Proceedings* (Iowa State University Digital Press, 2025).

[7] R. Gozalo-Brizuela and E. Garrido-Merch'an, "ChatGPT Is Not all You Need. A State of the Art Review of Large Generative AI Models" 2023, https://arxiv.org/abs/2301.04655.

[8] C. Saharia, W. Chan, S. Saxena, et al., "Photorealistic Text-to-Image Diffusion Models With Deep Language Understanding," *Advances in Neural Information Processing Systems* 35 (2022): 36479–36494.

[9] A. Ortis, G. M. Farinella, and S. Battiato, "An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges," in *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications, ICETE 2019* (SciTePress, 2019), 296–306, https://doi.org/10.5220/0007909602900300.

[10] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on Visual Sentiment Analysis," *IET Image Processing* 14, no. 8 (2020): 1440–1456, https://doi.org/10.1049/iet-ipr.2019.1270.

[11] M. Shukor, C. Dancette, A. Ram'e, and M. Cord, "Unified Model for Image, Video, Audio and Language Tasks" 2023, https://arxiv.org/abs/2307.16184.

[12] S. Zhang, Q. Fang, Z. Yang, and Y. Feng, "Llava-Mini: Efficient Image and Video Large Multimodal Models With One Vision Token" 2025, https://arxiv.org/abs/2501.03895.

[13] S. Toumaj, A. Heidari, R. Shahhosseini, and N. Jafari Navimipour, "Applications of Deep Learning in Alzheimer's Disease: A Systematic Literature Review of Current Trends, Methodologies, Challenges, Innovations, and Future Directions," *Artificial Intelligence Review* 58, no. 2 (2025): 44, https://doi.org/10.1007/s10462-024-11041-5.

[14] H. Wang, S. Toumaj, A. Heidari, A. Souri, N. Jafari, and Y. Jiang, "Neurodegenerative Disorders: A Holistic Study of the Explainable Artificial Intelligence Applications," *Engineering Applications of Artificial Intelligence* 153 (2025): 110752, https://doi.org/10.1016/j.engappai.2025.110752.

[15] A. Heidari, N. J. Navimipour, S. Zeadally, and V. Chamola, "Everything You Wanted to Know About ChatGPT: Components, Capabilities, Applications, and Opportunities," *Internet Technology Letters* 7, no. 6 (2024): e530, https://doi.org/10.1002/itl2.530.

[16] I. Amerini, M. Barni, S. Battiato, et al., "Deepfake Media Forensics: State of the Art and Challenges Ahead," in *International Conference on Advances in Social Networks Analysis and Mining* (Springer Nature Switzerland, 2025), https://doi.org/10.1007/978-3-031-85386-9_3.

[17] I. Amerini, M. Barni, S. Battiato, et al., "Deepfake Media Forensics: Status and Future Challenges," *Journal of Imaging* 11, no. 3 (2025): 73, https://doi.org/10.3390/jimaging11030073.

[18] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media+ Society* 6, no. 1 (2020): https://doi.org/10.1177/2056305120903408.

[19] D. M. Anstine and O. Isayev, "Generative Models as an Emerging Paradigm in the Chemical Sciences," *Journal of the American Chemical Society* 145, no. 16 (2023): 8736–8750, https://doi.org/10.1021/jacs.2c13467.

[20] R. F. Betzel and D. Bassett, "Generative Models for Network Neuro-Science: Prospects and Promise," *Journal of the Royal Society Interface* 14, no. 136 (2017): 20170623, https://doi.org/10.1098/rsif.2017.0623.

[21] K. Somoray and D. J. Miller, "Providing Detection Strategies to Improve Human Detection of Deepfakes: An Experimental Study," *Computers in Human Behavior* 149 (2023): 107917, https://doi.org/10.1016/j.chb.2023.107917.

[22] M. T. Soto-Sanfiel, A. Angulo-Brunet, and S. Saha, "Deepfakes as Narratives: Psychological Processes Explaining Their Reception," *Computers in Human Behavior* 165 (2025): 108518, https://doi.org/10.1016/j.chb.2024.108518.

[23] L. Guarnera, O. Giudice, F. Guarnera, et al., "The Face Deepfake Detection Challenge," *Journal of Imaging* 8, no. 10 (2022): 263, https://doi.org/10.3390/jimaging8100263.

[24] A. Kumar, D. Singh, R. Jain, D. K. Jain, C. Gan, and X. Zhao, "Advances in DeepFake Detection Algorithms: Exploring Fusion Techniques in Single and Multi-Modal Approach," *Information Fusion* 118 (2025): 102993, https://doi.org/10.1016/j.inffus.2025.102993.

[25] A. L. Pellicer, Y. Li, and P. Angelov, "PUDD: Towards Robust Multi-Modal Prototype-Based Deepfake Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2024), 3809–3817.

[26] O. Pontorno, L. Guarnera, and S. Battiato, "Deepfeaturex Net: Deep Features Extractors Based Network for Discriminating Synthetic From Real Images," in *International Conference on Pattern Recognition* (Springer Nature Switzerland, 2024), 177–193.

[27] M. Casu, L. Guarnera, P. Caponnetto, and S. Battiato, "GenAI Mirage: The Impostor Bias and the Deepfake Detection Challenge in the Era of Artificial Illusions," *Forensic Science International: Digital Investigation* 50 (2024): 301795, https://doi.org/10.1016/j.fsidi.2024.301795.

[28] M. A. Hoque, M. Ferdous, M. Khan, and S. Tarkoma, "Real, Forged or Deep Fake? Enabling the Ground Truth on the Internet," *IEEE Access* 9 (2021): 160471–160484, https://doi.org/10.1109/access.2021.3131517.

[29] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O'Brien, "Fake Images: The Effects of Source, Intermediary, and Digital Media Literacy on Contextual Assessment of Image Credibility Online," *New Media & Society* 21, no. 2 (2019): 438–463, https://doi.org/10.1177/1461444818799526.

[30] P. Wason, "Confirmation Bias," in *The SAGE Encyclopedia of Political Behavior* (SAGE Publications, Inc, 2017), https://doi.org/10.4135/9781483391144.n61.

[31] K. Goddard, A. Roudsari, and J. Wyatt, *Decision Support and Automation Bias: Methodology and Preliminary Results of a Systematic Review* (IOS Press, 2011), https://doi.org/10.3233/978-1-60750-709-3-3.

[32] T. Nguyen, "ChatGPT in Medical Education: A Precursor for Automation Bias?," *JMIR Medical Education* 10 (2024): e50174, https://doi.org/10.2196/50174.

[33] A. Potasznik, "ABCs: Differentiating Algorithmic Bias, Automation Bias, and Automation Complacency," in *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* (IEEE, 2023), 1–5, https://doi.org/10.1109/ETHICS57328.2023.10155094.

[34] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *International Journal of Aviation Psychology* 8, no. 1 (1998): 47–63, https://doi.org/10.1207/s15327108ijap0801_3.

[35] D. Holz, "Midjourney— Independent Research Lab" 2024, https://www.midjourney.com/home.

[36] J. Lopez and E. Nicolas, "Magnific AI—The Magic Image Upscaler & Enhancer" 2024, https://magnific.ai.

[37] A. J. Adetayo, "Reimagining Learning Through AI Art: The Promise of DALL-E and MidJourney for Education and Libraries," *Library Hi Tech News* (2024): https://doi.org/10.1108/LHTN-01-2024-0005.

[38] T. P. Aliya, A. T. R. Aurelia, N. S. Najmi, F. A. Apsarini, and N. A. Rakhmawati, "Observation of AI Text to Image Usage on the Credibility of Visual Artworks," *Journal of Information System, Informatics and Computing* 7, no. 2 (2023): 387, https://doi.org/10.52362/jisicom.v7i2.1270.

[39] S. Tanugraha, "Review Using Artificial Intelligence-Generating Images: Exploring Material Ideas From MidJourney to Improve Vernacular Designs," *Journal of Artificial Intelligence in Architecture* 2, no. 2 (2023): 48–57, https://doi.org/10.24002/jarina.v2i2.7537.

[40] S. Göring, R. R. Ramachandra Rao, R. Merten, and A. Raake, "Analysis of Appeal for Realistic AI-Generated Photos," *IEEE Access* 11 (2023): 38999–39012, https://doi.org/10.1109/ACCESS.2023.3267968.

[41] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, "Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection," *IEEE Transactions on Information Forensics and Security* 19 (2024): 401–413, https://doi.org/10.1109/TIFS.2023.3324739.

[42] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deep-Fake Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2024), 28130–28139.

[43] R. Xia, D. Liu, J. Li, L. Yuan, N. Wang, and X. Gao, "Mmnet: Multi-Collaboration and Multi-Supervision Network for Sequential Deepfake Detection," *IEEE Transactions on Information Forensics and Security* 19 (2024): 3409–3422, https://doi.org/10.1109/TIFS.2024.3361151.

[44] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending Forgery Specificity With Latent Space Augmentation for Generalizable Deepfake Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2024), 8984–8994.

[45] S. Concas, G. Perelli, G. L. Marcialis, and G. Puglisi, "Tensor-Based Deepfake Detection in Scaled and Compressed Images," in *2022 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2022), 3121–3125, https://doi.org/10.1109/ICIP46576.2022.9897606.

[46] C. T. Doloriel and N.-M. Cheung, "Frequency Masking for Universal Deepfake Detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), 13466–13470, https://doi.org/10.1109/ICASSP48485.2024.10446290.

[47] O. Pontorno, L. Guarnera, and S. Battiato, "On the Exploitation of DCT-Traces in the Generative-AI Domain," in *2024 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2024), 3806–3812, https://doi.org/10.1109/ICIP51287.2024.10648013.

[48] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake Style Transfer Mixture: A First Forensic Ballistics Study on Synthetic Images," in *International Conference on Image Analysis and Processing* (Springer International Publishing, 2022), 151–163, https://doi.org/10.1007/978-3-031-06430-2_13.

[49] S. Xie, R. Girshick, P. Doll'ar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVF, 2017), 1492–1500.

[50] L. Guarnera, O. Giudice, and S. Battiato, "Mastering Deepfake Detection: A Cutting-Edge Approach to Distinguish GAN and Diffusion-Model Images," *ACM Transactions on Multimedia Computing, Communications and Applications* 20, no. 11 (2024): 1–24, https://doi.org/10.1145/3652027.

[51] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proceedings of the IEEE*

*International Conference on Computer Vision* (CVF, 2015), 3730–3738.

[52] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115, no. 3 (2015): 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[53] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing* 28, no. 11 (2019): 5464–5478, https://doi.org/10.1109/TIP.2019.2916751.

[54] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision* (CVF, 2017), 2223–2232.

[55] W. Cho, S. Choi, D. K. Park, I. Shin, and J. Choo, "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 (CVPR)* (CVF, 2019), 10639–10647.

[56] K. Li, T. Zhang, and J. Malik, "Diverse Image Synthesis From Semantic Layouts via Conditional IMLE," in *Proceedings of the IEEE International Conference on Computer Vision* (CVF, 2019), 4220–4229.

[57] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation" 2018, https://arxiv.org/abs/1710.10196.

[58] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVF, 2018), 8789–8797.

[59] Y. Choi, Y. Uh, J. Yoo, and J. W. Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2020), 8188–8197.

[60] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2019), 4401–4410.

[61] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2020), 8110–8119.

[62] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation With Clip Latents" 2022, https://arxiv.org/abs/2204.06125.

[63] A. Q. Nichol, P. Dhariwal, A. Ramesh, et al., "Glide: Towards Photorealistic Image Generation and Editing With Text-Guided Diffusion Models," in *International Conference on Machine Learning* (PMLR, 2022), 16784–16804.

[64] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVF, 2022), 10684–10695.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision* 128, no. 2 (2020): 336–359, https://doi.org/10.1007/s11263-019-01228-7.

[66] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An Overview of Artificial Intelligence Ethics," *IEEE Transactions on Artificial Intelligence* 4, no. 4 (2023): 799–819, https://doi.org/10.1109/TAI.2022.3194503.

[67] J. A. Nasello and J. M. Triffaux, "The Role of Empathy in Trolley Problems and Variants: A Systematic Review and Meta-Analysis," *British Journal of Social Psychology* 62, no. 4 (2023): 1753–1781, https://doi.org/10.1111/bjso.12654.

[68] R. Leotta, O. Giudice, L. Guarnera, and S. Battiato, "Not With My Name! Inferring Artists' Names of Input Strings Employed by Diffusion Models," in *International Conference on Image Analysis and Processing* (Springer Nature Switzerland, 2023), 364–375, https://doi.org/10.1007/978-3-031-43148-7_31.

[69] Y. Apolo and K. Michael, "Beyond a Reasonable Doubt? Audiovisual Evidence, AI Manipulation, Deepfakes, and the Law," *IEEE Transactions on Technology and Society* 5, no. 2 (2024): 156–168, https://doi.org/10.1109/TTS.2024.3427816.

[70] C. F. Lyon, "Fake Cases, Real Consequences: Misuse of ChatGPT Leads to Sanctions. NYSBA NYLitigator 28" 2023, https://www.tinyurl.com/23qjl75b.