Machine-Mediated Rediscovery: A Case Study in Novelty, Priority, and Self-Correction

Takuya Nemoto
“Independent Researcher” is a provisional title used for clarity; in fact the author is an undergraduate (ZEN University, Faculty of Social Informatics).
ORCID 0009-0000-1406-0547

Abstract

A language model supplied the mathematical content of a short series of preprints that were presented as a new "framework". Less than a year later, the same author, now working with a different language model, established that the central operator was a special case of the Friedkin–Johnsen model of opinion dynamics, a standard model that the first preprint had itself named as a neighbouring field without citing any of its literature. The route from the first publication ran through nine recorded stages, the last three recorded contemporaneously, and ended in that identification. This paper uses the case to ask what existing frameworks in the philosophy and sociology of science can and cannot say about novelty claims when the generating step is performed by a machine. I argue that (i) the case is neither a Mertonian multiple nor a rediscovery in the usual sense, because the arrival was neither independent nor human; (ii) when generation moves outside the human agent, the context of justification must absorb a task that was previously distributed across both contexts, namely the search for priority; and (iii) the correction that followed is best read as epistemic iteration from a defective starting point, in which implementation, not reflection, exposed the error that mattered. The correction was itself carried out with a language model, and the record does not separate the human share from the machine's share at each step. The paper makes no claim of mathematical novelty.

1. Introduction

Published claims of novelty that turn out to restate known results are not new. A widely cited example is a 1994 paper in *Diabetes Care* that presented, as a new "mathematical model", a method for the total area under metabolic curves (Tai 1994). A letter in the same journal later that year identified the method as the trapezoidal rule (Monaco & Anderson 1994).

What is new is the route by which such claims can now arise. Large language models can return standard constructions without indicating their provenance, and their users may publish those constructions as their own ideas. Mossel frames the question of what such models do in science as a contrast between reasoning and plagiarism, and argues that, because the training data and the interaction transcript remain opaque, claims of reasoning do not satisfy Popper's refutability principle (Mossel 2026). A study of problems listed as open in the Erdős problems database reports resolving 13 of them, 4 with seemingly novel solutions and 9 by locating solutions

already in the literature. It raises "the risk of 'subconscious plagiarism' by AI", that is, reproducing knowledge acquired in training without attribution (Feng et al. 2026; the counts are those of version 3, and the first version reported 5 and 8). Among the known solutions, the authors include cases in which the system produced a solution itself that then turned out to be in the literature, which they call independent rediscoveries. That is close to the configuration examined here.

Those studies examine expert settings in which the question is whether a model's output is already known. This paper examines a different configuration: a non-expert published a model-generated construction as a new framework, later established, while working with a second language model, that it was known, and kept a record of the route. One model produced the claim of novelty; work with another model withdrew it. The record is unusually well documented. Frozen preprints fix what was claimed and when. A version-controlled history fixes the later stages. An errata file fixes what was retracted and why.

The aim is modest. I do not propose a new theory of discovery. I ask how far four existing frameworks reach into this case, and where they stop: the distinction between the contexts of discovery and justification (Reichenbach; Duede), Merton's distinction between multiples and singletons, recent accounts of credit for generative AI outputs (Khosrowi, Finn & Clark), and Chang's epistemic iteration.

2. The case

2.1 The first version

In October 2025 the author published a preprint introducing a "framework" under the notation Ⅲ∞ (doi:10.5281/zenodo.17173703). The notation fixed a setting and nothing more: three quantities, iterated to the limit. It did not fix a map, a space, a metric, the existence or uniqueness of a limit, or a rate of convergence. The map, the fixed points, the convergence claims and the theorems were produced by a language model (ChatGPT, OpenAI) and published largely as generated. The preprint claimed applications in game theory, mathematical logic, mechanical engineering, AI alignment and climate policy.

The first section of that preprint named consensus dynamics in distributed systems as a field in which the idea "recurs". None of the three preprints in the series cites any work from that field.

2.2 The route to identification

The later history is recorded in nine stages (Table 1). Stages 1–6 are reconstructed from frozen preprints; stages 7–9 were recorded as they happened. From stage 2 onward the work was done with Claude Code (Anthropic), a second language model. The 2026 revisions of the three preprints were drafted, formulated and computationally checked with its help, and 145 of the 151 commits in the two repositories that record stages 7–9 and the later identification carry a co-authorship

line naming it. When this paper says "the author" for stages 2–9, it means the author working with that model.

Table 1. The route. "Contemporaneous" means recorded at the time in a version-controlled repository.

#	Date	Event	Kind	Record
1	2025-10	Model output published as a new framework	Generation	Reconstructed
2	2026-08	Operator written as `x ← DQx + (I − D)p`	Known form, not recognised as known	Reconstructed
3	2026-08	Convergence by the Banach fixed-point theorem, correctly cited	Use of known result	Reconstructed
4	2026-08	Contraction constant $\max_i a_i$ for coordinate-wise rates	Re-derivation	Reconstructed
5	2026-08	The revision shows that `n = 3` plays no role	Refutation of own claim	Reconstructed
6	2026-08	Unprovable and unsupported claims withdrawn	Retraction	Reconstructed
7	2026-09-06	Implementation outside the paper's assumptions: convergence is governed by the spectral radius, not the operator norm	Collision with known result	Contemporaneous
8	2026-09-07	A norm in which the operator contracts is constructed via a Lyapunov equation	Re-derivation	Contemporaneous
9	2026-09-08	Operator identified as a stationary iteration; the series declared a mathematical rediscovery	Identification	Contemporaneous

Two stages deserve comment.

Stage 5 is the first point at which later work removed content that the first model's output had supplied. The notation's "three" was shown to have no mathematical role: the proof goes through for every $n \geq 2$.

Stage 7 exposed an error that could not have been seen on paper. Under the paper's assumptions (a cyclic permutation `Q` and rates in [0, 1)), both the operator norm and the spectral radius of `A = DQ` are below one, so the distinction between them never changes the verdict on convergence. When all rates are equal, `A` is normal and the two coincide; with coordinate-wise rates the norm is $\max_i a_i$ and the spectral radius is the geometric mean of the rates, but both remain below one. Outside those assumptions the distinction matters. With `Q = [[1, 40], [0, 1]]` and every rate equal to 0.5, the matrix `A = DQ` has spectral radius 0.5 but operator norm about 20.01. The iteration still converges, yet the error first grows: for the initial error used in the demonstration, it rises from 5.0 to about 59.0, 11.8 times its initial size, by the second step before it decays. The worst-case amplification is about 20.

2.3 The identification with Friedkin–Johnsen

On 13 September 2026, in a working session with Claude Code, the operator was identified as a special case of the Friedkin–Johnsen model of opinion dynamics. Two formulations need to be distinguished. The 1990 paper writes $Y_t = \alpha W Y_{t-1} + \beta X B$, with scalar weights α on the endogenous and β on the exogenous conditions (Friedkin & Johnsen 1990). The scalar version of the operator has this form with $\beta = 1 - \alpha$ and the anchor p in the place of the exogenous term. The form now standard in the literature writes $x(k + 1) = \Lambda W x(k) + (I - \Lambda) u$, with a row-stochastic influence matrix W, a diagonal matrix Λ of individual susceptibilities, and anchoring opinions u (Proskurnikov & Tempo 2017). The version of the operator with coordinate-wise rates has this form. The operator restricts the influence matrix to a cyclic permutation, that is, to a ring in which each agent listens to exactly one neighbour.

The standard form is traced in the literature to Friedkin and Johnsen (1999). That paper has not been consulted, and nothing here rests on it. Later literature reports that the 1999 model sets the anchor to the initial opinions and ties the diagonal of W to the susceptibilities. If so, a cyclic permutation, whose diagonal is zero, would force every susceptibility to 1 and remove the anchor. The claim of this paper is therefore limited to the 1990 form and the standard form.

The size of the restriction was then measured. The question was whether a cyclic operator can reproduce a Friedkin–Johnsen equilibrium map $p \mapsto x^*$ for every anchor p. For $n \geq 3$, the set of equilibrium maps realisable with a cyclic permutation has dimension n, against $n(n - 1)$ for a general influence matrix, so the special case occupies a set of measure zero. At the same time, a single equilibrium cannot tell the two apart: every one of 8,000 equilibria generated from general influence matrices satisfied the condition implied by the cyclic form, a condition that is necessary and, outside degenerate cases, sufficient: namely that for some cyclic permutation σ each coordinate satisfies $x^*_i = a_i \, x^*_{\sigma(i)} + (1 - a_i) \, p_i$ for some rate a_i in [0, 1]. From a single equilibrium, the special case and the general model are indistinguishable. In the literature, the correspondence was available from the start.

2.4 What the record can and cannot show

The dates of stages 1–6 are fixed by the archive; what the author was thinking at those dates is inferred from what was published. The record fixes outcomes, not mental states. I return to this in Section 7.

3. Contexts of discovery when the generator is a machine

Reichenbach introduced the terms "context of discovery" and "context of justification" to mark the distinction between a thinker's way of finding a result and the way of presenting it to a public, and held that epistemology is occupied only in constructing the context of justification (Reichenbach 1938, §1). The separation has allowed philosophers to set aside the question of who, or what, produced a hypothesis.

Duede has used the same distinction to argue that the opacity of deep learning need not undermine discoveries made with it, because opacity bears on the context of discovery while justification proceeds by other means. His two cases, one from low-dimensional topology and one from the forecasting of earthquake aftershocks, place the network's output in the context of discovery and the final justification elsewhere, in a proof or in improved physical theory (Duede 2023). The present case agrees with that argument about justification and adds something about priority.

In the ordinary division of labour, part of the check for priority happens during discovery. A researcher who arrives at a construction usually arrives from somewhere in a literature, and that position already rules out some duplications. When a model supplies the construction, this part of the check disappears. The user receives a result without the literature that would normally surround it. The whole burden of establishing that the result is not already known moves into the context of justification.

The case shows what happens when that transfer is not noticed. The first preprint justified its claims internally (by fixed-point arguments) and never performed the external check. The generated text even named the relevant field: in effect, the model supplied a pointer to the literature that the user did not follow.

4. Multiples, singletons, and the delayed arrival

Merton called independent arrivals at the same result "multiples", in contrast to "singletons" (Merton 1961). The argument below needs only that distinction.

The case fits neither category cleanly. A multiple is an independent arrival at the same result, and this arrival was not independent in the relevant sense: the construction came from a model trained on text that may well include that literature. It also came decades after the Friedkin–Johnsen papers, although the delay alone would not settle the question. Nor is it a rediscovery in the ordinary sense, in which a person reaches a known result without knowing it is known. The person here did not reach the result at all; they received it.

I suggest that such cases need their own description: a known construction surfaced by a machine and asserted as new by a human. The description matters for the sociology of priority, because Merton's account of why priority disputes are fierce rests on the value placed on originality: in the institution of science, originality is at a premium, and recognition of priority is its reward (Merton 1957, 639–640). In the present case there is no originality to dispute, but there is a claim of originality to withdraw.

5. Priority search as an epistemic duty

Return to the trapezoidal-rule paper of 1994. What made it a failure was not the duplication itself. It was that the paper presented a known method as new and did not say that it was known.

If that is right, the relevant norm is a duty to search for priority before asserting novelty. A language model changes the cost of discharging that duty in two opposite ways. It makes the search cheaper, because the model can often name the field, as it did here. It also makes the duty easier to skip, because the output arrives in finished form.

Recent work on credit for generative AI outputs proposes that such outputs are created by collectives in the first instance, and that claims to creatorship come in degrees, depending on the nature and significance of individual contributions (Khosrowi, Finn & Clark 2023). Their examples are chiefly image-generating systems, but the view is stated generally. The present case suggests a companion point about responsibility. Whatever the distribution of credit, the duty to search for priority cannot be distributed to the model. It falls on the person who asserts novelty, because only that person makes the assertion.

The case adds a further twist. The claim of novelty came from work with one model, and its withdrawal came from work with another. The second model did not discharge the author's duty; it made the duty cheap enough that it was finally discharged. That is the first of the two effects just described, observed after the second had already done its damage.

6. Iteration from a defective starting point

Chang defines epistemic iteration as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (Chang 2004, 45–46). Inquiry starts by adopting an existing system of knowledge without any firm assurance that it is correct, and its results lead to the refinement and even the correction of that starting point.

Table 1 has that shape. The starting point, a notation that fixed only a setting together with the model's output built on it, was adopted without any assurance that it was correct. Each stage used the results of the previous one to remove part of the starting point: the role of "three" at stage 5, the unsupported applications at stage 6, the reliance on the operator norm at stage 7, and the claim of novelty at stage 9. Chang distinguishes two outcomes of iteration, enrichment of the starting system and its self-correction. The route in Table 1 is almost entirely of the second kind.

Two features of the iteration stand out.

First, the order in which errors were found was not the order of their difficulty. The meaninglessness of "three" was found on paper. The confusion between norm and spectral radius could not have been found on paper at all, because under the paper's assumptions it never changes the verdict on convergence. It appeared only when the operator was implemented outside those assumptions.

Second, the final correction came from the literature, not from further iteration. The measurement in Section 2.3 shows why. From any single equilibrium the special case

was indistinguishable from the general model, so computing equilibria on the author's own examples would not have revealed the identification. The iteration corrected the starting point up to the limit of what the author's own materials could show. The literature supplied the rest.

7. Limits of the case

There is a single case, and its author is also its subject. Self-report is exposed to the usual biases, and the author has an interest in presenting the correction favourably.

Two-thirds of the route is reconstructed. For stages 1–6 the archive fixes what was published and when, but not what the author understood at the time.

The first model's contribution is known only in outline. No transcript of the 2025 interaction survives, so it cannot be established which parts of the first preprint were generated, which were edited, and whether the model's output drew on the Friedkin–Johnsen literature.

The second model's contribution is recorded but not apportioned. The repositories show that Claude Code took part in nearly every commit from stage 7 onward and in the identification. They do not show, step by step, which suggestions came from the author and which from the model. Where this paper says "the author" for those stages, the claim is about a human working with a model, not about a human alone.

The identification with the Friedkin–Johnsen model was made before either original paper was read. The 1990 formulation has since been checked against the original, through a model-assisted reading reported with quotations. The 1999 paper has not been consulted, and the correspondence in Section 2.3 is stated only for the 1990 form and the standard form.

For these reasons the paper does not generalise. What it can show is where four existing frameworks reach in one well-documented case, and where they stop.

8. Conclusion

A known model was produced by a machine, published as new by a person, and later identified by the same person working with another machine, with a record of each step. The case is not a Mertonian multiple, and it is not a rediscovery in the usual sense. The discovery–justification distinction survives, but the check for priority, which used to be shared between the two contexts, now falls entirely on justification. The correction that followed has the shape of epistemic iteration, with one limit that the record makes visible: iteration on one's own materials could not reveal what only the literature could.

Disclosure of AI use

The case. The mathematical content of the 2025 preprint (the map, fixed points, convergence claims and theorems) was generated by ChatGPT (OpenAI) from the author's notation. No record of the version or the interaction survives.

The later stages. Claude (Anthropic), used through Claude Code, took part as follows.

- The 2026 revisions of the three preprints were drafted, formulated mathematically and checked computationally with Claude's help (stages 2–6).
- The implementation outside the paper's assumptions, the counterexample of stage 7, the Lyapunov construction of stage 8, and the measurement reported in Section 2.3 were written and run with Claude Code.
- The identification with the Friedkin–Johnsen model (13 September 2026) was made in a working session with Claude Code, by literature search.
- The errata file and the record of the route were written with Claude Code.

Of the 151 commits in the two repositories concerned, 145 carry a co-authorship line for Claude. Those lines record that Claude took part. They do not record how the work was divided at each step.

This paper. Claude Code drafted almost all of the text from the author's records, located the candidate literature by search, and translated between Japanese and English. Because the drafting environment could not reach the sources, the literature was checked through reports from three other models asked to open each source and quote it: ChatGPT (OpenAI), Grok (xAI), and Claude used outside Claude Code. The text was corrected where the reports disagreed with it, and a claim was treated as checked only where the reports agreed and at least one reached the original. The author chose the venue and the topic, approved the outline, and decided which claims to keep. At the time of writing, the author has not checked every quotation against the original sources; the quotations rest on the model-assisted readings described above. The author is responsible for every sentence. No AI system is an author.

References

Checked through a model-assisted reading of each source (see the disclosure), except where marked.

- Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.
- Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, 90(5), 1089–1099. doi:10.1017/psa.2023.8
- Feng, T., Trinh, T., Bingham, G., Kang, J., Zhang, S., et al. (2026). Semi-autonomous mathematics discovery with Gemini: A case study on the Erdős problems. arXiv:2601.22401, version 3.
- Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4), 193–205 (the publisher's record gives 193–206). doi:10.1080/0022250X.1990.9990069

- Friedkin, N. E., & Johnsen, E. C. (1999). Social influence networks and opinion change. *Advances in Group Processes*, 16, 1–29. (Not consulted; cited only for the attribution made in later literature.)
- Khosrowi, D., Finn, F., & Clark, E. (2023). Diffusing the creator: Attributing credit for generative AI outputs. *AIES '23*, 890–900. doi:10.1145/3600211.3604716
- Merton, R. K. (1957). Priorities in scientific discovery. *American Sociological Review*, 22(6), 635–659.
- Merton, R. K. (1961). Singletons and multiples in scientific discovery. *Proceedings of the American Philosophical Society*, 105(5), 470–486.
- Monaco, J. H., & Anderson, R. L. (1994). Tai's formula is the trapezoidal rule. *Diabetes Care*, 17(10), 1224–1225. doi:10.2337/diacare.17.10.1224
- Mossel, E. (2026). LLMs, reasoning and plagiarism. arXiv:2601.02380, version 5.
- Proskurnikov, A. V., & Tempo, R. (2017). A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control*, 43, 65–79. doi:10.1016/j.arcontrol.2017.03.002
- Reichenbach, H. (1938). *Experience and Prediction*. University of Chicago Press.
- Tai, M. M. (1994). A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. *Diabetes Care*, 17(2), 152–154. doi:10.2337/diacare.17.2.152