Computers in Human Behavior 176 (2026) 108876

Contents lists available at ScienceDirect

Computers in Human Behavior

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

The illusion of reality: How AI-generated images (AIGIs) are fooling social media users

Márk Miskolczi

Corvinus University of Budapest, Institute of Sustainable Development, Közraktár u. 4-6., H-1093, Budapest, Hungary

ARTICLE INFO

Keywords:
AI-Generated images [AIGIs]
Social media
Grounded Theory [GT]
Cognitive biases
User deception

ABSTRACT

This study investigates the deceptive impact of AI-generated images [AIGIs] on social media, focusing on content features, user engagement, and underlying psychological mechanisms. By analyzing Facebook pages and user reactions [n = 8922], this research addressing a significantly underexplored research area, identifies key AIGI attributes and examines the most common user reactions. Findings reveal that cognitive biases play a remarkable role in amplifying the deceptive power of AIGIs, fostering users' trust in fabricated content. The study explains a novel method for detecting AIGIs and AI bots in social media environment, and with the extension of Grounded Theory [GT] coding process, a robust framework for future research is also presented. Highlighting the growing threat of AIGIs in suppressing reality perception, this study calls for rapid regulatory measures, increased digital literacy and systematic monitoring to tackle the spread of the phenomenon.

1. Introduction

In recent years, the rapid progress of generative Artificial Intelligence [genAI] has enormously changed the way visual content is produced. AI-generated images [AIGIs] seem to become the part of our everyday digital environment, especially on social media, where they are easily accessible and widely shared (DiResta – Goldstein, 2024). Facebook, as one of the largest social media websites of Meta Platform Inc., provides a breeding ground for the dissemination of such images, due to algorithms that favor content that elicits an emotional response (Bucher, 2019). Despite Meta's announcement in early 2024, that AI-generated content on social media platforms would be labeled (Aboutfb.com, 2024), this remains unrealized, and the proliferation of such content is escalating. The spread of AI-generated images calls into question the credibility of social media (DiResta – Goldstein, 2024), as users often do not realize that they are encountering artificially produced content. Previous research also raised concerns that genAI tools have several negative impacts (e.g., misuses – Ferrara, 2024; phishing content – Basit et al., 2021; or using it for misleading political propaganda – Perach et al., 2023; Goldstein et al., 2024). Furthermore, manipulation techniques that trigger emotional reactions may contribute to the spread of disinformation and a loss of trust in online platforms (Iosifidis & Nicoli, 2020). To date, little is known about the mechanisms by which AIGIs evoke engagement on mainstream social platforms. The deceptive power of AIGIs is increasingly worrying, especially as they often appear alone or with misleading captions, making their impact stronger and still not fully understood.

Based on these, this study aims to uncover the mechanisms driving the spread of AI-generated images on Facebook, focusing on the deceptive power of such content from the users' perspective, highlighting key concerns and implications. This paper's primary contribution lies in offering empirically grounded insights into the thematic, emotional, and deceptive patterns of AI-generated images, by inductively constructing a typology based on visual and symbolic analysis. For this, a multi-step [extended Grounded-theory] analysis has been conducted. The study also discusses the dangers of the uncontrolled spread of AIGIs, such as the potential increase in distrust of AI tools, and the long-term socio-psychological impacts of disinformation. The study thus draws attention to an increasingly pressing issue that has received limited attention in scientific research and underscores the urgent need for its comprehensive investigation and resolution.

The study is structured as follows: Section 2 reveals the background of AI-content creation business (content farms), highlighting the motivations and ethical issues behind the phenomenon. This section also interprets the most relevant theories on consumer deception that may influence perceptions of AIGIs, along with the key research gaps identified in the current literature. Section 3 introduces the research methodology and research questions (RQs) raised. Section 4 summarizes the key findings along with the main steps of the analysis, and Section 5 concludes the key insights and raises concerns related to the

E-mail address: mark.miskolczi@uni-corvinus.hu.

https://doi.org/10.1016/j.chb.2025.108876
Received 12 June 2025; Received in revised form 7 November 2025; Accepted 20 November 2025
Available online 24 November 2025

phenomenon from both theoretical and practical aspects.

2. Background

2.1. AI-generated contents – technological background

The development of Artificial Intelligence [AI] over the last decade has disrupted digital content creation, especially with the emergence of genAI models. AIGIs rely on solutions that can create content that did not exist before, such as text, images, videos or musical compositions (Chiu, 2023). Two main approaches to generative modelling are generative adversarial networks [GANs] and large language models [LLMs], such as GPT-based systems (Kasneci et al., 2023). These technologies open new perspectives in content production, but also raise social, ethical and economic issues. LLMs, such as GPT (Generative Pre-trained Transformer) (Zhao et al., 2023) are based on large amounts of textual data. These models can produce natural language text, generate answers to questions, or even write coherent texts that are almost indistinguishable from human writing. Such systems use deep learning [DL] algorithms that analyze the input data in context and generate new content based on it (Zhao et al., 2023). The main attractions of AI-generated content are efficiency and cost reduction. These technologies have revolutionized the creative industries, marketing and content production by enabling the rapid production of personalized content (Wahid et al., 2023). One of the largest areas of use is the social media; they are used to produce clickbait content, creative posts or even "viral" campaigns. Increasingly, AIGIs are being seen on Meta platforms to increase user interaction. Scholars (e.g., DiResta – Goldstein, 2024) highlight that AI-generated content can easily mislead users, especially if it is intended to spread disinformation or manipulate them. For example, deepfake technologies can create realistic but fake images and videos, which pose serious ethical and security risks. In addition, content automation can contribute to the devaluation of digital content, as mass-produced, low-cost products can overshadow high-quality, human creations (Westerlund, 2019).

Among the different forms of AI-generated contents, visual stimuli stand out due to their basic nature. Images are processed faster and more holistically than text or audio (Barry, 1997; Messaris, 1994), making them more effective in capturing attention and triggering emotional responses. Images also activate affective decision pathways more directly than verbal stimuli (Brady, Konkle, Alvarez, & Oliva, 2008). In addition, image-based posts tend to receive remarkably more engagement on social media than textual content (Eslami et al., 2015). This facilitates rapid dissemination, especially where visuals can be shared without language barriers or literacy. While emotionally charged clickbait, influencer content, or sensational videos also exploit users' affective responses, these are typically grounded in real-world events (Wang et al., 2025). In contrast, AI-generated images operate without such constraints: they can be synthetically produced at scale, often detached from reality, and yet visually optimized to trigger strong emotional responses. They are typically characterized by hyperrealistic textures, symbolic exaggerations (e.g., idealized cities), visual inconsistencies (e.g., unnatural light), and emotionally loaded composition that collectively enhance their persuasive potential (DiResta – Goldstein, 2024). This makes them particularly effective in environments where emotional resonance overrides fact-checking or contextual awareness (DiResta – Goldstein, 2024). Furthermore, unlike AI-generated text or audio, which typically require more effort to be convincing, AI-generated images often appear credible even in isolation or when paired with misleading captions (Vosoughi et al., 2018). Given these concerns, it is particularly important to examine AI-generated images through the lens of their persuasive power on social media platforms, which has so far received limited scholarly attention.

2.2. Content farms – the engine of AIGCs

AI-generated images are usually created on content farms; however, the scientific literature on this topic remains notably limited. Consequently, this study also utilizes grey literature to provide a more comprehensive understanding of the phenomenon. AI-driven content farms (other terms: clickbait farm, fake content farm, ad farm, engagement bate site, content mill, spam content site) are websites that create AI-generated content, such as text and images, typically in high volumes and low quality. The main goal of these farms is to maximize advertising revenue through advertising systems that automatically place ads on different websites (Chen et al., 2022). According to the MIT Technologyreview.com (2023), these sites are often full of incoherent, or even misleading content that relies solely on search engine optimization (SEO) and ad serving. The promotion of such farms can contribute to the spread of disinformation and a loss of trust in social media and online content. With the help of AI tools such as ChatGPT (https://chatgpt.com/), DALL-E (https://askaichat.app/) or Midjourney (https://www.midjourney.com), these farms generate keyword-oriented articles and visual content. The main purpose of content production is to drive traffic to pages where ads are displayed. These systems, such as Google AdSense, place ads automatically, often without advertisers being aware of where their ads are appearing (Chen et al., 2022). AI-generated contents thus attempt to generate interaction through emotional manipulation, clickbait titles and engagement bait tactics (e.g., Facebook post with title such as "You will not believe what happened to this family!") to encourage users to click. Content farms often use low-cost labor or fully automated systems and typically operate in countries where digital work is cheaper and less regulated (e.g., India, Bangladesh, Philippines) (MIT Technology Review, 2023). However, with the spread of AI technology, they have also spread to more developed countries where content automation is carried out with a high level of technological support. Content farms often run multiple websites and social media accounts in parallel to distribute their content more widely. A single popular article or image is shared across multiple platforms to maximize reach (Chen et al., 2022). For example, a farm may run several Facebook pages and distributing the same AI-generated content. Compounding the problem, such content is often misleading and contributes to the spread of disinformation.

As such, while content farms have been examined in the academic literature (e.g., Chen et al., 2022), their evolving role, especially in the context of genAI remains significantly underexplored. The rapid progress in genAI have led to a remarkable increase in content production (Raj et al., 2024), yet the mechanisms driving this growth, and its broader social consequences have received limited scholarly attention. This gap underscores the necessity for more in-depth research to uncover the adverse effects behind this accelerated phenomenon.

2.3. AIGIs in social media

As AIGIs have become widespread in social media, some recent research has begun to study the responses of platforms and their algorithms to AIGI detection mechanisms. For example, platforms like Facebook, Instagram, Twitter, and TikTok employ a combination of algorithmic tools and human moderation to identify and manage synthetic content (Ara et al., 2024). On Twitter, one study found that 0.05 % of the profile photos were AI-generated. This reveals an existence of AI-generated visual content that is noticeable but still underestimated (Ricker et al., 2024). Other studies revealed that AI-generated works of art and images in portrait style have changed users' viewing patterns and their aesthetical point of reference (Zhou & Lee, 2024; Wei & Tyson, 2024, October). Although these contributions bring a wealth of valuable information to our understanding around AIGIs, they tend either to focus on policy issues from a platform point of view or else concentrate completely on content detection. They do not therefore directly address the affective and deceptive power of AIGIs in users' emotional

responses. In addition, previous research (e.g., Romele, 2022) has generally seen AIGIs as just a technical or ethical problem rather than examining their social effect. Overall, based on the current literature, we know little about how emotionally charged AI-generated images influence public perception and weaken critical judgment on mainstream platforms.

2.4. Cognitive biases – the drivers of deception

Heuristics and cognitive biases have a broad literature, they play a crucial role in shaping human decision-making under conditions of uncertainty or limited information (Acklin, 2015). Heuristics are mental shortcuts that individuals rely on to make complex decisions more manageable (Shad & Olukemi, 2024). While these strategies often lead to efficient decision-making, they can also introduce systematic errors, known as cognitive biases, which skew perception and judgment (Krawiec & Śliwowski, 2024). Understanding these biases is particularly relevant in the context of AI-generated content analysis, where such content is designed to exploit emotional and cognitive tendencies to influence user responses.

Scholars (McConnell et al., 1994; Murata et al., 2015; Felfernig et al., 2018; Fiedler et al., 2022; French et al., 2023) distinguish several cognitive biases, but six has a distinctive importance in decision-making. Confirmation bias (1) refers to the tendency to seek, interpret, and recall information in a way that confirms pre-existing beliefs or values (French et al., 2023). This bias reinforces subjective perspectives, often at the expense of objectivity, leading individuals to overlook contradictory evidence. Anchoring (2) occurs when individuals rely on the first piece of information they encounter (the "anchor") when making decisions (Felfernig et al., 2018). The familiarity effect (3) describes the preference for stimuli that evoke past experiences (Whittlesea, 1993). Repeated exposure to familiar themes or visuals fosters a sense of comfort and trust, reducing the likelihood of critical thinking. Novelty bias (4) reflects the human attraction to new and unique experiences. These biases trigger enhanced attention and emotional reactions to unusual or unexpected stimuli (Horstmann & Herwig, 2016). Groupthink (5) refers to the tendency of individuals within a group to conform to the prevailing opinions or behaviors, often at the expense of critical analysis (Murata et al., 2015). Illusory correlation (6) occurs when individuals perceive a connection between two unrelated elements due to their co-occurrence (McConnell et al., 1994).

While these cognitive biases are well-documented in the context of decision-making and media interpretation, their specific activation by AI-generated images remains largely unexplored. Studies (Romele, 222, Simões & Caldeira, 2024) focusing on AIGIs typically emphasize detection or ethical concerns, without systematically addressing the psychological mechanisms that make these images persuasive or misleading. As a result, there is limited understanding of how heuristics might be exploited by visual synthetic content on social media platforms.

2.5. Research gap

Based on the literature, research gaps are identified as follows:

- Although AI-generated content has attracted growing academic interest, the deceptive power of AIGIs especially within social media environments remains underexplored. While earlier studies have addressed AI-generated texts or multimodal content in general (e.g., Vosoughi et al., 2018), images have received less attention. DiResta – Goldstein (2024) acknowledged the presence of AIGIs on platforms like Facebook, but without a deep analysis of their psychological impact.
- Despite the growing spread of AIGIs, especially content produced and distributed by AI-based content farms, current scientific research has not yet systematically examined which visual themes dominate this phenomenon and how these themes trigger emotional responses that can override users' critical thinking.
- Furthermore, although there is a growing body of work on cognitive biases in the context of misinformation (e.g., confirmation bias, emotional reasoning), these theories have rarely been applied to AI-generated visual content. The unique psychological dynamics that unfold when users respond to synthetic, yet emotionally potent imagery have not been sufficiently theorized or empirically tested.
- Another underexamined dimension is the long-term consequence of deceptive AIGIs on user trust and AI–human interaction. When users later discover that emotionally moving content was in fact artificially generated, it may damage not only trust in digital content but also in AI technologies more broadly. To date, no study has explicitly addressed this erosion of epistemic trust from a user-centered, platform-specific perspective.

This study therefore addresses a critical gap by:

- developing a visual typology of frequently recurring AIGI themes;
- analyzing user responses to these themes through the lens of emotional engagement and cognitive heuristics;
- and proposing a detection framework that, unlike black-box AI detectors, offers interpretability and educational potential.

By combining visual analysis with socio-psychological interpretation, this research lays the foundation for both academic reflection and practical intervention. The typology and findings may not only guide future studies on synthetic media deception but also support user-oriented digital literacy initiatives aimed at mitigating the risks of AI-generated misinformation.

3. Data and methods

Building on the research gaps identified, the following research questions (RQs) are formulated:

- **RQ1:** *What are the main thematic categories of AI-generated images (AIGIs) shared on Facebook, based on their visual motifs and symbolic aesthetics (i.e., recurring visual content types)?*
- **RQ2:** *How do social media users emotionally and cognitively respond to different thematic categories of AI-generated images (AIGIs)?*
- **RQ3:** *Which deception mechanisms [cognitive biases] can be inferred from social media user interactions with AI-generated images?*
- **RQ4:** *Which thematic categories of AI-generated images are most and least effective in deceiving social media users?*

For answering RQs, data collection and analysis were carried out according to the guidelines of the Grounded Theory [GT] method developed by Corbin-Strauss (1990), with additions (Fig. 1).

3.1. PH.1 data collection

Data collection has been carried out in two phases:

- *PH.1.1: Immersion in data*: Social Media Observation [SMO]
  - o *1.1.A: Randomized search:* As a first step, a social media observation has been conducted. This implies a randomized Facebook Feed monitoring, i.e. multiple refreshes of the Feed to detect unknown, presumably AI-generated content [Page and posts]. The process led to the identification of an image displaying characteristics suggestive of AI generation (*see Attachment A1 for details*). To define the content's nature, a two-step validation procedure has been developed (*details in section: PH.1.2.A: Theoretical sampling*).
  - o *1.1.B – Systematic search*: Once an AI-generated content was successfully identified, a keyword search based on the identified picture is conducted on Facebook's public search platform using

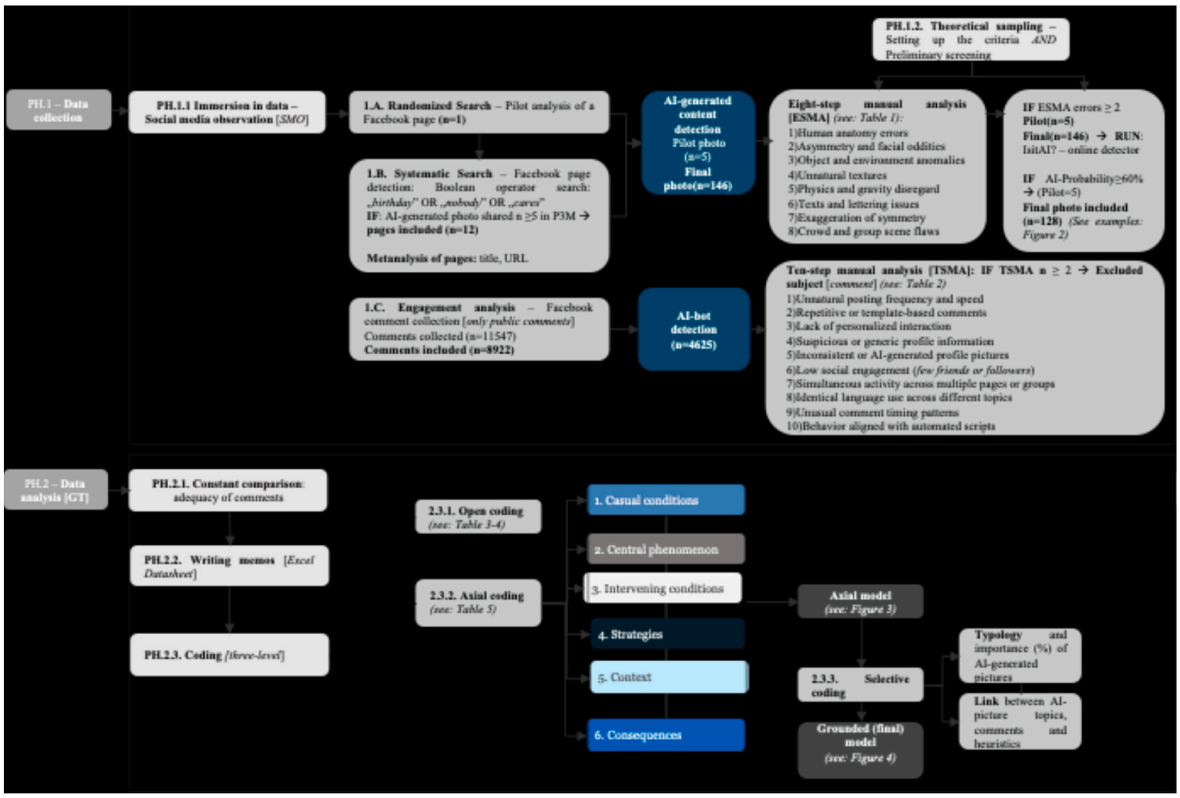

Fig. 1. Methodological background.
Source: Author's own editing

the following keywords and Boolean operators: "*birthday*" OR "*no one*" OR "*wish*". The search revealed several different Facebook pages with very similar posts. From this list, pages with different themes were selected to ensure a diverse data set. Pages (n = 12) with at least five posts in past three months [P3M] including AI-generated images were collected for further analysis (*list of pages: see attachments: A2*). In line with Grounded Theory principles (Charmaz, 2006), the sampling process followed an iterative and theoretically informed approach. New pages were included until no novel types of AI-generated images or visual distortions emerged during manual coding, indicating that theoretical saturation had been achieved. To reduce sampling bias and increase thematic diversity, pages were intentionally selected to cover a wide range of topics (e.g., rural nostalgia, elderly care, spirituality, craft culture, everyday tragedies). This purposive selection strategy ensured that no single thematic content would dominate the sample. Instead of aiming for representativeness in a statistical sense, the goal was to capture a variety of recurring AIGI motifs that circulate in Facebook's public ecosystem. Thematic categories later used in the coding phase were not predefined but emerged inductively through open and axial coding procedures, consistent with Grounded Theory principles (Charmaz, 2006).

- *PH.1.2.A: Theoretical sampling*: Once the pages were selected, additional AI-generated images were manually collected in two phases:
  - o *Eight-step manual analysis [ESMA]:* The Eight-step Manual Analysis (ESMA) was developed following Grounded Theory principles, where analytical categories and tools emerge inductively from the data rather than being predefined. This iterative approach ensured that the coding framework remained sensitive to the visual and contextual nuances of AI-generated image content on social media. While the ESMA was developed by the researcher during the analytical process, these eight steps were later formalized and applied consistently across the dataset. This emergent design aligns with the GT's emphasis on theoretical sensitivity and methodological reflexivity (Charmaz, 2006; Corbin & Strauss, 2015). Recurrent visual distortions, such as unnatural anatomy, implausible textures, or logical inconsistencies were categorized into eight main signal types. These categories align with established visual communication and perception theories that emphasize how humans detect anomalies in imagery (Barry, 1997; Messaris, 1994). Thus, the framework is both empirically grounded and theoretically informed. If at least two of the below-mentioned 8 attribute-categories are detected, the picture is considered AI-generated (Table 1). To ensure transparency, a full demonstration of the ESMA process is included in the supplementary material (Appendix A1, A3), and its theoretical grounding is supported by established theories of visual communication and perception (Barry, 1997; Messaris, 1994).

  - o *Further AI-detection*: Once a suspicious attribute was discovered during ESMA, the publicly available image was downloaded and tested on an online AI detector platform [IsitAI? - (IsitAI.com, 2024)]. If both the manual analysis and the detector confirm AI-origin (Test result: probability≥ 60 %), the image is utilized for further analysis (n = 146). To further validate the ESMA method's robustness, it was also applied to a human-created image (*Details in the attachment: A3*), thus reinforcing the method's ability to differentiate real from synthetic visuals.

1.2.B: Theoretical sampling: Once AI-generated images were collected, a database of related comments has been generated to provide a basis for further analysis.

- o *Preliminary screening*: This includes the saving of public comments and has resulted in a robust database (n = 11547).
- o *Ten-step manual analysis* [*TSMA*]: Since comments may also be AI-generated [AI-bots], another step of AI-detection has been performed along with a ten-step manual analysis (Table 2). In case of

Table 1
Eight-step manual analysis [ESMA] for detecting AI-generated images.

N	AI-generated content signals	Definition	Examples
1.	Human anatomy inaccuracies	This category refers to biologically implausible depictions of the human body, including incorrect numbers of fingers or teeth, distorted limbs, or other anatomical errors.	(1) unusual number of fingers; (2) fingers are bent at an unnatural angle; (3) dysmorphia (too long, short limbs); (4) unnatural number of teeth or only a continuous "white stripe"
2.	Asymmetry and facial oddities	This includes unnatural irregularities in facial structure or balance, such as mismatched eyes, misaligned pupils, or missing ears.	(5) different size or shape of eyes; (6) pupils look in different directions; (7) unnatural hairlines; (8) ear may appear distorted or disappear
3.	Object and environment anomalies	This covers inconsistencies in how objects interact with their surroundings, like floating items, incorrect reflections, or overlapping elements that defy spatial logic.	(9) grasping object incorrectly; or (10) objects seem to float; (11) reflections in glass, mirrors or water surfaces are incorrect or missing; (12) object overlaps other object in an unnatural way
4.	Unnatural textures	This refers to skin or surface patterns that appear overly smooth, hyper-detailed, or feature repetitive, irregular motifs uncharacteristic of real-life materials.	(13) too smooth skin, unnaturally perfect or too detailed (e.g., all pores appear to be evenly distributed); (14) patterns on clothes/surfaces are irregular/repetitive
5.	Physics and gravity disregard	This includes visual elements that violate basic physical principles, such as levitating objects, static liquids, or mismatched shadows.	(15) hanging objects appear to float; (16) unnatural liquids (e.g., drink in glass appears static or solid); (17) direction of shadows does not match the light source
6.	Texts and lettering issues	This category involves malformed or nonsensical text on visual surfaces, often appearing as distorted characters or unreadable signs and posters.	(18) incorrect lettering: distorted letters, meaningless characters; (19) text on books, signs or posters are unreadable or blurred
7.	Exaggeration of symmetry	This refers to unrealistic levels of symmetry in faces or objects, often resulting in unnatural and overly perfect appearances.	(20) exaggeratedly symmetrical faces/objects
8.	Crowd and group scene flaws	This includes errors in group depictions, such as distorted background figures or animals with anatomical deformities, commonly seen in AI-generated crowd scenes.	(21) people in the background are blurred/ distorted; (22) animals with anatomical flaws (e. g., extra legs, missing tails)

Note: The eight attribute categories were inductively developed by the author based on recurring visual patterns observed during the Social Media Observation phase.
Source: Author's own collection and editing

number of suspicious activities≥2, comment is excluded from the dataset. This process resulted in a 20.12 % reduction in the number of comments used for further analysis (final sample n = 9082).

- The validity of the ESMA and TSMA frameworks follows qualitative research standards rather than quantitative benchmarking. Both frameworks emphasize procedural transparency, internal consistency, and interpretive traceability (Lincoln, 1985; Tracy, 2010). The coding logic was refined through iterative comparison and researcher memoing to ensure that criteria were applied consistently. This form of internal validation supports methodological reliability and allows for transferability to similar datasets, while remaining sensitive to contextual differences across platforms.

Table 2
Ten-step manual analysis [TSMA] for detecting bot users.

N	AI-generated content signals	Examples
1.	Unnatural posting frequency and speed	Bots often post at an unnaturally high frequency, sometimes multiple times per minute or with no realistic delay between posts. *E.g.,* more than 5–10 comments or posts per minute.
2.	Repetitive or template-based comments	Bots tend to use the same or slightly modified comment templates across different posts. *E.g.,* "*Amazing work*!" multiple times across different pages.
3.	Lack of personalized interaction	Bot accounts rarely engage in personalized conversations, avoiding direct replies to user comments or specific context-related discussions.
4.	Suspicious or generic profile information	Profiles with only a few details (e.g., "*Lives in Earth*" *or* "*Works at Freelancer*") and no clear timeline activity are common among bots.
5.	Inconsistent or AI-generated profile pictures	Many bots use AI-generated images, which has attributes as described in Table 1.
6.	Low social engagement (*few friends or followers*)	Bot accounts typically have very few friends or followers and rarely receive meaningful engagement on their posts. *E.g.,* fewer than 10 friends
7.	Simultaneous activity across multiple pages or groups	Identical posts in more than 5 groups or pages within a short timeframe (e.g., 30 min).
8.	Identical language use across different topics	Repeated phrases like "So true!" or "*Great idea*!" on diverse subjects (e.g., technology, fashion, and politics) highlight AI-origin.
9.	Unusual comment timing patterns	Bot accounts are active 24/7, often posting at times when typical users are inactive, such as early morning hours or late at night.
10.	Behavior aligned with automated scripts	Bots generally lack contextual awareness, often posting irrelevant or out-of-place comments that do not align with the content of the post. *E.g.,* replying "*That's delicious*!" on a political news article indicates a lack of contextual understanding.

Source: Author's own collection and editing

3.2. PH.2 data analysis

Once the dataset is constructed, a three-steps analysis was performed. Phase 2 builds directly on the image dataset established during Phase 1, using only those images that met strict validation criteria (manual detection + algorithmic probability ≥60 %). This ensures that the thematic and user-response analysis is grounded in a systematically selected and verified corpus of AI-generated visual content.

- *PH.2.1: Constant comparison:* This involves reading and interpreting comments. The aim is to filter out comments that are not meaningful (e.g., meaningless text, incomplete words). The process resulted in the exclusion of 160 comments.
- *PH.2.2: Writing memos:* It aims to identify patterns in the comments. Memos also include research ideas and key insights created during the process.
- *PH.2.3: Coding:* Based on the methodology, a three-level coding process was employed as follows:
  - *2.3.1: Open coding:* the process of creating categories from comment without limitations.
  - *2.3.2: Axial coding:* the process of defining the correlations among categories and create higher-level categories as suggested by Corbin and Strauss (1990) and to create an axial model:
- *Causal conditions:* factors (AI-generated images) that cause the central phenomenon (users' reactions).

- *Central phenomenon:* the core of the theory (users' reactions to AI-generated images).
- *Strategies:* the topic of comments that realized by the central phenomenon (users' reactions).
- *Context:* the trigger of the central phenomenon, i.e., the reason of posting a certain comment.
- *Intervening conditions:* some condition that affect strategies (users' reactions). In this context, based on the wording and content of the comments, the presumed cognitive biases and heuristics that might support deception are identified.
- *Consequences:* means the outcome of the actions: users' engagement (writing a comment) and presumed deception (believing the content).
  - o *2.3.3: Selective coding:* it involves the set-up of a grounded theory model that explains the different types of AIGIs (typology), the nature of user reactions and the underlying patterns (heuristics).

4. Results

Results are summarized along with the GT-coding process.

4.1. Open coding

During this phase, comments are examined on a sentence-by-sentence basis to ensure the identification of significant comment categories. Following the instructions outlined by Charmaz (2006), researchers aim to identify as many topics (categories) as possible. In practice, this phase involves reading through the comments collected while manually collecting key themes and researchers' impressions in an MS Excel table.

4.2. Axial coding

The axial coding phase revealed that certain thematic categories (identified in RQ1) systematically correspond to specific user reactions, supporting the interconnected structure of the research questions. Table 3 highlight the key themes and emotional stimuli employed by AI-generated images. Although the analysis primarily focused on the visual properties of the selected images, brief captions appearing directly above the images (e.g., emotionally framed or suggestive texts such as "*Would you like to sit down with them for a coffee?")* were also considered during interpretation. These short textual cues were not analyzed independently but were considered as part of the visual-affective framing of the posts.

The most frequent categories (Examples: Fig. 2), such as "*Emotion and nostalgia*" and "*Arousing empathy and compassion*" reflect a strategic use of sentimental and emotionally evocative content, particularly focusing on themes like anniversaries, rural lifestyles, and compassion-inducing scenarios. For example, images of "*Old couples, anniversary*" (freq.: 14 %) and "*Would you like to sit down with them for a coffee?*" (freq.: 10 %) suggest a deliberate effort to resonate with users' nostalgic and empathetic tendencies. In contrast, less frequent themes, such as "*Reference to inter-ethnic conflicts*"" (freq.: 1 %) and "*A little child who has a birthday, but no one congratulates them*" (freq.: 1 %), indicate a niche focus, potentially designed to provoke specific reactions such as outrage or shock [outrage bait]. The category "*Extraordinary achievements, hobbies*" demonstrates the use of novelty and surprise to capture attention, albeit with a moderate average frequency (6–6.5 %), suggesting a balanced approach to engaging curiosity without oversaturation. The data also underline the importance of emotionally charged narratives in driving user engagement, with religious and spiritual themes, like "*The face of Jesus hidden in a landscape*"" (freq.: 5 %), leveraging symbolic recognition and personal beliefs.

Table 4 provides summary of how cognitive heuristics influence the interpretation of AIGIs. Each topic can be linked to a cognitive bias that might shape user reactions. Nostalgic themes (e.g., anniversaries or rural lifestyles) leverage confirmation bias to evoke positive associations, while anchoring magnifies the initial emotional impact. Similarly, empathetic content (e.g., images of neglected children) capitalizes on users' immediate emotional responses, often leading to oversimplified or unquestioned acceptance of the image's narrative. The examples of user comments in Table 4 showcasing the powerful role of visual stimuli in shaping perceptions and amplifying emotional engagement.

After the typology of AIGIs, the nature of comments is interpreted. Comments are analyzed according to the six axial coding categories, creating a logical connection between concepts. Fig. 3 represents the categories and the correlations between axial codes. As seen in the axial model, several categories of AI-generated images are distinguished (causal conditions n = 5) that lead to user reactions (phenomenon n = 1). User reactions are affected by psychological bias/heuristics and manifest in different types of comments (strategy n = 11). Comments

Table 3
Analysis of AIGIs as causal conditions of the axial model.

Category	ID	Sub-category	Freq.	Occurrence
Emotion and nostalgia	1	Old couples, anniversary (e.g., together for 50–60 years)	14.00	Frequent
	2	Showing the good old days, especially the benefits of a rural lifestyle and habits (e.g., pig slaughter, country style wedding)	11.50	Frequent
	3	Showing the wealth of a bygone era (e.g., financial security, self-sufficiency, home growing, animal breeding in the old rural communities)	8.00	Medium
	4	Anniversaries: mother and children (baby photo vs. adults with old mother)	7.50	Medium
	5	Twins recreate a childhood picture	8.00	Medium
Arousing empathy and compassion	6	Would you like to sit down with them for a coffee? - compassionate pictures of elderly and/or people in difficult financial circumstances	10.00	Medium
	7	A sad fate, loss of a parent, only grandmother takes care of a child	8.50	Medium
	8	The derogation of agricultural work (e.g., if you don't look down on us because we're farmers, say hello)	7.00	Medium
	9	Reference to inter-ethnic conflicts	1.00	Less common
Extraordinary achievements, hobbies	10	Astonishing needlework (e.g., weaving, sewing, other creative) but no one congratulates	6.50	Medium
	11	An astonishing hobby for a man	6.00	Medium
Birthdays and social ills	12	Birthday, but no one congratulates, protagonist: worker with low-earning professional	6.00	Medium
	13	A little child who has a birthday, but no one congratulates them.	1.00	Less common
Religion and spirituality	14	The face of Jesus hidden in a landscape ("squint to see the wonder")	5.00	Less common

Note: „Freq." indicates the approximate proportion of analyzed AIGIs that fall into each thematic category. "Occurrence" refers to a qualitative categorization of these percentages into three bands: less common (≤5 %), medium (6–10 %), and frequent (>11 %) to support easier interpretation of thematic prevalence.
Source: Author's own editing

Fig. 2. Montage of AI-generated images (examples)[1]
[1]Numbering based on the categories presented by Table 3. AI probability test results: 1: 65 %; 4: 81 %; 5: 76 %; 6: 81 %; 8: 81 %; 10: 63 %; 12: 66 %; 13: 63 %; 14:79 % [based on: IsitAI? – online AI detector tool].
Source: Based on systematic search (1.1.B), Author's own editing

have many triggering topics (context n = 16) and the process overall leads to a high user engagement and deception.

Table 5 lists the codes presented on Fig. 3. The codes generated during the open coding are linked to each axial code. For determining the code importance, a command line was run [Explore →Word Frequency] in NVivo software, which aimed to search for keywords relating to the codes (Wong, 2008). This step aims to reveal the frequency of occurrence of the codes and thus to refine the construction of the model. Boolean operators [AND & OR] have been employed for the keyword search. During the process, Total count and the Weighted percentage (WP) [%] values were considered. Weighted percentages facilitate to decide how relevant a code in the documents analyzed (Jackson & Bazeley, 2019). Since the specific values of the thresholds strongly depend on the volume of the texts analyzed and the researcher's

Table 4
Examples of psychological bias/heuristics related to the sub-topics of AI-generated images.

ID	Topic	The hypothetical role of heuristics	Example comment that is evidence for the heuristic*
1	Old couples, anniversary (e.g., together for 50–60 years)	**Confirmation bias**: reinforce positive associations with the value of a long-lasting relationship or marriage. **Anchoring**: first impression (desire to experience the pleasure of a long relationship) distracts from inaccuracies in content.	**Confirmation bias:** *"This is what true love looks like [heart, flower]. My grandparents were together for 70 years—such a blessing! [black heart]"* **Anchoring**:*"60 years together! I hope my relationship lasts this long, such an inspiration! [crying emoji]"*
2	Showing the good old days, especially the benefits of a rural lifestyle and habits (e.g., pig slaughter, country style wedding)	**Confirmation bias**: images reinforce the cognitive bias association that used to be all right. **Familiarity effect**: some visual stimuli at first glance recall past experiences.	**Confirmation bias:** *"Life was so much simpler and better back then. These pictures really show how we've lost touch with our roots."* **Familiarity effect:** *"This reminds me of how my grandparents used to live on their farm—it brings back such memories [black heart]."*
3	Showing the wealth of a bygone era (e.g., financial security, self-sufficiency, home growing, animal breeding in the old rural communities)	**Confirmation bias**: images reinforce the cognitive bias association that used to be all right. **Familiarity effect**: some visual stimuli at first glance recall past experiences.	**Confirmation bias:** *"Back then, people were more self-reliant and didn't depend on anyone. We've lost that independence today … [angry face emoji]"* **Familiarity effect:** *"I remember visiting places like this as a kid; it just feels like home!!!"*
4	Anniversaries: mother and children (baby photo vs. adults with old mother)	**Confirmation bias**: images reinforce associations of a strong mother-child relationship. **Anchoring**: first impression (emotional reaction) distracts attention from inaccuracies in content.	**Confirmation bias:** *"This is what family is all about. A mother's love is forever."* **Anchoring:** *"Seeing this made me cry—it reminds me of my own mom and how much she means to me."*
5	Twins recreate a childhood picture	**Confirmation bias**: images reinforce associations of a strong relationship between siblings. **Anchoring**: first impression (emotional reaction) distracts attention from inaccuracies in content.	**Confirmation bias:** *"It's so heartwarming to see siblings stay close after all these years. Family is everything [heart, heart, heart]"* **Anchoring:** *"This picture just hits—it's such a beautiful memory brought back to life."*
6	Would you like to sit down with them for a coffee? - compassionate pictures of elderly and/or people in difficult financial circumstances	**Confirmation bias**: reinforce associations with the lower social class and the loneliness of the elderly.	**Confirmation bias:** "*It's so sad to see older people left alone like this. Sure, I would sit with them!!!"*
7	A sad fate, loss of a parent, only grandmother takes care of a child	**Anchoring**: first impression (emotional reaction) distracts attention from inaccuracies in content. **Confirmation bias**: images reinforce associations with the importance of protecting the child and the family bond.	**Anchoring:**"*This just breaks my heart—imagine a child growing up without parents.*" **Confirmation bias:**"*Children should always be cared for no matter what.*"
8	The derogation of agricultural work (e.g., if you don't look down on us because we're farmers, say hello)	**Confirmation bias**: reinforcing the association with the poor perception of physical work.	**Confirmation bias:**"*Farmers work harder than anyone, but they never get the respect they deserve. This needs to change!!*"
9	Reference to inter-ethnic conflicts	**Anchoring**: first impression (shock, outrage at racism) distracts from inaccuracies in content.	**Anchoring:**"*You should not listen to the haters*!"
10	Astonishing needlework (e.g., weaving, sewing, other creative) but no one congratulates	**Novelty bias**: surprise at the sight of a unique work of art overshadows the questioning of the product. **Groupthink**: the high number of posts/comments reinforces the sense of uniqueness.	**Novelty bias:**"*Incredible! You are very talented*!" **Groupthink:**"*Joining others: it deserves so much more recognition!!! Congrats! [heart]*"
11	An astonishing hobby for a man	**Novelty bias**: surprise at the sight of a unique work of art overshadows questioning the product. **Groupthink**: the high number of posts/comments reinforces the sense of uniqueness.	**Novelty bias:**"*Super! Keep it up and do not listen to others …"* **Groupthink:**"*Everyone is right—this is truly amazing work. Hats off to him!"*
12	Birthday, but no one congratulates, protagonist: worker with low-earning professional	**Anchoring**: first impression (emotional reaction, regret, pity) distracts attention from inaccuracies in content.	**Anchoring:** *"It is your day! Happy birthday, happy life! [cake emoji]"*
13	A little child who has a birthday, but no one congratulates them.	**Anchoring**: first impression (indignation that the child is not getting enough love, attention) distracts attention from inaccuracies in content.	**Anchoring:** *"How could people forget this child's birthday? It's heartbreaking to see such neglect."*
14	The face of Jesus hidden in a landscape ("squint to see the wonder")	**Illusory correlation**: strength of religiosity and surprise might contribute the acceptance of spurious correlations.	**Illusory correlation:** *"God bless everyone, it is beautiful! AMEN! [praying emoji]."*

Notes:
"ID" numbers refer to the thematic categories listed in Table 3, allowing easier cross-referencing between image types and related psychological mechanisms.
*Translations of original comments identified during data collection.
Source: Author's own editing

preferences, there are no generally accepted standard values of weighted percentages. As the number of words in the database is relatively high [47.159], the weight percentage values are low. For utilizing the values for further analysis, the following thresholds was defined: Category 1:

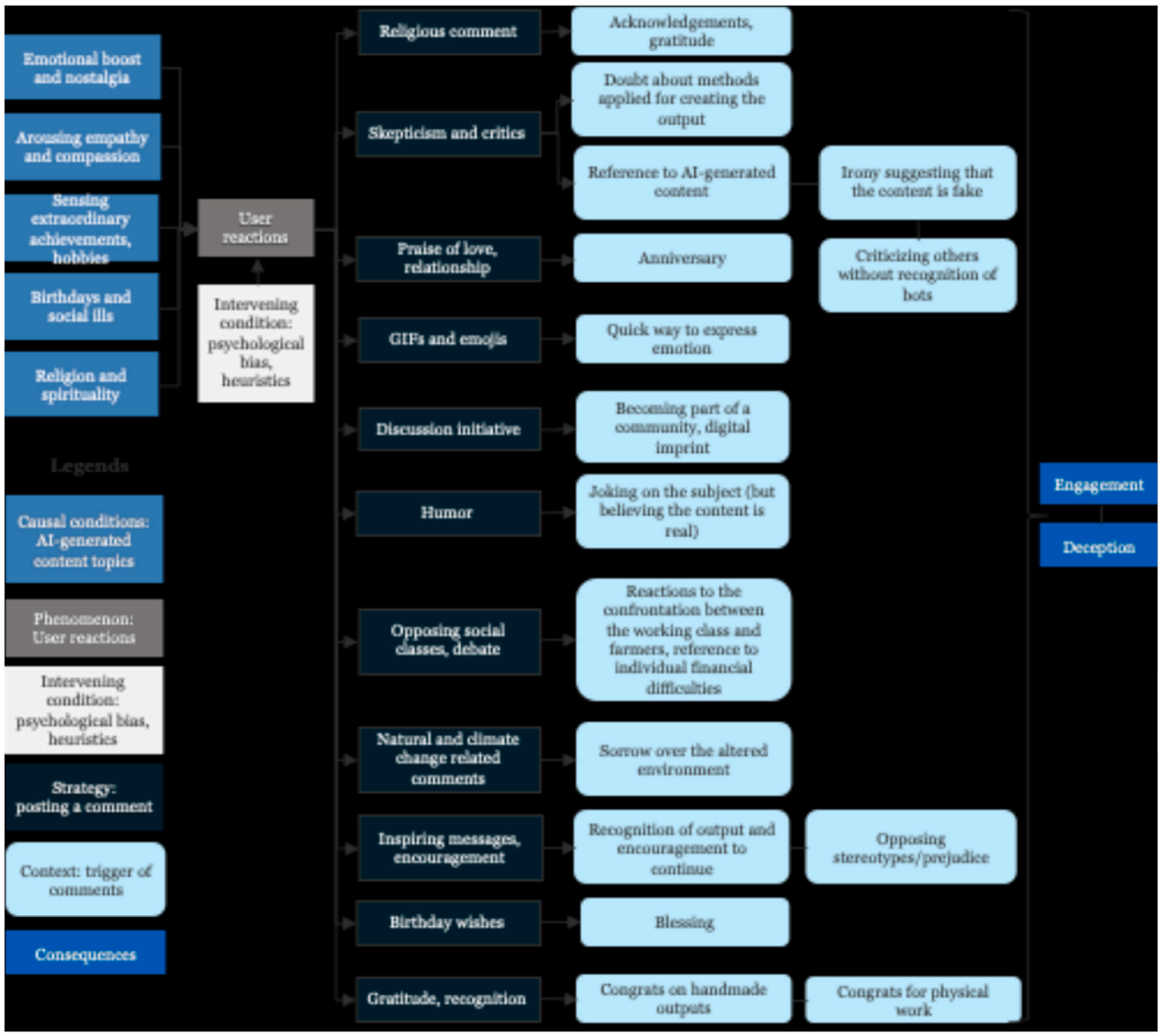

Fig. 3. Results of Axial coding.
Source: Author's own editing

2.13–4.20; Category 2: 1.08–1.85; Category 3: 0.03–0.89.

The significance of the axial codes presented in Table 5 lies in their ability to group similar reactions and behaviors, highlighting the dominant patterns in how users respond to specific AI-generated images. Codes such as "*Inspiring messages, encouragement*" (Category 1) and "*Birthday wishes*" realize a strong prevalence, with the highest counts and share percentages (20.197 % and 22.19 %, respectively). These codes underline themes that elicit positive reinforcement and emotional connections, making them central to user engagement.

Codes like "*Religious content*" and "*Gratitude, recognition*" (Category 2) showcase significant user interaction, but slightly lower emphasis compared to high-importance themes. These categories reflect behaviors tied to cultural and social norms, such as expressing appreciation or invoking blessings, which are deeply rooted in religious convictions. By leveraging these faith-based expressions, AIGIs exploits the emotional resonance of religious beliefs, enhancing its deceptive power and fostering trust in the authenticity of the message, even in the absence of factual accuracy. Within Category 2, one notable code highlights that a significant proportion of users initiate interaction by commenting, often out of a desire to engage socially. This phenomenon underscores how AI-generated deception creates an illusion of community connection, while paradoxically reinforcing social isolation. In several cases, users unknowingly direct their efforts toward interacting with automated accounts, further exacerbating the lack of genuine human connection [dead internet theory – Walter, 2024, pp. 1–2].

A variety of codes, including "*Skepticism and critics*", "*Nostalgia*", and "*Natural and climate change*"-related comments, fall into Category 3. These reflect more niche behaviors, such as skepticism towards AI-generated images or lamentation of environmental changes. The skepticisms code also reveals that among those who question the authenticity of the images, there is a high proportion of people who are fooled by the bot comments, thus being partly deceived by the AI-generated content.

4.3. Selective coding

In the open coding process (Fig. 4), the variable reduction of the axial model is the goal and the interpretation of the path to the central phenomenon.

Based on the selective coding, three main stimuli can be identified. Witnessing seniority (1) and triggering nostalgia (2) evoke positive emotional reactions where critical thinking is suppressed. Images representing ignorance (3) of a person enhance tension [outrage bait], thus contributes to a high engagement and the ignoring signs of deception. This phenomenon is amplified mainly by four heuristics considered. Confirmation bias (1) reinforces existing beliefs, ensuring users are less likely to question the content's authenticity. As a result of anchoring (2), emotional first impressions distract users from inconsistencies. Familiarity effect (3) uplifting themes resonate deeply with users, enhancing

Table 5
Axial coding – qualitative validation (Total counts, average weight percentages (%)) of codes (comments).

Axial code	Total count[a]	Share (%)	Open codes	X_k	X_l	Boolean operators (Keywords)	Typical example	x^i [b]	Importance Category[c]
Religious content	1490	16.70	"Amen"	1	510	"*Amen*"	"*AMEN. Praise be to God*"	1.0814	Category 2 (Medium Importance)
			"God blesses!"	2	870	"*God*" and "*Lord*" OR "*bless*" OR "*blesses*"	"*May the Lord bless you and your family*."	1.8448	Category 2 (Medium Importance)
			"God be with you!"	3	110	"*God*" and "*Lord*" OR "*you*" OR "*everyone*" OR "*mercy*" AND "*praise*"	"*The Lord be with you always, I praise you!*""	0.2332	Category 3 (Low Importance)
Inspiring messages, encouragement	1802	20.197	Recognition of results and encouragement to continue [I made it myself, but nobody cares]	4	1004	"*wish*" OR "*continue*" OR "*keep it going*" OR "*future*" OR "*admire*"	"*May you always keep your enthusiasm and hope!*"	2.12896	Category 1 (High Importance)
			Opposing stereotypes/ prejudice [reaction to I would be happy if you would appreciate it!]	5	798	"*not*" OR "*matters*" OR "*care*" OR "*valuable*"	"*It's not the color of the skin, it's what it covers. Behind the color is a feeling heart*" "*You are beautiful, don't care about anything*"	1.6921	Category 2 (Medium Importance)
Birthday wishes	1980	22.19	Birthday wishes, blessings	6	1980	"*happy*" OR "*birthday*" OR "*God*" OR "*bless for your birthday*" OR "*love*"	*"Happy Birthday with God's blessing!"* *"With much love, I wish you a very happy birthday."*	4.1986	Category 1 (High Importance)
Gratitude, recognition	1080	12.10	Congratulations on a handmade creation	7	570	"*respect*" OR "*congratulations*" [and syn] OR "*keep*" or "*work*"	*"Tough, wonderful work, keep it up and congratulations"*	1.2087	Category 2 (Medium Importance)
			Congratulations for doing physical work	8	510	"*beautiful*" [and syn] OR "*congratulations*" [and syn] OR "*success*"	*"This is a beautiful cake, congratulations and good luck."*	1.0814	Category 2 (Medium Importance)
Discussion initiative	720	8.07	A comment made for the purpose of having a conversation with anybody.	9	720	"*Good morning/evening*" OR "*have a nice day"* [and syn]	*"Have a nice day."* *"Good evening to you all."* *"Good work and good evening!"*	1.5267	Category 2 (Medium Importance)
Skepticism and critics	520	5.28	Doubt about how the product was made (not that it was man-made)	10	84	"*How*" OR "*made*" OR "*technique*"	*"What technique was used? That's not the way to do it."*	0.1781	Category 3 (Low Importance)
			Reference to AI-generated image	11	269	"*attention*" OR "*fake*" OR "*AI*" OR "*artificial intelligence*" "*made*" OR "*scam'*	*"STOP AI!!! ATTENTION! THIS PICTURE IS NOT REAL!"*	0.5704	Category 3 (Low Importance)
			Criticizing commenters but without recognizing bots	12	151	"*stupid*" [and syn] OR "*fake*" OR "*crying laughing emoji*" OR "*bullshit*"	*Is everyone here stupid? These are not real pictures, you brainless […] [emojis].*	0.3202	Category 3 (Low Importance)
			Humor and irony, suggesting that the subject has found out about the scam	13	16	Ironic phrases were collected separately during manual collection.	*"Do potatoes come out of the ground washed?"* *Mercy! Don't cut up this AI pastry chef's masterpiece!* *"There's a dinosaur missing from the village landscape."*	0.0339	Category 3 (Low Importance)
GIFs and emojis	420	4.70	GIFs and emojis	14	420	Only GIFs or emojis without text has been collected.	*Most common: Smiling emojis, flowers, hearts, moving hearts, moving pictures with congratulations captions, dancing emojis*	0.8906	Category 3 (Low Importance)
Nostalgia	330	3.69	Recalling childhood memories	15	185	"*when*" OR "*child*" OR "*past*"	*"Reminds me of my grandmother's old tablecloths!* *"I had a yard like that myself, even though I was in state care, unfortunately."*	0.3923	Category 3 (Low Importance)
			Sorrow over the past	16	130	"*sad*" [and syn] OR "*past*" OR "*memories*"	*"It's getting rarer and rarer nowadays"* *"That was the good old days of sledding for laughs."*	0.2757	Category 3 (Low Importance)
			Specific places, mention of personal life stories	17	15	Specific comments were collected separately during manual collection.	*"Sliding down the hill in [City], it was so much fun!"*	0.0318	Category 3 (Low Importance)
Praise of love, relationship	280	3.13	Anniversary congratulations to couples, siblings, family members	18	280	"*congratulations*" [and syn] OR "*love*" OR "*anniversary*"	*"They used to be a beautiful couple, just like now!"* *"This is truly beautiful and rare love!"*	0.5937	Category 3 (Low Importance)

(*continued on next page*)

Table 5 (*continued*)

Axial code	Total count[a]	Share (%)	Open codes	X_k	X_l	Boolean operators (Keywords)	Typical example	X^j [b]	Importance Category[c]
Humor	190	2.13	Joking on the subject, but believing the content is real	19	190	Specific comments were collected separately during manual collection.	*"Very nice, how can I get there to help you store (eat) it?"* *"Nice cake! My grandma used to make meatloaf."* *"Sure-sure, I can do that too"*	0.4029	Category 3 (Low Importance)
Natural and climate change related comments	60	0.67	Sorrow over the altered environment	20	60	"*Global warming*" OR "*climate*" OR "*destroy*" [and syn]	*"Mankind is ruining everything, global warming is here, never again will there be such wonderful winter times!!!"* *"It was snowing then!"*	0.1273	Category 3 (Low Importance)
Opposing social classes, debate	50	0.56	Reactions to the confrontation between the working class and farmers, reference to individual financial difficulties	21	50	Specific comments were collected separately during manual collection.	*"If the clothing factory didn't produce, the farmers would walk around naked."* *"When will you ever be able to pay for this kind of pesticide?"*	0.1060	Category 3 (Low Importance)

a = Frequency of occurrence of the axial code in the comments analyzed.

b = Average of weight percentage (WP) of open codes. The evaluation value is x_l^k, where k is the number of open codes [1-21] and l is the number of occurrences of open codes. To determine weight percentage (X^j) of a code, the x_i^j values are divided by the total words of comments analysed (47.159), multiplied by 100 (eq. (1)). $X^j = \left\lfloor \frac{x_l^k}{\sum m} \right\rfloor x100$

c = Categorization based on the WP values.

Source: Author's own editing

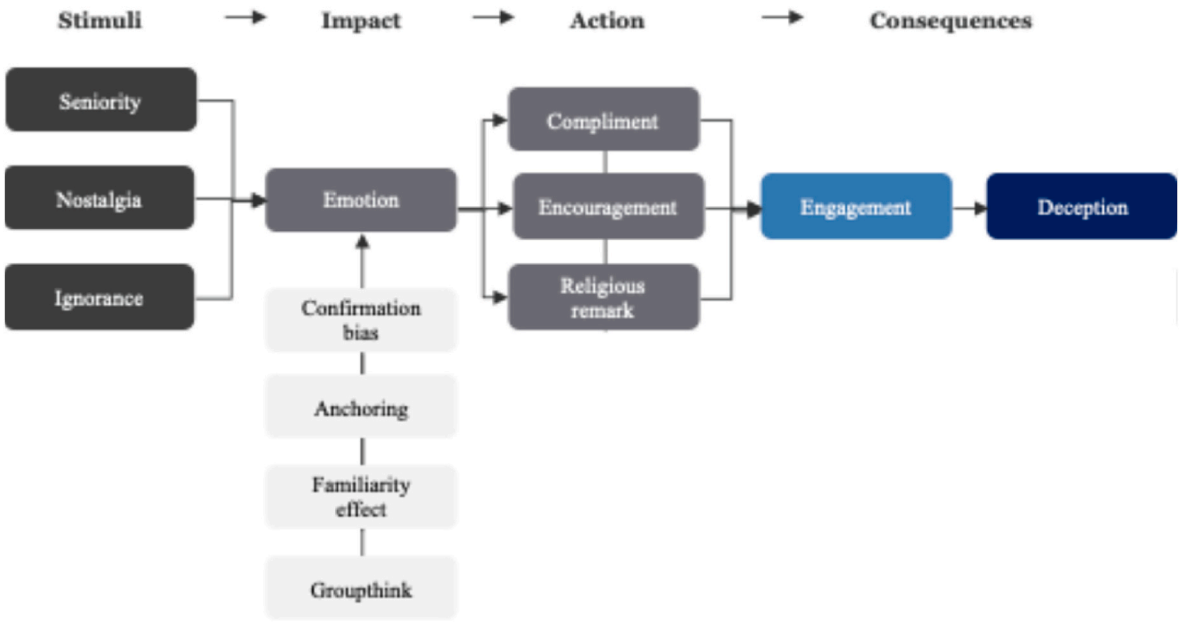

Fig. 4. Results of Selective coding.
Source: Author's own editing

believability, and groupthink (4), i.e., the high volume of engagement (comments) reinforces trust in the AIGIs validity. All of this realizes action: users become emotionally involved, offer comfort and seek to become involved (e.g., in a celebration), which are largely framed by religious elements. This leads to high engagement and misleading users who encounter AI-generated content.

5. Discussion and conclusion

This study examined how AI-generated images [AIGIs] on Facebook exploit users' vulnerabilities, revealing their remarkable potential for digital deception. Using Grounded Theory, this research identified dominant AIGI themes and showed how these images elicit affective responses, including empathy, longing, and admiration. Emotional engagement was often accompanied by cognitive biases which together reduced users' critical reflection. These findings highlight the psychological mechanisms that make AIGIs persuasive and draw attention to an escalating problem: the unregulated and ethically questionable use of AI-generated content on social media. If left unaddressed, the spread of synthetic images and AI-driven interactions may undermine platform credibility, erode user trust, and lead to broader public skepticism toward AI technologies in the long run. As such, the study underscores the need for platform-level safeguards, improved digital literacy, and transparent detection frameworks to mitigate the risks of synthetic media.

5.1. Answers to RQs

- *RQ1:* What are the main thematic categories of AI-generated images (AIGIs) shared on Facebook, based on their visual motifs and symbolic aesthetics (i.e., recurring visual content types)?

The analysis identified five main thematic categories framing AIGIs: emotion and nostalgia, arousing empathy and compassion, extraordinary achievements and hobbies, birthdays and social ills, and religion and spirituality. Among these, the most prevalent were the emotionally nostalgic and empathy-inducing themes, often centered around elderly individuals, rural life, and intergenerational bonds. This typology complements prior literature that describes the rise of emotionally charged synthetic media (DiResta – Goldstein, 2024) but goes further by identifying recurring thematic formulas that appear optimized for engagement. Importantly, the categorization was not predefined but constructed through constant comparison and saturation during the coding process, increasing its empirical grounding.

- *RQ2: How do social media users emotionally and cognitively respond to different thematic categories of AI-generated images (AIGIs)?*

The most prevalent emotional responses were empathy and care (expressed through affectionate comments or personal stories), nostalgia and longing (especially toward idealized pasts or spiritual imagery), and moral support or admiration in response to extraordinary achievements. Many users accepted the images as authentic without questioning their origin, especially when captions or comment sections reinforced their emotional interpretations. This confirms earlier findings that emotionally triggering content is more likely to be perceived as credible (Iosifidis & Nicoli, 2020), especially when social proof (e.g., likes, supportive comments) accompanies it.

- *RQ3: Which deception mechanisms (cognitive biases) can be inferred from social media user interactions with AI-generated images?*

The study found that AIGIs often deceive users by simultaneously triggering multiple cognitive biases such as confirmation bias, anchoring, and familiarity effect, which together reduce critical thinking. These mechanisms illustrate how affect-driven content can override cognitive vigilance, aligning with recent work on emotional misinformation (Goldstein et al., 2024; Perach et al., 2023). The study shows that AIGIs use several biases at once, creating a more powerful and harder-to-resist persuasive effect than typical misinformation.

- *RQ4: Which thematic categories of AI-generated images are most and least effective in deceiving social media users?*

Representations of seniority, nostalgic themes, and outrage baits (e.g., rural memory scenes, long-term couples, or forgotten birthdays) proved highly effective in suppressing critical thinking due to their strong emotional resonance. In contrast, themes such as inter-ethnic conflict or a neglected child's birthday elicited relatively lower engagement or more skeptical responses. These cases indicate that emotionally charged content does not uniformly lead to deception, and user sensitivity or perceived authenticity may moderate its effectiveness.

5.2. Theoretical and practical contributions

In terms of *bridging the theoretical research gap*, the most relevant contributions are listed as follows:

- The identification and *typology of common AI-generated image themes* offer new insights into AI-content classification and provide frameworks for detecting AI-originated media stimuli. This typology serves as a steppingstone for theoretical advancements in understanding AI-based deception.
- The study pioneers the *analysis of user reactions to AIGIs*, filling a critical gap in empirical evidence and offering a nuanced understanding of how such content manipulates user perceptions.
- The research *demonstrates how cognitive biases contribute to user engagement* and the effectiveness of deception. This supports building theories on how mental shortcuts affect user engagement with AI.
- The study shows how emotional overload and familiar symbols can mislead people, extending earlier research on disinformation to AI-generated images. As such, this is the *first study to highlight this escalating issue from this perspective*, laying the groundwork for a broader scholarly discourse on AIGIs.

From methodological aspects:

- This study introduces a *two-step verification framework for detecting AI-generated images* (see: PH.1.2.A – ESMA), outlining systematic manual evaluation criteria that are cost-effective and scalable. This methodological tool created by the author provides a sound basis for researchers seeking to deepen the exploration of AI-generated content.
- The proposed two-step verification framework and bot detection procedure were not designed to outperform automated detection tools in terms of accuracy or scalability, but rather to provide *transparent, interpretable, and researcher-guided methods for identifying suspicious content and minimizing selection bias.* These procedures serve as complementary tools, particularly useful in exploratory or qualitative investigations where algorithmic solutions may lack contextual nuance.
- A novel *bot detection framework* has also been developed (see PH.1.2.B – TSMA) that employs multiple analytical criteria to increase the accuracy of social media content analysis and distinguish real user reactions from automated interactions.
- By adapting the Grounded Theory [GT] methodology to the analysis of social media and AI-generated content, this research establishes a *replicable approach to similar studies* and thus promotes methodological rigor in this emerging field.

From a practical aspect:

- The proposed frameworks for AI origin detection and bot identification have direct application for users to filter social media content, helping to reduce the spread of deceptive AI-generated material and *enhance users' AI/digital literacy*.
- By highlighting the deceptive potential of AIGIs, this research *underscores the importance of consumer education campaigns* designed to strengthen digital literacy and reduce the chance of emotional manipulation.
- The study provides actionable insights for developing social media strategies that counteract the exploitation of cognitive biases, fostering a more transparent and trustworthy online environment. Furthermore, the research *validates the implementation of AI-generated content labelling on all social media platforms at short notice.*
- While the proposed framework is not intended to replace automated AI-detection systems, it offers complementary advantages. Unlike most black-box detection tools that output probability scores without explanation, the ESMA provides transparent, interpretable criteria that *users can learn and apply themselves.* This human-readable format may enhance digital literacy and promote more conscious content engagement, especially in environments where algorithmic moderation is limited or delayed. Therefore, the framework also has a *pedagogical function*, contributing to user empowerment and platform resilience against visual misinformation.
- More broadly, the findings serve as a *wake-up call for platform developers, policymakers, and educators.* If emotional deception via AIGIs

becomes normalized, users may begin to distrust not only digital content but also ethical applications of AI. Without transparent detection tools and user-centered literacy efforts, the persuasive strength of synthetic content could erode public confidence in technological innovation.

5.3. Limitations and further research directions

As an exploratory study, this research has certain limitations. First, the analysis was limited to public Facebook content, which may not reflect platform-specific or culturally embedded differences in how users engage with AI-generated images. Second, while the ESMA and TSMA frameworks were applied consistently, they have not yet been externally validated. These limitations point to important directions for future research, including cross-platform comparisons and independent testing of the proposed frameworks.

Declaration of competing interest

The author declares no conflict of interest.

Appendix

Attachments.

A1: Example of AI-origin detection

The first picture detected during SMO: Caption: "I'm a doctor, it's my birthday today, but no one wished me a happy birthday!"; Signals of AI-origin: (4) Unnatural textures – background and subject's skin; (7) Exaggerating the symmetry of the subject's aesthetics; AI-test result: Highly likely AI-generated (Probability test result: 80 %).

Fig. A1. AI-generated picture identified during SMO
Source: URLA2: List of publicly available Facebook pages with AI-generated content

Table A2
List of Facebook pages analyzed during the GT process

N	Translated title [ENG]	URL	Availability	Fieldwork period
1	The Gospel of Grace	URL	Public	01.03.2025–01.05.2025.
2.	Old House	URL	Public	01.03.2025–01.05.2025.
3.	The love of the Virgin Mary	URL	Public	01.03.2025–01.05.2025.
4.	Love of God	URL	Public	01.03.2025–01.05.2025.
5.	Jesus loves us	URL	Public	01.03.2025–01.05.2025.
6.	In memory of our loved ones	URL	Public	01.03.2025–01.05.2025.
7.	Catholic Prayer Group	URL	Public	01.03.2025–01.05.2025.
8.	Best wishes	URL	Public	01.03.2025–01.05.2025.
9.	Join us	URL	Public	01.03.2025–01.05.2025.
10.	The World of Roses	URL	Public	01.03.2025–01.05.2025.
11.	Let's Praise Jesus Together	URL	Public	01.03.2025–01.05.2025.
12.	God Wishes	URL	Public	01.03.2025–01.05.2025.

Source: Author's own editingA3: Validation of ESMA Framework with a Human-Created Image

To demonstrate the robustness of the ESMA framework, the eight-step manual analysis was applied to an image known to be human-created (source: Pixabay). The IsItAI detection tool also classified this image as "Highly Confident – Human" with a 99 % confidence level. The content safety score confirmed no synthetic or unsafe elements.

Fig. A2. Human-created picture – analysis (ESMA and AI detection tool validation).
Source: URL

Table A3
ESMA analysis of the validation image (people-9185852.jpg)

Step	Category	Observations
1	Human anatomy inaccuracies	No abnormalities in fingers, limbs, or facial proportions; all features appear biologically plausible.
2	Asymmetry and facial oddities	No mismatched eyes, missing ears, or unnatural hairlines. Facial structure appears natural and realistic.
3	Object and environment anomalies	No floating or overlapping objects; environment is spatially coherent.
4	Unnatural textures	Textures (skin, clothing, street) appear naturally varied and realistic.
5	Physics and gravity disregard	No violations of physics; shadows and object positions are consistent with light source and gravity.
6	Texts and lettering issues	No visible textual elements that could be misrendered.
7	Exaggeration of symmetry	Natural human asymmetry present; no artificial balance or mirroring.
8	Crowd and group scene flaws	All background figures appear proportionate and undistorted.

The ESMA framework, supported by the IsItAI tool, accurately identifies this real-life image as non-AI-generated, demonstrating the method's effectiveness in distinguishing authentic content from synthetic ones.

Data availability

Data will be made available on request.

References

Aboutfb.com. (2024). Our approach to labeling AI-Generated content and manipulated media. URL: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/Downloaded. (Accessed 11 January 2024).

Ara, A., Alam, M. S., & Mifa, K. A. F. (2024). A comparative review of ai-generated image detection across social media platforms. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology, 3*(1), 11–22. https://orcid.org/0009-0002-1704-8388.

Barry, A. M. (1997). *Visual intelligence: Perception, image, and manipulation in visual communication*. Suny Press.

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems, 76*, 139–154. https://doi.org/10.1007/s11235-020-00733-2

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Bucher, T. (2019). The algorithmic imAIGInary: Exploring the ordinary affects of Facebook algorithms. In *The social power of algorithms* (pp. 30–44). Routledge.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Chen, Y. C., Huang, P. Y., Lin, C., Huang, Y. T., & Chen, M. C. (2022). Headline diagnosis: Manipulation of content farm headlines. *arXiv preprint arXiv:2204.11408* https://doi.org/10.48550/arXiv.2204.11408.

Chiu, T. K. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2253861

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3–21. https://doi.org/10.1007/BF00988593

DiResta, R., & Goldstein, J. A. (2024). How spammers and scammers leverage ai-generated images on Facebook for audience growth. *arXiv preprint arXiv:2403.12838*. https://doi.org/10.48550/arXiv.2403.12838

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., … Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]" Rrasoning about Iivisible Algorithms in News Ffeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 153–162). https://doi.org/10.1145/2702123.2702556

Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 1–21. https://doi.org/10.1007/s42001-024-00250-1

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus, 3*(2), pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Horstmann, G., & Herwig, A. (2016). Novelty biases attention and gaze in a surprise trial. *Attention, Perception, & Psychophysics, 78*, 69–77. https://doi.org/10.3758/s13414-015-0995-1

Iosifidis, P., & Nicoli, N. (2020). The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation. *International Communication Gazette, 82*(1), 60–81. https://doi.org/10.1177/1748048519880729

IsitAI. Online Text and image detector. URL: https://isitai.com. (Accessed 10 December 2024).

Jackson, K., & Bazeley, P. (2019). *Qualitative data analysis with NVivo*. Sage.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Krawiec, J., & Śliwowski, P. (2024). Heuristics and biases in the post-truth era—a piece of advice for policymakers. In *Evaluation in the post-truth world* (pp. 76–96). Routledge.

Lincoln, Y. S. (1985). *Naturalistic inquiry* (Vol. 75). Sage.

Messaris, P. (1994). *Visual" literacy": Image, mind, and reality*. Westview Press.

Perach, R., Joyner, L., Husbands, D., & Buchanan, T. (2023). Why do people share political information and misinformation online? Developing a Bottom-Up descriptive framework. *Social media + Society, 9*(3). https://doi.org/10.1177/20563051231192032

Raj, R., Sharma, C., Uttara, R., & Animon, C. R. (2024). A literature review on clickbait detection techniques for social media. In *2024 11th international conference on reliability, Infocom technologies and optimization (Trends and future directions) (ICRITO)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICRITO61523.2024.10522359.

Ricker, J., Assenmacher, D., Holz, T., Fischer, A., & Quiring, E. (2024). AI-generated faces in the real world: A large-scale case study of Twitter profile images. In *Proceedings of the 27th international symposium on research in attacks, intrusions and defenses* (pp. 513–530). https://doi.org/10.1145/3678890.3678922

Romele, A. (2022). Images of artificial intelligence: A blind spot in AI ethics. *Philosophy & Technology, 35*(1), 4. https://doi.org/10.1007/s13347-022-00498-3

Shad, R., & Olukemi, A. (2024). *The impact of cognitive biases on consumer decision-making*.

Simões, J. M., & Caldeira, W. (2024). Ethics concerns in the use of computer-generated images for human communication. *Journal of Ethics in Higher Education*, (4), 169–192. https://doi.org/10.26034/fr.jehe.2024.5962

Technologyreview.com. (2023). Junk websites filled with AI-generated text are pulling in money from programmatic ads. URL: https://www.technologyreview.com/2023/06/26/1075504/junk-websites-filled-with-ai-generated-text-are-pulling-in-money-from-programmatic-ads. (Accessed 10 October 2024).

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wahid, R., Mero, J., & Ritala, P. (2023). Written by ChatGPT, illustrated by Midjourney: Generative AI for content marketing. *Asia Pacific Journal of Marketing and Logistics, 35*(8), 1813–1822. https://doi.org/10.1108/APJML-10-2023-994

Walter, Y. (2024). *Artificial influencers and the dead internet theory*. AI & SOCIETY. https://doi.org/10.1007/s00146-023-01857-0

Wang, Y., Hu, B., Tang, C., & Yang, X. (2025). Decoding clickbait: The impact of clickbait types and structures on cognitive and emotional responses in online interactions. *Cyberpsychology, Behavior, and Social Networking, 28*(1), 18–27. https://doi.org/10.1089/cyber.2024.0295

Wei, Y., & Tyson, G. (2024). Understanding the impact of ai-generated content on social media: The pixiv case. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 6813–6822). https://doi.org/10.1145/3664647.3680631

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review, 9*(11).

Whittlesea, B. W. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1235. https://doi.org/10.1037/0278-7393.19.6.1235

Wong, L. P. (2008). Data analysis in qualitative research: A brief guide to using NVivo. *Malaysian Family Physician: The Official Journal of the Academy of Family Physicians of Malaysia, 3*(1), 14.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*. https://doi.org/10.48550/arXiv.2303.18223

Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus, 3*(3), pgae052. https://doi.org/10.1093/pnasnexus/pgae052