When Artificial Intelligence Participates in Creation: Aesthetic Authorship and Critical Frameworks for Generative Art

Haoran Wang [1]

[1] Faculty of Humanities and Arts, Macau University of Science and Technology; Macau, China

Abstract

The maturation and large-scale commercial deployment of generative artificial intelligence (Generative AI) is systematically challenging the attribution of artistic subjectivity, the cognitive foundations of aesthetic judgment, and the evaluative standards of critical discourse. Taking Roland Barthes's "death of the author" and Michel Foucault's concept of the "author function" as theoretical points of departure, and integrating George Dickie's institutional theory of art, Nelson Goodman's symbolic aesthetics, and Helmut Leder's model of aesthetic information processing, this paper examines the theoretical challenges posed by generative art across three core dimensions: (1) the deconstruction of aesthetic authorship — how subjectivity is redistributed in human-machine co-creation; (2) the cognitive effects on aesthetic judgment — the systematic influence of "human-created" labeling on aesthetic evaluation; and (3) the reconstruction of critical frameworks — the applicability and limitations of existing art criticism tools when confronted with AI art. Drawing on Grba's (2022) critical framework for AI art and Bellaiche et al.'s (2023) cross-cultural experimental study as empirical support, the paper proposes a Stratified Attribution Model (SAM) that distinguishes three strata of subjectivity in AI art: an Intentional Layer (the human prompter/curator), a Procedural Layer (the algorithmic system and training data), and an Output Layer (the receptive community of the generated result). On this basis, a four-dimensional critical framework for generative art is constructed: intentional depth, algorithmic transparency, training data ethics, and receiver co-construction. The paper's central argument is that the aesthetic challenges of generative art do not announce the termination of artistic subjectivity, but rather compel art criticism to theoretically reconstruct subjectivity from the myth of individual genius toward distributed creative processes — a reconstruction that carries direct normative implications for contemporary arts education, copyright legislation, and critical practice.

Keywords: *generative art, aesthetic authorship, death of the author, stratified attribution model, AI art critical framework, human-machine co-creation*

1 Introduction: An Epistemological Crisis of "Who Creates"

In October 2018, Christie's sold Edmond de Belamy, a portrait generated by the Paris-based collective Obvious using a generative adversarial network (GAN), for USD 432,500. The work was famously signed with the loss function formula used to generate it. This event quickly became one of the most controversial topics in contemporary art[1]: what exactly was purchased in this transaction? Who is the author of the work—the human members who designed the prompts, the thousands of historical painters embedded in the training dataset, the GAN itself, or the operator who ultimately pressed the "generate" button?

With the large-scale commercialization of text-to-image models such as Midjourney, DALL·E 2, and Stable Diffusion, this question has evolved from an isolated case into a systemic epistemological crisis[2]. The involvement of generative AI destabilizes the traditional structure of authorship in artistic production: when anyone can generate visually sophisticated images by simply inputting a text prompt, the boundaries of the "artist," the criteria of "originality," and the standards of "aesthetic value" are all fundamentally challenged by technological conditions.

The theoretical objective of this paper is to critically engage with existing aesthetic theories and propose a Stratified Attribution Model (SAM) alongside a four-dimensional critical framework applicable to generative art. These aim to provide operational conceptual tools for contemporary art criticism, while maintaining methodological awareness of the framework's limitations.

2 Theoretical Resources: From "The Death of the Author" to the "Author-Function"

2.1 Barthes' "Death of the Author" and the Subjective Vacuum in Generative Art

In his seminal 1967 essay "The Death of the Author," Barthes declared that the meaning of a text is not determined by the author's intention but is produced through the act of reading.[1] The death of the author gives birth to the reader—shifting the authority of meaning from creator to audience. While this claim was emancipatory in literary criticism, its full application to visual art has long faced resistance from traditions emphasizing artistic intention.

Generative AI literalizes Barthes' claim in an unexpected way: in many AI-generated works, human intention is highly ambiguous or even absent. The final outputs often exceed the expectations of prompt designers and contain complex statistical associations derived from training data. Here, the "death of the author" is no longer a critical stance but a technical fact: the randomness and high dimensionality of generative processes make it difficult to attribute any work to a single coherent human intention.

2.2 Foucault's "Author-Function" and the Reconstruction of Attribution in AI Art

In 1969, Foucault offered an important revision to Barthes' argument. Rather than abolishing the concept of the author, he redefined it as a discursive function—the author-function.[2] This function links particular discourses (works) to specific subjects (authors) as a classificatory operation, serving purposes such as legitimization, copyright attribution, and interpretive control.

Foucault's insight reveals that even if the myth of the individual author is deconstructed by technology, the social necessity of the author-function—a locatable agent of responsibility—persists. Generative art exposes the tension within this framework: while technological processes make individual authorship increasingly difficult to defend, institutional practices such as copyright law, auction records, and museum archives continue to require identifiable authors.This tension is not a theoretical contradiction but reflects the current instability of generative art at the institutional level[3].

2.3 Dickie's Institutional Theory and Goodman's Symbol Theory

Dickie's institutional theory of art argues that an object becomes "art" not because of intrinsic properties but because it is designated by the "artworld" as a candidate for appreciation[4]. This perspective provides an institutional lens for understanding the legitimacy of generative art: AI-generated images enter the art market and museum systems not due to their technical attributes alone, but because of recognition by curators, auction houses, and critics[5].

Goodman's symbol theory offers an ontological complement. He distinguishes between autographic arts (where identity depends on the history of production) and allographic arts (where identity depends on correct symbolic interpretation). Digital generative art aligns more closely with the latter: its value does not depend on a unique physical object but on conventions of symbolic interpretation[6]. This distinction has direct implications for authorship, copyright, and collection practices.

3 Empirical Basis: Attribution Effects in Aesthetic Judgment

3.1 Bellaiche et al. (2023): A Cross-Institutional Randomized Controlled Experiment

Bellaiche et al. (2023) [4]conducted one of the most rigorous randomized controlled experiments to date, examining the systematic impact of attribution labels on aesthetic evaluation[7]. The study presented 960 AI-generated images, randomly labeled as either "human-created" or "AI-created," and asked participants to rate them across four dimensions: liking, beauty, profundity, and monetary worth.

The key finding was that works labeled as "human-created" consistently received significantly higher ratings across all dimensions, even after controlling for visual features. More importantly, when participants explained their preferences, the two most cited reasons were perceived emotional investment and inferred creative effort—both unrelated to the actual visual content and purely triggered by attribution labels.

This finding has important methodological implications: it demonstrates that current aesthetic judgments contain significant non-aesthetic psychological biases. Any critical framework for generative art must explicitly distinguish between effort perception bias and the independent evaluation of aesthetic qualities.

3.2 Grba (2022): A Review of AI Art Criticism Frameworks

Grba (2022) [3]provides the most comprehensive review to date of AI art criticism frameworks, tracing developments from early experiments in the 1970s to contemporary diffusion-based models. He identifies key evaluative dimensions such as:Poetic complexity: the integration of conceptual depth and formal exploration;Epistemic honesty: transparency regarding technological processes and data sources[8]; Ethical reflexivity: critical awareness of the artwork's own technological conditions[9].

While highly influential, Grba's framework lacks a systematic account of the co-constructive role of the

audience[10]. This paper addresses that gap by incorporating audience participation as the fourth dimension of the proposed framework.

4 The Stratified Attribution Model (SAM)

4.1 Basic Structure of the Model

The Stratified Attribution Model (SAM) proposed in this paper divides the aesthetic agency of generative art into three analytical layers, each corresponding to distinct forms of agency and critical focus(see in Table 1 and Table 2):

Intentional Layer (IL): This layer refers to human participants—prompt engineers, model trainers, curators, or commissioners—who impose directional control over the creative process[11]. The key question is: to what extent do human agents shape the conceptual framework, aesthetic direction, and ethical boundaries of the generated work? While this aligns with traditional intention-based criticism, it must account for the inherent uncertainty of generative processes.

Procedural Layer (PL): This layer corresponds to the AI system itself[12], including model architecture (e.g., transformers, diffusion models), training datasets (with embedded aesthetic traditions and cultural biases), and generative algorithms. The critical focus here is: which aesthetic features emerge statistically from the system rather than directly from human intention? How does the cultural-geographic distribution of training data influence the generated output?

Output Layer (OL): This layer refers to the field of meaning generated when the artwork encounters its audience, including critics, institutions, and the public[13]. Echoing Barthes' reader-centered theory, this layer highlights that, due to incomplete intention and procedural opacity, the audience plays a significantly greater role in meaning construction in generative art than in traditional contexts.

Table 1. Three-layer analytical dimensions and critical foci of the Stratified Attribution Model (SAM)

Layer	Type of Agency	Core Question	Theoretical Reference	Critical Tools
Intentional Layer (IL)	Human intentional and directional control	What do human participants shape? Where are the boundaries?	Intentionalism; Institutional Theory of Art (Dickie, 1974)	Prompt analysis; examination of artist statements
Procedural Layer (PL)	Statistical transformation by algorithmic systems	Which features emerge from the algorithm? How do data biases influence style?	Symbol theory of art (Goodman, 1968); algorithmic criticism	Model auditing; dataset transparency reports
Output Layer (OL)	Meaning construction by the audience community	How do audiences co-construct meaning? How do attribution labels affect judgment?	Death of the Author (Barthes, 1967/1977); Aesthetic information processing model (Leder et al., 2004)	Audience experiments; discourse analysis

Note: The three layers are not discrete or isolated; rather, they are interrelated and nested analytical dimensions. In actual critical practice, analysis across these layers should be integrated rather than conducted separately.

4.2 Theoretical Propositions of the SAM Model

The core theoretical claim of the Stratified Attribution Model (SAM) is that the aesthetic evaluation of generative art should be simultaneously and stratified across three layers, rather than reduced to a single axis of attribution (e.g., "this is an AI work" or "this is an artist's work"). This claim entails three key implications:

First, plurality of attribution. A single generative artwork can be evaluated at the intentional layer as an expression of the curator's or designer's aesthetic vision[14]; at the procedural layer as a statistical projection of specific training data and aesthetic traditions; and at the output layer as the product of collective interpretation by its audience. These modes of evaluation are all legitimate and mutually complementary rather than mutually exclusive.

Second, stratification of responsibility. Ethical responsibility in generative art should be distributed across layers. Human participants at the intentional layer bear primary responsibility for the ethical consequences of the final work, while developers at the procedural layer (e.g., model trainers) bear institutional responsibility for issues such as copyright in training data and systemic bias in outputs. The SAM model thus offers a layered alternative to "all-or-nothing" approaches to responsibility in copyright legislation.

Third, layer-specific critical focus. Critical tools developed for traditional art—such as intentionalist analysis of artist statements, biography, and intellectual background—are primarily applicable to the intentional layer and should not be uncritically extended to the entire generative artwork. Evaluation of the procedural and output layers requires new methodological tools, including algorithmic criticism, data ethics analysis, and audience studies.

5 Construction of a Four-Dimensional Critical Framework

Building on the SAM model, this paper proposes a four-dimensional critical framework for generative art (see Table2), in which the four dimensions correspond to the central evaluative concerns in contemporary generative art criticism:

Table 2. Integrated structural diagram of the four-dimensional critical framework for generative art and the Stratified Attribution Model (SAM) Four-Dimensional Critical Framework for Generative Art (with SAM Layer Correspondence)

Critical Dimension	Core Evaluation Question	Corresponding SAM Layer	Theoretical Reference
D1 Intentional Depth	How clear and profound are the human participant's conceptual vision, aesthetic choices, and ethical boundaries?	Intentional Layer (IL)	Intentionalism; Institutional Theory of Art
D2 Algorithmic Transparency	To what extent does the creator disclose the model architecture, generative parameters, and technical processes with epistemic honesty?	Procedural Layer (PL)	Grba (2022); Epistemic Honesty
D3 Training Data Ethics	How are issues of copyright compliance, cultural representation, and generative bias in training data addressed and reflected upon?	Procedural Layer (PL)	Data Ethics; Symbol Theory of Art (Goodman)
D4 Receiver Co-construction	Does the work provide meaningful space for interpretive participation, and how diverse is the audience community involved in meaning-making?	Output Layer (OL)	Barthes (1967); Bellaiche et al. (2023)

Core Proposition of SAM:The aesthetic value of generative art should be evaluated simultaneously and in a stratified manner within an integrated framework that combines three layers of attribution (IL/PL/OL) and four dimensions of criticism (D1–D4).
Note: The four-dimensional framework is not a scoring scale but an analytical tool. The relative weight of each dimension may vary depending on the type of artwork, exhibition context, and critical purpose.

Taking Obvious's Edmond de Belamy as an example, the proposed framework can be briefly applied as follows. At the level of D1 (Intentional Depth), the collective clearly stated that its creative intention was to explore the application of GANs within the art-historical tradition of portraiture, giving the work a certain degree of conceptual clarity. At the level of D2 (Algorithmic Transparency), Obvious disclosed the GAN architecture it used, but did not fully reveal the training parameters, meaning that its transparency remained limited. At the level of D3 (Training Data Ethics), the training data came from the publicly available WikiArt dataset, providing a basic level of copyright legitimacy; however, there was almost no reflection on the cultural and geographical bias of the dataset, which was dominated by European male artists. At the level of D4 (Receiver Co-construction), the loss function formula printed as the signature provided an interpretive entry point for informed viewers, but for audiences without a technical background, the space for co-construction remained quite limited.

6 Reflection on Existing Critical Paradigms and Methodological Awareness

6.1 Recontextualizing the Aesthetic Information Processing Model

Leder et al. (2004) [7]proposed an aesthetic information processing model that divides aesthetic appreciation into five stages: perceptual analysis, implicit memory integration, explicit classification, cognitive mastering, and affective evaluation.The experimental results of Bellaiche et al. (2023) [4]suggest that attribution labels mainly interfere with the stages of "cognitive mastering" and "affective evaluation." When viewers attribute a work to AI, their sense of self-efficacy in cognitive mastering decreases ("this is not the kind of art I can understand"), which in turn leads to systematically lower affective evaluation scores.

This mechanism reveals a key task for critical education in generative art: to cultivate aesthetic literacy that helps

audiences consciously distinguish between projective judgments about the creator and independent judgments about the visual and conceptual qualities of the work itself. This does not mean denying the reality of attribution effects; rather, it means bringing them into the scope of critical awareness so that viewers can reflectively regulate aesthetic bias in an informed manner.

6.2 Limitations of the SAM Model and Future Research

The SAM model and four-dimensional critical framework proposed in this paper remain at a preliminary theoretical stage and have several limitations that require further research.

First, the dynamic interaction between the three layers—especially how the intentional layer constrains the procedural layer through prompts, and how effective this constraint is within high-dimensional generative space—needs to be examined through empirical research in computational aesthetics.

Second, the relative weight of each dimension in the four-dimensional framework still lacks cross-cultural validation. Different understandings of the concept of "author" in East Asian aesthetic traditions—for example, the intertwining of authorship and reception in Chinese poetry and painting—may produce culturally specific weightings of these dimensions.

Third, the rapid iteration of generative models requires the framework to remain open to technological change. Analytical dimensions that are currently suitable for diffusion models may need systematic revision when the next generation of generative architectures emerges.

7 Conclusion: The Reconstruction of Attribution Rather Than the Disappearance of Subjectivity

Through theoretical construction and empirical integration, this paper argues that the aesthetic challenge posed by generative art does not signal the end of artistic subjectivity. Rather, it compels critical discourse to move away from the myth of the singular individual genius and toward a multilayered analysis of distributed creative processes. This shift responds theoretically to the classic arguments of Barthes and Foucault, is supported empirically by the studies of Bellaiche et al. (2023)[4] and Grba (2022)[3], and is operationalized through the proposed SAM model and four-dimensional critical framework.

This shift has direct implications for three practical fields. In art education, algorithmic literacy and data ethics should be introduced so that learners can develop critical capacities at the intentional, procedural, and output layers, rather than relying on wholesale affirmation or rejection of AI art. In copyright legislation, SAM's layered model of responsibility offers a more refined alternative to the current "all-or-nothing" model of copyright attribution: human contributions at the intentional layer may be protected, while mandatory disclosure requirements should be imposed regarding copyright compliance in training data at the procedural layer. In art criticism, the four-dimensional framework provides critics with analytical tools that move beyond the pseudo-question of "whether this is real art," enabling independent and arguable value judgments regarding intentional depth, algorithmic transparency, training data ethics, and receiver co-construction.

The era of generative AI in artistic creation does not mark the end of art criticism. Instead, it offers an opportunity for art criticism to advance toward a higher level of methodological self-awareness.

References

[1] Logie, J. (2013). 1967: The birth of "The Death of the Author". College English, 75(5), 493-512.

[2] Foucault, M. (1969). Qu'est-ce qu'un auteur?.

[3] Grba, D. (2022). Deep else: A critical framework for ai art. Digital, 2(1), 1-32.

[4] Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., ... & Seli, P. (2023). Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. Cognitive research: principles and implications, 8(1), 42.

[5] Dickie, G. (1974). Art and the aesthetic: An institutional analysis.

[6] Goodman, N. (1976). Languages of art: An approach to a theory of symbols. Hackett publishing.

[7] Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. British journal of psychology, 95(4), 489-508.

[8] Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. arXiv preprint arXiv:1706.07068.

[9] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2), 3.

[10] Rinkerman, G. (2023). Artificial Intelligence and evolving issues under US copyright and patent law. Interactive Entertainment Law Review, 6(2), 48-65.

[11] Di Dio, C., Ardizzi, M., Schieppati, S. V., Massaro, D., Gilli, G., Gallese, V., & Marchetti, A. (2023). Art made by artificial intelligence: The effect of authorship on aesthetic judgments. Psychology of Aesthetics, Creativity, and the Arts.

[12] Wollheim, R. (1980). Art and Its Objects (2nd ed.). Cambridge: Cambridge University Press.

[13] McCormack, J., Gifford, T., & Hutchings, P. (2019, April). Autonomy, authenticity, authorship and intention in computer generated art. In International conference on computational intelligence in music, sound, art and design (part of EvoStar) (pp. 35-50). Cham: Springer International Publishing.

[14] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.