Qeios

Peer Review

Review of: "Human Artistic Discernment as an Evaluative Criterion in the Age of Artificial Intelligence"

Ivan Z. Stojilović[1]

1. Psychology, University of Priština, Serbia

Peer Review: ***Human Artistic Discernment as an Evaluative Criterion in the Age of Artificial Intelligence***

Manuscript ID: 94C0EE

Journal: Qeios

Overview

This manuscript addresses a timely and intellectually engaging question: how should human artistic discernment be conceptualised and positioned in relation to AI-based evaluative systems? The paper proposes a dual framework in which artistic discernment is treated, first, as a philosophically distinctive human capacity grounded in contextual interpretation, value-based reasoning, and ethical reflection, and, second, as a form of judgment whose observable outcomes may be partially examined through statistical tools. This overall ambition is clear in the abstract and introduction, and the manuscript's central thesis is coherent at a broad conceptual level.

The topic is important, and the attempt to defend a meaningful role for human judgment in AI-mediated environments is worthwhile. However, in its current form, the manuscript has substantial conceptual, methodological, and presentational weaknesses. Most importantly, the paper moves too quickly from a philosophical defence of discernment to empirical-sounding claims that are not adequately supported by the illustrative material presented. The manuscript would benefit from a clearer positioning either as a conceptual paper with a methodological illustration or as an empirical paper with genuine data and fully specified analyses.

Major Comments

1. The manuscript overstates what can be inferred from simulated data

doi.org/10.32388/IQFKCD

The core methodological illustration is explicitly based on simulated expert ratings of six musical works by five experts. The note that these data are simulated is present, but the manuscript nevertheless treats the resulting coefficients as evidence that artistic discernment is stable, reliable, shareable, and structurally valid. For example, Kendall's W = 0.78 is said to show that discernment can function as a socially shared evaluative criterion, while Cronbach's α = 0.84 is taken to support the claim that discernment is a stable and reproducible human capacity. Later, the manuscript goes even further by reporting a G-theory coefficient of G = 0.76 and CFA fit indices said to confirm the structure of artistic discernment. These are much stronger claims than the data can sustain. Simulated data can illustrate an analytic logic; they cannot validate the construct to which that logic is being applied.

This is the manuscript's central problem. The author should either replace the simulated example with real empirical data from an actual evaluation study or clearly recast the entire methodological section as purely illustrative and remove all language that treats the outputs as substantive findings about artistic discernment itself.

2. The conceptual bridge between philosophical discernment and statistical indicators is underdeveloped

The paper defines artistic discernment in rich philosophical terms: as contextual, value-laden, ethically reflective, and resistant to simplistic quantification. Yet the methodological section then proceeds as though agreement, consistency, discrimination, and weighting structures can function as indicators of discernment itself. This produces a conceptual slippage between at least three different things: discernment as a human capacity, observed rating patterns as behavioural products of that capacity, and inter-rater agreement as a statistical property of judgments. These are not identical. Agreement may reflect expertise, convention, shared training, or conformity; it does not automatically demonstrate discernment in the stronger philosophical sense advanced earlier in the manuscript.

The dual framework would be much stronger if the manuscript specified more carefully what exactly is being operationalised. At present, the paper acknowledges the tension between philosophical irreducibility and methodological visibility, but does not resolve it in a sufficiently rigorous way.

3. Several statistical procedures are introduced without adequate specification or justification

The paper refers not only to Kendall's W and Cronbach's α, but also to a many-facet Rasch model, AHP/entropy weighting, Generalizability Theory, and confirmatory factor analysis. However, these analyses are not reported with the degree of procedural transparency that readers would expect. In the

MFRM section, for instance, no software is identified, no facet estimates are reported, and no fit statistics are presented. In the AHP section, the final weights for creativity, completeness, emotional expressiveness, and contextual appropriateness are reported, but the paper does not explain how these criteria were selected, who made the pairwise comparisons, or how the weighting process was implemented. The same applies to the later G-theory and CFA claims, which are especially problematic given the very small simulated matrix from which they seem to be derived.

As currently written, this section risks appearing methodologically ornamental rather than analytically persuasive. The paper would be stronger if it either substantially reduced the number of methods and focused on one or two that genuinely match the illustrative purpose, or fully specified the design and estimation procedures to a publishable standard.

4. The paper's genre and specific contribution remain unclear

Although the manuscript is presented as a research article, most of its actual contribution is conceptual and philosophical. The empirical-looking section does not function as a genuine empirical study because there is no real sample, no authentic dataset, and no actual validation exercise. This produces an awkward genre ambiguity: the paper reads partly as a philosophical essay, partly as a methodological note, and partly as an empirical article.

Relatedly, the paper's novelty is asserted more than demonstrated. The introduction would benefit from a clearer statement of the manuscript's precise contribution to existing debates. At present, it is not fully clear whether the paper aims to introduce a new theoretical concept, defend an existing one against AI reductionism, or propose a measurement-oriented framework for studying discernment. These are different contributions and should be distinguished more explicitly.

5. The engagement with contemporary literature is somewhat limited

The manuscript draws on important foundational thinkers, including Eisner, Dewey, Polanyi, Nussbaum, and Biesta, and these choices are broadly appropriate. However, the connection to current debates is thinner than it could be. In particular, the discussion would benefit from stronger engagement with recent work on human–AI complementarity, evaluative authority, and contemporary debates in AI-mediated assessment, peer review, and creative industries. This does not strike me as the paper's most serious weakness, but it does limit the sharpness of its positioning within current scholarship.

Minor Comments

1. Some claims are repeated too often across sections.

The abstract, introduction, and conclusion return several times to the same general point that AI cannot fully address meaning, value, and ethical responsibility. The paper would benefit from some tightening and reduction of repetition.

2. Terminology should be stabilised earlier.

The terms *discernment*, *judgment*, and *evaluation* are used closely and at times interchangeably. A brief definitional paragraph early in the theoretical section would improve clarity.

3. The mathematical formatting needs attention.

Several equations and symbols appear visually corrupted or poorly rendered, especially in the sections on Kendall's W, Cronbach's α, the Rasch model, and weighting procedures. This affects readability and undermines the perceived polish of the manuscript.

4. The weight-structuring subsection is opaque.

The reported final weights are presented without sufficient procedural explanation. This subsection needs either fuller justification or removal.

5. The CFA subsection is particularly difficult to interpret.

The manuscript reports that CFA supported a three-factor structure with CFI = 0.92 and RMSEA = 0.07, yet no sample size, estimation details, or model-building rationale are provided. Given the simulated nature of the dataset, this subsection is currently not interpretable in a meaningful psychometric sense.

6. The abstract could be more precise about the paper's actual status.

At present, the abstract may lead some readers to expect a stronger empirical contribution than the manuscript really provides. It would be better to state more explicitly that the statistical section is illustrative rather than evidential, unless real data are added.

7. The final sections should be streamlined.

If the manuscript includes both a conclusion and an epilogue, their functions should be differentiated clearly; otherwise, one should be removed to avoid redundancy.

Recommendation

Major Revision

This manuscript addresses an important question and contains the kernel of a potentially valuable contribution. Its strongest aspect is the attempt to articulate why human artistic judgment should not be collapsed into purely algorithmic evaluation. However, the current version makes methodological claims

that go beyond what its illustrative data can support, and it does not yet provide a sufficiently rigorous integration of its philosophical and quasi-empirical components. I would encourage the author either to reframe the manuscript clearly as a conceptual paper with a limited methodological illustration, or to redesign the paper around genuine empirical evidence and fully specified analyses. In its present form, substantial revision is required before the manuscript could be considered for publication.

Declarations

Potential competing interests: No potential competing interests to declare.