AI AS CRITIC: ETHICAL DILEMMAS IN MACHINE-BASED LITERARY INTERPRETATION

Swapnil Subhash Patil
Department of English, Shri Seth Murlidharji Mansingka Arts, Science and Commerce College Pachora, Dist. Jalgaon
swapnilspatil002@gmail.com

Dr. Dinesh Prakash Patil
Professor and HoD English, Appasaheb R.B. Garud Arts, Commerce and Science College Shendurni Tal. Jamner, Dist. Jalgaon
dr.di.puresearch@gmail.com

Abstract
The rapid integration of Artificial Intelligence (AI) into literary studies has generated both enthusiasm and unease within the academy. AI-driven tools in natural language processing and machine learning now allow large-scale textual analysis, sentiment mapping, and intertextual pattern recognition, yet they simultaneously challenge the foundations of humanistic inquiry. This paper, "AI as Critic: Ethical Dilemmas in Machine-Based Literary Interpretation", interrogates whether machines can meaningfully interpret literature, particularly metaphor, symbolism, irony, and cultural nuance—features resistant to reduction into data. The study situates AI within the broader field of digital humanities, acknowledging its contributions to distant reading, comparative literature, and archival research, while also examining its limitations in addressing historical context, cultural memory, and the affective dimensions of texts. Ethical dilemmas such as algorithmic bias, the displacement of human creativity, and the uncertainty of authorship when interpretive agency is shared with machines are central to the discussion. Drawing on poststructuralist theory, reader-response criticism, and debates on authorship, the paper argues that while AI may enrich scholarship by offering new methodological lenses, it cannot supplant the interpretive authority of human critics. Instead, an ethically responsible integration of AI is necessary to preserve the imaginative and cultural richness of literature.

Key Words: *Artificial Intelligence in Literary Studies; Ethical Criticism; Digital Humanities; Algorithmic Bias; Hermeneutics and Interpretation; Authorship and Authority; Machine-Based Criticism.*

1. Introduction

The integration of Artificial Intelligence (AI) into the humanities has initiated one of the most significant debates in contemporary literary scholarship. While AI-driven tools such as natural language processing and machine learning have opened unprecedented avenues for analyzing texts, they have also unsettled the very foundations of interpretation. Tasks such as sentiment mapping, topic modeling, and intertextual comparison can now be performed at speeds and scales unimaginable to traditional scholarship. These developments have generated enthusiasm for new methodologies but have also provoked unease, particularly concerning the ethical implications of entrusting machines with interpretive labor. Literature, unlike other forms of data, embodies cultural memory, metaphorical richness, irony, and symbolic complexity—features that resist easy reduction into computational patterns. The possibility of AI functioning as a "critic" therefore demands critical scrutiny.

At the center of this debate lies a pressing question: can machines meaningfully interpret literature? While algorithms are capable of identifying stylistic patterns or recurring motifs, they often falter when confronted with ambiguity, metaphor, or cultural nuance. Moretti's idea of "distant reading" demonstrated the value of scale-based analysis, yet it also shifted critical practice away from the deep interpretive engagement traditionally associated with the humanities (Moretti 48). The current use of AI intensifies this tension by raising questions about authorship, interpretive agency, and critical responsibility. If machines are credited with interpretive authority, the role of the human scholar risks displacement, and the criteria of literary criticism may be reduced to patterns rather than meanings.

Equally important are the ethical dilemmas that accompany the adoption of AI in literary studies. Issues such as algorithmic bias, lack of transparency, and the uncertainty of authorship have profound consequences for how texts are read and taught. The training data of AI models frequently reflects dominant cultural narratives, potentially silencing marginalized voices and reinforcing preexisting hierarchies. Furthermore, the opacity of machine learning systems challenges the humanist expectation that interpretations be accountable, transparent, and rooted in textual evidence. As Cathy O'Neil observes, algorithms are never neutral but "opinions embedded in code"

(O'Neil 21), and their extension into the domain of criticism magnifies these risks.

2. Literature Review

1. Digital Humanities and Computational Criticism

The digital humanities first opened the door for large-scale approaches to literature. Franco Moretti's idea of "distant reading" shifted attention from a few canonical texts to broad literary systems (48). Matthew Jockers expanded this in *Macroanalysis*, showing how themes and styles could be tracked across thousands of works (32). Ted Underwood's *Distant Horizons* also demonstrated the use of computational evidence to understand long-term literary change (14). These studies revealed the strengths of large-scale analysis but also raised questions about what might be lost when interpretation is reduced to patterns and numbers.

2. Debates on Authorship and Interpretation

Questions of authorship and meaning remain central to evaluating AI as a critic. Roland Barthes's "death of the Author" emphasized the role of the reader in shaping meaning (148), while Michel Foucault's "author function" highlighted the cultural authority attached to authorship (112). Stanley Fish added that interpretation itself is shaped by "interpretive communities" (15). These theories remind us that criticism is not neutral and that AI-generated readings cannot escape questions of accountability or cultural positioning.

3. Ethical Concerns of AI in Humanities

Recent scholarship warns of the risks of adopting AI without caution. Cathy O'Neil describes algorithms as "opinions embedded in code" (21), while Safiya Umoja Noble shows how biased data can reproduce social inequalities (85). Luciano Floridi also stresses the dangers of opaque "black box" systems that limit accountability (19). Together, these works argue that AI in the humanities should be used critically, with awareness of bias, transparency, and ethical responsibility.

3. AI and the Transformation of Literary Criticism

The entry of Artificial Intelligence into literary studies marks a turning point in the evolution of critical practice. Earlier developments in digital humanities had already expanded the field by introducing computational tools for large-scale analysis. Scholars such as Franco Moretti, Matthew Jockers, and Ted Underwood demonstrated that patterns of genre, theme, and stylistic change could be identified through the study of massive corpora rather than a few canonical texts (Moretti 48; Jockers 32; Underwood 14). AI builds upon this foundation, yet it does more than extend the reach of statistical models. Through natural language processing and machine learning, it offers interpretive gestures that appear to mimic human reading, from mapping the emotional trajectory of a novel to generating thematic paraphrases of poetry.

These possibilities have altered the perception of what constitutes literary criticism. Traditionally, the critic's role has been defined by close reading, cultural contextualization, and interpretive judgment. The arrival of AI blurs these boundaries by producing outputs that resemble interpretive claims. For example, an algorithm may identify alienation as a central theme in a modernist poem, or frame a Woolf passage as a meditation on memory. Such outputs do not emerge from conscious understanding but from correlations in data. The resemblance to criticism, however, forces scholars to ask whether machines are engaged in interpretation or in the simulation of interpretive practices.

The enthusiasm surrounding AI stems largely from its ability to process texts at scales unavailable to individual scholars. Vast archives of novels, periodicals, or letters can be mined in minutes, directing attention to overlooked texts and enabling new historical perspectives. This capacity supports Moretti's call to move beyond the "small canon" and explore the systemic tendencies of literary history (Moretti 54). Properly applied, such tools may help diversify syllabi, recover neglected voices, and expose structures of cultural circulation invisible to close reading alone.

At the same time, limitations quickly become evident. AI systems frequently misinterpret irony, satire, and metaphor—features central to literary language. A sentiment analysis model might misread Swift's *A Modest Proposal* as an endorsement of utilitarian economics rather than a satire, while a metaphor-detection tool may struggle to recognize culturally specific figures of speech. These errors highlight a fundamental divide between statistical recognition and interpretive judgment. As N. Katherine Hayles notes, meaning in literature is not reducible to information transfer but is situated in embodied, cultural, and historical contexts (Hayles 28).

Equally transformative are the questions AI raises about authorship and interpretive authority. Roland Barthes once declared the "death of the author," emphasizing the role of the reader in creating meaning (Barthes 148). Yet when an AI generates a reading, responsibility becomes diffused: does authorship belong to the designers of the system, the human operator who prompted it, or the corpus

that shaped its responses? Michel Foucault's concern with the "author function" as a cultural category acquires renewed urgency in an age when the "critic" may be partly or wholly machinic (Foucault 112).

4. Core Ethical Dilemmas in Machine-Based Interpretation

The integration of Artificial Intelligence into literary studies has not only expanded methodological possibilities but also surfaced pressing ethical concerns. Unlike earlier tools in digital humanities, which largely provided quantitative insights, AI appears to offer interpretive claims, creating tensions around authority, responsibility, and academic integrity. These tensions manifest in several interconnected dilemmas—algorithmic bias, hermeneutic opacity, figurative misreadings, authorship and originality, and the risks to pedagogy.

4.1 Algorithmic Bias and Canon Formation

One of the most significant ethical challenges concerns bias in training data. AI systems are trained on vast corpora, which frequently overrepresent dominant cultural traditions while marginalizing voices from non-Western or historically excluded communities. This imbalance risks reinforcing the very hierarchies that literary criticism often seeks to challenge. As Safiya Umoja Noble argues in *Algorithms of Oppression*, digital systems are not neutral: "They are reflective of the social, cultural, and political norms of the people who create them" (85). In literary contexts, such bias means that AI-driven interpretations may privilege mainstream traditions while overlooking marginalized texts, thus narrowing rather than expanding the canon.

4.2 Hermeneutic Opacity

Another dilemma is the opacity of AI-generated interpretations. Machine learning models, particularly neural networks, operate through complex layers of computation that are often inaccessible even to their designers. While a critic is expected to defend interpretive claims through evidence and reasoning, an AI output offers little explanation for its conclusions. This raises fundamental questions about the standards of evidence in literary studies. Luciano Floridi warns that "black-box systems challenge the very notion of accountability" in scholarly practice (Floridi 19). When an AI identifies "alienation" as a theme in a novel, on what grounds does it make this claim? Without transparency, such assertions cannot be meaningfully debated, which undermines the dialogic and argumentative nature of criticism. Literary interpretation thrives on disagreement and evidence-based reasoning; AI, by contrast, risks presenting opaque conclusions that resist scholarly scrutiny.

4.3 Figurative Misreadings: Irony, Satire, and Metaphor

The limits of AI become especially visible in the domain of figurative language. Literature relies on irony, metaphor, and cultural allusion to generate meaning, but algorithms often misinterpret or ignore these features. A sentiment analysis system may label Swift's *A Modest Proposal* as a text endorsing rational economic policy rather than recognizing its satirical critique of English colonialism. Similarly, metaphor-detection tools trained primarily on Western corpora may fail to grasp non-Western symbolic traditions.

As Hayles reminds us, "literary meaning emerges through the interplay of textual patterns and cultural contexts" (29). When AI overlooks context, it risks producing readings that are not simply inadequate but misleading. This dilemma underscores the irreplaceable role of human judgment in interpreting nuance and cultural specificity.

4.4 Authorship, Originality, and Responsibility

The rise of AI also unsettles debates about authorship. Roland Barthes declared that the "birth of the reader must be at the cost of the death of the Author" (148), emphasizing that meaning emerges in the act of reading. Yet with AI, meaning is no longer produced solely by human readers but also by machine outputs. This disperses responsibility: is an AI-generated interpretation authored by the programmer, the user who framed the prompt, or the corpus that shaped the model? Michel Foucault's notion of the "author function" as a system of classification is newly relevant in a world where interpretive authority may be partly machinic (112).

Questions of originality also arise in classrooms and publishing. If a student submits an essay generated partly by AI, does it constitute plagiarism, collaboration, or a new form of authorship? Without clear protocols for attribution, the integrity of scholarly practice is at risk. Cathy O'Neil's warning that algorithms are "opinions embedded in code" (21) reminds us that AI interpretations are not autonomous creations but products shaped by human and institutional decisions.

4.5 Risks to Pedagogy and Critical Thinking

Perhaps the most immediate ethical concern lies in pedagogy. Students increasingly turn to AI for summaries, interpretations, and even full essays. While such tools may support accessibility and comprehension, overreliance risks eroding the very skills that literary education is meant to cultivate:

close reading, argumentative writing, and critical engagement. If AI is treated as a substitute for analysis rather than as an aid, it threatens to hollow out the interpretive practices that form the foundation of the humanities.

Moreover, the use of AI in classrooms raises issues of disclosure and fairness. Should students be required to cite AI assistance in essays, much as they would cite a critical source? What guidelines should instructors adopt to distinguish between legitimate use and academic dishonesty? These questions are not ancillary but central to the ethical integration of AI into literary studies.

5. Responsible Integration of AI in Literary Studies

The dilemmas surrounding Artificial Intelligence in criticism do not imply that these technologies must be rejected. Rather, they call for frameworks that position AI as a supplementary tool, one that can broaden scholarly horizons without displacing human judgment. Responsible integration requires transparency, accountability, and a recognition of the limits of machine-based interpretation.

5.1 AI as Tool, Not Critic

At its best, AI can enhance scholarship by enabling new forms of discovery. Text-mining programs, for example, can reveal thematic or stylistic continuities across vast corpora, pointing researchers toward connections they might otherwise overlook. Such uses align with what Moretti envisioned as a "collective" project of literary history, in which computational tools assist scholars in mapping large-scale patterns (Moretti 56). Yet the temptation to treat machine outputs as interpretive authority must be resisted. As Ted Underwood cautions, "computational models are not substitutes for humanistic reasoning but provocations that direct us back to the texts themselves" (Underwood 22).

This distinction between assistance and authority is crucial. AI is most productive when it generates questions rather than answers, serving as a heuristic device that prompts further critical inquiry. For instance, an algorithm that detects recurring motifs across nineteenth-century novels should be understood as identifying potential sites of interpretation, not as delivering conclusive readings.

5.2 Transparency and Attribution

Responsible use also demands transparency in documenting the role of AI in scholarly work. Just as critics cite secondary sources, they should also disclose the prompts, models, and tools employed in their analyses. The Modern Language Association has recently emphasized that students and researchers must "acknowledge the use of generative AI when it contributes substantively to the ideas, wording, or structure of their work" (MLA 9th Handbook, sec. 5.97). Such disclosure not only maintains academic integrity but also allows others to evaluate the reliability of AI-assisted claims.

5.3 Guarding Against Bias and Exclusion

As discussed earlier, the biases inherent in training data pose significant risks to equitable scholarship. Responsible integration requires deliberate efforts to counteract these tendencies. One approach is to supplement mainstream corpora with marginalized texts, ensuring that computational analyses reflect a more inclusive literary history. Another is to pair AI-driven insights with critical traditions—postcolonial, feminist, or queer theory—that foreground questions of power and representation.

Safiya Umoja Noble's warning that algorithms often "reinforce oppressive social relations under a veneer of neutrality" (88) underscores the importance of this step. Literary studies must therefore not only critique algorithmic bias but also model practices that diversify and democratize the materials on which AI systems operate.

5.4 Pedagogical Protocols

In educational settings, AI can serve as both a resource and a challenge. When integrated thoughtfully, it can support learning by providing quick textual summaries, translation aids, or exploratory prompts that help students engage with difficult texts. However, these benefits must be balanced against the risk of intellectual dependency. To preserve the value of critical thinking, instructors should encourage students to compare AI outputs with their own interpretations, treating the machine as a conversation partner rather than an authority.

Practical guidelines can reinforce this balance. Students might be required to append any AI-generated material they consulted in an appendix, along with a short commentary evaluating its usefulness and limitations. Such practices transform AI into an object of critique rather than a shortcut, strengthening rather than weakening critical skills. As Stanley Fish once argued, "interpretive authority is not about arriving at the correct answer but about justifying one's reading in a community of argument" (Fish 15). AI can have a place in this community only if its role is transparent and subject to scrutiny.

5.5 Toward an Ethical Protocol

The responsible integration of AI ultimately calls for the development of explicit ethical protocols for "AI-assisted criticism." Such protocols might include six basic principles:

1. **Attribution**: Always disclose the use of AI, citing tools, versions, and prompts.
2. **Transparency**: Archive outputs alongside interpretive arguments for review.
3. **Bias Awareness**: Actively diversify corpora and interrogate exclusions.
4. **Alignment**: Use AI in ways consistent with the goals of literary criticism, avoiding overextension into interpretive authority.
5. **Pedagogical Safeguards**: Teach students to critically evaluate AI outputs.
6. **Accountability**: Treat all AI-assisted claims as provisional, requiring human verification.

By adopting such principles, scholars can harness the advantages of AI while safeguarding the values of humanistic inquiry. The integration of AI into criticism need not undermine interpretation; rather, it can prompt renewed reflection on the ethical and methodological commitments of the humanities.

6. Conclusion

The incorporation of Artificial Intelligence into literary studies has generated both opportunities and challenges, forcing the field to confront questions that extend beyond methodology to the ethical and philosophical foundations of interpretation. On the one hand, AI provides tools of remarkable scope: it can process archives at a scale beyond the reach of individual scholars, highlight patterns invisible to close reading, and democratize access to literary history. Properly used, these capabilities enrich the humanities by opening new directions of inquiry and inviting greater inclusivity in the canon.

On the other hand, the dilemmas that accompany AI cannot be dismissed as technical limitations. Algorithmic bias, hermeneutic opacity, figurative misreadings, and uncertainties of authorship pose profound risks to the integrity of criticism. As Safiya Umoja Noble reminds us, algorithms are never neutral but "embed the values of those who design and deploy them" (88). Left unexamined, these values may reproduce exclusions and distortions that the humanities have long struggled to resist.

What emerges from this study is a clear imperative: AI must remain a tool in the service of criticism, not a replacement for it. Interpretation, unlike information processing, requires accountability, contextual awareness, and the willingness to defend claims within a community of readers. As Ted Underwood observes, computational models should be understood not as conclusions but as "provocations that direct us back to the texts themselves" (22). This recognition safeguards the interpretive depth that defines the humanities while acknowledging the potential of new technologies.

The future of literary studies therefore depends on balance. Scholars must engage with AI critically, embracing its capacity to extend research while resisting the temptation to attribute to it interpretive authsority it cannot legitimately hold. This balance is achievable only through transparent protocols of attribution, ethical awareness of bias, and pedagogical practices that foreground human judgment. Instructors and researchers alike have a responsibility to model these practices, ensuring that students view AI not as a substitute for critical thought but as an object of analysis and reflection.

Ultimately, the encounter between AI and literary criticism should be seen not as a crisis but as an invitation. It compels the humanities to reexamine their methods, reaffirm their values, and articulate the distinctiveness of interpretation in an era increasingly dominated by data. The question is not whether machines can read but how human critics will respond to their presence. The answer, as this paper has argued, lies in embracing AI as a catalyst for dialogue while preserving the interpretive agency that makes literature—and its study—indispensable.

References

1. Barthes, R. (1977). *Image, music, text* (S. Heath, Trans.). Hill and Wang.
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922
3. Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities*. Harvard University Press.
4. Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology, 33*(1), 1–3. https://doi.org/10.1007/s13347-019-00379-8
5. Foucault, M. (1977). What is an author? In D. F. Bouchard (Ed.), *Language, counter-memory, practice: Selected essays and interviews* (D. F. Bouchard & S. Simon, Trans., pp. 113–138). Cornell University Press.
6. Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. University of Chicago Press.
7. Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

8. Modern Language Association of America. (2021). *MLA handbook* (9th ed.). Modern Language Association of America.
9. Moretti, F. (2013). *Distant reading*. Verso.
10. Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
11. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
12. Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. University of Chicago Press.
13. Modern Language Association. (2023). Guidelines for ethical use of artificial intelligence in research and teaching. *MLA Commons*. https://mla.hcommons.org/ethical-ai-guidelines/
14. Zalta, E. N. (Ed.). (2023). Artificial intelligence and ethics. In *Stanford encyclopedia of philosophy* (Spring 2023 ed.). Stanford University. https://plato.stanford.edu/entries/ethics-ai/