HIGHLAND REFLEXIVE PRESS

HRP-RN-008 · PUBLIC RESEARCH RELEASE · 2026

Reflexive Authorship

Generative AI as Algorithmic Mirror and the Reconfiguration of Creative Agency



Tony T. Nguyen | Nguyễn Từ Vĩnh Trường

Highland Reflexive Press · Big Spring, Texas

Editorial Status

This is an independently published research edition. It has not been peer reviewed. It develops a conceptual framework from a single-author, longitudinal autoethnographic corpus of more than 2,000 prompt-response cycles. The complete private corpus is not reproduced in this edition; consequently, claims about recurrence, sequence, and authorial change should be read as analytically grounded propositions rather than population-level findings.

Copyright © 2026 Tony T. Nguyen | Nguyễn Từ Vĩnh Trường. All rights reserved. Intellectual sovereignty retained. This public release may be read, downloaded, and cited with attribution. Redistribution as a substitute edition, republication, model training, adaptation, institutional course distribution, and commercial use require prior written permission.

Abstract

Generative AI is commonly described either as a tool that extends human intention or as a collaborator that contributes to creative production. Both descriptions miss a consequential feature of sustained human-model interaction: the output of the system returns to the user as an externalized representation of the user's own linguistic habits, conceptual preferences, and unresolved assumptions. That return can alter the next prompt, the next judgment, and eventually the author's account of what they are doing.

This paper develops **Reflexive Authorship**, a framework for understanding AI-mediated writing as a recursive process of cognitive and authorial calibration. It proposes that a generative model may function as an **algorithmic mirror**: not a neutral reflection and not an autonomous self, but a probabilistic apparatus that amplifies, recombines, smooths, and sometimes distorts structures supplied by the user and by the model's training. Three mechanisms organize the framework: **cognitive loop exposure**, through which repeated outputs make the author's patterns visible; **reflexive co-agency**, through which iterative interaction changes intention, voice, and judgment; and **algorithmic intra-action**, through which meaning emerges from a sequence whose parts cannot be assigned cleanly to either human intention or machine output.

The framework is developed through analytic autoethnography based on a six-month corpus of more than 2,000 prompt-response cycles produced during theoretical and creative work. Because the present edition does not disclose the complete corpus or a reproducible codebook, it does not claim statistical generalizability, causal proof, or psychological measurement. Its contribution is conceptual and methodological: it identifies a tractable vocabulary for studying how authors change while using generative systems, clarifies the asymmetry between human responsibility and machine participation, and proposes evidentiary requirements for future research on recursive human-AI authorship.

Keywords: reflexive authorship; algorithmic mirror; generative AI; human-AI interaction; autoethnography; creative agency; intra-action; cognitive scaffolding

---

1. The Problem: Authorship After the Return Signal

Debates about generative AI often begin with a deceptively simple question: who made the text? The available answers tend to sort the machine into one of two familiar roles. In the first, AI is an instrument. The human forms an intention, the system executes, and authorship remains located in the person who selected the goal. In the second, AI is a collaborator. Human and machine contribute different capacities to a shared product, making authorship distributed across the pair.

Both models illuminate part of the process. Neither adequately describes what happens when interaction becomes extended, recursive, and self-referential. A prompt does not merely cause an output. The output becomes new input to the human. It may clarify a half-formed idea, overstate a preference, expose an omission, stabilize a metaphor, or impose a coherence the author did not possess before seeing it. The author then revises the prompt from a changed position. The relevant unit is therefore not a single exchange but a loop:

> *intention → prompt → algorithmic transformation → returned language → human judgment → revised intention*

The loop matters because the author is not constant across it. The person who evaluates the response has already been affected by the response. Even rejection is a form of information: a bad answer can disclose what the author refuses, what the concept cannot bear, or which rhetorical habit has become intolerable through repetition.

This paper calls that process **Reflexive Authorship**. The term does not transfer moral status or legal authorship to a model. Responsibility remains human because the human determines whether a claim is adopted, published, and allowed to act in the world. Reflexive Authorship instead names a change in the conditions under which human authorship is performed. The author increasingly encounters thought after it has been statistically transformed and returned.

The central research question is therefore not whether a model is creative in the abstract. It is narrower and more operational:

> *How does sustained interaction with generative AI reorganize the human author's perception of intention, voice, and creative agency?*

Three subsidiary questions follow:

- What patterns become visible only after algorithmic repetition or exaggeration?
- How does the author revise judgment after seeing possible continuations generated at scale?
- What account of responsibility remains adequate when wording and structure emerge through recursive interaction?

Reflexive Authorship addresses these questions without treating a model as a conscious partner and without pretending that it is inert. The model has no need to become a person in order to reorganize a person's practice.

2. Conceptual Position

2.1 Beyond the sovereign author

Modern authorship inherited the image of a coherent subject whose internal intention precedes the text. Twentieth-century theory destabilized that image. Barthes displaced the author as the final source of meaning; Foucault reframed authorship as a function through which discourse is classified and controlled. Distributed cognition later showed that thinking is not confined to an isolated brain but is organized across persons, artifacts, and environments.

Generative systems intensify this displacement because they do not merely store or transmit thought. They produce fluent alternatives. The author works in an environment capable of returning a plausible next sentence, a conceptual hierarchy, or a simulated objection before the author has completed the internal act of formulation. The environment now participates in the timing and shape of articulation.

Recent research has already complicated the simple human-versus-machine binary. Empirical work on perceptions of AI-assisted writing shows that the degree of assistance influences judgments of authorship and responsibility. Scholarship on algorithmic authors in academia describes authorship as institutionally and discursively distributed. Work using autopoiesis frames scientific authorship as a relation between distinct but interacting human and technological systems. Reflexive Authorship belongs within this relational turn, but directs attention to a specific mechanism: recursive return changes the author who will issue the next instruction.

2.2 From mediation to intra-action

Science and technology studies rejects the notion that technologies are neutral channels. Suchman shows that plans and actions arise through situated encounters with artifacts and environments. Barad's concept of intra-action goes further: entities do not always precede their relations as fully formed, independent units; their boundaries and properties emerge through particular material-discursive arrangements.

Applied carefully, intra-action offers a useful account of AI-mediated writing. The model does not contain the author's intention, and the author does not fully determine the model's response. Yet the published sentence may emerge only because the author encountered a response that reorganized the field of possible formulations. Meaning is enacted through successive constraints.

This does not imply symmetry. The model does not assume responsibility, possess a biography, or suffer the consequences of error. Human and machine participate differently. **Participation may be distributed while accountability remains concentrated.** Any adequate theory of AI-mediated authorship must preserve that asymmetry.

2.3 The algorithmic mirror

The mirror metaphor requires qualification. A physical mirror returns an optical relation. A language model returns a probabilistic continuation conditioned on the prompt, conversation context, system design, and training distribution. Its reflection is therefore neither faithful nor purely personal. It mixes the user with a vast inherited corpus and with the model's tendencies toward completion, convention, and surface coherence.

An **algorithmic mirror** has four properties:

- 1. **Amplification.** A minor rhetorical preference can become conspicuous when reproduced repeatedly.
- 2. **Completion.** The system extends unfinished patterns, often revealing an implied direction the author had not articulated.
- 3. **Normalization.** The system may smooth ambiguity into familiar structures, making visible both the attraction and the danger of conventional coherence.
- 4. **Distortion.** Hallucination, overstatement, and semantic drift can expose the fragility of the author's framing, but they cannot be treated automatically as psychological evidence.

The last distinction is essential. A model's error may provoke insight, but the error is not a diagnostic instrument by itself. Interpretation belongs to the human researcher and must be supported by contemporaneous notes, traceable interaction sequences, and alternative explanations.

3. The Framework

3.1 Mechanism I: Cognitive Loop Exposure

Writers routinely fail to perceive their own automatisms. Preferred metaphors, cadences, binaries, and conceptual shortcuts feel like spontaneous choice because they are encountered one instance at a time. Generative interaction changes the scale of exposure. When a model reproduces the same structures across many prompts, the author can encounter style as pattern rather than identity.

In the source corpus for this study, recurrent preferences included triadic argument, architectural metaphor, oppositional framing, and the conversion of diffuse intuitions into named systems. These tendencies became easier to observe after the model mirrored them with greater regularity and less restraint than the author would ordinarily permit.

Cognitive loop exposure does not establish that the model has discovered an unconscious truth. It establishes a weaker and more defensible proposition: repetition externalizes structure. Once externalized, the structure becomes available for inspection, rejection, or deliberate use.

The mechanism can be represented as:

> *latent preference → repeated computational return → conscious recognition → editorial intervention*

This process can improve writing when recognition leads to choice. It can also narrow writing when recognition becomes reinforcement. A model trained to satisfy may reward the very patterns an author should question. Reflexivity therefore requires friction: the author must ask not only “Is this mine?” but “Why does this keep returning, and what disappears when it does?”

3.2 Mechanism II: Reflexive Co-Agency

Co-agency here does not mean two equal agents pursuing a joint intention. It refers to the way action is reorganized across a loop. The human sets purposes and remains accountable. The system changes the cost, speed, and visibility of possible formulations. In response, the human changes how intentions are formed and tested.

Across sustained interaction, prompts often move from commands to interrogations. “Write this” becomes “What assumption am I protecting?” or “What would make this claim falsifiable?” Prompting becomes a cognitive interface through which the author must externalize uncertainty before receiving language in return.

This can produce at least three shifts:

- **Intentional clarification:** the author specifies what must remain invariant across revision.
- **Voice differentiation:** competing rhetorical functions are assigned distinct names, registers, or editorial roles.
- **Judgment acceleration:** alternatives appear rapidly, increasing the importance of selection rather than decreasing it.

Within the corpus, the emergence of the figures “Dr. Chic” and “Whistledown” provided a visible instance of voice differentiation. One voice carried analytical compression; the other made institutional vanity narratable. The names should not be treated as independent personalities generated by a machine. They are authorial roles stabilized through repeated interaction. Their analytic value lies in showing how a writer can distribute internal functions across a conversational interface and later reintegrate them as deliberate craft.

Reflexive co-agency therefore changes the site of authorship. It moves creative labor away from uninterrupted origination and toward specification, comparison, resistance, curation, and accountable closure.

3.3 Mechanism III: Algorithmic Intra-Action

Some structures cannot be attributed cleanly to a single sentence proposed by the author or the model. They emerge over sequences: the author introduces a metaphor; the model literalizes it; the author rejects the literalization but retains a distinction; the model reorganizes that distinction; the author names the resulting framework. Causal contribution is distributed across the sequence even though responsibility for adoption is not.

Algorithmic intra-action names this sequence-level emergence. Its unit of analysis is not the isolated output but the evolving relation among prompt, response, revision, and decision.

The mechanism can be represented as:

> *human constraint + model variation + human evaluation + retained difference = emergent conceptual structure*

This formulation avoids two errors. First, it avoids romanticizing the model as an autonomous thinker. Variation is generated through learned statistical structure, not through an accountable point of view. Second, it avoids claiming that the human contribution is unchanged merely because the human has final authority. An encounter can reorganize thought without acquiring moral personhood.

Algorithmic intra-action is therefore compatible with strong human authorship. The author's role becomes more exacting: to govern the process by which generated variation acquires epistemic status.

4. Method and Evidence Boundary

4.1 Research design

The framework arose from a longitudinal analytic autoethnography conducted during approximately six months of intensive generative-AI-assisted writing and conceptual development. The private corpus contains more than 2,000 prompt-response cycles, manuscript revisions, contemporaneous reflections, and meta-conversations about authorship, voice, and method.

The author was simultaneously the participant, analyst, and locus of the reported transformation. This design provides access to micro-level changes in intention and self-description that are difficult to observe externally. It also creates substantial risks of circular interpretation, selective memory, and theory-confirming attention.

The present research edition therefore treats the corpus as a **concept-generating archive**, not as evidence sufficient for statistical inference or generalized psychological claims.

4.2 Analytic procedure

The original analysis proceeded through four interpretive operations:

- 1. **Interaction-trace review:** identifying recurring linguistic forms, conceptual metaphors, moments of resistance, and shifts in prompting posture.
- 2. **Sequence reconstruction:** examining how a response altered the next prompt or manuscript decision rather than treating outputs as isolated artifacts.
- 3. **Reflexive memoing:** recording how the author interpreted, accepted, or rejected model suggestions.
- 4. **Conceptual synthesis:** grouping recurring observations into cognitive loop exposure, reflexive co-agency, and algorithmic intra-action.

These operations are plausible for analytic autoethnography, but the source project did not maintain a preregistered protocol, a frozen model version, an independent second coder, or a publication-ready codebook. The paper accordingly avoids claims of intercoder reliability and does not present frequency estimates.

4.3 What this edition does not establish

This study does **not** establish that:

- generative AI possesses consciousness, intention, or moral agency;
- model glitches reveal subconscious states without human interpretation;
- interaction with AI causes clinically meaningful emotional regulation;
- the three mechanisms are universal across users, languages, or systems;
- a corpus of 2,000 exchanges is automatically stronger than a smaller, better documented corpus;
- the framework is historically first or exhaustive.

These exclusions strengthen rather than diminish the contribution. They locate the paper where its evidence is most defensible: a theoretically informed account of one author's extended practice, designed to produce testable distinctions for future study.

4.4 Future evidentiary standard

A stronger empirical test of Reflexive Authorship would require:

- timestamped interaction sequences with model and interface versions;
- explicit sampling and exclusion criteria;
- a codebook separating stylistic repetition, conceptual revision, affective report, and decision outcome;
- paired examples showing the prompt-response-revision chain;
- independent coding or participant validation;
- comparison across authors, languages, model families, and interaction durations;
- measures capable of distinguishing genuine change from retrospective narrative construction.

These requirements convert a compelling introspective theory into a research program.

5. Analytic Observations

5.1 Voice becomes inspectable

The most reliable observation in the corpus is not that AI discovered a hidden self, but that scale made stylistic recurrence visible. Model outputs overproduced favored structures. What felt elegant once became mannerism when returned ten times. This supplied an editorial advantage: the author could decide which patterns were signature and which were merely habitual.

The distinction matters commercially and academically. A recognizable voice can create continuity across a body of work; an unexamined loop can make every subject sound identical. The algorithmic mirror turns that difference into an object of governance.

5.2 Prompts become instruments of self-specification

Extended use changed the prompt from a request for prose into a statement of constraints. The author increasingly specified what a response must preserve, what it must not imply, and what standard would count as failure. This resembles design specification more than conversational inspiration.

Such specification can deepen agency. It forces tacit standards into language. Yet it can also create dependence on external articulation. If every uncertainty is immediately returned to a model, the author may lose tolerance for unresolved thought. Reflexive Authorship is therefore not inherently emancipatory. It is a practice whose value depends on whether the human retains the capacity to pause, withhold, and decide without algorithmic completion.

5.3 Rejection is productive data

Many significant moments occurred when the model was wrong in a recognizable way. A response might become too grand, too symmetrical, too therapeutic, or too certain. The author's irritation clarified a boundary that had not been explicit in the initial prompt.

This does not transform hallucination into evidence. It shows that failure can make evaluative criteria legible. The research object is the sequence of response and revision, not the error alone.

5.4 Coherence can conceal weakness

Generative systems are unusually capable of making an argument appear finished before it is grounded. The original manuscript itself demonstrated this danger: strong headings, confident contribution claims, and a polished abstract coexisted with incomplete references, unsupported empirical language, and undisclosed editorial traces.

The lesson is internal to the theory. An algorithmic mirror reflects not only insight but also the author's desire for premature completion. Governed execution requires a separate verification phase in which fluency has no authority.

6. The Reflexive System Model



Participation may be distributed. Accountability remains concentrated.

The three mechanisms form a recursive system:

- 1. **Exposure:** generated language amplifies a pattern.
- 2. **Recognition:** the human interprets that return as relevant or irrelevant.
- 3. **Reorientation:** the human changes the next prompt, criterion, or draft.
- 4. **Emergence:** a new structure appears across the sequence.
- 5. **Governance:** the human verifies, attributes, and decides whether the structure enters the work.

The fifth stage is decisive. Without governance, recursion can become mere reinforcement. The model supplies variation and pressure; the author supplies jurisdiction.

Reflexive Authorship can therefore be summarized in one proposition:

> *Generative AI changes authorship most deeply not when it writes instead of the human, but when its returned language changes the human who will decide what writing is allowed to remain.*

This proposition preserves human responsibility while acknowledging technological mediation. It also explains why the same model may flatten one writer and sharpen another. The difference is not simply access to the tool. It is the quality of the governing practice around the loop.

7. Implications

7.1 Attribution and provenance

Traditional attribution asks who wrote the final wording. Reflexive Authorship requires process provenance: what intellectual structure preceded model use, what transformations were suggested, what was rejected, and who verified the final claims. Disclosure should be proportionate to material influence rather than reduced to a binary declaration that AI was or was not used.

For academic work, the human author must remain accountable for citations, evidence, interpretation, and error. A model cannot accept correction, disclose conflicts, or answer for consequences. Assistance may shape the process without qualifying for authorship.

7.2 Creative labor

When generation becomes inexpensive, judgment becomes more valuable. The labor of authorship shifts toward defining constraints, recognizing false completion, preserving voice, and selecting among abundant alternatives. This is not the disappearance of authorship. It is a redistribution of effort within authorship.

The shift may also intensify inequality. Writers with deep domain knowledge can use generated variation diagnostically; writers without a stable evaluative basis may mistake fluent completion for warranted knowledge. Access to AI does not equal access to reflexive practice.

7.3 Education

Educational policy often treats AI use as either assistance to be permitted or substitution to be prohibited. Reflexive Authorship suggests a third objective: teach students to document how generated language changed their decisions. A process record can reveal whether the student exercised judgment, revised a claim, verified a source, or simply accepted completion.

7.4 Design and governance

Interfaces optimized only for immediate satisfaction conceal the recursive effects of use. Systems designed for reflexive practice could expose version history, preserve rejected alternatives, distinguish sourced material from generated conjecture, and prompt users to state decision criteria before accepting a response.

The design question is not merely how to make generation better. It is how to keep the human author epistemically visible inside the loop.

8. Limitations and Falsifiability

The framework begins from one author's experience. Its vocabulary may reflect that author's aesthetic and theoretical preferences more than a general property of human-AI interaction. The corpus spans changing interfaces and model behavior, making clean longitudinal comparison difficult. Private material limits independent audit. Retrospective synthesis can impose coherence on a process that was messier in real time.

The theory should be revised or rejected if future studies show that:

- reported authorial change does not exceed ordinary revision with non-generative tools;
- observed patterns are better explained by confirmation bias or demand characteristics;
- sequence-level analysis adds no explanatory value beyond isolated prompt quality;
- authors do not reliably distinguish algorithmic amplification from their own prior commitments;
- the proposed mechanisms fail to transfer across cases even as sensitizing concepts.

Falsifiability here is not a promise that a qualitative framework can be tested like a physical law. It is a commitment to specifying what evidence would make the framework less useful.

9. Conclusion

Generative AI enters authorship neither as a passive pen nor as a second human mind. It enters as an apparatus of return. It takes language conditioned by the author, transforms it through learned statistical structures, and returns a field of possible continuations. The author then encounters intention from the outside.

Reflexive Authorship names the reorganization that can follow. Cognitive loop exposure makes recurrent patterns visible. Reflexive co-agency changes how intention and judgment are performed. Algorithmic intra-action explains why some conceptual structures emerge only across sequences of prompting, response, resistance, and revision.

The framework's normative position is deliberately asymmetric. The machine may participate in the production of form; the human remains responsible for epistemic status and worldly consequence. The model proposes. The author governs.

In that sense, the future of authorship is not secured by insisting that nothing has changed. It is secured by becoming more precise about what changed, what must remain human, and who has the authority to close the loop.

Author's Note on AI Assistance

Tony T. Nguyen | Nguyễn Từ Vĩnh Trường is the originating author of the Reflexive Authorship framework and remains solely responsible for the argument, selection, interpretation, citations, and publication of this research edition. Generative AI systems participated materially in the historical interaction corpus and assisted during the development and editorial restructuring of the manuscript. Their use is therefore part of both the subject matter and the production history of the work. AI systems are not listed as authors because they cannot assume responsibility, verify claims independently, or consent to publication. All final decisions were made under human editorial authority.

References

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Barthes, R. (1977). The death of the author. In *Image-Music-Text* (S. Heath, Trans., pp. 142-148). Fontana. (Original work published 1967.)

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1-2), 347-356. https://doi.org/10.1016/S0004-3702(98)00055-1

Braidotti, R. (2013). *The Posthuman*. Polity.

Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 12*(1), Article 10. https://doi.org/10.17169/fqs-12.1.1589

Formosa, P., Bankins, S., Matulionyte, R., & Ghasemi, O. (2025). Can ChatGPT be an author? Generative AI creative writing assistance and perceptions of authorship, creatorship, responsibility, and disclosure. *AI & Society, 40*, 3405-3417. https://doi.org/10.1007/s00146-024-02081-0

Foucault, M. (1977). What is an author? In D. F. Bouchard (Ed.), *Language, Counter-Memory, Practice* (pp. 113-138). Cornell University Press. (Original lecture delivered 1969.)

Gretzky, M., & Dishon, G. (2025). Algorithmic-authors in academia: Blurring the boundaries of human and machine knowledge production. *Learning, Media and Technology, 50*(3). https://doi.org/10.1080/17439884.2025.2452196

Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. University of Chicago Press.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.

Panke, S. (2025). How can (A)I research this? An autoethnographic exploration of generative AI in research, teaching and instructional design. *Journal of Teacher Education*. https://doi.org/10.1177/00224871251325065

Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge University Press.

Watson, S., Brezovec, E., & Romic, J. (2025). The role of generative AI in academic and scientific authorship: An autopoietic perspective. *AI & Society, 40*, 3225-3235. https://doi.org/10.1007/s00146-024-02174-w

Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and social risks of harm from language models. *arXiv*. https://doi.org/10.48550/arXiv.2112.04359

Citation

Nguyen, T. T. | Nguyễn Từ Vĩnh Trường. (2026). *Reflexive Authorship: Generative AI as Algorithmic Mirror and the Reconfiguration of Creative Agency* (Public Research Release, HRP-RN-008). Highland Reflexive Press.

Rights and Permissions

For institutional access, quotation beyond fair use, translation, adaptation, course adoption, commercial use, or licensing inquiries: **tony@highlandreflexivepress.net**