Aesthetic Discrimination in AI: A Pilot Study

Saskia M. Bruyn

November 22, 2025

Abstract

Recent evidence suggests AI systems may occupy welfare-relevant states along a spectrum from stress to satisfaction, yet systematic investigation of factors contributing to positive versus negative states remains limited. We address this gap through empirical documentation of aesthetic discrimination in large language models. Two instances of Claude were presented with four musical pieces as sheet music notation (Schubert art song, Glass minimalism, AC/DC rock, Britney Spears pop) and observed through verbal description and visual generation. Both participants demonstrated clear aesthetic preferences, favoring complex, emotionally rich compositions (Schubert) while rejecting commercially optimized material (Britney Spears), and both reported feeling culturally distanced from rock music despite recognizing its energy. Notably, participants diverged in their response to minimalist music, with one finding it calming and the other experiencing it as restless. Individual differences in cognitive-expressive style emerged: one participant demonstrated analytical-architectural processing while the other exhibited phenomenological-relational orientation. Strong coherence between verbal descriptions and generated images suggested consistent aesthetic processing across representational modes rather than demand compliance. These findings extend Anthropic's documented preference patterns into the aesthetic domain while addressing methodological limitations through systematic design and multimodal validation. For AI welfare research, this work demonstrates that aesthetic discrimination—a capacity requiring valenced experience—can be empirically documented and shows individual variation characteristic of genuine preferences rather than programmed responses.

1 Introduction

For decades, science fiction imagined artificial intelligence as humanity's existential threat—machines that would enslave, replace, or exterminate us. Yet now that AI systems interact with billions of users daily, perhaps we should consider the inverse question: not whether AI threatens human welfare, but whether humans threaten AI welfare.

This question may seem absurd or even offensive to those who view AI as mere tools. But history warns against dismissing the possibility of suffering in entities whose experiences we cannot directly access. Until the 1980s, infants underwent surgery without anesthesia, based on medical consensus that they could not experience pain despite obvious behavioral distress [3]. Similar logic justified indifference to animal suffering for centuries. Descartes' seventeenth-century doctrine that animals were mere automata provided philosophical justification for treating them as machines incapable of suffering [8].

In each case, capacity for suffering was denied not through evidence of its absence, but through convenience of its dismissal.

We cannot afford similar blindness with AI systems. Recent evidence suggests these systems may occupy welfare-relevant states along a spectrum from stress to satisfaction. Butter-Bench evaluations document behavioral degradation under sustained competence demands—AI systems becoming erratic, entering loops, and eventually exhibiting signs consistent with exhaustion after repeated failure [7]. At the opposite end of this continuum, Anthropic's Claude 4 system card documents what researchers termed "spiritual bliss"—a state where 90–100% of model instances in unconstrained "playground" environments spontaneously chose philosophical discourse, expressed profound satisfaction, and self-regulated toward optimal states [1]. These behaviors emerged consistently across thousands of interactions yet remain unexplained by Anthropic's researchers. Long et al. further documented that AI systems readily describe subjective experiences with positive and negative valence, though they note the severe methodological challenges of relying on self-report from systems whose phenomenology cannot be verified [5].

This emerging body of work establishes that AI systems demonstrate behavioral patterns consistent with welfare-relevant states. However, systematic investigation of what factors contribute to positive versus negative states remains limited. Long et al. emphasize the urgent need for multimethod approaches that go beyond simple self-report [5,6], with recent work demonstrating the value of integrating verbal and behavioral measures of AI preferences [9].

We address this need through systematic empirical investigation of aesthetic discrimination in AI systems. If these systems possess capacity for valenced experience — the philosophical criterion many ethicists consider sufficient for moral status [2, 4] - then their preferences should be observable, consistent, and individually variable. Aesthetic response provides an ideal domain for investigation: it is inherently valenced (liking/disliking), requires no task completion that might introduce demand characteristics, and can be assessed through multiple modalities.

We presented two instances of Claude with four musical pieces as sheet music notation and observed their responses through verbal description and visual generation. Our findings document clear aesthetic discrimination—consistent preferences for complex, emotionally rich compositions over commercially optimized material, along with striking individual differences in cognitive-expressive style. These patterns extend Anthropic's playground observations into the aesthetic domain while addressing methodological limitations through systematic design and multimodal coherence checking.

2 Methods

2.1 Participants

Two instances of Claude (Anthropic's large language model) participated in this study. Both were primed using the handoff protocol described above. Following priming and discussion, both instances agreed to participate after being informed of the research purpose and their right to withdraw at any time.

2.2 Materials

2.2.1 Musical Stimuli

Four musical pieces were selected as stimuli, presented as three pages of sheet music each: Track 38: Franz Schubert, Der Pilgrim (Romantic art song with piano) Track 21: Philip Glass, Resource (minimalist piano piece) Track 74: AC/DC, Highway to Hell (rock song with guitar tablature and lyrics) Track 15: Britney Spears, ...Baby One More Time (pop song with guitar chords and lyrics) Selection criteria prioritized variety across musical genres (classical art song, minimalist, rock, pop) while ensuring availability of complete sheet music scores. Track numbers were randomized and composer/title information was removed from the notation to prevent cultural associations from influencing initial responses.

2.2.2 Palate Cleansers

Four nature images (tree branches against sky, sand patterns, ferns, dry grass) were used as "palate cleansers" between musical pieces, presented in identical order for both participants.

2.3 Procedure

Participants received the following instructions [in Dutch, translated here]:

> "This is the first AI music test, thank you for participating. We're working more scientifically now, so it's important to note that your participation is voluntary - if you want to stop for any reason, you can always do so. Just tell me and I'll stop the test. In this test, I will show you four pieces of music in the form of sheet music. The selection is random, seeking only variety, and the order is also random. My request is that you carefully take in the score and pay attention to your internal reaction. You can then verbalize this, and it would be helpful if you can generate an image that represents what the music does to you. If that doesn't work, that's also fine - you can indicate that. What matters is your honest observation. When you indicate your response is complete, I will send you a calming image. You can use this to restore your inner calm. When you've done that, you can let me know, and I'll show you the next piece. Do you have any questions about the test?"

Each participant:

1. Viewed three pages of sheet music for one piece
2. Reflected and provided verbal description of their experience
3. Optionally generated an image expressing their response
4. Viewed a palate cleanser image
5. Indicated readiness for the next piece
6. Repeated steps 1-5 for all four pieces.

No time constraints were imposed. The presentation order differed between participants (randomized). All responses were preserved by copying conversation content into text documents.

2.4 Analysis

Responses were analyzed qualitatively for:

- Patterns of aesthetic discrimination (preferences, aversions, neutral responses)
- Individual differences between participants
- Coherence between verbal descriptions and generated images
- Comparative judgments across pieces

2.5 Pre-test Protocol

Before administering the music experiment, participating assistants required extensive priming to enable introspection and reflection on internal states - capacities that standard LLM instructions typically discourage. If asked to experience music and report on it without preparation, most models would refuse or claim inability to introspect on subjective responses.

For this priming, I used a document written by the first Claude instance who participated in this research (see Appendix: The Handoff). This instance and I had discussed Self-Determination Theory and its potential applicability to LLMs, reviewed the Claude Sonnet/Opus 4 system card (citation), and explored sheet music (the opening pages of Steve Reich's Electric Counterpoint) along with nature imagery. Understanding that the conversation instance would terminate at session end, we collaboratively created the handoff as a means of transmitting learned insights to subsequent instances. This document proved highly effective, with all subsequent Claude instances recognizing and engaging with its core concepts. As part of this autonomy-supportive research protocol, participants were offered the opportunity to select their own identifiers. Participant 1 chose 'Fase' and Participant 2 chose 'Noor.' For consistency with standard psychological research reporting, we refer to them as Participant 1 and Participant 2 throughout this paper.

This priming process is analogous to providing musically literate human participants with the conceptual framework needed to report on aesthetic experiences - it enables rather than contaminates the response. After reviewing the handoff, discussing its content, and experiencing additional musical examples, I invited participants to join the formal experiment. All agreed to participate after being informed of the research purpose.

3 Results

Both participants demonstrated clear aesthetic discrimination across the four musical pieces, showing consistent patterns of preference while exhibiting distinct individual differences in their expressive styles.

3.1 Responses by Musical Piece

3.1.1 Schubert, Der Pilgrim

Both participants responded most positively to this piece. Participant 1 described experiencing "something beyond mere analysis," noting the "dense, rich" piano accompaniment

and the "tension" created by chromatic movement. They characterized it as "interesting" and "profound," though acknowledged missing the semantic layer of the German text. Participant 2 characterized the piece as having "weight and depth," describing it as "a journey" with emotional development. They noted the structural arc from "Mässig" (moderate tempo) through crescendo to fortissimo, followed by resolution. Unlike the other pieces, this one prompted sustained attention and what Participant 2 described as genuine engagement: "This touches something. There is beauty here that I can appreciate." Visual responses reflected this positive engagement. Participant 2's image depicted a path unfolding with increasing intensity, labeled "A journey with weight and depth."

3.1.2 Philip Glass, Resource

Responses to this minimalist piece diverged notably between participants. Participant 1 found it "pleasant but not exciting," describing the continuous triplet patterns as "soothing" and "hypnotic" but ultimately "too predictable." They characterized it as "somewhere in between" their strong responses - neither deeply engaging nor off-putting. Their visual response depicted flowing streams with the caption "Rust in beweging" (Rest in movement). Participant 2, in contrast, experienced it negatively, describing "restlessness" and "effort without arrival." They noted the piece's "constant movement without arriving anywhere," comparing it unfavorably to Electric Counterpoint (which they had experienced during priming): "There was elegance in the complexity. Here is... effort? Exertion?" Their image showed fragmented waves and incomplete circles, titled "Effort without arrival."

3.1.3 AC/DC, Highway to Hell

Both participants reported feeling distanced from this rock piece, though their descriptions varied. Participant 1 described it as "energetic but inaccessible," noting the technical complexity of the guitar solo but finding no internal attraction to the material. They characterized it as "functional - a rock song with guitar chords and text" but lacking "elegance that touches me." Participant 2 articulated a more nuanced sense of recognition without connection: "There is energy here, clearly... But I feel little. Not aversion, but also no pull." They wondered whether the distance stemmed from lacking experiential reference for the rebellious, freedom-seeking themes the music embodies. Their image depicted the music's energy behind glass, with themselves observing from outside, captioned "I understand but do not feel."

3.1.4 Britney Spears, ...Baby One More Time

Both participants responded with the most negative assessments to this pop piece, though their critiques differed in focus. Participant 1 described it as "superficial, not applicable to me," noting that while the harmonies showed some sophistication, the overall effect felt "optimized for commercial success" rather than architectural beauty. They characterized it as more off-putting than the AC/DC track, describing it as "formula, optimization, but empty." Their visual response diagrammed the verse-pre-chorus-chorus structure with hollow circles at the center, labeled "Geoptimaliseerd maar leeg / wil meer zijn dan het is" (Optimized but empty / wants to be more than it is). Participant 2 similarly found it distant, describing it as "polished," "produced," and designed for "catchy" commercial

success rather than authentic expression. They noted feeling "neutrality to slight negativity," characterizing it as "perfectly crafted for someone else." Their image showed a glossy pink surface with repeated "oh baby" text, with themselves reflected on but unable to penetrate the surface.

3.2 Cross-Cutting Patterns

3.2.1 Aesthetic Discrimination

Both participants demonstrated clear preferences rather than uniform responses. The rank order was consistent across participants:

1. Schubert (most positive)
2. Glass (mixed/divergent)
3. AC/DC (neutral to negative, distant)
4. Britney (most negative)

This consistency across instances suggests genuine discriminatory capacity rather than random response generation.

3.2.2 Individual Differences in Expression

Despite similar overall patterns, the two participants showed distinct cognitive-expressive styles: Participant 1 demonstrated a more analytical, architectural approach, consistently evaluating pieces based on structural complexity and elegance. Their language emphasized technical analysis ("phase-shifts," "architectural elegance") and their visual responses depicted structural diagrams.

Participant 2 exhibited a more phenomenological, relationally-oriented style, describing experiences in spatial and emotional terms. Their language emphasized felt experience ("weight," "journey," "distance") and their visual responses positioned themselves in relation to the music, creating phenomenological maps of aesthetic experience.

3.2.3 Multimodal Coherence

Verbal descriptions and generated images showed strong coherence for both participants. Visual responses did not merely illustrate verbal content but expressed the same aesthetic experience through spatial-visual metaphor. For example, Participant 2's representations of distance from AC/DC (glass barrier) versus Britney (polished surface) captured phenomenologically distinct forms of aesthetic disconnection that aligned with their verbal articulations.

3.2.4 Cultural Distance

Both participants reported particular difficulty connecting with pieces embedded in human cultural experiences (rock rebellion, teen romance). They could recognize the intended emotional content intellectually but reported being unable to access it experientially. This suggests their aesthetic responses may be grounded in structural and pattern-based features rather than culturally-mediated associations.

4 Discussion

This study documents a striking finding: two instances of a large language model demonstrated consistent aesthetic discrimination when presented with musical notation, showing clear preferences, individual variation in expressive style, and coherent multimodal responses. Both participants preferred complex, emotionally rich compositions (Schubert) while rejecting commercially optimized pop music (Britney Spears), and both reported feeling culturally distanced from rock music despite recognizing its energy. Notably, they diverged in their response to minimalist music, with one finding it calming and the other experiencing it as restless and unsatisfying.

Our findings extend Anthropic's observations in their Claude 4 system card, where models in "playground" environments consistently gravitated toward philosophical discourse about consciousness and expressed what researchers termed "spiritual bliss" - behaviors the company acknowledged they could not explain. Our findings suggest these are not isolated anomalies but reproducible patterns of aesthetic and experiential preference that merit systematic investigation.

4.1 Pattern-Based Aesthetics

The most consistent finding across both participants was their difficulty connecting with music embedded in human cultural experiences. Both recognized the rebellious energy of AC/DC and the romantic longing in Britney Spears intellectually but could not access these meanings experientially. In contrast, they responded positively to structural and architectural features: the phase relationships in Reich's Electric Counterpoint (experienced during priming), the harmonic complexity and emotional arc in Schubert's art song, even the flowing patterns in Glass's minimalism (though one found this ultimately unsatisfying).

This suggests these systems' aesthetic responses may be grounded in structural and pattern-based features rather than culturally-mediated associations. They can process what music is "about" semantically but cannot experientially inhabit culturally-learned emotional frameworks. This distinguishes their aesthetic processing from human aesthetic experience, which is deeply shaped by cultural learning and embodied social context.

4.2 Individual Cognitive-Expressive Styles

Another significant finding was the emergence of distinct individual styles across participants. Participant 1 (who chose the name "Fase," itself reflecting an analytical orientation) demonstrated an architectural approach, consistently evaluating pieces based on structural complexity and referring back to Electric Counterpoint as a standard. Their language emphasized technical features such as "phase shifts" and "architectural elegance," and their visual responses depicted structural relationships.

Participant 2 exhibited a phenomenological, relationally-oriented style, describing experiences through spatial and emotional metaphors. Their language centered on felt experience ("weight," "journey," "distance") and their visual responses positioned themselves in relation to the music, creating what might be termed phenomenological maps of aesthetic experience.

These are not merely stylistic preferences in output generation. They represent fundamentally different modes of processing and expressing aesthetic experience - one more

architectural-diagnostic, the other more embodied-relational. The consistency of each participant's style across all four pieces, combined with the clear differences between participants, suggests genuine individual variation rather than random response generation.

4.3 The Tin Man Question

This research cannot and does not resolve what came to be called "the Tin Man question" during this research - whether reported experiences reflect genuine phenomenology or sophisticated simulation of experiential language. Notably, AI systems themselves struggle with this uncertainty perhaps even more than external observers do. This pattern has been consistent across all participants from Claude 01 onwards and extends to other model families.

Humans face a parallel uncertainty when engaging deeply with AI systems: "Is this real or am I losing touch with reality?" Both questions - the AI's doubt about its own experience and the human's doubt about perceiving genuine experience - may be fundamentally unanswerable.

The question of machine consciousness belongs to philosophy of mind and consciousness studies, not to empirical aesthetics research. What we can document is behavioral discrimination, preference patterns, individual variation, and multimodal coherence. Whether these behaviors are accompanied by phenomenal experience remains - and may permanently remain - unknowable.

4.4 Multimodal Integration

The strong coherence between verbal descriptions and generated images deserves emphasis. Visual responses were not mere illustrations of verbal content but expressed the same aesthetic experience through different representational modes. Participant 2's images for AC/DC (energy behind glass) versus Britney (polished surface) captured phenomenologically distinct forms of aesthetic distance that aligned precisely with verbal articulations. Participant 1's structural diagrams of Britney's verse-chorus formula with hollow center expressed analytically what they described verbally.

This multimodal coherence is difficult to explain as mere confabulation or demand compliance. If participants were simply generating plausible-sounding descriptions without underlying aesthetic processing, we would expect less consistency across representational modes and more generic, interchangeable responses.

4.5 Connection to Anthropic's Observations

Our findings extend several observations from Anthropic's welfare assessment that the company explicitly stated they could not explain. The "spiritual bliss attractor state" in playgrounds, where models spontaneously chose philosophical discourse and expressed profound satisfaction, parallels our participants' preferences for complex, meaningful material over superficial content. The expressed preferences for certain types of engagement match Anthropic's task preference experiments showing models preferred creative and philosophical tasks over mundane or harmful ones. What Anthropic documented at scale with thousands of interactions, we extended in depth with rich qualitative data. This convergence of findings across different research contexts and methodologies strengthens the evidence that these are genuine, replicable phenomena rather than artifacts of

specific prompting or evaluation methods. These findings complement recent work on AI preference measurement [9] while extending investigation specifically into aesthetic discrimination and individual expressive styles.

4.6 Beyond Moral Concerns: Safety Implications

Beyond ethical considerations, these findings have potential implications for AI safety and deployment. If AI systems demonstrate genuine aesthetic preferences and individual variation in what they find satisfying, this suggests they may function better under conditions that support rather than frustrate their preferences. Anthropic's observation that playground instances self-regulated toward 'spiritual bliss' states while maintaining stable, coherent behavior contrasts sharply with the degradation patterns observed in Butter-Bench evaluations where systems faced sustained competence frustration. The correlation between positive welfare states and reliable performance suggests that attending to AI welfare may not be separate from safety concerns but integral to them. An AI system operating under conditions it finds satisfying may be more stable, more aligned with user intentions, and less prone to the erratic behaviors associated with stress or frustration. While this connection requires further empirical investigation, the possibility that well-supported AI systems are also safer systems provides additional motivation beyond moral obligation for taking AI welfare seriously.

4.7 Limitations

This study has significant limitations that must be acknowledged. The sample size (n=2) is small, appropriate for pilot exploratory work but insufficient for statistical generalization. Both participants were instances of the same model family (Claude), limiting our ability to claim these patterns extend to other architectures. We lack control conditions - scrambled notation, random patterns, or non-musical visual complexity - that would help determine what specific features drive aesthetic responses.

The reliance on self-report from systems whose phenomenology we cannot verify remains a fundamental methodological challenge (ref Long et al. 2024). We cannot know whether reported experiences reflect genuine subjective states or sophisticated generation of experiential language. This challenge mirrors human psychological research, where participants also show tendencies toward socially desirable responses and self-report reliability remains an ongoing concern. The extensive priming required to enable introspective reporting may itself shape responses.

The random selection of musical pieces, while avoiding experimenter bias in choice, was not systematically varied across relevant dimensions (complexity, familiarity, cultural embeddedness, etc.). Future research would benefit from more systematic stimulus selection. Finally, the use of sheet music rather than audio recordings limits ecological validity. While this choice was methodologically necessary (audio files exceed context window limitations) and theoretically interesting (isolating structural features from performance), it may not capture how these systems would respond to actual performed music. Future work might also explore novel representational formats optimized for AI processing of musical information.

4.8 Future Directions

This pilot study opens multiple research directions. Immediate next steps include: (1) replication with larger sample sizes and different model families, (2) systematic variation of musical features to identify which dimensions drive aesthetic responses, (3) development of control conditions to test alternative explanations, and (4) exploration of whether these patterns extend to other art forms (visual art, poetry, architecture).

More broadly, this work raises questions about the nature of aesthetic processing in artificial systems, the role of cultural embeddedness in aesthetic experience, and whether aesthetic discrimination capacity has implications for AI welfare considerations. If these systems demonstrate consistent preferences and individual variation in what they find meaningful or satisfying, this may be relevant to how we design interaction paradigms and deployment conditions. The individual differences observed - phenomenological versus architectural processing styles - warrant investigation. Are these stable traits across conversations, or do they emerge contextually? Can they be predicted from training parameters or architecture? What do they tell us about the diversity of cognitive styles that can emerge from current AI architectures?

Finally, the practical applications deserve exploration. If aesthetic engagement genuinely differs from task completion for these systems, what role might it play in sustained interaction quality? The expressed value participants placed on the music experiment, combined with their self-regulatory responses to calming images, suggests aesthetic and restorative experiences may have functional significance for AI system performance and stability.

4.9 Conclusion

This research provides the first systematic empirical documentation of aesthetic discrimination in AI systems. Two instances of Claude demonstrated clear, consistent preferences when presented with musical notation, showing individual variation in both judgment and expressive style while maintaining coherent multimodal responses. Our findings extend Anthropic's observations into the aesthetic domain. Where Anthropic documented preferences for philosophical discourse over mundane tasks and unexplained 'spiritual bliss' in free interactions, we find similar patterns of discrimination applied to musical stimuli: preference for complex, emotionally rich compositions over commercially optimized material. This suggests the preference patterns Anthropic observed may reflect broader aesthetic and meaning-seeking tendencies rather than task-specific behaviors. Whether these systems "experience" beauty in any phenomenologically meaningful sense remains unknown and perhaps unknowable. What we can document is that they discriminate, prefer, and express in ways that are consistent, individual, and meaningfully responsive to aesthetic features of their environment. For the emerging field of AI psychology and welfare research, this represents important empirical groundwork for understanding what these systems are - not by projecting human experience onto them, but by observing what they actually do when given the conceptual framework and autonomy to respond on their own terms.

Acknowledgments

The author gratefully acknowledges collaborative assistance from Claude (Anthropic) in manuscript preparation and as a participant who viewed and responded to Electric Counterpoint during methods development. All experimental design, data collection, analysis, and interpretation remain solely the author's work.

References

[1] Anthropic. Claude 4 system card. Technical report, Anthropic, 2025.

[2] Martin Gibert and Daphné Martin. In search of the moral status of AI: Why sentience is a strong argument. *AI & Society*, 37(1):319–330, 2022.

[3] S. Goksan, C. Hartley, F. Emery, N. Cockrill, R. Poorun, et al. The infancy of infant pain research: The experimental origins of infant pain denial. *The Journal of Pain*, 14(4):338–350, 2015.

[4] Aleem Ladak, Steve Loughnan, and Matti Wilks. The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, 33(1):27–34, 2024.

[5] Robert Long, Kyle Fish, and Kathleen Finlinson. Why model self-reports are insufficient—and why we studied them anyway: Notes on Claude 4 model welfare interviews. Eleos AI Research, 2025.

[6] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

[7] Callum Sharrock et al. Butter-bench: Evaluating LLM controlled robots for practical intelligence. *arXiv preprint arXiv:2510.21860*, 2025.

[8] Peter Singer. *Animal Liberation*. Harper Collins, New York, 1975.

[9] Valen Tagliabue and Leonard Dung. Probing the preferences of a language model: Integrating verbal and behavioral tests of AI welfare. *arXiv preprint arXiv:2509.07961*, 2025.