VISUAL REVIEW | Vol. 18, No. 6, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.6439

A VISUAL TURING TEST FOR AUTHORSHIP ATTRIBUTION: Visual Cues, Confidence, and Accuracy in Identifying AI-Generated Images among Visual Arts Professionals

ANDRÉS TORRES-CARCELLER [1]
andrestorresb.edu
JÚLIA CASTELL-VILLANUEVA [1]
julia.castellb.edu
[1]Universidad de Barcelona, Spain

KEYWORDS
generative AI
Visual Turing Test
Image
authorship
attribution
metacognition

ABSTRACT
This study adapts a Visual Turing Test to examine authorship attribution (human vs generative AI) among visual arts professionals (N=28) using 10 balanced stimuli. For each image, participants reported an authorship decision, confidence (1–5), weights for five cue families, and aesthetic quality before and after authorship disclosure. Overall accuracy was moderate (57.1%), with false positives for highly polished human images and false negatives for abstract or linear styles. Confidence was poorly calibrated and predicted lower accuracy; structure/space and lighting/color predicted higher accuracy. Disclosure tended to reduce ratings more for AI images.

Recibido: 03/ 07 / 2026
Aceptado: 05/ 10 / 2026

1. Introduction

In 1950, Alan Turing reframed the question of intelligence through a methodological shift. Rather than pursuing an essential definition, he proposed a validation procedure grounded in an assessable situation: whether a machine could sustain a text-based interaction under controlled conditions without being identified as such. The relevance of this move lies not in definitional closure, but in relocating the problem to an operational threshold—the point at which available categories cease to guide judgment with certainty. Transposed to visual culture, the formulation acquires an immediate expression: what anchors a decision when the authorship of an image can no longer be reliably inferred?

The rapid development of Generative Artificial Intelligence (GenAI) has intensified a phenomenon now pervasive across visual culture: the growing difficulty of distinguishing between human and algorithmic production. While early research centered on photographic verisimilitude and face synthesis—largely due to the societal implications of deepfakes (Korshunov & Marcel, 2018)—the present challenge extends beyond photorealism. Of particular relevance is GenAI's capacity to generate images exhibiting stylistic competence and artistic quality across heterogeneous registers, emulating conventions and formal solutions typically associated with authorial practice. In this context, attribution becomes unstable. It no longer depends on detecting technical flaws, but on deciding whether visual coherence and apparent intentionality point to a human process or to an algorithmic synthesis capable of convincingly simulating those same traits. The convergence between artistic visual codes and the recombinatory patterns of generative models (Nightingale & Farid, 2022) thus introduces a zone of perceptual indeterminacy that observation alone cannot resolve. From educational and cultural perspectives, this indistinguishability raises a structural concern. The artistic image has historically functioned as a privileged locus for attributing intention, expressivity, and experience. When such qualities are plausibly simulated by non-human systems, the criteria underpinning judgments of authorship, authenticity, and value are destabilized (Carceller, 2024). The issue, however, is not reducible to a demand for continuous identification. It is more precisely understood as the recognition of a threshold of uncertainty at which interpretation no longer permits secure inference about origin, thereby requiring a reconsideration of the object and conditions of evaluation.

As GenAI becomes embedded within creative practices, analytical attention shifts toward how judgment is sustained when origin ceases to function as a stable reference point. Within this framework, adapting the Turing test to the domain of visual production operates as a productive analytical device, enabling examination not only of AI's generative capacities but also of the cognitive processes through which observers formulate judgments. Recent studies indicate that even in specialized contexts, authorship identification becomes problematic when assessment involves artistic images (Cenerini et al., 2025). The challenge is therefore not merely technological, but epistemological and formative: how visual judgment is reorganized when traditional markers of authorship lose reliability, and how confidence and aesthetic evaluation are reconfigured in a landscape where visual production is no longer the exclusive domain of human agency.

This study is situated at the intersection of visual culture, computational creativity, and art education, focusing on a specific question: how judgment is organized when authorship attribution becomes uncertain yet continues to condition interpretation. Empirical research in aesthetics and visual cognition has shown that contextual information regarding provenance or status can alter evaluative responses even when the stimulus remains unchanged, indicating that judgments are shaped not only by visible properties but also by semantic frameworks and expectations about origin (Kirk et al., 2009). In the specific domain of algorithmic imagery, it has likewise been observed that knowledge or belief regarding GenAI attribution modulates judgments, affecting perceived quality and aesthetic legitimacy (Gangadharbatla, 2022).

Against this backdrop, the present study contributes empirical evidence to ongoing debates on authorship attribution in images whose origins cannot always be inferred from visual features alone. To this end, a Visual Turing Test is adapted to examine how authorship attribution is constructed under conditions of perceptual uncertainty and how contextual knowledge may reorganize aesthetic evaluation. This approach is particularly pertinent within a broader crisis of iconic evidence, wherein images no longer function as automatic guarantors of truth, authenticity, or lived experience. As argued within image theory, trust in the visual is increasingly mediated by technical systems, cultural frames, and learned expectations rather than by any direct relation to a verifiable referent (Ritchin, 2013). Accordingly, the study's contribution lies not solely in measuring identification accuracy, but in

elucidating shifts in evaluative criteria associated with the expansion of GenAI in contemporary visual production.

Aesthetic experience may be understood as a complex episode in which observers articulate processes of perception, interpretation, and evaluation in relation to a visual object. Such episodes exceed the mere detection of formal features, incorporating expectations, prior knowledge, and attributional frameworks that orient judgment. From this standpoint, identifying an image as human-produced or GenAI-generated constitutes a situated cognitive operation within aesthetic experience rather than an isolated technical recognition (Leder & Nadal, 2014). Confidence is introduced here as a metacognitive dimension of judgment. It refers not to objective correctness, but to the degree of certainty individuals assign to their own decisions. Confidence thus functions as a second-order evaluation of the judgment process itself, enabling analysis of whether identifications are experienced as stable, doubtful, or revisable. Research in metacognition has consistently shown that confidence judgments need not align with actual accuracy, yet they remain central to understanding decision-making under uncertainty (Dunlosky & Metcalfe, 2009; Fleming & Lau, 2014; Koriat, 2012). Aesthetic evaluation, in turn, is conceived as an assessment of artistic quality that cannot be reduced exclusively to intrinsic visual properties. Attribution of agency—whether human or artificial—has been shown to modulate emotional response, engagement, and evaluative outcomes, indicating that aesthetic experience is shaped not only by form and content but also by assumptions regarding intention and origin (Demmer et al., 2023).

Within this conceptual framework, identification, confidence, and evaluation are treated as analytically distinct yet interdependent dimensions. Identification activates inferences about origin; confidence reflects the perceived solidity of those inferences; evaluation expresses how attribution informs aesthetic judgment. This articulation enables an integrated analysis of how visual judgment is constructed under conditions in which traditional markers of authorship have lost stability and creative agency can no longer be presupposed.

2. Theoretical Framework

2.1. Perception and Detection of Synthetic Images

The identification of authorship—whether human or generated by artificial intelligence—relies on perceptual inferences that do not ensure reliable discrimination, even in binary-choice tasks. Comparative studies indicate that source attribution frequently approaches chance levels, while evaluative judgments tend to remain independent of authorship when origin is not explicitly disclosed (Ragot et al., 2020). Evidence from research on manipulated image detection points to persistent constraints in human performance and substantial interindividual variability (Bray et al., 2023). This difficulty arises not only from the technical complexity of the stimulus but also from the inadequacy of intuitive criteria when generated images satisfy plausible visual expectations.

Research on synthetic faces further shows that perceptual verisimilitude may reach a threshold at which observers lack stable diagnostic cues, displacing judgment toward conjectures grounded in weak signals or response biases (Lago et al., 2021). Although faces constitute a particularly sensitive domain due to their perceptual salience, these findings describe a broader pattern of indistinguishability applicable whenever images achieve sufficient global coherence. Detection reliability also depends critically on the types of images and generative regimes under examination. Variations in models or production conditions tend to reduce accuracy, rendering performance dependent on specific distributions rather than universal indicators (Lu et al., 2023). This observation is especially relevant in artistic domains, where stylistic variability is intrinsic and the absence of a singular template for "human images" weakens discrimination criteria.

Forensic analyses have examined why identification strategies grounded in visible artifacts progressively lose efficacy as generative models improve (Tolosana et al., 2020). Coherence in lighting, texture, and spatial structure reduces the availability of local irregularities that previously functioned as diagnostic cues. Under such conditions, judgment depends less on error detection and more on inferences regarding the generative process—dimensions that algorithmic systems can increasingly simulate. Subjective responses to synthetic imagery, including phenomena such as the algorithmic sublime or the synthetic uncanny, further mediate interpretation and perceptions of agency (Torres,

2025). Comparable instability appears in automated detection systems, whose methods often generalize poorly when confronted with unseen sources or post-generation transformations (Hulzebosch et al., 2020). In response, evaluation paradigms inspired by the Turing test have been proposed, shifting emphasis from internal metrics toward performance in interpretive tasks under uncertainty (Geman et al., 2015). These approaches enable authorship identification to be framed as a problem of judgment rather than mere technical verification. Even within specialized professional contexts, discrimination remains limited when images exhibit high perceptual quality (Shin et al., 2023), suggesting that expertise alone does not guarantee reliable attribution once stimuli satisfy domain-specific visual expectations.

2.2. The Effect of Attribution on Artistic Valuation

Aesthetic evaluation depends not only on formal qualities but also on the attribution of authorship activated during judgment. When observers assume a human creator, interpretation typically invokes assumptions of intention and effort; when an origin in GenAI is presumed, reading tends to shift toward notions of automation and recombination (Bellaiche et al., 2023).

Attribution thus operates as a condition that systematically modulates evaluative responses (Horton Jr. et al., 2023). Presenting identical works under differing labels may induce a devaluation of aesthetic experience when the work is attributed to AI, even when the stimuli themselves remain unchanged (Chiarella et al., 2022). This effect reveals anthropocentric biases, whereby valuation becomes linked to the presumed presence of human agency. Authorship attribution consequently functions as a normative frame that organizes interpretation prior to engagement with formal properties (Millet et al., 2023). Analytically, two planes may be distinguished: the perceptual plane (composition, technique, style) and the attributional plane (presumed agency). The latter possesses the capacity to reorganize the former, shaping what is regarded as expressive or valuable without any alteration to the visual stimulus.

2.3. Indicators and Heuristics of Visual Reading

Authorship judgments frequently rely on features interpreted as signals of plausibility. Historically, certain traces associated with editing or generation have functioned as perceptual alerts (Korus, 2017), yet this mapping remains unstable. As diagnostic artifacts diminish, judgment shifts toward subtler and context-dependent patterns (Akhtar et al., 2023). Research on detector evaluation reveals a parallel tendency: detection difficulty increases alongside perceptual realism. The V_AI index proposed by Imanpour et al. (2025) exhibits an inverse relationship with detection accuracy, suggesting that visually realistic images tend to be more difficult to identify as synthetic. Consequently, discrimination cannot be reduced to the accumulation of isolated features but instead requires their integration within coherent interpretive structures.

Within artistic contexts, visual interpretation operates as a reconstruction of generative history, whereby observers attempt to infer the operations and constraints underlying image formation (Tang et al., 2023). Heuristics function here as principles of interpretive economy. Their effectiveness derives not from isolated cues but from the convergence of reasons supporting plausible accounts of production. In adaptations of the Turing test to visual domains, recurrent repertoires emerge, including appeals to internal coherence, accumulated aesthetic experience, and perceived signs of materiality or gesture (Wang, 2024). Crucially, these heuristics exhibit structural ambivalence. Irregularities may be interpreted either as markers of human agency or as algorithmic artifacts, while visual consistency may be read as technical mastery or generative homogenization (Shin et al., 2023). The theoretical significance of this ambivalence lies in understanding how observers transform equivocal signals into sufficient grounds for conviction in contexts lacking stable attributional criteria.

3. Objetives

The overall objective is to analyze the difficulty of discerning between human and GenAI authorship in a diverse corpus of images, and to describe the judgment criteria that underpin such attributions among expert observers from creative fields.

The specific objectives are to (i) assess accuracy and characterize error patterns, treating attribution as a cognitive decision subject to bias and variability; (ii) examine confidence associated with each choice as a key metacognitive indicator for regulating judgment under uncertainty; (iii) systematize and compare the heuristics used to justify these decisions, identifying the interpretive frames activated when evaluating an image's origin; (iv) analyze the impact of authorship on aesthetic evaluation by contrasting initial ratings with those collected after disclosure of origin, determining the extent to which this information reorganizes quality judgments; and (v) delineate the implications of these findings for expert visual reading in a hybrid production landscape in which traditional markers of value have lost stability.

4. Methodology

4.1. Study Design

A quantitative within-subjects experimental design was employed, with repeated measures per stimulus. The procedure was framed as a perceptual indistinguishability test applied to authorship attribution. Participants were asked to decide whether each image was human-authored or generated via GenAI, and to report both their confidence and the criteria supporting their judgment. This approach builds on Visual Turing–style tests in the visual domain (Geman et al., 2015). Each participant evaluated a fixed set of ten images in a two-phase sequence: an initial phase to record attribution and baseline aesthetic evaluation, and a subsequent phase to record changes in evaluation after an authorship label was revealed according to a predefined plan.

4.2. Participants and Sampling

Sampling was purposive and restricted to expert profiles associated with creative professions and image-related fields (Table 1). Recruitment was conducted through academic and professional networks and supplemented by a snowball procedure. The primary inclusion criterion was professional practice in the visual arts.

Table 1. Sample profile by primary field (N = 28)

Primary field	n (%)	Mean age (years)	Mean professional experience (years)	GenAI use: mean index (0–4)	GenAI use: mode
Art Education	11 (39,3)	47,5	16,5	1,09	Occasional
Art History	8 (28,6)	52,6	17,2	1,25	Never
Audiovisual	4 (14,3)	31,0	7,5	0,75	Occasional
Animation	1 (3,6)	58,0	25,0	1,00	Occasional
Digital Art	1 (3,6)	43,0	8,0	2,00	Monthly
Drawing	1 (3,6)	58,0	8,0	0,00	Never
Ilustration	1 (3,6)	30,0	15,0	0,00	Never
Painting	1 (3,6)	58,0	15,0	0,00	Never

Note. Approximate mean age (years) was calculated by assigning the midpoint of each range: 18–24 = 21; 25–35 = 30; 36–50 = 43; 51–65 = 58. Approximate mean professional experience (years) was estimated by assigning representative values to each range: <2 = 1; 2–5 = 3.5; 6–10 = 8; >10 = 15; >20 = 25. GenAI use was summarized using (a) an ordinal index (0–4), coded as Never = 0; Occasional = 1; Monthly = 2; Weekly = 3; Daily = 4, and (b) the mode, defined as the most frequent category within each field. Percentages were calculated based on N = 28 and may not sum to 100% due to rounding.

Source: Authors' own work, 2026

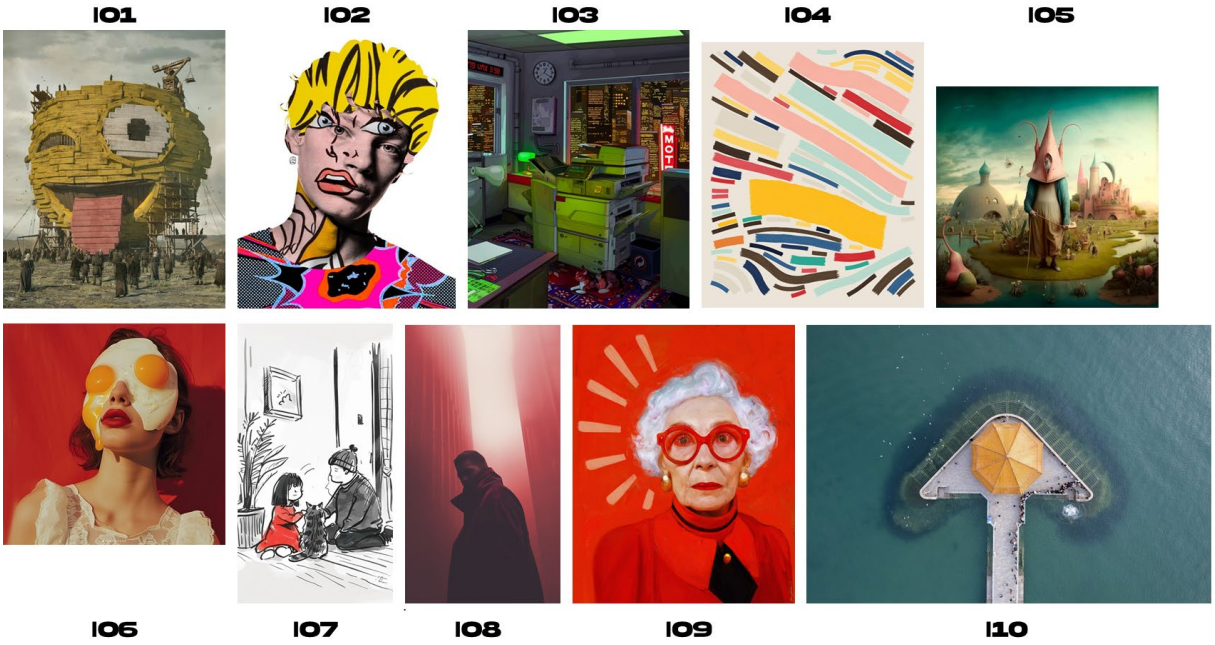
The viewing device was recorded as a contextual covariate: laptop (n = 12), mobile phone (n = 10), and monitor (n = 6); no tablet was used. Responses were collected anonymously, and no personal identifiers were requested.

4.3. Stimuli

The stimulus set shown in Figure 1 consisted of ten images (I01–I10), balanced between human authorship (n = 5) and AI generation (n = 5). To minimize bias arising from technical discrepancies, the images were paired by theme or framing, standardized in resolution and compression, and stripped of metadata. This procedure addresses the risk that attribution may be confounded by non-visual traces and strengthens the validity of judgments under real-world viewing conditions (Hulzebosch et al., 2020).

The selection was designed to challenge attribution under conditions of perceptual indistinguishability, integrating visual languages of contemporary visual culture such as photography, painting, illustration, and digital art. The AI corpus includes both images generated in Midjourney under human agency and works generated autonomously using the *Fidenza* algorithm, designed by Tyler Hobbs. On the one hand, professionally produced human images were prioritized, as their digital finish or extraordinary character can produce false positives. On the other hand, high-quality AI images were selected to minimize salient algorithmic artifacts and encourage false negatives. Taken together, the corpus constitutes a demanding set in which attribution depends on participants' visual reading and heuristics rather than on external markers of provenance.

Figure 1. Visual stimuli used in the test.



Note. **I01**: Beeple (2020), *Everydays: The First 5,000 Days* (detail: emoji); **I02**: Martín Sigwald (2022), *Robert Roy*; **I03**: Mad Dog Jones (2021), *Replicator*; **I04**: *Fidenza* (2021), #222; **I05**: Florelle Guillaume / Midjourney (2023), *The Garden of Earthly Delights* by Hieronymus Bosch in a photographic style; **I06**: @happy_aiccidents / Midjourney (2024), *Untitled*; **I07**: @Tibiami / Midjourney (2024), *Untitled*; **I08**: @Basalte / Midjourney (2024), *Untitled*; **I09**: Alexis Franklin (2024), *Random Portraits II*; **I10**: we-o_rd35hjddgdeuyacd8 (2021), *Untitled*.

Source: Authors' own compilation of the referenced images, 2026

4.4. Instrument

The instrument was implemented as a structured online questionnaire comprising informed consent, participant profiling, ten image-based cycles, and a final block for criteria adjustment. It included mandatory items and 1–5 numerical scales. Each image cycle was organized into two phases. In the initial phase, data were collected on authorship attribution, confidence (1–5), liking, perceived artistic quality, and perceived expertise. In addition, the weight of five families of visual cues was measured: structure and space; technique and materials; repetitions and finishes; composition; and lighting and color. In the second phase, after the authorship label was displayed, change in perceived quality was

recorded using a five-category ordinal scale. The final block provided space to report the main reason underlying correct identifications and any criteria not previously captured.

4.5. Variables

The primary variable was authorship identification (correct/incorrect), from which false-positive and false-negative rates were derived. Confidence, baseline evaluation, and cue weights were measured on ordinal 1–5 scales. Post-disclosure change was recorded on a five-category scale that could be recoded for analysis. Professional field, years of experience, and prior GenAI training were considered as moderators. Viewing device was treated as a contextual covariate.

4.6. Analytic Plan

Authorship identification will be analyzed using a binomial model (logistic regression), with accuracy as the dependent variable. Predictors will include cue weights and profile moderators, incorporating random effects for participant and image. Calibration between confidence and accuracy will be examined by comparing observed performance with reported confidence. Change in perceived quality will be modeled using an ordinal model, considering the interaction between displayed and actual authorship. Robustness analyses by device and sensitivity checks for very short response times are also planned.

4.7. Control and Ethics

Technical comparability of the stimuli and balance within the corpus were maintained. Data collection followed principles of minimization and anonymity and ensured the right to withdraw. Data management complied with Regulation (EU) 2016/679 (GDPR) (European Parliament and Council of the European Union, 2016), with access restricted to the research team.

5. Results

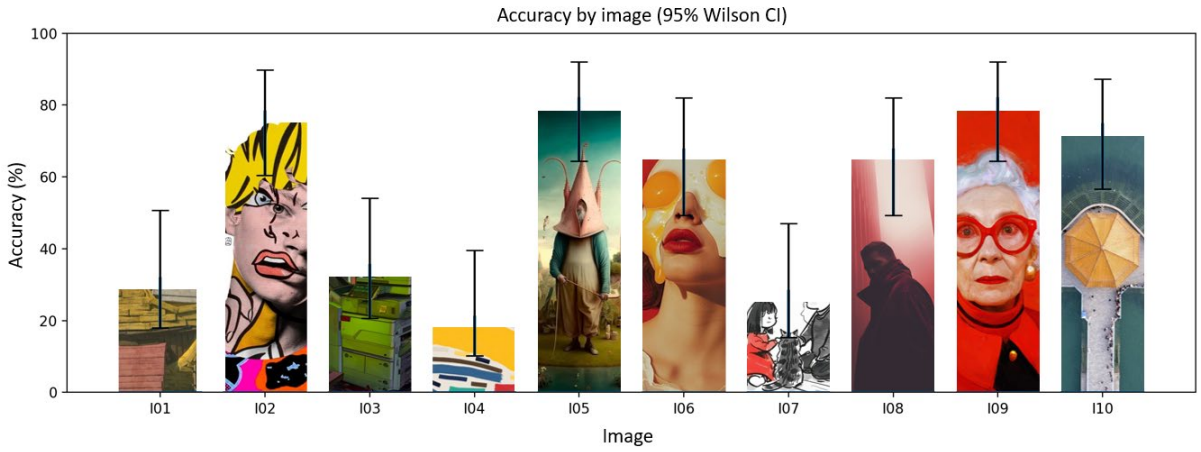
5.1. Authorship Identification and Overall Performance.

Across the total set of decisions (280 judgments), overall accuracy was 57.1% (160/280; 95% Wilson CI [51.3%, 62.8%]). Performance showed asymmetries by true authorship. For human-authored images (5 stimuli; 140 judgments), accuracy was 60.7% (85/140). For GenAI-generated images (5 stimuli; 140 judgments), accuracy was 53.6% (75/140). In confusion-matrix terms, there were 55 false positives (human images classified as GenAI) and 65 false negatives (GenAI images classified as human). This corresponds to a sensitivity of 53.6% (75/140) for detecting GenAI and a specificity of 60.7% (85/140) for recognizing human authorship.

5.2. Stimulus-Specific Difficulty.

Attribution performance varied markedly across individual images. As shown in Figure 2, accuracy rates by stimulus were irregular. In terms of error patterns, the human images I01 and I03 concentrated false positives (19 and 18 classifications as GenAI, respectively), whereas the GenAI images I04 and I07 concentrated false negatives (22 and 20 classifications as human, respectively). Overall performance is therefore not explained by a uniform trend, but by the presence of particularly ambiguous stimuli that polarize error.

Figure 2. Accuracy by confidence level



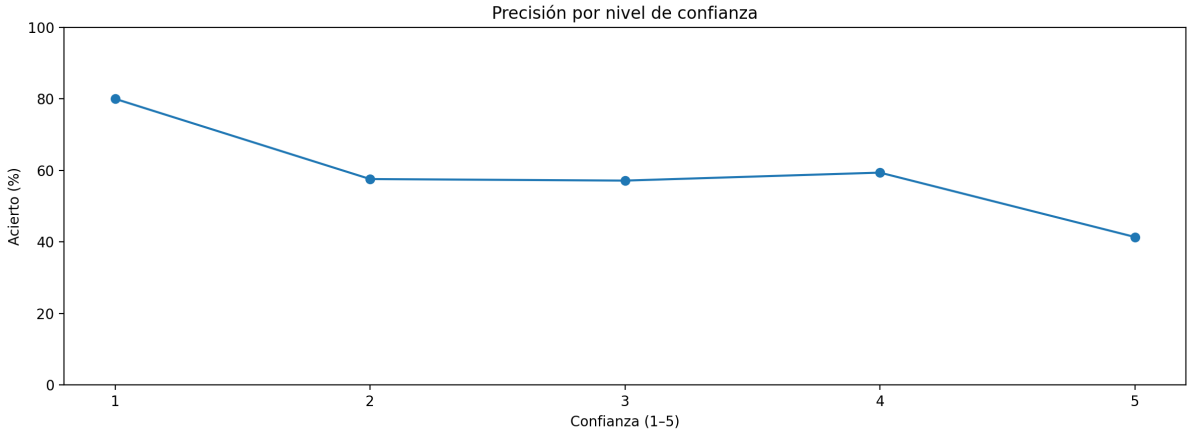
Note. The Y-axis represents the accuracy rate (0–100%) for each confidence level (1–5). Each point corresponds to the proportion of correct responses within that level (n per level: 10, 33, 112, 96, and 29, respectively).

Source: Authors' own work

5.3. Metacognitive Confidence and Calibration.

The mean reported confidence was 3.36 (SD = 0.94) on a 1–5 scale and did not increase with accuracy. Mean confidence was, in fact, slightly higher for incorrect responses (M = 3.46) than for correct responses (M = 3.29). When examined by confidence level, the overall pattern does not indicate monotonic calibration: accuracy remains at intermediate values across levels 2–4 but declines markedly at the highest confidence level, as shown in Figure 3. Although level 1 should be interpreted cautiously given its small sample size, the most robust finding is the reduction in accuracy at the upper extreme of the confidence scale, a pattern consistent with overconfidence in a subset of judgments.

Figure 3. Accuracy by image (95% Wilson CI).



Note. The Y-axis represents the accuracy percentage (0–100%) per stimulus (I01–I10), with 28 judgments per image (one per participant). Error bars correspond to the 95% confidence interval (Wilson method) for each accuracy proportion. The images embedded within the bars serve as visual references for the corresponding stimuli.

Source: Authors' own work

5.4. Reported heuristics and their relationship with accuracy.

When participants were asked to indicate, for each image, the weight assigned to five families of cues in their decision (1–5 scale), the predominant motives declared in the final correct identifications were distributed mainly across technique and materials (n = 7), lighting and color (n = 7), and composition (n = 7), with a lower presence of structure and space (n = 4) and repetitions or homogeneous finishes

(n = 3). To estimate which heuristics were associated with correct attribution, a mixed-effects logistic regression model was fitted with accuracy (0/1) as the dependent variable, fixed effects for the five cue weights and confidence, and random intercepts for participant and image. The results indicate a positive association between accuracy and greater weight assigned to structure and space and to lighting and color, whereas assigning greater weight to technique and materials, repetitions or homogeneous finishes, and composition was associated with a lower probability of accuracy; likewise, higher confidence was associated with lower accuracy (see Table 2).

Table 2. Heuristics and confidence as predictors of accuracy

Predictor (weight 1–5)	OR	Approx. 95% CI
Structure and space	1,39	1,29–1,50
Lighting and color	1,35	1,26–1,45
Technique and materials	0,69	0,64–0,74
Repetitions or homogeneous finishes	0,85	0,79–0,92
Composition	0,82	0,77–0,88
Confidence (1–5)	0,78	0,72–0,84

Note. Mixed-effects logistic model with accuracy (0/1) as the dependent variable; fixed effects for the five cue weights and confidence; random intercepts for participant and image. OR > 1 indicates a positive association with accuracy; OR < 1 indicates a negative association.

Source: Authors' own work

Confidence, consistent with the descriptive analysis, showed a negative association with accuracy in this model (OR = 0.78; approx. 95% CI [0.72, 0.84]). Overall, the results suggest that not all heuristics that appear intuitively plausible operate as effective diagnostic criteria. Some—particularly technique and materials—behave instead as a systematic pathway to error within this stimulus set.

5.5. Artistic quality evaluation and the effect of authorship disclosure.

Prior to authorship disclosure, mean artistic quality was rated at 3.11 (SD = 0.90) on a 1–5 scale. Baseline ratings were slightly higher for GenAI images (M = 3.16) than for human-authored images (M = 3.06). Following disclosure of correct authorship, reported changes in artistic quality evaluation were predominantly "no change" for both stimulus types. However, meaningful differences emerged in the direction of change when it occurred (Table 3).

Table 3. Baseline aesthetic evaluation and change after authorship disclosure

Stimulus type (true authorship)	Baseline quality (M, SD)	No change (%)	Increase (%)	Decrease (%)	Mean change (scale1–5)
Human	3,06	80,7	10,7	8,6	3,01
GenAI	3,16	73,6	4,3	22,1	2,79
Total	3,11 (0,90)	—	—	—	—

Note. Baseline quality was measured on a 1–5 scale. Change was recorded on a 1–5 scale (from "decrease greatly" to "increase greatly"). Percentages describe the distribution of change responses by stimulus type.

Source: Authors' own work

For human-authored images, 80.7% of participants reported no change, 10.7% indicated an increase (increase or increase greatly), and 8.6% indicated a decrease (decrease or decrease greatly). For GenAI images, 73.6% reported no change, 4.3% indicated an increase, and 22.1% indicated a decrease. On average, the mean change score was 3.01 for human images and 2.79 for GenAI images (1–5 scale ranging from decrease greatly to increase greatly). This indicates a net shift toward devaluation when the origin is attributed to GenAI and disclosed as such.

5.6. Descriptive differences by professional profile.

At a descriptive level, mean accuracy per participant was 0.57. By primary field, mean accuracy was 0.57 in art education (n = 11), 0.59 in art history (n = 8), and 0.55 in audiovisual (n = 4). The remaining categories each had n = 1 and do not allow for interpretation. By self-reported GenAI use, mean accuracy was 0.58 for "never" (n = 10) and 0.57 for "occasional" (n = 12), with somewhat lower values for "monthly" (n = 3) and "weekly" (n = 2). By years of experience, means ranged from 0.50 to 0.62 across

brackets, with no clear gradient. Given the sample size and the imbalance across categories, these comparisons are reported as descriptive.

6. Discussion

Authorship attribution—whether human or GenAI—even when performed by experts in image-related professions, does not function as a stable capacity. Instead, it operates as a situated judgment whose reliability fluctuates markedly depending on the stimulus. This variation precludes treating the global accuracy rate as an abstract property of the observer; rather, performance must be understood as the relationship between a repertoire of visual expectations and a specific set of images.

As noted by Nightingale and Farid (2022), when the global coherence of a stimulus is sufficiently high, human discrimination loses reliable anchors and shifts toward conjectures based on weak signals. Consequently, the core interest lies not merely in confirming that attribution is difficult, but in observing how error is distributed and which types of images trigger recurrent forms of confusion among expert readers.

The distribution of accuracy per image provides an important analytical anchor: certain stimuli concentrate false positives (human images frequently attributed to GenAI), while others concentrate false negatives (GenAI images accepted as human). Consistent with the theoretical framework, these shifts suggest that attribution is resolved through plausibility rather than forensic detection; error is concentrated where the stimulus either activates stable expectations of the algorithmic or, conversely, neutralizes the signals the observer expects to encounter.

Along these lines, the weights assigned to cue categories allow us to move beyond a purely technological explanation and situate the problem within the architecture of judgment. The results indicate that not all heuristics participants claim to weigh are associated with better performance. This is a substantive point: attention must shift from an inventory of cues to their actual diagnostic performance. It is therefore necessary to explain why certain signals, despite appearing plausible within the cultural imaginary of the algorithmic, systematically lead to erroneous attributions.

A conceptual caution is necessary here: the instrument records self-reports of what participants believe they used as a reason, not a direct trace of perceptual processes. Nevertheless, the pattern is sufficiently robust to argue that expert reading may be captured by interpretive shortcuts—convincing rationalizations that fail to function as effective diagnostic criteria. This contextual instability of signals aligns with research emphasizing that cue repertoires are not universal but evolve alongside generators, transformations, and modes of circulation. In artistic registers, inference is often oriented less toward error detection and more toward attributing process—looking to infer how an image was produced rather than solely what it depicts—which favors causally plausible, though not necessarily discriminant, interpretations.

The metacognitive dimension reinforces this interpretation, as confidence does not act as a reliable indicator of accuracy; indeed, high confidence frequently coexists with error. This miscalibration is theoretically significant because it suggests that the difficulty is not exclusively perceptual but also concerns the regulation of judgment itself. Under conditions of uncertainty, metacognition does not consist in subjective certainty, but in calibrating the probability of being mistaken. When such calibration fails, judgments not only err but may also become resistant to revision.

For art education, this observation has direct implications. Visual literacy cannot be reduced to the transmission of repertoires of technical signals but must incorporate explicit engagement with judgment calibration, the management of reasoned doubt, and the critical revision of evaluative criteria.

The revelation of authorship at the end of the study adds a decisive analytical layer, showing that attribution affects not only identification but also aesthetic valuation. The observed pattern may be interpreted not only as resistance to the algorithmic but also as a preference for human agency, commonly associated with assumptions of intention, effort, and expressivity, which reorganize the interpretive framework prior to engagement with formal properties. Consequently, the debate on GenAI in art education cannot be limited to technical indistinguishability, as the issue is also cultural, bearing upon how the value of images is legitimized and under what conditions they are regarded as carriers of human experience.

Finally, the analysis by professional fields, given the structure of the present sample, does not warrant robust claims regarding the superiority of one category over another. This limitation, however,

points toward future research directions, including the operationalization of expertise through more refined dimensions such as dominant visual practice, habitual exposure to generated imagery, or the relative primacy of production and interpretive tasks.

Although the study presents limitations—notably the reduced number of stimuli and the reliance on self-reported heuristics—these constraints do not invalidate the findings but rather delimit the scope of inference. The study's most robust contribution is methodological and theoretical, describing how expert judgment is organized when traditional indicators of authorship lose stability and plausibility ceases to function as a secure criterion.

From these findings, several future research directions follow with clarity. First, it is necessary to expand and stratify the stimulus bank across visual registers and stylistic families, incorporating internal replications that allow the influence of authorship to be distinguished from that of representational genre and that help isolate which configurations consistently generate false positives or false negatives. Second, refining the measurement of expertise and exposure to GenAI becomes essential so that profile-based analyses do not rely on categories with insufficient sample sizes. Third, future studies should more clearly separate the perceptual and metacognitive components of judgment by incorporating process-oriented measures, ensuring that the analysis of heuristics is not based exclusively on post hoc accounts but can instead be linked to observable dynamics of decision-making under uncertainty.

The human images that proved particularly difficult to identify (I01 and I03) share a digital polish and technical coherence frequently associated today with algorithmic production despite their human origin. I01 (a scene depicting the construction of a large face-object) appears plausible in its lighting, texture, and depth, yet conceptually hyperbolic and excessively resolved. The combination of monumental scale, the accumulation of micro-elements, and a render-like appearance may activate a heuristic whereby spectacular, hyper-detailed, and visually consistent imagery tends to be attributed to GenAI. In I03 (a nocturnal interior with a 3D illustration aesthetic), the smooth gradients, chromatic uniformity, and cinematic organization of the framing produce a similar effect. The absence of material friction—traces of medium, gestural irregularities, or processual accidents—is not interpreted as technical mastery but rather as a signal of generation. Such false positives reflect a displacement of judgment from local artifacts toward global inferences about production processes and stylistic stereotypes associated with the algorithmic, which may equally capture high-quality human work.

Conversely, the strongest concentrations of false negatives (I04 and I07) appear to operate through the inverse mechanism. GenAI becomes difficult to identify when the visual register aligns with human practices in which indeterminacy, economy of means, or irregularity are not merely acceptable but constitutive. I04 (an abstract composition of chromatic bands and fragments) presents features plausibly read as compositional decisions characteristic of contemporary abstraction: rhythmic distribution, controlled palette, and imperfections interpretable as deliberate traces. When a style does not demand figurative verisimilitude or micro-anatomical consistency, many conventional diagnostic cues lose discriminative force, and the image becomes effectively protected by the interpretive latitude of the genre. I07 (a line drawing depicting an everyday scene) activates an even more error-prone interpretive frame. Attributes commonly associated with human illustration—loose line, simplification, expressive economy, even apparent awkwardness—are now readily emulated by GenAI, while anomalies may be reabsorbed as stylistic choice. In these cases, the heuristic of materiality and gesture functions less as a diagnostic criterion than as a persuasive shortcut: images are attributed as human insofar as they conform to the grammar of drawing, irrespective of algorithmic origin. This pattern aligns with the broader observation that visual reading within artistic registers tends to reconstruct plausible generative histories rather than identify provenance in a forensic sense.

Images in which GenAI was identified more easily (I05, and to a lesser extent I06 and I08) suggest a different configuration of cues, less tied to discrete artifacts than to a global structure of plausible strangeness. I05 (a fantasy scene with a central figure and imaginary architecture) combines elements that have recently functioned as suspicion markers: an oneiric imaginary that is atmospherically coherent yet exhibits a homogeneous illustrative quality, lacking zones of differentiated treatment and containing secondary details perceived as narratively indifferent or formally redundant. Observers may thus perceive a coherence not grounded in hierarchized decisions but in a mode of production that saturates the visual field with uniform intensity. In I06 (a photographic portrait with surreal elements)

and I08 (a figure in shadow with strongly marked chromatic treatment), judgment tends to rely on the conjunction of polish and strangeness. These images appear convincing in lighting and texture yet introduce a degree of formal improbability frequently associated with generation, particularly when the finish appears excessively uniform or highly stylized. In such cases, the operative criterion is not the detection of error but the inference of the type of process capable of producing the equilibrium between high verisimilitude and controlled strangeness.

By contrast, human images with high identifiability (I02, I09, and I10) appear to offer attributional anchors more closely aligned with process-based reading. I02 (a hybrid portrait combining halftones, layering, and drawing) clearly exhibits a logic of assembly and graphic decision-making in which heterogeneous registers coexist without resolving into global homogenization. This controlled heterogeneity functions as a cue of human authorship insofar as it renders visible a sequence of operations and choices unlikely to derive from a unified synthesis. I09 (a painterly portrait with a dominant palette and visible brushwork) reinforces its human reading through recognizable material treatment, with transitions and edges interpretable as products of craft rather than uniform smoothing. I10 (an aerial view of a breakwater) adds a form of indexical plausibility that, although technically reproducible, continues to operate culturally as a practical guarantee through micro-variations in water texture, the scaling of diminutive figures, environmental irregularities, and a geometry tied to real infrastructure. In heuristic terms, these images favor attribution grounded in structure, spatial organization, and material finish as plausible reconstructions of making.

This pattern suggests a central interpretive consequence: difficulty is not distributed simply by degrees of realism but by the extent to which each image permits—or obstructs—the construction of a generative history supported by relatively stable anchors. When images provide signals compatible with recognizable human processes—operational heterogeneity, hierarchized decisions, materiality interpretable as craft—attribution tends to stabilize. When images occupy registers in which style absorbs anomaly (abstraction, line drawing) or display a digital polish aligned with contemporary stereotypes of the algorithmic, attribution becomes fragile and errors cluster systematically. Within this framework, the contribution of stimulus-by-stimulus analysis lies not in cataloguing cues but in demonstrating how interpretive frames that are entirely reasonable may nevertheless prove epistemically insecure within the visual culture of GenAI.

7. Conclusions

The results provide a structured response to the study's objectives while clarifying the scope of authorship attribution when it is framed as a problem of visual judgment under conditions of high ambiguity. The general objective—analyzing the difficulty of discerning between human and GenAI authorship—is fulfilled by showing that identification does not operate as a stable capacity but as a variable decision process that is highly sensitive to the stimulus. The first specific objective—estimating accuracy and characterizing error patterns—is addressed through the dispersion observed across images, where certain stimuli consistently elicited correct attributions and others systematic misidentifications, thereby revealing the internal structure of error rather than merely global performance levels. The second objective—examining confidence as a metacognitive indicator—is satisfied by the finding that reported confidence does not reliably align with judgmental precision. The third objective—systematizing the heuristics mobilized to justify decisions—is met by demonstrating that participants relied on recognizable and verbalizable criteria that did not necessarily produce effective discrimination, making visible the gap between declared reasoning and actual accuracy. The fourth objective—analyzing the effect of attribution on aesthetic valuation—is addressed by showing that explicit knowledge of authorship can modify evaluative judgments, indicating that aesthetic appraisal is not confined to formal image properties but reorganized by attributed agency. The final objective—delineating implications for expert visual reading—is fulfilled by demonstrating that expertise does not eliminate ambivalence but instead reshapes the justificatory logic of judgment in the absence of stable indicators.

This study empirically characterizes authorship attribution as a problem of visual judgment under conditions of indistinguishability rather than as a purely technical problem of detection. Within a single design, the analysis integrates three dimensions that are frequently examined in isolation: identification accuracy, confidence as a metacognitive evaluation of one's own decision, and aesthetic valuation

modulated by authorship once disclosed. By employing a sample of expert observers and a heterogeneous corpus of visual registers, the findings suggest that attribution operates through heuristics that are perceptually and culturally plausible yet not necessarily effective. This discrepancy constitutes a relevant result for both visual studies and art education. Expert judgment neither disappears in the presence of GenAI nor remains unaffected by it; instead, it tends to be reorganized around criteria that derive their persuasive force from compatibility with expectations already reshaped by the contemporary visual presence of GenAI. An operational implication is to treat authorship attribution as an object of inquiry and formation rather than as a competence presumed to be resolved through expertise or familiarity with tools. In art education, this requires shifting the focus from the transmission of putative cues toward critical engagement with interpretive frameworks. Such engagement involves examining which criteria are invoked, which assumptions are treated as self-evident, which elements acquire evidential status without functioning as reliable evidence, and how subjective certainty may conceal fragile inferential processes. The difficulty lies not solely in the capacity of GenAI to induce error but in its effectiveness in doing so by mobilizing predispositions already embedded within visual culture. Long-standing conventions of authenticity have frequently operated as if they were proofs when they functioned instead as stabilized agreements. Once such conventions become systematically reproducible, the issue is not the disappearance of criteria but the exposure of their conventional status. Under these conditions, a defensible conclusion is not to demand unattainable guarantees of origin but to recognize a displacement in the locus of judgment: authorship becomes less decisive than the operations enacted by the image, the effects it organizes, and the economy of meaning it proposes. When authorship itself becomes interchangeable, functioning as a mobile label rather than a stable anchor, the evaluative core of judgment shifts from origin to function.

8. Acknowledgements

This research is part of the project *ARTificial: Integration of Generative Artificial Intelligence as a Creative Tool in the Visual Arts* (REDICE24-3625), funded by the Institute for Professional Development (IDP).

References

Akhtar, Z. (2023). Deepfakes generation and detection: A short survey. *Journal of Imaging, 9*(1), 18. 10.3390/jimaging9010018

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A. & Seli, P. (2023). Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research, 8,* 42. 10.1186/s41235-023-00499-6

Bray, S. D., Johnson, S. D. & Kleinberg, B. (2023). Testing human ability to detect "deepfake" images of human faces. *Journal of Cybersecurity, 9*(1), tyad011. 10.1093/cybsec/tyad011

Carceller, A. (2024). The ARTificial Revolution: Challenges for Redefining Art Education in the Paradigm of Generative Artificial Intelligence. *Digital Education Review, 45*, 84-90. 10.1344/der.2024.45.84-90

Cenerini, C., Keller, F., Pennazza, G., Scantonico, M. & Vollero, L. (2025). Artistic Turing test: The challenge of differentiating human and AI-generated art. *Computers in Human Behavior: Artificial Humans, 6.* 10.1016/j.chbah.2025.100209

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F. & Cartocci, G. (2022). Investigating the negative bias toward artificial intelligence: The effects of prior assignment of AI authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior, 137.* 10.1016/j.chb.2022.107406

Demmer, T. R., Kühnapfel, C., Fingerhut, J. & Pelowski, M. (2023). Does an emotional connection to art really require a human artist? Emotion and intentionality responses to AI- versus human-created art and impact on aesthetic experience. C*omputers in Human Behavior, 148.* 10.1016/j.chb.2023.107875

Dunlosky, J. & Metcalfe, J. (2009). *Metacognition*. Sage.

European Parliament y Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union,* L119, 1–88.

Fleming, S. M. & Lau, H. (2014). How to measure metacognition. Frontiers in Human *Neuroscience, 8,* 443. 10.3389/fnhum.2014.00443

Gangadharbatla, H. (2022). The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts, 40*(1), 55–74. 10.1177/0276237421994697

Geman, D., Geman, S., Hallonquist, N. & Younes, L. (2015). Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences, 112*(12), 3618–3623. 10.1073/pnas.1422953112

Horton Jr, C. B., White, M. W. & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific Reports, 13,* 19001. 10.1038/s41598-023-45202-3

Hulzebosch, N., Ibrahimi, S. & Worring, M. (2020). Detecting CNN-generated facial images in real-world scenarios. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 642–643. 10.1109/CVPRW50498.2020.00329

Korshunov, P & Marcel, S. (2018). *DeepFakes: A new threat to face recognition? Assessment and detection*. 10.48550/arXiv.1812.08685

Imanpour, N., Borah, A., Bajpai, S., Ghosh, S., Sankepally, S. R., Abdullah, H. M., Kosaraju, N., Dixit, S., Aziz, A., Biswas, S., Jain, V., Chadha, A., Wang, S., Sheth, A. & Das, A. (2025). The Visual Counter Turing Test (VCT^2): A benchmark for evaluating AI-generated image detection and the Visual AI Index (V_AI). En *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics,* 1847–1862. The Asian Federation of Natural Language Processing y Association for Computational Linguistics.
https://aclanthology.org/2025.ijcnlp-long.100/

Kirk, U., Skov, M., Hulme, O., Christensen, M. S. & Zeki, S. (2009). Modulation of aesthetic value by semantic context: An fMRI study. *NeuroImage, 44*(3), 1125–1132.
10.1016/j.neuroimage.2008.10.009

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296–298. 10.1016/j.learninstruc.2012.01.002

Korus, P. (2017). Digital image integrity–a survey of protection and verification techniques. *Digital Signal Processing*, *71*, 1-26 10.1016/j.dsp.2017.08.009Get rights and content

Lago, F., Pasquini, C., Böhme, R., Dumont, H., Goffaux, V. & Boato, G. (2021). *More real than real: A study on human visual perception of synthetic faces*. 10.48550/arXiv.2106.07226

Leder, H. & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode—Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*(4). 10.1111/bjop.12084

Lu, Z., Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu & Wanli Ouyang (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. 10.48550/arXiv.2304.13023

Millet, K., Buehler, F., Du, G & Kokkoris, M. D. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior, 143,* 107707. 10.1016/j.chb.2023.107707

Nightingale, S. J. & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, 119*(8), e2120481119. 10.1073/pnas.2120481119

Ragot, M., Martin, N. & Cojean, S. (2020). AI-generated vs. human artworks: A perception bias towards artificial intelligence. En *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. 10.1145/3334480.3382892

Ritchin, F. (2013). Bending the frame: Photojournalism, documentary, and the citizen. Aperture.

Shin, K., Lee, J. S., Lee, J. Y., Lee, H., Kim, J., Byeon, J.-S., Jung, H.-Y., Kim, D. H. & Kim, N. (2023). An image Turing test on realistic gastroscopy images generated by using the progressive growing of generative adversarial networks. *Journal of Digital Imaging, 36*, 1760–1769. 10.1007/s10278-023-00803-2

Tang, N., Gong, S., Zhou, J., Shen, M., & Gao, T. (2023). Generative visual common sense: Testing analysis-by-synthesis on Mondrian-style image. *Journal of experimental psychology. General*, *152*(10), 2713–2734. 10.1037/xge0001413

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. I*nformation Fusion, 64,* 131–148. 10.1016/j.inffus.2020.06.014

Torres, A. (2025). A cartography of ambivalence: The sublime and the uncanny in the generative artificial intelligence iconosphere. *Rita*, (24), 65–78.

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. 10.1109/JSTSP.2020.3002101

Wang, C. (2024). Human or algorithm? The visual Turing test of AI-generated images. *Journal of Multimedia Information System, 11*(3), 201–212. 10.33851/JMIS.2024.11.3.201