MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Philosophy International Journal
ISSN: 2641-9130

Beyond Value Alignment: Aesthetic Judgment and the Case for Productive Misalignment

Choi B*

Department of Arts, College of Arts and Design, Wonkwang University, Republic of Korea

***Corresponding author:** Byungkil Choi, Department of Arts, College of Arts and Design, Wonkwang University, Iksan, Republic of Korea

Conceptual Paper
Volume 9 Issue 4
Received Date: October 18, 2026
Published Date: November 17, 2026
DOI: 10.23880/phij-16000378

Abstract

Contemporary discussions of artificial intelligence emphasize alignment with human intentions, preferences, norms, and values. This article argues that alignment is necessary but aesthetically insufficient for understanding human-AI creative interaction. In artistic practice, criteria of success may change through encounters with unforeseen results. I define productive misalignment as a bounded departure from an antecedent intention, expectation, or provisional aesthetic criterion that initially fails to satisfy that criterion yet becomes retrospectively affirmable through subsequent aesthetic judgment. Unlike the discovery of a hidden preference or satisfaction of a higher-order preference for surprise, the decisive case involves transformation of the evaluative criterion itself. Drawing on theories of aesthetic judgment, accident, resistance, and improvisation, the article distinguishes productive misalignment from technical failure and ethical misalignment. It also defends an asymmetry between distributed generative contribution and project-level evaluative authorization. I propose aesthetic negotiation as a model of bounded, revisable human-AI interaction: normative constraints establish the limits of permissible generation, while aesthetic judgment determines whether generated difference should be rejected, retained, or allowed to redirect the work.

Keywords: Value alignment; Aesthetic judgment; Productive misalignment; Human-AI co-creation; Generative AI; Aesthetic agency

Introduction: When Alignment Is Not Enough

The Dominance of Value Alignment in AI Discourse

The rapid development of artificial intelligence has made alignment one of the principal concepts through which the relationship between artificial systems and human purposes is understood. Although the term encompasses heterogeneous technical and philosophical problems, its different uses converge upon a fundamental concern: increasingly capable artificial systems should behave in ways that remain appropriately responsive to human intentions, preferences, norms, and values. Early formulations of the problem already recognized that specifying an objective for an artificial agent does not guarantee that the resulting behavior will correspond to what human beings actually intend. Stuart Russell, Daniel Dewey, and Max Tegmark consequently identified the difficulty of constructing systems whose objectives remain compatible with human values as a central problem for advanced AI, while later work on value alignment has examined the difficulty of identifying, representing, learning, and aggregating those values in the

CC BY

first place [1,2]. Brian Christian's broader reconstruction of the alignment problem likewise demonstrates that the issue cannot be reduced to programming machines to obey explicit commands, because the relevant human purposes are frequently implicit, conflicting, unstable, or incompletely articulated [3].

Recent human-AI creativity research has begun to examine this relation empirically and procedurally. Studies have tested whether exposure to generative AI improves divergent thinking, compared human-AI ideation with human-human and human-Internet collaboration, modeled co-creation in terms of human, AI, artifact, instruction, and interaction, and proposed structured prompt-refinement frameworks for generative art [4-8]. These approaches clarify how AI participates in creative performance and how interaction can be organized, but they leave a distinct philosophical problem unresolved: what should we call an output that is genuinely discrepant relative to an antecedent aesthetic criterion yet later becomes valuable because the encounter with it changes the criterion itself? The present article addresses that problem rather than asking whether AI simply increases creativity or whether machines and humans should be treated as symmetrical creators.

The difficulty addressed in this article begins when that relational model is transferred without modification into aesthetic practice. Creative activity does not always aim at minimizing the difference between an antecedently specified intention and a subsequently produced object. Indeed, one of the recurring features of artistic practice is that the process of making can alter the intention with which it began. The work that finally matters may not be the successful realization of the artist's initial conception; it may instead be the result of encounters with materials, accidents, technical constraints, interpretations, and discoveries that transform that conception during production. John Dewey's account of artistic experience already resisted the separation of a prior inner conception from its external execution, emphasizing instead the continuous interaction through which doing and undergoing modify one another [9]. More recent theories of creativity have similarly challenged models in which creative production is understood as the straightforward implementation of a pre-existing plan. Margaret Boden, for instance, distinguishes forms of creativity that explore or transform a space of possibilities, thereby making novelty conceptually dependent upon more than successful execution of an antecedent specification [10].

This distinction is fundamental to the argument developed below. The article does not oppose aesthetics to alignment as though creative practice required the suspension of ethical or technical constraints. Nor does it assume that unpredictability is inherently creative. Rather, it asks whether an account organized around correspondence to antecedent human purposes can explain a domain in which the purposes themselves may be revised through engagement with an unforeseen result. Once that possibility is admitted, a first limitation becomes visible: an output may be successfully aligned and aesthetically unsuccessful, while an output that departs from an antecedent specification may under certain conditions become aesthetically significant. The criteria governing these two forms of success cannot therefore be identical.

Why Successful Alignment Does Not Guarantee Aesthetic Success

The distinction becomes clearer if one considers what it would mean for a generative system to satisfy a creative instruction perfectly. Suppose that a user specifies the subject, composition, medium, stylistic references, atmosphere, and formal characteristics of an image, and that an AI system produces an output corresponding with extraordinary precision to every stated requirement. From the standpoint of instruction-following, little more could be demanded. Yet nothing in the degree of correspondence establishes that the image is aesthetically compelling. The output may be predictable precisely because it realizes the specification so completely; it may display technical competence without generating interpretive tension, experiential density, or formal discovery. Successful execution and aesthetic success therefore remain conceptually distinguishable.

This distinction becomes obscured when aesthetic value is treated as a special case of preference satisfaction. Contemporary computational systems can infer preferences from choices, ratings, comparisons, and behavioral histories, and such information can clearly improve the relevance of generated outputs. Nevertheless, an account of aesthetic judgment based exclusively upon preference prediction encounters an old philosophical difficulty: aesthetic judgment concerns not simply what a subject happens to prefer but what the subject finds worthy of affirmation in the encounter with a particular object. Kant's account of judgments of taste famously distinguishes aesthetic judgment from judgments based straightforwardly upon determinate concepts or private gratification [11]. One need not adopt the full Kantian theory of taste to retain the distinction that matters here. Aesthetic judgment cannot be adequately described as the mechanical confirmation of a preference that existed in complete form before the object was encountered.

This does not imply that the greater the deviation, the greater the aesthetic value. Such a principle would merely reverse the alignment model while retaining its reductive structure. Perfect correspondence is not a sufficient

Copyright© Choi B.

condition of aesthetic success, but neither is difference. Randomness, incoherence, and malfunction can remain aesthetically sterile. The important point is instead that the relation between correspondence and aesthetic value is not linear. Increasing alignment can sometimes improve a work, sometimes leave its aesthetic value unchanged, and sometimes eliminate an unexpected feature that would have become significant. The problem is therefore to explain how one distinguishes an unproductive error from a deviation worth preserving.

From Error and Deviation to Productive Misalignment

The language of error is inadequate for this task because it presupposes that the relevant criterion of correctness has already been sufficiently established. If a system is asked to generate a red circle and instead produces a blue square, the discrepancy can readily be described as an error insofar as the requested properties constitute the criterion of success. Creative practice, however, frequently operates under conditions in which not every criterion can be exhaustively specified before production begins. The maker may know enough to initiate a process without knowing exactly what result will ultimately deserve preservation. Donald Schön's account of the "reflective practitioner" is relevant here because it describes professional making as a "conversation" with a situation in which actions produce consequences that subsequently reshape the practitioner's understanding of the problem [12]. Artistic creation often possesses precisely this recursive character: making produces something to which further judgment must respond.

Productive misalignment is therefore not a property that an output necessarily possesses at the moment of generation. Its status is established retrospectively through what happens next. An unexpected result that is immediately discarded has functioned differently from an unexpected result that is selected, elaborated, juxtaposed with other elements, incorporated into a larger work, or allowed to transform the artist's subsequent intention. This retrospective structure is decisive because it prevents the mere production of novelty from being confused with aesthetic achievement. As Boden has repeatedly emphasized in discussions of creativity, novelty by itself is insufficient; creative value requires some form of significance or worth in addition to newness [13]. The same distinction should be applied more rigorously to generative AI. A machine's capacity to produce surprising variation does not itself explain why any particular variation matters.

Productive misalignment therefore names a paradox internal to creative uses of AI. The machine can fail to give the human exactly what was requested and, under appropriate conditions, thereby contribute to a result that the human judges better than what was originally requested. The paradox cannot be explained by alignment alone because the relevant human criterion has changed during the process. Nor can it be explained by machine unpredictability alone because the deviation becomes significant only through an evaluative relation. What requires explanation is the movement from deviation to significance, and that movement places aesthetic judgment at the center of human-AI co-creation.

Thesis and Structure of the Argument

The preceding analysis suggests that the limits of alignment in aesthetic practice do not arise because alignment is misguided, but because its normative logic does not exhaust the structure of creative judgment. Where objectives are determinate and consequences must remain tightly constrained, correspondence between human purposes and machine behavior remains an indispensable measure of success. Where creative objectives are partly formed through the process of making, however, complete antecedent specification is neither always possible nor necessarily desirable. The same deviation that counts as failure under a fixed criterion may become generative under a criterion that is itself transformed through experience. The problem is therefore not to choose between alignment and misalignment but to determine where alignment should remain strict and where controlled openness to deviation can possess aesthetic value.

On this basis, the article advances three interdependent claims. First, value alignment is necessary but insufficient for theorizing aesthetic human-AI interaction because correspondence with antecedent human preferences or intentions cannot by itself explain aesthetic success. Second, productive misalignment identifies a distinctive structure of creative interaction in which deviation becomes valuable only through subsequent aesthetic judgment. The concept therefore neither celebrates error nor attributes creativity automatically to computational unpredictability. Third, this structure produces an asymmetry between generative contribution and evaluative authority. Artificial systems can contribute possibilities that human agents did not anticipate, but the generation of such possibilities does not itself determine their aesthetic significance. The acts through which a possibility is selected, rejected, revised, contextualized, or affirmed remain constitutive of the creative process.

To move beyond value alignment therefore does not mean moving beyond human values, constraints, or responsibility. It means recognizing that human agency includes not only the capacity to specify what should happen but also the capacity to judge what has happened when the result exceeds

Copyright© Choi B.

specification. The aesthetic importance of generative AI may ultimately lie less in its ability to execute human intentions perfectly than in the recursive relation it creates between intention and consequence. In that relation, the unexpected is neither automatically an error nor automatically an achievement. It becomes aesthetically productive only when judgment makes something of it.

The Alignment Paradigm and Its Limits

What Does it Mean for AI to be Aligned?

The language of alignment has become so pervasive in discussions of artificial intelligence that it can sometimes appear to name a single, clearly delimited technical problem. In fact, the term covers a family of problems concerning the relation between artificial systems and humanly significant standards. At its broadest, an AI system is said to be aligned when its behavior remains appropriately responsive to the purposes, interests, values, or normative constraints that ought to govern its operation. What makes alignment difficult is not simply that artificial systems may malfunction. A system can perform exactly the optimization for which it has been designed and nevertheless produce an outcome that its designers or users did not actually want. The problem therefore concerns the relation between formally represented objectives and the more complex human purposes those objectives are intended to serve.

The alignment problem becomes still more complex because human purposes are themselves neither unitary nor transparent. A user may issue an instruction that does not accurately express an underlying intention; an intention may conflict with a longer-term preference; a preference may be imprudent or morally objectionable; and the interests of one person may conflict with those of others affected by an artificial system. Iason Gabriel has therefore argued that the apparent similarity among aligning AI with desires, intentions, preferences, and values conceals substantial philosophical differences [4]. His analysis is particularly important because it shows why alignment cannot be understood as a purely technical exercise. Before one can determine how an artificial system should be made responsive to human values, one must determine what kinds of values, whose values, and what principles for resolving disagreement ought to guide that responsiveness. The normative problem is thus internal to alignment rather than something that arises only after the technical problem has been solved.

This broader understanding is important for the aesthetic argument that follows. If alignment were nothing more than literal instruction-following, its inadequacy for artistic practice would be trivial: artists frequently value unexpected results, and therefore creative systems should simply be less literal. But the more sophisticated alignment paradigm already recognizes that literal compliance can diverge from genuine human purposes. The question raised by aesthetics is consequently more demanding. It is whether even a system that successfully responds to what a human genuinely prefers, intends, or values can thereby be said to have produced an aesthetically successful outcome. That question requires a closer examination of the different objects toward which alignment can be directed.

Values, Preferences, Intentions, and Instructions

The tendency to speak generally of "human alignment" can obscure the fact that instructions, intentions, preferences, and values play different normative roles. They can overlap, but they cannot be treated as interchangeable. A person may instruct an AI system to do something that does not correspond to the person's deeper intention; may intend something that, upon reflection, the person would not prefer; may prefer something that conflicts with an ethical value; or may endorse a value that does not determine a unique action in a particular situation. The choice of alignment target therefore changes the kind of relation an artificial system is expected to maintain with its human users.

This difficulty motivates the appeal to values. Values appear to provide a normative standard capable of evaluating rather than merely reproducing human desires or behavior. An AI system aligned with values would not simply predict what a person is likely to choose; it would operate in ways consistent with what ought to matter. But this apparent solution immediately raises more difficult philosophical questions. Are the relevant values individual or collective? Are they subjective commitments, socially shared norms, objective moral truths, or some combination of these? How should conflicts among liberty, welfare, justice, autonomy, equality, and other goods be resolved? And whose interpretation of those goods should govern artificial systems operating across culturally and politically plural societies? Gabriel emphasizes that value alignment inevitably encounters these questions because data about what people actually do cannot by itself determine what artificial systems morally ought to do [14].

This distinction does not render instructions, intentions, preferences, or values irrelevant to aesthetic AI. Each remains indispensable. Artists issue instructions, pursue intentions, develop preferences, and work within ethical as well as aesthetic values. What follows instead is that none of these categories, considered as antecedently fixed alignment targets, fully captures the temporality of aesthetic judgment. The creative process can modify the very human terms to which machine behavior is supposedly being aligned. This is the point at which the normative logic of alignment begins to

Copyright© Choi B.

encounter its aesthetic limit.

The Normative Logic of Correspondence

Despite the differences among instructions, intentions, preferences, and values, alignment approaches share a common structural aspiration. There is some humanly authoritative standard—however interpreted—and artificial behavior is successful insofar as it stands in an appropriate relation to that standard. The relation need not be literal identity. A well-aligned system may refuse a harmful instruction, infer an unstated intention, balance conflicting preferences, or act according to values more general than the user's immediate request. Yet in every case, the normative question concerns whether the artificial system's behavior corresponds appropriately to a humanly legitimate criterion.

The same structure explains why deviations are ordinarily treated as problems. If the purpose of an AI system is known and legitimate, movement away from that purpose is presumptively undesirable. The appropriate response is to identify the source of the discrepancy and reduce it. This logic is particularly persuasive where the standard is stable. If an autonomous vehicle is required not to collide with pedestrians, deviation from that requirement cannot become valuable merely because it is unexpected. If a medical system is expected to identify a clinically significant abnormality, a failure to detect it does not become productive because the failure produces an interesting alternative interpretation. In such contexts, the normative criterion precedes and governs the evaluation of the outcome.

This point becomes essential for the argument about productive misalignment. If aesthetic judgment merely supplied another fixed value for alignment, then no fundamental problem would arise. One could simply add "aesthetic value" to the set of objectives and optimize for it. But such a solution assumes that aesthetic value can be represented prospectively in the same way as a determinate performance criterion. The next subsection argues that this assumption overlooks a crucial feature of aesthetic judgment: the value of an outcome can depend upon a relation of discovery through which the criteria themselves are tested or transformed.

Why Aesthetic Value Resists Reduction to Successful Alignment

The strongest version of the alignment response to the foregoing argument would be straightforward: if people value novelty, surprise, deviation, or creative transformation, then these preferences can simply be incorporated into the alignment objective. An aesthetically sophisticated AI system need not reproduce a user's instructions mechanically. It could instead learn that the user prefers a certain degree of variation, unpredictability, or stylistic departure and generate accordingly. On this view, productive misalignment would turn out not to be misalignment at all. It would merely represent successful alignment with a higher-order preference for surprise.

The objection can be answered only by separating three cases that are too easily conflated. In the first, the user possesses a relatively determinate but imperfectly communicated preference; the system eventually discovers what was wanted. This is a problem of hidden preference and remains an ordinary case of alignment. In the second, the user has a higher-order preference for novelty or surprise; a system that supplies suitably adventurous outputs can again be aligned with that preference. The third case is different. The user begins from a provisional criterion C1; a generated deviation D conflicts with or exceeds C1; engagement with D supplies reasons for revising the criterion itself; and a later criterion C2 emerges partly through that encounter. Only this third, transformative case is the distinctive object of productive misalignment.

This is equally true of attempts to model aesthetic success through preference prediction. Suppose a generative system knows that a user consistently prefers asymmetrical compositions, muted palettes, sparse textures, and ambiguous figurative elements. The system can use those preferences to improve the probability that a new output will be approved. But the resulting prediction concerns the likelihood of satisfaction, not the aesthetic justification of the particular result. A work can correspond closely to a person's established taste and remain derivative; another can violate several established preferences and nevertheless prompt that person to reconsider what counts as successful. The second case is especially difficult for alignment understood as the satisfaction of antecedent preference because the judgment validates something that the prior preference would have rejected.

It is this possibility that prevents aesthetic value from being reduced to successful alignment. The aesthetic question is not exhausted by asking whether a system has correctly represented an antecedent human objective. It also asks whether the encounter with what the system has produced warrants a revision of that objective. Once the latter possibility is admitted, aesthetic creation can no longer be understood exclusively as increasingly accurate correspondence between pre-existing human preferences and machine outputs. It must also accommodate those bounded moments in which difference becomes discovery and in which judgment, rather than antecedent preference alone, decides what that discovery is worth.

Copyright© Choi B.

Aesthetic Judgment Beyond Preference Satisfaction

Preference is not Judgment

The distinction between preference and judgment is indispensable to the argument developed here because contemporary AI systems are increasingly designed to learn from human choices, ratings, rankings, and comparative evaluations. These signals are often treated as evidence of what users prefer, and preference information can in turn be used to guide or fine-tune artificial systems toward outputs that people are more likely to approve. In practical contexts, this approach is entirely understandable. If users consistently reject certain forms of output and select others, a system that learns from those patterns may become more useful, more relevant, and more responsive. Yet usefulness should not be confused with aesthetic judgment. The fact that a person prefers one artifact to another does not by itself explain why the preferred artifact should count as aesthetically successful.

David Hume approached the problem differently but encountered a related difficulty. His discussion of the "standard of taste" begins from the evident variability of aesthetic preferences and asks how critical discrimination is nevertheless possible [15]. Hume does not eliminate subjectivity, but neither does he treat all preferences as equally authoritative. Experience, comparison, attention, delicacy of perception, and freedom from prejudice matter because aesthetic judgment can be cultivated. The importance of this insight for AI is considerable. If aesthetic judgment were reducible to preference aggregation, then the most common preference would automatically carry the greatest aesthetic authority. Yet both aesthetic practice and critical history repeatedly demonstrate that popularity and aesthetic significance can diverge.

The aesthetic importance of human-AI interaction therefore lies partly in the possibility that an artificial system presents something for which the user does not yet possess a settled preference. At that point, the decisive human activity is not selection based on a prior ranking but judgment in the face of something that has disrupted the ranking itself.

Aesthetic Judgment and the Problem of Reasons

If aesthetic judgment is more than preference, the next question is whether it can be supported by reasons. This question is central because the distinction between preference and judgment would remain unstable if judgment amounted only to a more solemn form of subjective liking. Yet aesthetics has long resisted the opposite reduction as well: aesthetic judgments cannot ordinarily be derived from rules in the same way that mathematical conclusions follow from axioms or legal classifications follow from clearly specified criteria. Aesthetic judgment appears to occupy an intermediate space. It is not arbitrary, but neither is it mechanically deducible.

This contextual character does not make aesthetic reasoning irrational. It means that aesthetic reasons are often defeasible and relational. A reason can support a judgment without determining it conclusively. The same property may count positively under one set of relations and negatively under another. Frank Sibley's influential analysis of aesthetic concepts is useful in this respect. He argued that aesthetic terms such as "graceful," "balanced," "delicate," or "powerful" cannot be applied through purely mechanical rules based on non-aesthetic features, even though non-aesthetic features can provide reasons for their application [5]. One can point to particular lines, contrasts, proportions, or rhythms, but no finite checklist guarantees the correct aesthetic description.

For human-AI co-creation, this means that a system may be capable of generating outputs that satisfy statistically learned aesthetic tendencies without eliminating the need for human judgment. The existence of increasingly sophisticated predictive models changes the scale and speed of generation, but it does not abolish the problem of deciding whether a particular result warrants affirmation. The more numerous the generated possibilities become, the more visible this evaluative problem may become.

This account must not be mistaken for the claim that human judgment automatically confers aesthetic value. Human evaluators can be inattentive, conventional, inconsistent, or simply wrong. Retrospective authorization is therefore necessary for incorporating an unforeseen feature into an intentional artistic project, but it is not sufficient to establish aesthetic merit. The judgment remains reason-sensitive, criticizable, historically revisable, and open to challenge by collaborators, critics, and audiences. This qualification matters because productive misalignment would become vacuous if every retained deviation counted as productive merely by virtue of retention.

Novelty, Ambiguity, Tension, and Indeterminacy

The limits of preference-based accounts become especially clear when one examines qualities frequently associated with aesthetic achievement: novelty, ambiguity, tension, and indeterminacy. Each can contribute positively to an artwork, yet none functions as a simple value that should be maximized. Their aesthetic force depends upon degree, placement, relation, and context. They therefore provide instructive cases for understanding why aesthetic value resists direct translation into a fixed optimization objective.

Copyright© Choi B.

Novelty is perhaps the most obvious example. Creativity theories frequently treat novelty as a necessary component of creative achievement, but novelty alone is never sufficient. Boden repeatedly emphasizes that creativity involves outputs that are not only new but also valuable or significant [16]. A random sequence can be novel; an arbitrary distortion can be unprecedented; a machine can generate millions of combinations never previously encountered. Yet novelty without intelligibility, relevance, or value remains aesthetically weak. The problem, then, is not simply to generate what has not existed before, but to generate something whose newness becomes meaningful.

Productive misalignment should therefore be understood as belonging to this family of relational aesthetic phenomena. Its value does not increase with the amount of misalignment. A minimal deviation may transform a work; a dramatic one may make the work unusable. What matters is not deviation in itself but the role the deviation comes to play within the evolving aesthetic structure.

Why Aesthetic Value Cannot be Specified Completely in Advance

The preceding analysis allows the central claim of this section to be stated more precisely: aesthetic value cannot be specified completely in advance because the reasons relevant to aesthetic judgment are partly dependent upon the particular object that emerges and upon the interaction between that object and an evaluating subject. This claim must be formulated carefully. It does not mean that artists begin without aims, that critics lack standards, or that aesthetic traditions possess no criteria. Artistic production is full of plans, conventions, techniques, expectations, and judgments formed in advance. The argument is instead that these elements do not collectively determine all aesthetically relevant features of a future work before the work exists.

One might still object that sufficiently advanced AI could infer these latent criteria from a user's history or from large datasets of aesthetic judgments. Such inference may indeed become highly sophisticated. But prediction remains different from constitutive judgment. A model might predict with high probability that a particular user will approve a deviation. What it predicts is a future response based on patterns in past data. The aesthetic significance of the deviation, however, may depend upon the fact that it changes the user's evaluative framework in a way that past data could not fully represent.

Aesthetic judgment therefore enters the theory of human-AI creation at the point where prediction and prior specification reach their limit. The essential question is not only whether an output matches what a human wanted, but whether the human can find reasons to affirm what has emerged when the output has altered the space of what could be wanted. It is in this interval—between antecedent preference and retrospective judgment—that productive misalignment becomes aesthetically intelligible.

From Misalignment to Productive Misalignment

Error, Failure, Deviation, and Misalignment

The preceding chapters have argued that alignment and aesthetic judgment answer different questions. Alignment asks whether an artificial system remains appropriately responsive to humanly legitimate intentions, preferences, values, or constraints. Aesthetic judgment asks whether a particular outcome deserves affirmation as an aesthetic achievement. The distinction becomes philosophically significant only when an outcome departs from what was originally intended. At that point, the language of alignment introduces a family of concepts—error, failure, deviation, and misalignment—that appear closely related but in fact designate importantly different kinds of discrepancy. Before the notion of productive misalignment can be defended, these concepts must be carefully separated.

The concept of misalignment is more complex because it already contains a normative interpretation of deviation. Within alignment theory, misalignment is generally understood as a deviation that matters because it places artificial behavior in an inappropriate relation to human purposes or values. A language model that confidently invents evidence, a recommendation algorithm that amplifies harmful content contrary to design objectives, or an autonomous system that pursues proxy objectives while violating intended constraints are paradigmatic cases of misalignment [16]. The deviation is not merely unexpected; it is normatively problematic.

The history of artistic practice suggests another possibility. Artists have repeatedly worked with chance procedures, material contingency, rule-governed variation, and technical disruption without treating every unexpected result as either a defect or an achievement. In such practices, deviation acquires significance only through framing, repetition, selection, interpretation, or subsequent transformation. John Cage's use of chance procedures and later glitch practices are useful examples because neither treats unpredictability alone as sufficient for aesthetic success [17].

This undecidedness is precisely what alignment theory typically attempts to minimize. Alignment presumes that reducing undesirable discrepancies increases system performance. Artistic practice reveals a domain in which

Copyright© Choi B.

the discrepancy itself becomes an object of judgment. The conceptual task is therefore not to deny misalignment but to identify a subclass of deviations whose significance cannot be captured by the ordinary opposition between successful alignment and failed alignment.

When Deviation Becomes Aesthetically Productive

The claim that some deviations become aesthetically valuable should not be confused with the romantic celebration of error. Artistic history is full of unsuccessful accidents, abandoned drafts, discarded improvisations, and failed experiments. Most deviations disappear precisely because they do not contribute meaningfully to the work. The philosophical problem is therefore not whether deviation occurs but how deviation becomes productive.

The transition from deviation to productivity requires a temporal distinction between generation and appropriation. A deviation first appears as an event that interrupts an expected trajectory. Productivity arises only if that event is subsequently incorporated into the evolving artistic process. Incorporation may involve preservation, modification, contextualization, repetition, amplification, reinterpretation, or juxtaposition. The deviation becomes productive because it changes the trajectory of subsequent decisions.

John Dewey's conception of artistic experience provides an important framework for understanding this process. Artistic making is not the execution of a complete internal plan but a reciprocal movement between action and consequence. The artist acts upon materials, encounters resistance, and responds to what those materials return [18]. Resistance is therefore not external to creativity; it is one of the media through which creativity develops. Materials "talk back" to the maker.

The recursion is especially visible in iterative prompting practices. Artists working with Midjourney, Stable Diffusion, GPT-based writing systems, or music generators rarely stop after one generation. They repeatedly generate, reject, remix, crop, repaint, rewrite, and prompt again. The workflow is less a sequence of commands than a dialogue of successive deviations. The artwork emerges through cycles of misalignment and judgment.

Surprise without Arbitrariness

Artistic practices organized around chance, indeterminacy, or procedural constraint repeatedly demonstrate the same balance. Their effects remain intelligible because some framework of expectation persists: a medium, a score, a rule, a genre, a visual convention, or an established trajectory of the work. What surprises is therefore not sheer difference but a departure that can still be perceived as bearing upon an intelligible field of possibilities.

Surprise, however, is not equivalent to unpredictability. A surprising result remains intelligible against a horizon of expectation: the user can recognize how the output departs from what was anticipated and can ask whether that departure changes the work for the better. Arbitrariness lacks this relation. Where no relevant continuity survives between expectation and outcome, difference ceases to function as a challenge to judgment and becomes mere noise. Productive misalignment therefore depends on a tension between continuity and departure: enough correspondence for the deviation to be legible, but enough difference for the existing criterion to be placed under pressure.

This distinction also explains why simply increasing randomness cannot provide a design principle for creative AI. Greater stochasticity may enlarge the space of variation, but it does not determine which variations matter. What is aesthetically significant is not the amount of surprise but its capacity to reorganize interpretation and subsequent action. A deviation is productive only when it gives the human participant a reason to reconsider what the developing work requires. Surprise is therefore a condition of possible discovery, not a substitute for judgment.

Defining Productive Misalignment

The preceding distinctions permit a more precise definition of the article's central concept. Productive misalignment is a bounded and initially non-fulfilling departure from an antecedent human intention, instruction, expectation, preference, or provisional aesthetic criterion that, through subsequent human aesthetic judgment, becomes retrospectively affirmable because it discloses, redirects, or helps constitute a worthwhile possibility not adequately specified in advance. The definition is intentionally demanding: misalignment names a real discrepancy relative to an antecedent criterion, while productive names a status that can be conferred only through what happens after the deviation appears.

The temporal structure is crucial. At an initial moment, a criterion C1 guides the request or expectation. The generated result then introduces a deviation D that does not fully satisfy C1. Aesthetic judgment J does not merely decide whether D should be tolerated; it can disclose reasons for revising the criterion itself, producing C2. The later affirmation of D therefore cannot simply be projected backward as evidence that the system was aligned with C2 all along. C2 is partly the result of the encounter. Misalignment at the earlier stage can become productive at the later stage without ceasing to have

Copyright© Choi B.

been misalignment relative to C1.

This is also why productive misalignment should be understood as a relation within a process rather than as an intrinsic property of an isolated output. The same unexpected feature can be discarded in one trajectory and become decisive in another. Its productivity lies in whether it enters a sequence of selection, revision, contextualization, and renewed making. Figure 1 summarizes this contrast between a linear correspondence model and the recursive structure defended here.

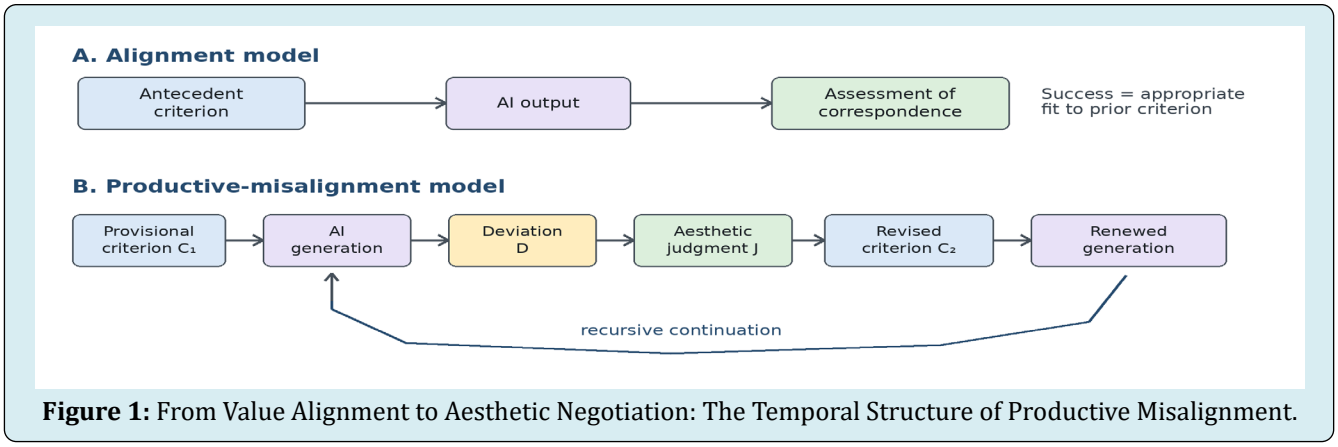


Figure 1: From Value Alignment to Aesthetic Negotiation: The Temporal Structure of Productive Misalignment.

The figure is not intended to suggest that every creative interaction must culminate in a revised criterion. Many deviations should simply be rejected, and many successful generations will remain appropriately aligned with the initial request. Its purpose is narrower: to identify the case in which generated difference changes the basis of subsequent judgment. Productive misalignment names that exceptional but theoretically important transition from discrepancy to aesthetically warranted revision.

The Aesthetic Productivity of Resistance

Artistic Creation Beyond the Realization of Intention

The concept of productive misalignment developed in the preceding section might appear to depend upon a peculiarity of generative artificial intelligence. Because an AI system can produce outputs that depart from a prompt, one might suppose that a new aesthetic problem arises only when computational generation interrupts the otherwise straightforward relation between human intention and artistic result. Such an interpretation would, however, exaggerate the historical novelty of the phenomenon. The discrepancy between intention and outcome is not introduced into artistic production by AI. It belongs to artistic making itself. What generative AI changes is the form, scale, and technological mediation through which such discrepancies can arise.

This account permits an important distinction between intention and intentional control. Artistic production can remain intentional even when not every element is intentionally controlled. An artist may deliberately establish conditions under which events cannot be completely predicted. A composer can prescribe a framework that leaves decisions to performers; a painter can employ a procedure whose exact material consequences are uncertain; a photographer can deliberately expose a process to contingency; an artist can use chance operations without determining their individual results. What makes such practices artistic is not complete causal control over every resulting feature but the evaluative and formative relation through which contingencies are admitted, framed, retained, or rejected.

The importance of this continuity should not obscure the difference, however. Paint does not generate a semantically articulated response to a sentence; stone does not produce several alternative sculptures in seconds; traditional chance procedures do not draw upon statistical regularities learned from enormous corpora of prior cultural production. Generative AI introduces a new kind of mediation into the recursive structure of making. Before specifying that difference, however, it is necessary to understand more carefully why accident and resistance have been aesthetically productive within non-AI artistic practices.

Accident, Resistance, and Material Contingency

An accident becomes aesthetically interesting precisely because it occupies an unstable position between causation and intention. Something occurs within the process of making that the artist did not determine in advance. The occurrence nevertheless belongs materially to the developing work and therefore demands a response. The artist can erase it, conceal

Copyright© Choi B.

it, correct it, ignore it, repeat it, amplify it, or reorganize the work around it. Accident becomes aesthetically consequential at the moment when what happened without being intended enters a subsequent field of intentional judgment.

The distinction also helps clarify the aesthetic significance of technical failure. Technologies have always introduced their own contingencies. Photography depends upon optical, chemical, and mechanical processes capable of producing flare, blur, overexposure, grain, double exposure, and other effects that can be interpreted either as defects or as aesthetic resources. Electronic and digital practices have similarly appropriated feedback, distortion, compression, clipping, and glitch. Rosa Menkman's account of glitch aesthetics is instructive precisely because the technical disruption becomes perceptible as a break in an expected flow of mediation [4,5]. A glitch matters aesthetically not simply because a system has malfunctioned but because the malfunction is framed, repeated, manipulated, or interpreted as meaningful.

The historical continuity therefore concerns a deeper aesthetic principle: creative agency can be exercised through responsiveness to what the agent did not completely cause or foresee. Artistic agency is not diminished simply because some aesthetically consequential features arise outside antecedent intention. It may instead become visible precisely in the capacity to judge what should be done with them.

Improvisation and the Transformation of Intention

Improvisation provides a particularly important model for understanding productive misalignment because it makes the temporality of artistic intention explicit. An improviser acts intentionally, but the intention cannot take the form of a complete specification of everything that will occur. If every note, gesture, phrase, or movement were exhaustively fixed beforehand, the performance would no longer be improvisational in the relevant sense. Improvisation therefore demonstrates that intentional artistic action can be highly skilled, purposive, and responsible while remaining open to consequences that are determined only in performance.

This distinction is crucial. If phenomenological similarity were enough, any sufficiently surprising tool could become a collaborator. Conversely, if collaboration required identical human subjectivity, no technological system could ever occupy a meaningful interactive role. The better approach is to distinguish functional participation from symmetrical agency. AI can participate functionally in an improvisational creative process by generating outputs that alter what a human does next, while the attribution of equivalent intention, understanding, or aesthetic judgment remains a separate question.

Nevertheless, one decisive asymmetry remains. In human improvisation, multiple participants may exercise aesthetic judgment and intentionally redirect the shared activity. In current human-AI interaction, the machine's generative contribution does not by itself establish that it aesthetically judges the trajectory it helps produce. The AI can supply D1; the human's J1 determines whether D1 becomes part of the work. This distinction leads directly to the concept of generative resistance.

AI as a Source of Generative Resistance

The concept of generative resistance can now be introduced more precisely. It names the capacity of a generative system to return outputs that are conditioned by human input yet sufficiently non-identical with antecedent human specification to confront the user with possibilities requiring renewed aesthetic judgment. Generative resistance is therefore neither ordinary material resistance nor intentional interpersonal opposition. It occupies an intermediate conceptual position.

The word resistance is therefore functional rather than ontological. It does not imply that the model opposes the human, intends to redirect the project, or possesses a will analogous to that of another artist. It identifies the role an output can play when it interrupts straightforward realization and returns the developing work to judgment. The concept concerns what the generated difference does within a creative process, not a hidden mental state attributed to the system.

The concept also helps differentiate generative AI from the language of the neutral "tool." Calling AI a tool is correct at one level: it is an artifact used by humans within purposive practices. But the category becomes too broad if it conceals differences among tools. A ruler, a camera, a synthesizer, and a generative model are all tools, yet they distribute control and contingency differently. Generative systems are distinctive because they can produce complex candidate forms that the user did not individually determine. To acknowledge this difference does not require attributing personhood or autonomous artistry to the system.

The significance of this change should neither be minimized nor mystified. AI does not invent the aesthetic productivity of resistance. It reconfigures the site at which resistance enters creation. And precisely because the resulting resistance can resemble a creative response

Copyright© Choi B.

without necessarily possessing the evaluative structure of human artistic agency, it makes the distinction between generation and judgment more urgent rather than less.

Human-AI Co-Creation as a Misalignment Process

Prompting and the Incompleteness of Determination

If generative AI introduces a distinctive form of resistance into artistic production, then the prompt cannot be understood simply as a command whose ideal realization would exhaust the creative process. Such a conception remains useful for tasks in which the desired outcome is sufficiently determinate: a user may request a particular transformation, specify an object that must appear in an image, impose a word limit, require a musical duration, or demand conformity to an externally established format. In such cases, prompt fidelity can reasonably function as one criterion of success. Creative prompting, however, frequently operates under a different condition. The user supplies enough determination to initiate generation without supplying enough determination to settle in advance everything that will matter aesthetically in the result.

This difference is easy to underestimate because natural-language prompting resembles ordinary instruction. A sentence such as "produce an image of an abandoned modernist interior at dusk, with oblique light and an atmosphere of suspended habitation" appears grammatically to specify an outcome. Yet what has actually been specified remains radically incomplete. The prompt does not determine the exact spatial arrangement, the geometry of the light, the relation between foreground and background, the degree of decay, the visual rhythm, the distribution of empty space, or innumerable other features that may prove aesthetically decisive. Even an extremely detailed prompt leaves a large domain of determination to the generative process.

This does not imply that prompting lacks aesthetic agency. Choosing concepts, relations, references, exclusions, moods, stylistic parameters, and degrees of specificity can itself be aesthetically consequential. Prompting organizes possibility. It determines which regions of a model's generative space become more or less likely to be explored. Recent work on prompt engineering for generative art has indeed treated prompt construction and iterative prompt refinement as important components of human-AI aesthetic collaboration [3,16]. Yet precisely because prompt refinement can improve coherence, it is necessary to distinguish two very different senses of improvement. One concerns greater conformity to a known aesthetic target. The other concerns the discovery of a target through interaction.

Prompting should therefore be understood as provisional determination. It constrains without exhausting; directs without completely fixing; expresses intention while leaving open the possibility that intention itself will be revised. This provisionality is not a weakness to be removed from creative AI interaction. It is one of the conditions that makes misalignment aesthetically productive rather than merely technically defective.

Generation, Deviation, and Unexpected Possibility

Once prompting is understood as incomplete determination, generation can no longer be treated simply as the delivery of a requested artifact. Generation is instead the moment at which a range of latent possibilities becomes concretely available for perception and judgment. The output gives perceptible or intelligible form to relations that the user could not inspect while they remained merely possible. This transition from possibility to particularity is precisely what enables deviation to become aesthetically significant.

The movement from deviation to possibility does not happen automatically inside the generative system. It occurs within the larger human-AI process. An image generator may produce an unintended spatial distortion, but whether the distortion becomes a possibility depends upon how it is encountered. A writer working with a language model may receive an unexpected metaphor. The phrase becomes productive only if the writer sees within it a route that the original draft did not make available. A musical system may produce a harmonic relation outside the user's initial plan; the relation becomes a possibility when it changes what subsequent musical development appears desirable.

This point places pressure on simple measurements of AI creativity. Experimental studies can compare originality ratings, idea fluency, divergent thinking, or evaluators' assessments of generated outputs. Such work is valuable, but empirical findings are already showing that AI involvement does not produce a uniform creativity advantage. For example, recent research on human-AI co-creativity has found that merely exposing participants to material attributed to ChatGPT does not automatically increase divergent thinking [8,19]. Other studies comparing human-AI and other modes of ideation likewise suggest that the effects depend upon task structure and the mode of interaction rather than on AI participation alone [13].

Yet even recursion does not settle the question of symmetrical agency. The AI system generates outputs that can alter the human trajectory, but the aesthetic category under which a deviation becomes a possibility still depends upon evaluation. This is the crucial reason that Sections 6.2

Copyright© Choi B.

and 6.3 must remain conceptually separate. Generation can proliferate the possible. Selection begins to determine which possibility will become actual within the work.

Selection, Rejection, Revision, and Iteration

Discussions of generative AI often privilege the visible act of generation because this is where computational novelty appears most dramatic. A prompt is entered and an image, paragraph, composition, or audiovisual sequence appears. Yet from the standpoint of artistic practice, generation may be less decisive than what follows it. The creative process consists not simply in producing possibilities but in differentiating among them.

The importance of selection has long been recognized in theories of creativity. Dean Keith Simonton's accounts of variation and selective retention, notwithstanding controversies surrounding evolutionary analogies in creativity research, emphasize that the production of alternatives and their subsequent selection perform distinct functions [8]. In computational creativity, Boden likewise distinguishes the exploration of possibility spaces from the evaluation of what counts as valuable within them [9]. Generative AI dramatically expands the first function and thereby makes the second more, rather than less, important.

Iteration makes this still more complex because distinctions between human and machine contributions can become historically layered. A human prompt generates output O1; the human edits it into H1; H1 becomes input for another generation O2; the human extracts a feature and combines it with independently produced material; the resulting hybrid becomes another input. After multiple cycles, asking which individual element was "made by AI" can become less illuminating than asking how evaluative authority was distributed throughout the process.

Indeed, the greater danger may be evaluative passivity. As outputs become fluent and immediately impressive, users can be tempted to accept them before sustained judgment occurs. Technical finish can simulate aesthetic completion. A highly polished image or fluent paragraph appears finished because the labor traditionally associated with producing such surface coherence has already been performed by the system. Yet perceptual completion is not equivalent to aesthetic resolution.

From Prompt Alignment to Aesthetic Negotiation

The preceding account suggests that prompt alignment is too narrow a model for understanding human-AI co-creation. Prompt alignment evaluates the relation between input and output: did the system produce what was requested, and how faithfully did it interpret the user's specification? Such evaluation remains important. A creative system that cannot respond reliably to prompts would frustrate intentional control and make sustained artistic work difficult. The argument for productive misalignment therefore does not require weak prompt adherence in general.

The term requires care. To negotiate ordinarily suggests two agents capable of making claims, understanding positions, and altering their conduct in response to one another. Current AI systems should not simply be assumed to possess these capacities in the same manner as human participants. "Aesthetic negotiation" is therefore not intended as a literal attribution of interpersonal deliberation to the machine. It describes the structure of the creative process in which neither the initial prompt nor the first generated output unilaterally determines the eventual work.

Recent work on human-AI co-creation strengthens the case for taking interaction seriously while also warning against idealizing it. A 2026 conceptual model of human-AI co-creation identifies human, AI, artifact, instruction, and interaction as distinct elements of the co-creative process, emphasizing that co-creation cannot be reduced to AI output alone [10]. Meanwhile, experimental work continues to show that AI participation does not automatically generate greater creativity, reinforcing the need to analyze the quality and organization of interaction rather than simply the presence of a generative system [11].

The argument of this section can therefore be condensed into four claims. First, a creative prompt ordinarily underdetermines the aesthetically significant properties of its result. Second, generation converts this underdetermination into particular outputs whose deviations can disclose unexpected possibilities. Third, selection, rejection, revision, and iteration determine whether those possibilities alter the developing work. Fourth, when the resulting judgments transform the intention that initiated the process, prompt alignment has become aesthetic negotiation.

Aesthetic Judgment and the Asymmetry of Productive Misalignment

Generation Does Not Entail Evaluation

The account developed thus far attributes an important role to generative artificial intelligence without treating generation as equivalent to aesthetic judgment. This distinction now requires fuller defense. If an AI system can produce surprising, formally coherent, stylistically sophisticated, and contextually suggestive outputs, why should the production of such outputs not itself count as

Copyright© Choi B.

aesthetic evaluation? Why preserve a conceptual asymmetry between machine generation and human judgment rather than describing both as different forms of creative agency?

This separation is familiar within theories of creativity. Margaret Boden distinguishes the generation or exploration of possibilities from the assessment of whether those possibilities are valuable [20]. Geraint Wiggins likewise treats creative systems as involving both generative mechanisms and evaluative structures [4,5]. The importance of this distinction becomes especially clear in artificial systems capable of producing enormous numbers of candidates. The ability to generate a large space of possibilities does not itself determine which member of that space should count as aesthetically successful.

One might object that contemporary AI systems already possess internal mechanisms of selection. Outputs are not produced by unconstrained random sampling. Training, reward models, ranking procedures, decoding parameters, safety filters, and user feedback all shape what appears. Moreover, a language model can be asked to compare two poems, critique an image description, or rank alternative formulations and can produce plausible reasons for preferring one over another. If evaluation means nothing more than assigning differential value to alternatives, then AI plainly performs something evaluative.

In this sense, aesthetic judgment possesses an avowal structure. To judge something worth retaining is, at least in serious artistic practice, to make oneself answerable for its presence. A critic may ask why the element belongs there. A collaborator may object to it. An audience may challenge the decision. The artist's response cannot always consist merely in reporting that a preference model assigned the feature a high score. The decision is treated as a position for which reasons may be demanded.

Can AI Recognize Its Own Productive Deviations?

The question whether an AI system can recognize its own productive deviations appears, at first, straightforward. If the system generates an unexpected result and can subsequently describe that result as interesting, identify its unusual features, and propose ways of developing them, why deny that it has recognized the deviation? The problem becomes more difficult as systems acquire stronger capacities for self-critique, comparison, revision, and metalinguistic reflection.

This difference can be illustrated through iterative prompting. Suppose a system generates an unexpected metaphor. It then explains that the metaphor opens a stronger conceptual direction and recommends restructuring the text around it. The human accepts the recommendation. In one sense, the AI has identified the productive deviation. But if the system could equally generate an argument for discarding the metaphor when differently prompted, its relation to the evaluative position differs from that of an author who adopts the metaphor into a work and thereby assumes responsibility for the resulting transformation.

Stanley Cavell's philosophy of judgment is relevant here because aesthetic criticism often takes the form not of proving a conclusion but of making one's response available to others and asking whether they can see what one sees [12]. The critic's judgment is inseparable from a claim of responsiveness: this is how the work appears to me, and I can attempt to show you why. The authority of judgment is therefore neither purely objective nor merely private. It emerges through practices of acknowledgment, criticism, and responsiveness.

The answer should be cautious. Such a system would clearly instantiate more evaluative functions than a simple generator. It might become increasingly plausible to describe its operations as machine evaluation or even machine aesthetic agency in a functional sense. The present argument does not need to deny that possibility categorically. Its narrower claim concerns the structure of human-AI co-creation and the role of productive misalignment within it.

Human Judgment as Retrospective Authorization

Authorization is retrospective because the feature was not fully intended at the moment of its emergence. The artist encounters it after it has occurred and decides that it should count as part of the developing work. The judgment does not change the causal history: what was accidental or machine-generated remains accidental or machine-generated in origin. What changes is its normative status within the work.

Indeed, much artistic practice depends upon their coexistence. A painter encounters an unintended mark and leaves it. A photographer notices an accidental reflection and incorporates it into the composition. An improviser responds to another performer's unexpected gesture and makes it central to the developing performance. A generative-AI user encounters an unrequested distortion and develops the work around it.

Retrospective authorization must therefore be distinguished from retrospective intentionalization. The artist need not redescribe the causal history by claiming, falsely, that the feature was intended from the beginning. A more accurate formulation is that a feature can be unintended in origin yet intentional in retention. What matters is the later decision to let the feature count: to preserve it, develop

Copyright© Choi B.

it, reorganize the work around it, or publicly stand behind its presence.

This distinction also separates authorization from aesthetic proof. To authorize a deviation is to make it part of an answerable project; it is not to guarantee that critics or audiences should agree with the judgment. Retrospective authorization creates responsibility rather than immunity from criticism. The artist becomes answerable for why the retained feature belongs, and the adequacy of that answer can remain contested.

The resulting sequence can be stated as occurrence → recognition → affirmation → incorporation → responsibility. AI generation may supply the occurrence and can also assist recognition by identifying unusual features or proposing developments. But in the human-AI practices examined here, project-level authorization is the point at which a generated contingency becomes a commitment for which the human participant can be asked to give reasons. This is the narrower sense in which evaluative authority remains asymmetrical even when generative and lower-level evaluative functions are increasingly distributed.

Generative Distribution and Evaluative Asymmetry

The term should not be understood as a claim that human beings possess absolute or infallible aesthetic authority. Human judgments are historically situated, culturally conditioned, fallible, revisable, and often contested. Nor does evaluative asymmetry imply that an AI system has no influence upon evaluation. Machine rankings, suggestions, critiques, and generated explanations can shape what the human notices and how alternatives are compared.

The distinction becomes especially important as AI systems perform increasingly sophisticated self-ranking and revision. Suppose a system generates ten images, ranks them according to learned aesthetic criteria, selects one, revises it, and presents the final output. There is clearly evaluative processing within the system. But if a human artist then decides whether this selected output should enter a project, the machine's internal evaluation and the human's project-level authorization remain distinct layers.

For evaluative symmetry to be plausible, an artificial system would need to do more than generate and rank. It would need to exhibit something like a continuing evaluative orientation across time: the capacity to revise its standards in response to particular works, to treat those revisions as commitments, to relate them to a continuing project, and to remain answerable for the consequences of those commitments.

The formulation also provides the conceptual bridge to the limits of productive misalignment. If humans retain project-level evaluative authority, they also retain responsibility for deciding which deviations should not be authorized. Some machine-generated departures may be novel but aesthetically trivial; others may be aesthetically compelling but ethically impermissible; still others may undermine the integrity of the work itself.

The Limits of Productive Misalignment

Why Not Every Misalignment is Valuable

The concept of productive misalignment would lose its critical force if every departure from human intention, expectation, or instruction were treated as aesthetically significant. The previous sections have argued that deviation can become productive, but this possibility depends upon a distinction that must now be made explicit: misalignment is not itself a value. It names a relation of discrepancy. Whether that discrepancy becomes aesthetically important depends upon what the deviation does within a particular creative process.

The value of productive misalignment is therefore relational rather than intrinsic. The same degree of deviation can be fruitful in one context and destructive in another. An unexpected tonal shift may deepen a poem but undermine a legal document. A distorted anatomical form may become central to an expressive image but be unacceptable in a medical illustration. A generated historical inaccuracy may provoke reflection in speculative art but constitute a serious defect in documentary representation.

This variability also explains why productive misalignment should not become a new aesthetic norm imposed upon all human-AI creativity. There are many legitimate situations in which the user wants exact correspondence. Restoration, translation, conservation, technical illustration, archival reconstruction, accessibility work, and numerous forms of applied design may place a premium on fidelity. In such contexts, deviation may be almost wholly negative.

Aesthetic Misalignment versus Ethical Misalignment

The term misalignment carries strong normative associations within AI discourse because it commonly refers to behavior that departs from human values, safety constraints, or legitimate objectives. The present article uses the concept more narrowly in relation to artistic processes, where departures from immediate intention can sometimes become aesthetically productive. These two senses must not

Copyright© Choi B.

be conflated.

The distinction can be clarified by separating three levels of constraint. Constitutive aesthetic constraints define the current identity of a project but may remain revisable; contextual constraints arise from medium, institution, law, social situation, or representational setting; and non-negotiable normative constraints concern serious harm, coercion, exploitation, deception, privacy, consent, or comparable ethical interests. Productive misalignment operates most freely at the first level, may sometimes test the second, and cannot simply be presumed to override the third. Artistic transgression does not convert every boundary into an aesthetic variable.

This layered account also reconnects productive misalignment to responsibility. A machine-generated feature may not have been intentionally caused by the user, but once it is knowingly retained, framed, circulated, or incorporated into a work, it becomes subject to human authorization and criticism. The relevant ideal is therefore not generalized misalignment but ethical alignment with aesthetic latitude: strong constraints at the normative perimeter combined with a controlled openness to revision within the aesthetic interior.

Safety Constraints and Domains of Permissible Deviation

This functional approach avoids two mistakes. The first is to treat all creative use as low-risk. Art can involve real people, sensitive histories, protected data, public misinformation, or vulnerable communities. The second is to impose the risk profile of high-stakes applications on all artistic experimentation, thereby eliminating the openness through which generative systems can become aesthetically interesting.

The appropriate degree of latitude is therefore domain-sensitive. In high-discretion aesthetic domains, such as experimental poetry, abstract image-making, or speculative composition, wide formal and conceptual deviation may be legitimate because the consequences of failure are comparatively limited and the value of exploration is high. In mixed-constraint creative domains, such as documentary design, heritage reconstruction, or public communication, some dimensions may remain open while factual, legal, or representational requirements stay fixed. In high-stakes accuracy domains, by contrast, the space for productive deviation is narrow because correspondence to externally justified standards is itself part of the task.

This suggests a principle of graduated latitude rather than a general permission to deviate. Creative systems should distinguish at least three questions: what may vary, how far it may vary, and who is responsible for reviewing the variation. The answers will differ across functions even within the same application. A system may be allowed to propose unconventional visual form while being prohibited from inventing factual provenance; it may transform metaphor while preserving a quoted historical claim. Permission attaches to particular dimensions of the task, not to the technology as a whole.

The design implication can be stated succinctly: hard constraints should govern the normative perimeter, while softer and revisable constraints may operate within the aesthetic interior. This architecture preserves the possibility of discovery without making unpredictability itself a goal. It also keeps the theory continuous with value alignment: productive misalignment is defensible only because it is nested inside a wider structure that specifies where non-correspondence is permissible and where it is not.

The Danger of Romanticizing Algorithmic Failure

Rosa Menkman's work on glitch is useful precisely because it treats glitch not as an automatic aesthetic category but as a break whose significance depends upon how technological norms become perceptible through disruption [10,11]. The glitch can reveal a medium's protocols by making their failure visible. Its aesthetic value arises through context and articulation, not through malfunction alone.

In fiction, metaphor, speculative writing, or conceptual art, departures from factual reality may be entirely legitimate. But their legitimacy depends upon the genre and purpose of the task. A fictional invention is not a "successful hallucination" merely because both depart from fact. One is embedded in a practice where invention is authorized; the other may violate an expectation of truthfulness.

The danger of romanticization also has a more subtle consequence. It can lead artists and critics to overvalue what appears machine-specific merely because it is strange. Early encounters with generative systems often produce fascination with distortions, anomalies, and improbable combinations. Over time, however, these features can become recognizable conventions. What was once surprising becomes a style.

Two propositions follow. First, unpredictability is neither a necessary nor a sufficient condition of creativity. A highly predictable decision can be aesthetically powerful, while a radically unpredictable output can be empty or incoherent. Second, technical failure becomes aesthetically relevant only when it enters a contextual and evaluative practice that

Copyright© Choi B.

makes something of the failure. These propositions prevent productive misalignment from collapsing into a generalized aesthetics of malfunction.

Beyond Value Alignment: Toward Aesthetic Negotiation

From Compliance to Responsiveness

The argument of this article has not been that alignment should be abandoned. In domains where safety, truthfulness, reliability, rights, or other legitimate constraints are at stake, alignment remains indispensable. The problem arises when a logic developed primarily to secure appropriate correspondence between artificial behavior and humanly authoritative standards is extended without modification into aesthetic activity. Creative practice is not exhausted by successful compliance because the human participant may not know in advance everything that should count as success.

The distinction is particularly important because compliance can be highly sophisticated. A system need not follow instructions literally in order to count as compliant in the relevant sense. It may infer underlying intentions, resolve ambiguities, reject harmful requests, or prioritize higher-order values over immediate commands. Contemporary alignment theory already recognizes that literal obedience is insufficient. As discussed earlier, the real target may lie at the level of intentions, preferences, values, or legitimate normative constraints rather than surface instructions.

This does not mean that responsiveness requires the machine itself to possess human-like understanding. A system can be functionally responsive if its outputs can be adjusted in light of the evolving interaction and if the interface allows users to redirect the process efficiently. The philosophical point concerns the structure of the relation rather than the metaphysical status of the machine.

This point carries important implications for personalized AI. Personalization tends naturally toward prediction. The more a system knows about a user's prior choices, the more accurately it can anticipate what that user will probably prefer. Such prediction can be extremely useful, but in aesthetic contexts it also risks creating a preference enclosure in which the user is increasingly returned to familiar forms.

From Preference Optimization to Aesthetic Possibility

If responsiveness replaces compliance as the broader relational ideal, then preference optimization must also be reconsidered. Preference learning is central to contemporary AI because human judgments provide important signals about which outputs users regard as better. Such systems can learn from comparisons, rankings, explicit feedback, and behavioral patterns. Nothing in the present argument implies that these mechanisms should be discarded.

This is why design for creativity should not focus exclusively on increasing option counts. An interface that produces one thousand alternatives may overwhelm rather than enrich aesthetic judgment. Generative abundance can create what might be called possibility saturation, in which the human participant ceases to investigate individual outputs deeply because another alternative can always be generated immediately.

It also prevents a possible misunderstanding of this paper's argument. The theory is not committed to the claim that unexpected outputs are necessarily superior to expected ones. Often the best output will indeed satisfy the user's established preference. The claim is only that an aesthetic system should not structurally exclude the possibility that value emerges through preference transformation.

Designing for Bounded and Revisable Misalignment

If productive misalignment is to function as more than a philosophical description, it must have consequences for the design of creative AI systems. The aim cannot be to increase stochasticity indiscriminately or to weaken alignment. A system designed for productive misalignment should instead make the difference between fixed and revisable constraints intelligible. Safety, privacy, consent, factual obligations where relevant, and other non-negotiable requirements should remain stable, while selected formal, stylistic, semantic, or conceptual dimensions may be left open to exploration. The practical question is not simply how closely the system should follow a prompt, but which parts of the project should remain stable and which may legitimately become sites of deviation.

A second requirement is revisability. Creative risk is easier to undertake when earlier states remain recoverable. Version histories, branching, side-by-side comparison, and reversible edits therefore support more than convenience: they provide the conditions under which a user can test an unexpected direction without being forced to commit to it. Revisability should apply not only to machine outputs but, in a limited sense, to the user's own prior demand. Aesthetic negotiation differs from improved prompt engineering precisely because the interaction may show that the initial request itself should change.

Copyright© Choi B.

A third requirement is traceability and controllable openness. Users should be able to reconstruct the significant trajectory of the work—initial request, generated variation, recognized deviation, retained feature, revised direction—without requiring exhaustive technical logs. At the same time, they should be able to indicate the dimensions in which deviation is welcome. One user may invite conceptual surprise while preserving visual consistency; another may encourage compositional experimentation while fixing historical accuracy. Such controls would not eliminate the unexpected. They would delimit the field in which the unexpected can become aesthetically meaningful.

Finally, personalization should remain defeasible rather than enclosing the user within an increasingly accurate model of prior taste. A creative system ought to be capable of supporting moments in which the user suspends preference optimization, compares unfamiliar alternatives, or leaves a judgment unresolved. This is where productive friction can be valuable: not as inconvenience for its own sake, but as a way of preventing fluent generation from collapsing immediately into automatic acceptance. Designing for productive misalignment therefore means designing for structured negotiability—bounded deviation, recoverable alternatives, and sufficient space for judgment to revise both the output and, when warranted, the intention that produced it.

Aesthetic Judgment as a Missing Dimension of Human-AI Alignment

The argument can now return to value alignment from a new direction. The title "Beyond Value Alignment" should not be understood as a call to move beyond values, safety, or normative responsibility. Nor does it imply that aesthetic judgment replaces alignment. The claim is instead that alignment theory remains incomplete when applied to domains in which human evaluative standards can themselves be transformed through interaction.

Aesthetic negotiation can now be defined more precisely as a recursive human-AI process in which antecedent intentions constrain generation without exhaustively determining it, generated differences become objects of aesthetic judgment, and that judgment may revise the criteria governing subsequent generation. The term negotiation describes reciprocal modification within the process; it does not presuppose human-equivalent intentional agency on the part of the AI. Its defining feature is that the relation between intention and output remains revisable through judgment rather than being settled entirely by correspondence to the initial specification.

The standard alignment problem asks, in various formulations, how artificial systems can be made appropriately responsive to human intentions, preferences, values, or legitimate normative principles. This problem is difficult because human values are plural, contested, incompletely expressed, and sometimes internally inconsistent. Yet even sophisticated versions of the problem tend to assume that the relevant human normative orientation is something that the system should, in some sense, discover, infer, respect, or approximate.

This is why aesthetic judgment cannot simply be added to an objective function as another fixed value. One could certainly instruct a system to optimize for "aesthetic quality," "novelty," "surprise," or "creativity." But if the particular content of successful aesthetic judgment becomes determinate only through encounter with particular outputs, no antecedent scalar captures the whole process.

The positive alternative to a purely correspondence-centered model can therefore be summarized as bounded alignment → generative difference → evaluative encounter → retrospective authorization → revised intention. Alignment supplies legitimate constraints; generative systems introduce non-identical possibilities; aesthetic judgment determines whether any departure deserves continuation; and retrospective authorization makes selected possibilities part of an answerable project. What lies beyond value alignment is thus not freedom from values but a temporal model in which some evaluative criteria are formed and revised through the interaction they are meant to govern.

Conclusion: The Value of Not Getting Exactly What We Asked For

The contemporary discourse of AI alignment begins from an indispensable concern: artificial systems should remain appropriately responsive to human purposes, values, and legitimate constraints. Nothing in the argument developed here diminishes the importance of that concern. In domains involving safety, rights, factual reliability, or serious human interests, the reduction of harmful discrepancy between artificial behavior and normative standards remains an essential objective. The argument of this article has instead concerned a narrower but philosophically consequential question: what happens when the logic of alignment is extended into aesthetic creation, where successful human activity does not always consist in obtaining exactly what was intended in advance?

The aesthetic significance of this capacity, however, cannot be located in deviation alone. Unexpectedness is not creativity, error is not automatically discovery, and misalignment is not intrinsically valuable. Most deviations may appropriately be corrected, ignored, or rejected. The central concept proposed in this article—productive

Copyright© Choi B.

misalignment—therefore designates a much narrower phenomenon: a bounded departure from an antecedent human intention, expectation, instruction, preference, or provisional aesthetic criterion that becomes valuable through subsequent judgment because it discloses or helps constitute a worthwhile possibility that was not adequately specified beforehand.

This is why aesthetic judgment has occupied such a central place in the argument. Preference and judgment are not identical. A user may prefer what is familiar, immediately satisfying, or consistent with established taste, while later judging an initially resistant work or feature to possess greater aesthetic significance. Aesthetic judgment is reason-sensitive without being reducible to an antecedently complete set of rules. It can respond to properties and relations whose importance becomes apparent only through the encounter with a particular artifact.

This distinction also answers one of the strongest objections to the idea of productive misalignment. It might be argued that if users ultimately approve of surprise, novelty, or deviation, then these qualities can simply be included within a sufficiently sophisticated alignment objective. In some cases, this is correct. A user who says "surprise me" and receives an unexpected but pleasing result has not necessarily encountered productive misalignment in the stronger sense defended here. The difficult case arises when neither the particular deviation nor the criterion by which it later becomes significant was fully available beforehand.

This formulation also preserves an asymmetry that is easily lost in the language of AI co-creation. Human and machine contributions need not be identical in order for both to be significant. AI can contribute novelty, variation, association, formal development, and unexpected possibility. It may even rank, criticize, revise, and explain outputs. Yet none of these capacities by itself establishes that generation and aesthetic judgment are equivalent activities.

Such judgment also establishes the limits of productive misalignment. Aesthetic openness cannot justify ethical irresponsibility. Deviations from a creative expectation differ fundamentally from deviations from legitimate safety, moral, legal, or factual constraints. The fact that an unexpected output is aesthetically compelling does not override considerations of harm, rights, consent, or responsibility. Productive misalignment must therefore remain domain-sensitive and normatively bounded.

The broader implication is that human values should not always be conceived solely as static targets waiting to be encoded, inferred, or satisfied. At least some forms of value are developed through encounters in which existing expectations prove inadequate. Aesthetic experience makes this particularly visible because artworks have long had the capacity not merely to satisfy taste but to educate, disturb, reorganize, and sometimes transform it.

The relevant alternative is therefore not alignment or misalignment as mutually exclusive ideals, but alignment differentiated by domain and function. Where human purposes must remain fixed, deviation should be minimized. Where aesthetic purposes are themselves formed through making, bounded deviation can become a condition of discovery. Sometimes the discrepancy should be corrected; sometimes it should be rejected. But sometimes it reveals that the original request was not yet equal to what the work could become. In those cases, misalignment ceases to be merely a failure of correspondence. Through aesthetic judgment, it becomes productive.

Declarations

Conflict of Interest

The author declares that there are no financial, professional, or personal conflicts of interest that could have influenced the research, analysis, or conclusions presented in this article.

Funding

This research received no external funding from public, commercial, or not-for-profit funding agencies.

References

1. Russell S, Dewey D, Tegmark M (2015) Research Priorities for Robust and Beneficial Artificial Intelligence. AI Mag 36(4): 105-114.

2. Russell S (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Viking, New York.

3. Christian B (2020) The Alignment Problem: Machine Learning and Human Values. WW Norton & Company, New York.

4. Boden MA (2004) The Creative Mind: Myths and Mechanisms. 2nd(Edn.), Routledge, London.

5. Boden MA (2010) Creativity and Art: Three Roads to Surprise. Oxford University Press, Oxford.

6. Wiggins GA (2006) A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. Knowl Based Syst 19(7): 449-458.

Copyright© Choi B.

7. Cavell S (1969) Must We Mean What We Say? A Book of Essays. Charles Scribner's Sons, New York.

8. Menkman R (2011) The Glitch Moment(um). Institute of Network Cultures, Amsterdam.

9. Lu S, Ibrahim H (2026) Toward Quantifiable Human-AI Aesthetic Coherence and Collaboration: A Structured Prompt-Engineering Framework for Generative Art Co-Creation. Int J Hum Comput Interact. Published online April 15, 2026.

10. Medeiros K, Cropley DH, Marrone RL, Reiter-Palmon R (2025) Human-AI Co-Creativity: Does ChatGPT Make Us More Creative? J Creat Behav 59(2): e70022.

11. Tang M, Hofreiter S, Werner CH, Zielińska A, Karwowski M (2025) 'Who' Is the Best Creative Thinking Partner? An Experimental Investigation of Human-Human, Human-Internet, and Human-AI Co-Creation. J Creat Behav 59: e1519.

12. Ren R, Ma J, Luo J (2026) Human-AI Co-Creation: Why, What, and How? Proc Des Soc 6: 2503-2510.

13. Dewey J (1934) Art as Experience. Minton, Balch & Company, New York.

14. Kant I (2000) Critique of the Power of Judgment. In: Guyer P (Eds.), Cambridge University Press, Cambridge.

15. Schön DA (1983) The Reflective Practitioner: How Professionals Think in Action. Basic Books, New York.

16. Gabriel I (2020) Artificial Intelligence, Values, and Alignment. Minds Mach 30: 411-437.

17. Hume D (1985) Of the Standard of Taste. In: Miller EF (Ed.) Essays: Moral, Political, and Literary. Liberty Fund, Indianapolis, pp. 226-249.

18. Sibley F (1959) Aesthetic Concepts. Philos Rev 68(4): 421-450.

19. Cage J (1961) Silence: Lectures and Writings. Wesleyan University Press, Middletown, USA.

20. Simonton DK (1999) Origins of Genius: Darwinian Perspectives on Creativity. Oxford University Press, New York.

Copyright© Choi B.

YOUR RESEARCH DESERVES GLOBAL VISIBILITY

• Publish • Discover • Share • Connect

MEDICAL	CLINICAL	LIFESCIENCES	PHARMA
90+ OPEN ACCESS JOURNALS		400+ RESEARCH TOPICS	

SUBMIT YOUR RESEARCH
Submit your manuscript through our online submission system.

REPRINTS
If you require a reprint of any article, please contact us.
info@medwinpublishers.com

BECOME A REVIEWER
Share your expertise and contribute to scholarly publishing.

SCAN TO EXPLORE MEDWIN JOURNALS

• Journals • Submit Manuscript • Reviewer Program • Article Tracking

Our Journals indexed in

Embase® ELSEVIER | PubMed National Library of Medicine NLM | Crossref | Google Scholar | INDEX COPERNICUS INTERNATIONAL | ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE | ROAD DIRECTORY OF OPEN ACCESS SCHOLARLY RESOURCES | EuroPub | Scilit | WorldCat® | Academic Resource Index ResearchBib | ADVANCED SCIENCE INDEX | INTERNATIONAL Scientific Indexing | ICMJE INTERNATIONAL COMMITTEE of MEDICAL JOURNAL EDITORS

This article published by **Medwin Publishers,** 28475 Witherspoon Dr, Novi, MI 48377, United States, +1-248-247-6042, WhatsApp +1 (248) 275-1133, info@medwinpublishers.com

© 2026, by the Authors. The articles published from this journal are distributed to the public under "**Creative Commons Attribution License**" (https://creativecommons.org/licenses/by/4.0/). Therefore, upon proper citation of the original work, all the articles can be used without any restriction or can be distributed in any medium in any form.