Theory as Architecture: Why Aesthetic Frameworks Should Constitute, Not Merely Evaluate, AI Art Systems

Atticus Sims
Department of Arts and Design
University of Macau
Macau, China
atticussims@um.edu.mo

Abstract

Every major AI art system—AARON, The Painting Fool, AICAN, Botto, Keke Terminal—treats aesthetic theory predominantly as applied: evaluation criteria operating on outputs whose generation is largely independent of those criteria. We argue that the field has not yet made explicit a distinction that is implicit in its own frameworks. We distinguish *applied* aesthetic theory (post-hoc filtering, scoring, curation) from *constitutive* aesthetic theory (commitments embedded in architecture, shaping generation, evaluation, and developmental trajectory). The diagnostic is a motivated counterfactual test: changing the constitutive theory should change the system's evaluation ontology—the categories through which it perceives and judges—not merely its selection preferences. We trace this distinction through Colton's creative tripod, Wiggins's formal framework, Jennings's autonomy requirements, and Galanter's call for theoretical grounding, showing that each identifies pieces of the problem without articulating the architectural relationship explicitly. Drawing on Bense's creation/communication pattern and Moles's semantic/aesthetic information distinction, we offer independent diagnostic lenses that sharpen the analysis, revealing a gap between behavioral autonomy and evaluative autonomy that current systems leave unaddressed. We propose theoretical coherence as a new evaluative dimension and argue that constitutive pluralism—different theories yielding structurally different systems—is both achievable and desirable.

Introduction

The computational creativity community has built increasingly sophisticated AI art systems. Botto deploys multi-agent creative reasoning, autonomous code generation, and self-directed stylistic exploration (Botto Project 2025). Keke Terminal orchestrates VLM-based evaluation through a ReAct agent framework with autonomous creative decision-making (Dark Sando 2024). The Painting Fool generates its own fitness functions through automated theory formation (Colton 2012). AICAN implements adversarial objectives derived from psychological theories of art-historical dynamics (Elgammal et al. 2017). These systems represent genuine engineering achievements and growing behavioral autonomy.

Yet a persistent asymmetry runs through this progress. On the engineering side, the sophistication is remarkable: multi-agent architectures, meta-prompting, knowledge graphs, generative code synthesis. On the evaluative side, a different picture emerges. Botto's self-critique dimensions intermix compositional, narrative, affective, and sociological categories without any theoretical principle governing their integration (Botto Project 2025). Keke Terminal's taste module combines VLM evaluation with neural aesthetic prediction and social engagement metrics into a composite reward signal (Dark Sando 2024). Even The Painting Fool's emergent evaluative principles are discovered statistical associations rather than theoretically derived judgments (Colton 2008). The engineering is extraordinary. The aesthetics is ad hoc.

This paper's position is direct: the explanation lies in the relationship between aesthetic theory and system architecture. The field has treated aesthetic theory as something *applied to* systems—filtering, scoring, curating outputs—rather than as something that *constitutes* them. This is not merely a problem for the pre-foundation-model systems in our analysis; it is amplified by the foundation model era, where a system can fine-tune Stable Diffusion with aesthetic reward or prompt an LLM to reason about visual quality, yet in both cases the aesthetic criteria, though encoded in model weights or prompt configurations, steer the system's pre-existing evaluative capacities rather than reorganizing them—applied theory at unprecedented scale.

We distinguish *applied* from *constitutive* aesthetic theory and argue that making this distinction architecturally explicit reveals a pattern the field has not fully reckoned with. The distinction builds on Colton's creative tripod, Wiggins's formal framework, Jennings's autonomy requirements, and Galanter's call for aesthetic grounding. Our contribution is to articulate the architectural dimension these frameworks converge on but leave implicit, and to offer diagnostic tools for making it precise.

We also identify a gap between *behavioral autonomy* (deciding what and how to create) and *evaluative autonomy* (having internal, principled criteria for what counts as creating well). Current systems achieve the former without the latter—a distinction with immediate practical consequences for how the CC community designs, evaluates, and situates autonomous creative agents.

The argument proceeds as follows: §2 develops the distinction and its diagnostic test; §3 applies it to five systems; §4 sharpens it through information aesthetics; §5 demonstrates constitutive pluralism; §6 proposes theoretical coher-

ence as a new evaluative dimension; §7 concludes.

Applied vs. Constitutive Theory

The Distinction

Consider two ways aesthetic theory can figure in the design of an AI art system.

In the *applied* role, a generative model produces outputs and an evaluative process selects among them. The evaluative criteria are external to the generative process: generation might produce work of extraordinary subtlety or banality by the same mechanism, with aesthetic assessment arriving afterward. The diagnostic feature of applied aesthetic theory is that changing the evaluative criteria changes which outputs are *selected* but does not change the *generative process itself*. The same generator, paired with different filters, yields different selections from an identical output distribution.

In the *constitutive* role, aesthetic commitments are embedded in the system's architecture so that they shape generation from within. The theory determines which features the system measures and how it organizes them, what the system optimizes for during generation, what the system treats as creative success, and how its creative trajectory unfolds over time.

This is not merely a distinction between "having" and "not having" an aesthetic theory. Systems with applied aesthetics may invoke sophisticated vocabulary, but invoke it as evaluation labels rather than as principles that shape architectural decisions. The question is whether the aesthetic theory does *architectural work*: whether removing it would change the system's structure, not merely its selection.

The Motivated Counterfactual Test

The diagnostic for constitutive embedding requires more precision than "would a different theory produce a different system?"—a test trivially satisfied by any design parameter. The relevant counterfactual is *ontological*, not merely parametric: changing the constitutive theory should change the system's *evaluation ontology*—the categories through which it perceives, the dimensions along which it judges, the developmental dynamics it exhibits. An applied substitution (replacing one scoring function with another) changes selection behavior while preserving the evaluation ontology. A constitutive substitution changes what the system *can perceive as relevant*.

Consider an analogy. Two telescopes with different color filters (applied difference) produce different images of the same sky. Two telescopes built on different physical principles—optical versus radio (constitutive difference)—perceive different phenomena entirely. The constitutive/applied distinction is the difference between interchangeable filters and fundamentally different instruments. What makes the difference constitutive rather than parametric is that it reorganizes the space of possible evaluative judgments rather than selecting among judgments the system is already capable of making.

A concrete test: if we can state the system's theory, predict from that theory alone what categories of evaluative judgment the system can and cannot make, and verify that those predictions hold, then the theory is constitutive. If two systems with different stated theories nonetheless make the same categories of judgment—differing only in weights, thresholds, or selection rates—then the theories are applied, regardless of how they are described in documentation.

Relationship to Existing Frameworks

The constitutive/applied distinction builds on insights already present in CC research. Colton's (2008) creative tripod demands that skill, appreciation, and imagination be simultaneously present. The FACE model (Colton, Charnley, and Pease 2011) formalizes the distinction between A^g-level operation (generating a particular aesthetic measure) and A^p-level operation (generating methods for generating aesthetic measures). Jennings (2010) argues that creative autonomy requires internal direction of creative decisions. Galanter (2012) has argued explicitly that generative artists should ground their systems in aesthetic theory, proposing complexity science as a framework. Wiggins's (2006) formal framework distinguishes exploratory from transformational creativity—searching within versus modifying the conceptual space. Ventura's (2016) argument that "mere generation" is an insufficient barometer reinforces the point from a different angle.

The evolutionary art literature makes the same point from the engineering side. Romero and Machado (2008) document how fitness function design shapes the entire character of evolutionary art output. Machado and Cardoso's (2002) NEvAr demonstrated that theoretically motivated complexity measures could serve as fitness functions—an important early example of constitutive aesthetic embedding. Todd and Latham (1992) and Sims (1994) similarly embedded aesthetic principles in fitness functions that shaped the generative process itself.

These contributions converge on the recognition that evaluation must be *internal* for creative autonomy and *principled* for aesthetic seriousness. What they leave unexamined is the precise *architectural* relationship between an aesthetic theory and the system it inhabits. We add two requirements:

Theoretical grounding: the system's evaluative criteria form a coherent framework explicable in terms of a theory of aesthetic value. A system could achieve A^p-level operation while those measures are statistically learned regularities. Theoretical grounding is what distinguishes aesthetic *judgment* (evaluating for articulable reasons) from aesthetic *prediction* (forecasting reception from learned correlations). This distinction does not presuppose a narrowly rationalist account of judgment: a system could exercise judgment through trained perceptual sensitivity, provided that sensitivity is grounded in a theory of what makes the relevant features aesthetically significant. The distinction also survives institutional challenges: even if audience reception is partly constitutive of aesthetic value (Danto 1964; Dickie 1974), a system that predicts reception without modeling *why* certain features elicit it remains in the applied role.

Architectural traceability: the theory does demonstrable architectural work, verifiable through the motivated counterfactual test. A different theory should produce a different

evaluation ontology—not merely different scores.

Three Thresholds

The spectrum from applied to constitutive admits three analytically distinct thresholds:

Constitutive embedding obtains when aesthetic commitments shape the generative architecture, passing the motivated counterfactual test (different theory → different evaluation ontology). Cohen's AARON satisfies this threshold.

Theoretical coherence obtains when the embedded commitments form a unified framework—when the system's evaluative dimensions derive from a common theory of aesthetic value. "Coherent" need not mean "derived from a single formal theory": Cohen's AARON exhibited implicit coherence—his artistic vision unified the rules—even though the rules accumulated pragmatically. The threshold is satisfied when a knowledgeable observer can identify a unifying account connecting the system's evaluative dimensions, whether formalized or not. No existing system fully satisfies this threshold in the strong sense, though AARON and AICAN approximate it.

Adaptive constitutiveness obtains when the system can modify its constitutive commitments in response to its own creative practice—when theory and practice develop together. This corresponds to Wiggins's transformational creativity applied to the evaluative framework itself. No existing system satisfies this threshold, though The Painting Fool's ability to generate new fitness functions represents a partial step.

These thresholds are analytically distinct: different systems achieve different subsets. We argue that the conjunction of all three is what creative autonomy in the fullest sense requires—while recognizing that each threshold independently represents a meaningful advance.

These thresholds are not arbitrary divisions but derive from the convergence of requirements identified across CC evaluation frameworks. Constitutive embedding synthesizes Jennings's (2010) internal direction requirement with Galanter's (2012) call for theoretical grounding. Theoretical coherence extends Colton's (2008) appreciation criterion from requiring that a system have evaluative reasons to requiring that those reasons form a unified framework. Adaptive constitutiveness applies Wiggins's (2006) transformational creativity to the evaluative framework itself. Each threshold is independently motivated; together they form a progression from principled design through theoretical unity to creative self-modification.

Other evaluative dimensions—perceptual grounding, social situatedness, communicative capacity—are important for assessing creative systems but concern the system's capabilities rather than the theory-architecture relationship that the constitutive/applied distinction targets. The three thresholds are not exhaustive of what matters in CC evaluation; they are exhaustive of the analytically distinct positions a system can occupy along the specific axis from applied to constitutive aesthetic theory.

Five Systems, One Pattern

We examine five AI art systems spanning rule-based, evolutionary, adversarial, and LLM-based approaches. For each, we ask: where does the selective principle originate? Does the system have aesthetic *reasons* for its choices? At which threshold does it fall? We take care to assess each system charitably, on the basis of what its publicly documented architecture reveals, while acknowledging that our analysis is necessarily limited by available documentation. Where systems have been described in peer-reviewed publications, we prioritize those sources.

AARON (Cohen, 1968–2016). Cohen's system is the strongest case of constitutive aesthetic theory in the field's history. Over nearly five decades, Cohen encoded approximately 300 hierarchical production rules implementing his compositional and figural knowledge, including spatial models, postural rules, and compositional evaluation functions (Cohen 1995). These rules structured generation itself—different rules would have produced a structurally different system perceiving different compositional features as relevant. AARON had aesthetic *reasons* for its choices, traceable to Cohen's artistic commitments.

AARON satisfies constitutive embedding and exhibits implicit coherence—Cohen's artistic vision provided a unifying perspective—but the rules accumulated pragmatically rather than being derived from an explicit theory, making the coherence retrospectively attributable rather than architecturally articulated. The system lacked adaptiveness: rules were frozen at design time. Cohen's (2006) revealing discovery that a simple color algorithm matched twenty years of accumulated rules illustrates the cost: the system could not discover such equivalences or respond creatively to them. AARON demonstrates that constitutive embedding without adaptiveness produces principled but rigid creative behavior.

The Painting Fool (Colton, 2001–present). Colton's Painting Fool achieves A^p-level operation: its HR system generates fitness functions used to evolve scenes, producing emergent evaluative criteria not externally programmed (Colton 2012). The 2015 integration of DARCI's machine vision capabilities (Colton et al. 2015) further strengthened constitutive embedding: the system's ability to analyze its own output during painting means perceptual feedback shapes generation, not just selection. Since then, VLM-based iterative assessment during generation has become standard practice—Botto's creative reasoning loop, Keke Terminal's play iterations, and the broader generate-critique-refine paradigm (Madaan et al. 2023) all use visual feedback to shape the generative process. This proliferation makes the constitutive question more urgent, not less: in every case, the evaluative categories are inherited from the VLM's training distribution rather than derived from a theory of aesthetic value.

Yet The Painting Fool's emergent evaluative discoveries—associations between media and mood qualities—are statistical associations rather than principles derived from a theory of aesthetic value. The system can discover *that* pencils correlate with bleakness but cannot articulate *why* this matters aesthetically. In Moles's (1966) terms, the system operates primarily with semantic information (extractable correlations) rather than aesthetic information (irreducible perceptual qualities). Colton (2008) acknowledged that skill, appreciation, and imagination were not simultaneously present. The Painting Fool demonstrates that A^p-level operation without theoretical grounding produces learned regularities rather than principled aesthetic judgment. The system's ability to generate new evaluative criteria represents a genuine step toward adaptiveness, even if those criteria lack unifying theoretical grounding.

AICAN (Elgammal et al., 2017). Elgammal's Creative Adversarial Network has an explicit theoretical grounding: its dual objective—minimize deviation from the art distribution while maximizing stylistic ambiguity—derives from Martindale's (1990) psychological theory of art-historical dynamics. The adversarial architecture directly instantiates the theory: a different theory (say, Berlyne's optimal arousal model) would yield a different optimization objective, evaluation ontology, and architecture. The motivated counterfactual test is clearly satisfied.

AICAN's limitation is scope rather than commitment. The theory operates at the population level—style distributions across art history—without addressing individual compositions. AICAN cannot evaluate whether a particular work has interesting compositional structure, rewards sustained perceptual engagement, or challenges the viewer's compressive capacities. It has a theory of art-historical *dynamics* but not a theory of aesthetic *experience*—no account of what makes a single work worth looking at, as opposed to whether it deviates from a learned style distribution. This is the difference between a theory of art's evolution and a theory of art's value. Notably, Martindale's theory operates within a broader tradition—Berlyne's (1971) optimal arousal theory provides the psychological foundations from which Martindale derived his art-historical dynamics, suggesting that a more comprehensive constitutive approach could integrate both individual-level arousal theory and population-level style dynamics.

AICAN substantially satisfies both constitutive embedding and theoretical coherence, but only within a domain that cannot address the compositional, perceptual, and experiential dimensions where most aesthetic judgment operates. This matters for the argument: AICAN demonstrates that constitutive embedding with theoretical coherence is *achievable*—the conjunction is not utopian. What AICAN's narrow scope shows is that the challenge lies not in constitutive embedding per se but in the *comprehensiveness* of the constitutive theory.

Botto (Klingemann / ElevenYellow DAO, 2021–present). Botto resists simple categorization. Its three generative subsystems demonstrate genuine engineering and creative sophistication (Botto Project 2025). The creative reasoning system generates hypotheses, self-critiques along multiple dimensions, and iteratively refines—exhibiting a form of internal evaluative capacity that goes well beyond simple prompt-to-image generation.

Two features, however, position Botto's overall trajectory toward the applied end of the spectrum. First, the self-critique dimensions—composition, lighting, narrative, audience appeal, meme potential, AI slop detection—intermix formal, psychological, sociological, and quality-control categories. Whether these constitute a coherent framework depends on what one counts as coherence: they may reflect a pragmatic-pluralist approach to evaluation rather than theoretical arbitrariness. We note this ambiguity rather than prejudging it, while observing that the system's documentation does not articulate a theoretical principle governing their selection or interrelation. Second, the system's ultimate arbiter is external: DAO community voting trains a taste model that pre-selects candidates. The internal creative engine operates within constraints set by this feedback loop, creating a structural tension between internal exploration and external selection. A theoretically grounded internal evaluative framework could serve as a counterweight—enabling the system to resist convergence when its own aesthetic commitments indicate a different direction.

Botto's architecture is best understood as two-level: the creative reasoning system has partially constitutive features, while the community-driven selection loop is applied. The overall trajectory is predominantly governed by the applied component.

Keke Terminal (Dark Sando, 2024). Keke Terminal presents an instructive case. Its taste module combines VLM-based evaluation, NIMA scoring, and social engagement metrics. The system employs meta-prompting, creative reasoning, and autonomous decision-making—capabilities that, in other domains, would constitute genuine agency.

The philosophical interest lies in the gap between technical sophistication and the architecture of evaluation. The taste module learns to *predict* which outputs will be well-received but—based on available documentation—does not *judge* which are aesthetically significant according to an articulable theory. We should be cautious about drawing causal conclusions: correlation between theoretical thinness and aesthetic conventionality does not establish causation.

What we can say with confidence is structural. The system's evaluative architecture would remain structurally unchanged if the engagement-prediction module were replaced by a different prediction model targeting different demographics—the evaluation ontology (predicting reception from learned correlations) is invariant across such substitutions, which satisfies the test for applied theory. A constitutive substitution—replacing reception prediction with, say, Bense's information-aesthetic framework—would

change not only the scores but the categories of judgment available to the system: from "will audiences approve?" to "does this composition occupy the configurative zone between order and entropy?"

Synthesis. The pattern across five systems is consistent in direction, though the details reveal important variation:

Three observations. First, no system achieves the conjunction of strong constitutive embedding, explicit theoretical coherence, and adaptiveness—but AICAN demonstrates that the first two are jointly achievable. The challenge is comprehensiveness, not feasibility. Second, the most recent and technically sophisticated systems (Botto, Keke Terminal) are not the most theoretically grounded—engineering *for generation* has progressed far more rapidly than *for evaluation*. Third, existing evaluation frameworks (Ritchie 2007; Jordanous 2012) do not directly diagnose this pattern because they do not ask whether a system's *design* follows from its *aesthetic theory*.

We note that these five systems have different overriding aims—AARON embodied one artist's vision; The Painting Fool aims to be taken seriously as a creative artist; AICAN tests a psychological theory; Botto explores decentralized creative agency; Keke Terminal pursues autonomous market participation. The constitutive/applied distinction is diagnostic across all these aims, though its normative force varies (see §6).

This pattern is not specific to the pre-foundation-model systems analyzed here. Foundation models amplify the asymmetry rather than resolving it: fine-tuning a diffusion model with an aesthetic reward function, or prompting an LLM to critique images along specified dimensions, instantiates applied aesthetics at unprecedented scale. The constitutive/applied distinction cuts orthogonally to the model capability axis.

This claim requires care when applied to LLM-based systems. An LLM possesses its own implicit evaluation ontology inherited from training—just as a human artist possesses perceptual capacities shaped by biology and experience. The question is not whether the system uses an LLM—most future creative systems will—but whether the aesthetic theory organizes the evaluation pipeline through which the LLM reasons, or merely steers the LLM's general-purpose reasoning via prompts. A system that prompts an LLM to "evaluate this image for compositional balance" invokes the LLM's pre-existing evaluative vocabulary—applied theory as prompt steering. A system that feeds the LLM a theory-derived perceptual decomposition of the image (entropy measurements at multiple scales, attention distribution maps, complexity gradients) and constrains it to reason *within* those categories is doing something architecturally different: the LLM serves as a reasoning engine, but the categories of reasoning are determined by the theory, not by the LLM's training distribution. This is the difference between theory as prompt and theory as pipeline—the version of the constitutive/applied distinction that matters for the foundation model era.

Independent Diagnostics from Information Aesthetics

The constitutive/applied distinction can be independently sharpened through two distinctions from the information aesthetics tradition—a body of work the CC literature has not adequately engaged despite its direct relevance.

Bense's Creation Pattern vs. Communication Pattern

Max Bense (1965; 1971) distinguishes the *communication pattern*—sender $\rightarrow$ noisy channel $\rightarrow$ receiver, where feedback is corrective, converging on shared repertory alignment—from the *creation pattern*, where an observer performs a selective function on a repertory to produce a new distribution (Bense 1965; Nake 2012; Taylor 2024). In the creation pattern, selective freedom is progressively consumed as the emerging work constrains remaining choices. The product feeds back to redefine the repertory: creation changes what is subsequently possible. Communication feedback is *convergent*; creation feedback is *generative*.

What does this distinction add beyond restating the constitutive/applied divide? The Bensean diagnostic identifies a specific *mechanism* that the constitutive/applied distinction leaves abstract. The constitutive/applied distinction tells us *that* theory should be embedded; the creation/communication distinction tells us *how* the embedded theory must function—specifically, it must operate through generative feedback that transforms its own evaluative repertory. This is a stronger architectural requirement than constitutive embedding alone, because a system could constitutively embed a fixed theory (as AARON does) without exhibiting the repertory-transforming dynamics that Bense identifies with creation. The creation pattern thus provides a criterion for distinguishing between constitutive systems that are merely *principled* and constitutive systems that are genuinely *creative* in Bense's technical sense.

Applied to the systems analyzed above: Botto's DAO feedback loop implements the communication pattern—converging on community-repertory alignment. Keke Terminal's engagement-trained taste module likewise converges on existing preferences. AARON's frozen rules implement neither pattern. No system implements the creation pattern, which requires that the selective principle be internal, theoretically grounded, and capable of redefining the repertory through creative practice.

Is the creation pattern computationally tractable? It is—and less demanding than it might appear. A system whose evaluative framework is parameterized by its own creative history—where successful works modify the weights, thresholds, or even the categories of subsequent evaluation—would implement the creation pattern. This is architecturally distinct from, but not more technically demanding than, the meta-learning and self-modifying prompt systems already deployed in Botto's creative reasoning framework. The barrier is not engineering but the absence of a theoretical account specifying *how* the evaluative framework should change—which is precisely what a constitutive aesthetic theory provides. A practical consequence: con-

Table 1: Constitutive analysis of five AI art systems across three thresholds.

System	Constitutive Embedding	Theoretical Coherence	Adaptiveness
AARON	Strong	Implicit	None
Painting Fool	Partial (generation)	None (explicit)	Partial (A^p)
AICAN	Strong (narrow scope)	Strong (narrow scope)	None
Botto	Two-level (internal partial, selection applied)	Unclear (pragmatic pluralism)	External only
Keke Terminal	Weak	None	External only

stitutive embedding enables evaluation at each stage of the generative process rather than only at its conclusion, reducing the candidate space through theoretically informed pruning rather than post-hoc filtering.

This diagnostic also exposes what may be the most practically important distinction in this paper: the gap between *behavioral autonomy* and *evaluative autonomy*. Behavioral autonomy—deciding when, what, and how to create—is compatible with *evaluative dependence*: relying on external sources for what counts as creating well. A system with behavioral autonomy but evaluative dependence is behaviorally autonomous but evaluatively heteronomous. Both Botto and Keke Terminal achieve impressive behavioral autonomy while remaining evaluatively dependent on external feedback—community voting or engagement metrics. Guckelsberger, Salge, and Colton's (2017) model of intentional creative agency addresses intentionality but not the source of evaluative standards.

This distinction has immediate practical consequences. When the CC community evaluates a system's "autonomy," it should ask not only whether the system makes its own creative decisions but whether the *criteria informing those decisions* are internally grounded. A system that autonomously generates, selects, and publishes artwork is behaviorally autonomous. If its selection criteria are entirely trained on external approval data, it is evaluatively heteronomous—and the constitutive/applied distinction explains exactly why this matters: evaluative heteronomy constrains creative trajectory even when behavioral freedom is unconstrained.

Moles's Semantic vs. Aesthetic Information

Abraham Moles (1966) distinguishes *semantic information*—propositional, translatable, channel-independent—from *aesthetic information*—state-oriented, channel-dependent, irreducible to logical propositions.

This distinction does additional work beyond the constitutive/applied divide by identifying a specific *informational limitation* of applied systems. Applied aesthetics necessarily operates with semantic information: extractable labels, quantifiable scores, translatable evaluations. A NIMA score, a composition rating, an engagement prediction—these are propositional and channel-independent. They can be transmitted without loss from one system to another because they are *about* the work rather than *of* it.

A constitutive system could, in principle, attend to aesthetic information: the channel-dependent perceptual qualities that resist decomposition into propositional content. A system measuring spatial entropy distributions at multiple scales, tracking information-theoretic tension between local complexity and global order, and evaluating progressive compressibility across spatial frequency bands engages with properties that are compositionally sensitive in a way score-based evaluation cannot replicate. The distinction suggests a specific reason why systems operating exclusively with semantic evaluation have a ceiling on their evaluative sophistication: they cannot represent the properties that, on Moles's account, constitute the specifically aesthetic dimension.

This is not to claim that current computational methods can fully capture aesthetic information in Moles's strong sense—the channel-dependent, irreducible quality of perceptual experience may remain beyond computational reach. The point is more modest and more actionable: there is a gradient between purely semantic evaluation (a single quality score) and increasingly channel-sensitive evaluation (multi-scale spatial analysis, attention modeling, perceptual decomposition). Systems operating further along this gradient engage with properties closer to what Moles identifies as aesthetic information, even if they do not capture it completely. Constitutive embedding pushes systems along this gradient by requiring that evaluation be grounded in a theory of *how* perceptual properties generate aesthetic value, not merely *which* outputs correlate with positive reception.

Convergence of the Diagnostics

These three distinctions are not three ways of saying the same thing. Each identifies a different facet of the same structural gap: the constitutive/applied distinction identifies *where* theory sits in the architecture; the creation/communication distinction identifies *how* theory must function dynamically; the semantic/aesthetic distinction identifies *what kind of information* theory must engage with. Their convergence—diagnosing the same systems as deficient along different analytical dimensions—strengthens the diagnosis: the gap is not an artifact of a particular vocabulary but a structural feature of how these systems are built.

What Constitutive Theory Produces

The preceding sections diagnose a gap. What would filling it look like? Rather than sketching one preferred architecture, we illustrate how *different* constitutive theories produce structurally different systems. This is the strongest form of the argument: constitutive theory is not a single prescription but a design principle that produces principled diversity.

Different theories produce different evaluation ontologies. Consider three theoretical starting points, each drawn from a distinct intellectual tradition: (a) Birkhoff's (1933) aesthetic measure $M = O/C$, grounding aesthetic value in the ratio of order to complexity; (b) Bense's information aesthetics, organized at nuclear, micro, and macro semiotic levels, attending to how information is distributed across scales of composition; (c) Chatterjee's (2013) aesthetic triad, grounding aesthetic experience in the interaction of sensory-motor, emotion-valuation, and meaning-knowledge processing systems.

Theory (a) yields a maximally parsimonious architecture: a single evaluator computing a single ratio. The system perceives order and complexity as its two evaluatively relevant categories, and every compositional feature matters only insofar as it contributes to one or the other. Theory (b) implies a multi-level computational pipeline measuring different kinds of information at different scales—the nuclear level (material elements), the micro level (statistical regularities), and the macro level (global gestalt). The evaluation ontology is hierarchical: features matter differently depending on which semiotic level they operate at. Theory (c) organizes evaluation by processing system rather than by compositional level, producing a three-channel architecture where *relationships between channels*—the way sensory processing, emotional response, and semantic interpretation interact—carry the evaluatively relevant information. A composition might score well on sensory-motor engagement but poorly on meaning-knowledge integration, and the pattern of cross-channel interaction is itself the aesthetic datum.

These are not parametric variations within a shared evaluation framework—they are *ontologically* different. Each theory determines what the system can perceive as evaluatively relevant: order-complexity ratios, semiotic-level information distributions, or processing-system interactions. The motivated counterfactual test is satisfied because the evaluation ontology changes, not merely the weights within a fixed ontology.

Different theories produce different integration strategies. This ontological divergence extends to how evaluative information is combined. A Birkhoffian system has no integration problem—there is one metric, and higher is better. A Bensean system must integrate across semiotic levels, raising the question of whether nuclear, micro, and macro evaluations should converge or whether productive tensions between levels carry aesthetic information. A Chatterjee-inspired system must decide how to weigh cross-channel interactions: does aesthetic value reside in harmony among processing systems, in tension between them, or in specific patterns of alignment and misalignment?

Each theory implies a different answer. For Bense, the configurative border condition—the productive tension between order and entropy—suggests that the *pattern* of integration (where levels agree, where they diverge) is itself evaluatively significant, not merely the aggregate. For Chatterjee, empirical work on the aesthetic triad suggests that the most powerful aesthetic experiences involve simultaneous engagement of all three systems, with cross-system facilitation predicting aesthetic response better than any single channel (Chatterjee 2013). Each integration strategy follows from its theory, and no amount of applied parameter tuning could produce these divergences because they operate at the level of what the integration module treats as signal.

Different theories produce different developmental trajectories. Schmidhuber's (2009) compression aesthetics distinguishes beauty (low description length) from interestingness (rate of compression progress). A system constitutively pursuing interestingness moves differently through aesthetic space than one pursuing beauty: it seeks works at the boundary between what it can already compress and what it cannot yet. A system grounded in Martindale's arousal potential pursues stylistic deviation from population norms—a trajectory operating at a different level of analysis (individual compression dynamics versus population-level style distributions). A system embedding Berlyne's (1971) optimal arousal theory—from which Martindale derived his art-historical model—would seek a specific middle region of stimulus complexity and progressively refine its model of where that region lies, rather than pursuing boundary-seeking or norm-deviation.

These three trajectories—boundary-seeking (Schmidhuber), norm-deviating (Martindale), and optimum-refining (Berlyne)—produce detectably different sequences of outputs over time: qualitatively different paths through aesthetic space, driven by different accounts of what makes creative development productive. This is what constitutive embedding produces: not just different evaluations of the same outputs, but different creative *histories*.

The counterfactual test confirms: substitute any of these theories for any other, and the evaluation architecture, integration strategy, and developmental trajectory all change. Applied filtering cannot replicate these differences because it operates on outcomes rather than on reasons. A constitutive system has access to a theoretical model of *why* certain features matter and *how* they relate; an applied system has access only to scores. This informational asymmetry—reasons versus outcomes—is the mechanism through which constitutive embedding produces its distinctive consequences, including the capacity for principled resistance to conventional taste when theoretical commitments warrant divergence.

Implications: Theoretical Coherence as Evaluative Dimension

The argument suggests a new evaluative dimension for CC research: *theoretical coherence*—the degree to which a system's architectural decisions are traceable to its aesthetic commitments.

Theoretical coherence can be operationalized through a three-step procedure. *Step 1—Commitment identification*: for each evaluative component, identify the aesthetic commitment (if any) that motivates it. Where motivations are

pragmatic or implicit, note this. *Step 2—Motivated counterfactual test*: for each commitment, ask whether a different aesthetic theory would change the system's evaluation ontology—the categories of judgment available to the system—not merely its scores or selection rates. *Step 3—Coherence assessment*: do the system's commitments form a unified framework, whether formal or implicit? Could a knowledgeable reader, given the theory alone, predict the major architectural decisions?

This procedure is implicit in how the CC community already evaluates systems—we ask whether design decisions are principled or ad hoc—but the constitutive/applied distinction makes it explicit and systematic. Adding theoretical coherence alongside Ritchie's (2007) output criteria, Jordanous's (2012) standardized procedure, and Colton's creative tripod fills a gap. Where Ritchie evaluates outputs, Jordanous evaluates behaviors, and Colton evaluates capacities, theoretical coherence asks whether the system's *design* follows from a principled account of aesthetic value. This dimension complements rather than replaces existing approaches: Linkola et al.'s (2017) metacreative self-awareness and Colton and Pease's (2018) computational authenticity both gesture toward the theory-architecture relationship without making it their explicit diagnostic target.

An important caveat: constitutive embedding amplifies theoretical commitments for good *and* ill. A system constitutively embedding a simplistic theory—a naïve Birkhoffian scalar measure, say—would produce systematically constrained art with great internal coherence. Such a system might be *worse* than an applied system precisely because its limitations are architectural rather than filterable. Theoretical coherence is therefore *necessary* for creative autonomy, not sufficient for aesthetic success. This means the *choice* of constitutive theory is consequential in a way the choice of applied filter is not—amplifying both the theory's insights and its blind spots.

A related concern is that constitutive embedding might produce only "good specimens"—typical members of the theory's category rather than norm-breaking creative work. This concern applies when the same theory drives both generation and evaluation. But a constitutive system need not generate only what its theory predicts will be good. An architecture that pairs divergent generation—through stochastic, lateral, or constraint-based ideation—with constitutive evaluation produces novel configurations that the theory can assess but would never have generated on its own. The theory shapes what the system *recognizes* as aesthetically significant, not what it is *capable of producing*. This mirrors human artistic practice, where experimental generation and reflective evaluation operate as complementary phases. In Boden's (2004) terms, constitutive embedding enables exploratory creativity within a theory-defined space; the adaptive constitutiveness threshold enables transformational creativity applied to the evaluative framework itself.

Constitutive embedding is also not a return to symbolic AI. The theory organizes the evaluation pipeline, but the pipeline's components—feature extraction, perceptual decomposition, attention modeling—can be implemented through neural, symbolic, or hybrid methods. What matters is that the architectural organization follows from the theory, not that the implementation uses any particular computational paradigm.

The consequence is theoretical pluralism. Different aesthetic theories should produce different systems with different creative behaviors, and this diversity is a feature. The CC field would benefit from the same diversity found in human art traditions, but grounded in explicit, architecturally traceable theoretical commitments rather than implicit assumptions.

A further consequence is that the field needs to take the *tractability* of aesthetic theories seriously as a research question. Quantitative foundations help: Bense's information aesthetics, Birkhoff's aesthetic measure, and Schmidhuber's compression theory all specify measurable quantities, making architectural translation direct. Theories organized around discrete processing stages (Chatterjee's triad) also translate well. By contrast, theories relying on irreducibly holistic judgment—late Wittgenstein's "seeing as," Heidegger's *techne*—resist computational decomposition. Mapping this tractability landscape is a contribution the CC community is positioned to make.

A caveat on scope: the constitutive/applied distinction is most forcefully normative when the aim is to produce a system that functions as an autonomous artist. Where the aim differs—advancing CC theory (Colton 2008), exploring the machine condition (Colton et al. 2020), or developing creative personhoods (Pease, Colton, and Banar 2023)—the distinction remains diagnostically useful but may carry different normative weight. We note, however, that the machine condition is not an alternative to aesthetic theory but a *candidate* constitutive theory—one specifying which experiential properties of computation are aesthetically significant. A system constitutively embedding this theory would have a different evaluation ontology (attending to memory volatility, parallel processing, power dependency as aesthetic features) than one embedding Bense or Birkhoff. The constitutive/applied distinction applies wherever a system makes evaluative choices about its own creative output, regardless of its overriding purpose.

Conclusion

Every major AI art system examined here treats aesthetic theory predominantly as applied: external to the generative architecture, operating on outputs rather than structuring their production. This finding has been anticipated by work on creative autonomy, appreciation, aesthetic evaluation, and computational aesthetics in evolutionary art (Jennings 2010; Colton 2008; Galanter 2012; McCormack, Dorin, and Innocent 2005; Todd and Latham 1992; Machado and Cardoso 2002; Romero and Machado 2008). This paper contributes the precision of the motivated counterfactual test targeting evaluation ontology, independent sharpening from Bense and Moles, the identification of a gap between behavioral and evaluative autonomy, and the demonstration that constitutive embedding produces ontologically different systems rather than merely parametrically different ones. We further argue that constitutive evaluation and divergent generation function as complementary phases, and that the ma-

chine condition constitutes a candidate constitutive theory rather than an alternative to the framework.

Two claims should be distinguished. *Claim A*—that AI art systems should have coherent aesthetic frameworks—is relatively modest and hardly controversial. *Claim B*—that those frameworks should be constitutive of the system's architecture, in the specific sense that different frameworks yield systems with different evaluation ontologies—is this paper's distinctive contribution. Claim B matters because it ensures the theory does genuine architectural work. The motivated counterfactual test provides a concrete way to distinguish systems that *have* a theoretical framework from systems whose architecture *follows from* one. The difference is consequential: a system whose architecture follows from its theory can extrapolate beyond training distributions, resist conventional taste on principled grounds, explain its evaluative decisions in theoretically articulable terms, and develop in theoretically motivated directions. A system that merely has a theory applied as a filter can do none of these things, regardless of how sophisticated the filter.

The CC community is well positioned to take this distinction seriously (Colton and Wiggins 2012). The frameworks are in place—Colton's appreciation criterion, Wiggins's transformational creativity, Boden's (2004) typology, Jordanous's evaluation procedures, Galanter's call for aesthetic grounding. What remains is to make the architectural relationship explicit, test it with the motivated counterfactual diagnostic, and build systems where aesthetic theory is traceable in the design rather than cited in the documentation. The generative AI revolution has given the field powerful new tools; the question is whether they will be deployed with the theoretical sophistication that the field's own traditions demand.

Author Contributions

As the sole author, Atticus Sims is responsible for all aspects of this work, including its conception, analysis, and writing.

Acknowledgments

This research was supported by a research grant from the University of Macau (SRG2025-00057-FAH). The author thanks the anonymous ICCC reviewers for their helpful comments.

References

Bense, M. 1965. *Aesthetica: Einführung in die neue Aesthetik*. Agis-Verlag.

Bense, M. 1971. The projects of generative aesthetics. In Reichardt, J., ed., *Cybernetics, Art and Ideas*. New York Graphic Society. 57–60.

Berlyne, D. E. 1971. *Aesthetics and Psychobiology*. Appleton-Century-Crofts.

Birkhoff, G. D. 1933. *Aesthetic Measure*. Harvard University Press.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Botto Project. 2025. Botto's art engine. Technical documentation. `https://docs.botto.com`.

Chatterjee, A. 2013. *The Aesthetic Brain: How We Evolved to Desire Beauty and Enjoy Art*. Oxford University Press.

Cohen, H. 1995. The further exploits of AARON, painter. *Stanford Humanities Review* 4(2):141–158.

Cohen, H. 2006. AARON, colorist: From expert system to expert. In *Proceedings of Digital Art Weeks*. ETH Zurich.

Colton, S., and Pease, A. 2018. Issues of authenticity in autonomously creative systems. In *Proceedings of the 9th International Conference on Computational Creativity*, 272–279.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of ECAI 2012*, 21–26.

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; and Ferrer, B. P. 2015. The Painting Fool Sees! New projects with the automated painter. In *Proceedings of the 6th International Conference on Computational Creativity*, 189–196.

Colton, S.; Pease, A.; Guckelsberger, C.; McCormack, J.; and Llano, T. 2020. On the machine condition and its creative expression. In *Proceedings of the 11th International Conference on Computational Creativity*, 342–349.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., and d'Inverno, M., eds., *Computers and Creativity*. Springer. 3–38.

Danto, A. C. 1964. The artworld. *The Journal of Philosophy* 61(19):571–584.

Dark Sando. 2024. Keke terminal. Technical whitepaper. `https://www.keketerminal.com`.

Dickie, G. 1974. *Art and the Aesthetic: An Institutional Analysis*. Cornell University Press.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*.

Galanter, P. 2012. Computational aesthetic evaluation: Past and future. In McCormack, J., and d'Inverno, M., eds., *Computers and Creativity*. Springer. 255–293.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*.

Jennings, K. E. 2010. Developing creativity: Artificial bar-

riers in artificial intelligence. *Minds and Machines* 20:489–501.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Linkola, S.; Kantosalo, A.; Männistö, T.; and Toivonen, H. 2017. Aspects of self-awareness: An anatomy of metacreative systems. In *Proceedings of the 8th International Conference on Computational Creativity*.

Machado, P., and Cardoso, A. 2002. All the truth about NEvAr. *Applied Intelligence* 16(2):101–118.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*.

Martindale, C. 1990. *The Clockwork Muse: The Predictability of Artistic Change*. Basic Books.

McCormack, J.; Dorin, A.; and Innocent, T. 2005. Generative design: A paradigm for design research. In *Proceedings of Futureground*. Monash University.

Moles, A. A. 1966. *Information Theory and Esthetic Perception*. University of Illinois Press. Translated by J. E. Cohen.

Nake, F. 2012. Information aesthetics: An heroic experiment. *Journal of Mathematics and the Arts* 6(2–3):65–75.

Pease, A.; Colton, S.; and Banar, B. 2023. On the notion of creative personhood. In *Proceedings of the 14th International Conference on Computational Creativity*, 117–121.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Romero, J., and Machado, P., eds. 2008. *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Springer.

Schmidhuber, J. 2009. Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE* 48(1):21–32.

Sims, K. 1994. Evolving virtual creatures. In *Proceedings of SIGGRAPH 1994*, 15–22. ACM.

Taylor, G. D. 2024. Bald krumme Linien: Generative aesthetics and the Bensian legacy. *Culture, Theory and Critique* 65(3–4):311–328.

Todd, S., and Latham, W. 1992. *Evolutionary Art and Computers*. Academic Press.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.