Three Laws of Machine Discovery: Why AGI Must Be Plural

机器发现的三定律：为什么 AGI 必须是复数的

Position paper · Draft v0.1 · October 2026 Author: 无撷

First published: 9 October 2026 · Not peer reviewed.

Abstract

The dominant picture of artificial general intelligence is singular: one model that eventually surpasses everyone at everything. We argue that this picture misdescribes how new knowledge is produced. Defining AGI narrowly as the capacity to produce verifiable knowledge beyond one's training distribution, we start from three defeasible axioms (learning systems are bounded by their priors; new knowledge is anti-consensus at birth; generation is cheap while verification is costly) and derive three laws of machine discovery. (I) Cognitive inertia: a mind stays within its prior unless acted upon by a differently formed mind; discovery-capable AGI is therefore plural. (II) Conjecture: departure from consensus requires licensed, bold conjecture; suppressing all low-probability output suppresses discovery. (III) Reaction: every assertion must meet an equal and opposite test; discovery is bounded by verification throughput, and the more expensive verification is, the more diversity matters. We support these laws with evidence from the sociology of science, social epistemology, organization studies and social psychology, and with recent evidence that large language models are becoming homogeneous. We address six objections, state six falsifiable predictions, and describe a practitioner field protocol. The implication is that AI development should aim at an epistemic ecology of differently trained models and humans, and that diversity of training lineages should be treated as a public good.

摘要

主流的 AGI 想象是单数的：一个足够大的模型，最终在一切领域超过一切人。本文认为，这幅图景误解了新知识是如何产生的。我们把 AGI 狭义地理解为“在训练分布之外产生可验证新知识的能力”，并从三条可被反驳的公理出发——学习系统受其先验约束；新知识在诞生时是反共识的；生成廉价而检验昂贵——推出机器发现的三条定律。（一）认知惯性：一个心智若不受成形方式不同的心智作用，将停留在其先验之内，因此能发现的 AGI 必然是复数的。（二）猜想：对共识的偏离需要被许可的大胆猜想，抑制一切低概率输出等于取消发现。（三）反作用：每一个断言都必须遭遇相当的反向检验；发现速度受检验能力约束，而且检验越昂贵，多样性越重要。我们综合科学社会学、社会认识论、组织研究与社会心理学的证据，以及大模型日益同质化的最新实证，论证这三条定律的合理性；讨论六种反驳；提出六条可证伪的预测和一个实践者田野

观察方案。本文的含义是：AI 发展的目标应是一个由不同训练谱系的模型与人共同组成的认识生态，而不是一个单体；训练谱系的多样性应被当作公共品来维护。

Keywords: artificial general intelligence; social epistemology; division of cognitive labor; multi-agent systems; scientific discovery; cognitive monoculture

1 Introduction

The dominant picture of artificial general intelligence (AGI) is singular. A model is scaled until it outperforms everyone at everything, and what remains for humanity is to control it. This picture is sometimes described as a “god-like” intelligence.

The year 2026 seems to lend strong support to that picture. AI systems produced a string of research-level mathematical results, and on 8 September OpenAI announced that roughly ten thousand autonomous agents had found a singularity in the three-dimensional Navier–Stokes equations, with the result formally verified in Lean (Quanta Magazine, 2026). Three days later a group of Fields medalists published a joint statement, “A Severe Misalignment of AI in Mathematics,” which describes the mathematical community as a small-scale model of humanity: diverse in its approaches, bound by shared values, and dependent on human discussion and transmission (mathandai.org, 2026).

This paper does not ask whether AI is “smart enough.” It asks what kind of system produces new knowledge. Our claim is:

> **Intelligence that can produce new knowledge is not a property of any single mind, but of an organization of minds that differ in their priors, hold different stances, contest one another, and are disciplined by tests. Discovery-capable AGI must therefore be plural.**

Three clarifications are needed. First, “plural” refers to a plurality of priors (training lineages), not to the number of boxes: a thousand copies derived from the same pretraining remain singular; a machine containing several independently trained subsystems that genuinely disagree is a plural organization. Second, we define AGI narrowly in terms of discovery, as “the capacity to produce verifiable knowledge beyond one’s training distribution,” to set aside disputes over the definition of AGI. Third, this is a theoretical argument: we derive three “laws” from three defeasible axioms and use evidence from the social sciences and AI research to support the plausibility of their premises and implications. We do not claim that these laws have been empirically tested; Section 9 sets out ways to test them.

The word “law” is borrowed from Newton to organize the argument; these are not laws of physics. We choose Newton because of the following case.

2 A Historical Case and Its Contemporary Echo

2.1 1679–1687: Hooke, Newton, Halley and the Royal Society

The inverse-square law of gravitation is usually credited to Newton alone, but its discovery was a group process. In late 1679 Hooke wrote to Newton proposing that planetary motion be analyzed as a tangential inertial motion combined with an attraction toward the center, and in later letters he conjectured that the attraction falls off with the square of the distance. Hooke had the insight but not the proof. In 1684 Wren, Halley and Hooke discussed the same problem in London, and Wren offered a prize for a demonstration. That year Halley visited Newton in Cambridge to ask what orbit an inverse-square force would produce. Newton answered that it would be an ellipse, wrote De Motu, and expanded it into the Principia. When Hooke demanded recognition in 1686, Newton struck references to him and threatened to withhold the third book; Halley mediated and paid for the publication himself (Westfall, 1980; Jardine, 2003).

The episode contains every element of our argument. Hooke supplied a conjecture he could not test. Newton supplied a proof, that is, a severe test of the conjecture. Halley posed the decisive question and served as broker and financier. The Royal Society supplied the institutions of record, correspondence and publication. A point often overlooked: Westfall notes that several people defined the same problem for themselves at much the same time (Westfall, 1980). The inverse-square law was a multiple discovery, not the inspiration of a lone genius.

2.2 2026: The Navier–Stokes Dispute

According to Quanta Magazine, about twelve hours before OpenAI's announcement, Buckmaster and Alpöge announced that they had resolved several closely related problems with the help of AI models, including OpenAI's. Both teams built on groundwork by Córdoba and Martínez-Zoroa. OpenAI acknowledged that its work was inspired by rumors that the rival team had already achieved a breakthrough; it ceded priority on the three-dimensional Euler result but claimed the Navier–Stokes result, and the parties gave differing accounts of the timeline (Quanta Magazine, 2026). The result has not been peer reviewed, and whether it corresponds to the Clay Mathematics Institute's original formulation is also disputed.

Even the most singular-looking achievement of 2026 was therefore embedded in a plural network. Humans posed the problem and laid the groundwork; a rival team's signal steered the search; a formal verifier adjudicated correctness; and the result immediately triggered a dispute over credit. Three hundred and forty years after Hooke and Newton, the structure recurs almost unchanged. We take this to be a structural feature of how new knowledge is produced, not a coincidence.

3 Three Axioms

The following are working hypotheses with empirical support, not self-evident truths. Each could in principle be refuted.

Axiom 1 (Prior-boundedness). How a learning system generates and evaluates hypotheses is fixed by the inductive biases of its training (data, objective, architecture), and no inductive bias is optimal for all problems.

Support. No-free-lunch results show that, averaged over all possible problems, no learning algorithm outperforms another; a bias that helps on some problems must hurt on others (Wolpert, 1996). Empirically, large language models struggle to self-correct their reasoning without external feedback (Huang et al., 2024): errors consistent with a model's own prior are invisible from inside.

Axiom 2 (Novelty is anti-consensus at birth). A genuinely new idea has low probability under the prevailing prior when it is proposed; otherwise it would not be new.

Support. Kuhn (1962) described scientific revolutions as breaks with a governing paradigm. Barber (1961) documented systematic resistance by scientists to scientific discovery. Azoulay et al. (2019) found that after the unexpected death of star scientists, contributions to their fields by outsiders rise significantly: anti-consensus ideas enter once the guardians of the consensus leave.

Axiom 3 (Generation is cheap; verification is costly, and its cost varies by domain). Producing candidate ideas is far cheaper than establishing whether they are true. In formal domains verification can be automated; in experimental science it is slow and expensive.

Support. Where an automatic evaluator exists, loops of LLM generation and evaluation have already produced new results (Romera-Paredes et al., 2024). Meanwhile the bottleneck in mathematics has shifted to verification and digestion: Tao (2026) warns of "proof indigestion" when AI-generated proofs outpace human absorption, and the Leiden Declaration (2026) names plausible but unreliable arguments, hard to distinguish from correct proofs, as a primary risk.

4 Law I: Cognitive Inertia (Plurality)

> **A mind remains within the bounds of its training prior unless acted upon by a differently formed mind. Discovery-capable AGI is therefore plural.**

4.1 Derivation

By Axiom 1, a single system's hypothesis space and evaluation criteria are bounded by its prior. By Axiom 2, new ideas have low probability under the prevailing prior. Together these imply that a system trained on mainstream data and tuned to human preferences will systematically underweight new ideas, and cannot detect this by itself. Only an external, differently formed mind, or an external test (Law III), can push it beyond its prior.

4.2 Evidence from the Sociology of Science and Social Epistemology

Division of cognitive labor. Kitcher (1990) showed that if every member of a community rationally pursues the currently most promising theory, the community as a whole does worse than one that spreads its effort across alternatives. Strevens (2003) argued that science's priority rule is an incentive that achieves such spreading. In Weisberg and Muldoon's (2009) model of

epistemic landscapes, a minority of “mavericks” markedly improves how efficiently a population finds peaks.

Objectivity as a social achievement. Longino (1990) located scientific objectivity not in individual impartiality but in critical interaction among diverse perspectives, which requires venues for criticism, uptake of criticism, public standards, and a “tempered equality” of intellectual authority. Mill (1859) made the point earlier: “He who knows only his own side of the case, knows little of that.”

Multiple discovery and the collective brain. Ogburn and Thomas (1922) listed 148 discoveries made independently by different people; Merton (1961) concluded that multiples, not singletons, are the norm. Anthropology points the same way. Tasmanians lost technologies after isolation from the mainland (Henrich, 2004), and the complexity of toolkits across Oceanic islands tracks population size and contact (Kline & Boyd, 2010). Muthukrishna and Henrich (2016) conclude that the rate of innovation depends chiefly on population size, connectedness and transmission fidelity rather than on individual intelligence.

Teams and the burden of knowledge. Over the past half-century, teams have increasingly dominated knowledge production across fields and produce higher-impact work (Wuchty et al., 2007), partly because an expanding frontier forces individuals to specialize (Jones, 2009). Large teams tend to develop existing ideas, while small teams more often disrupt them (Wu et al., 2019): organizational structure shapes the kind of discovery that is made.

4.3 Evidence from Group Decision-Making and Social Psychology

Independence. The Condorcet jury theorem says that a majority of independent judges is more likely to be right than any one of them, but the advantage erodes quickly when judgments are correlated (Ladha, 1992). Page’s (2007) diversity prediction theorem states an identity: collective squared error equals average individual squared error minus the diversity of individual predictions. In the experiments of Lorenz et al. (2011), exposing participants to others’ estimates reduced the diversity of estimates without improving their accuracy.

Organizational learning. March (1991) showed that rapid socialization of individuals into an organizational code reduces exploration and, in the long run, the organization’s knowledge; retaining slow learners who deviate from the code helps. Zollman (2007, 2010) found, in network models of scientific communities, that on hard problems densely connected communities are more likely to converge prematurely on a wrong answer, and that transient diversity has epistemic value.

Authentic dissent. Nemeth et al. (2001) compared group members who genuinely dissented with members assigned to play devil’s advocate, and found that authentic dissent did more to stimulate divergent thinking. The implication for AI is direct: assigning personas to one model resembles appointing a devil’s advocate, while models from different training lineages come closer to authentic dissent.

4.4 Evidence from AI: Homogenization Is Under Way

An analysis of more than 350 large language models found that when two models both err, they give the same wrong answer about 60% of the time, and that larger, more accurate models have more correlated errors, even across architectures and providers (Kim et al., 2025). On open-ended prompts, different models produce strikingly similar responses, an “artificial hivemind” (Jiang et al., 2025). Among human writers, generative AI raised the rated creativity of individual stories but reduced the diversity of the stories collectively (Doshi & Hauser, 2024). In theory, when all decision-makers adopt the same algorithm, outcomes can worsen even if that algorithm is more accurate (Kleinberg & Raghavan, 2021). Models trained recursively on their own outputs lose the tails of the distribution (Shumailov et al., 2024), which is where new ideas live.

Evidence on multi-agent debate is mixed. Changing base models does not reliably help (M3MAD-Bench, 2026), and debate sometimes makes answers worse (Talk Isn’t Always Cheap, 2025). Under Law I this is not a counterexample: most current setups use the same or similar lineages and expose agents to each other’s answers in the first round, violating the conditions below.

4.5 Conditions

- **Independence:** participants’ priors come from different data, objectives, architectures or experience, and their first judgments are made independently.
- **Commitment:** participants hold persistent stances rather than temporary personas.
- **Parity:** no participant’s authority is so overwhelming that criticism stops working.

4.6 Two Corollaries

Corollary 1: There is no god-like singleton. Even an agent far more capable than all others disables a community’s error correction if its authority is overwhelming: Longino’s tempered equality fails, and humans systematically over-rely on automation (Parasuraman & Manzey, 2010). The problem with a god-like singleton is not that it is insufficiently intelligent. It is that no one can correct it.

Corollary 2: A weaker but different mind can be worth more than a stronger but similar one. By the diversity prediction theorem, adding a strong member whose errors are highly correlated with the group’s does little to reduce collective error, while adding a weaker member with uncorrelated errors can reduce it more. As frontier models converge (Section 4.4), humans, the most differently trained minds available, acquire epistemic value that is structural, not merely sentimental. Here our view meets that of the Fields medalists: humans are not legacy components awaiting replacement but the most distinct source of priors in a plural collective.

5 Law II: Conjecture

> **Departure from consensus grows with the force of conjecture and shrinks with the “mass” of the consensus, that is, the degree to which training and preference tuning have entrenched it.**

5.1 Derivation

By Axiom 2, new ideas have low probability under the prevailing prior, so any filter that passes only high-probability content removes new ideas along with errors. Lowering the sampling temperature, using human preferences to push outputs toward safe average answers, and penalizing all “hallucination” each reduce the force of conjecture. The more entrenched the consensus, the more force is required.

5.2 Evidence

Popper (1963) described the growth of knowledge as bold conjecture and severe refutation, and Campbell (1960) described creative thought as blind variation and selective retention, isomorphic to biological evolution. Historically, Maxwell’s displacement current first appeared in his 1861–62 mechanical model of the ether built from molecular vortices and idle wheels; the model was later abandoned, and the term that remained was correct. Experiments on sleep and creativity found that more than twice as many participants discovered a hidden rule after sleep as after wakefulness (Wagner et al., 2004), and that sleep onset is a creative sweet spot (Lacaux et al., 2021). The REBUS model proposes that certain states of consciousness change beliefs by relaxing the constraint that high-level priors place on lower-level signals (Carhart-Harris & Friston, 2019); we borrow this mechanism only as an analogy.

On the AI side, Kalai et al. (2025) argue that language models hallucinate partly because training and evaluation reward guessing over acknowledging uncertainty. On this account the root problem is that conjecture and assertion share one channel, not generation itself.

5.3 Conditions

- **Separate conjecture from assertion:** every output carries an epistemic status (conjecture, inference, verified), and conjectures never enter the assertion channel in the voice of fact.
- **Directed, long-range recombination:** good conjectures usually transplant structure from one domain to another, as Maxwell did from mechanics to electromagnetism, rather than random noise.
- **Law III as a precondition:** without tests, licensed conjecture only adds noise.

6 Law III: Reaction (Falsification)

> **Every assertion must meet an equal and opposite test; assertions that have not met one do not enter knowledge.**

6.1 Derivation

By Axiom 3, generation is cheap and verification costly, so the rate of discovery is bounded chiefly by verification capacity rather than by generative capacity. A system that can generate but cannot test will produce, at an ever-increasing rate, candidates whose truth it cannot distinguish.

6.2 Evidence

Lakatos (1978) described progress as competition among research programmes, much of which consists of abandoning degenerating ones. Most of this work is invisible: negative results are rarely published (Rosenthal, 1979), and in the social sciences most studies with null results are never written up (Franco et al., 2014). In AI, the success of systems such as FunSearch rests on an automatic evaluator (Romera-Paredes et al., 2024), and the 2026 Navier–Stokes result likewise depended on formal verification in Lean.

6.3 Corollary: Diversity Substitutes for Verification

Laws I and III interact. Where verification is cheap and perfect (board games, executable code, formal proof), a homogeneous population plus massive trial and error can break through: AlphaZero's self-play is, in effect, a homogeneous population with a perfect referee and exploration noise (Silver et al., 2018). Where verification is expensive, slow or ambiguous, as in frontier experimental science, not every candidate can be tested, so what to test first is decided by contest among different priors, and the value of diversity rises. Results in network epistemology agree: the effects of communication structure appear mainly on hard problems (Rosenstock et al., 2017). The corollary can also account for an observation. AI broke through first in mathematics and programming and has advanced more slowly in experimental science, not necessarily for lack of intelligence but for lack of the combination of Laws I and III.

6.4 Negative Results as a Public Good

In a plural system, refuted ideas should enter a shared record (a "falsification ledger") together with the reasons for their refutation, so that other members need not repeat the same trial and error. This is one of the things human science does worst, and perhaps one that plural AI systems could do better.

7 Institutions: Where the Three Laws Operate

The laws describe interactions among minds, and interaction needs a venue. In the seventeenth century the venue was the Royal Society and the Philosophical Transactions founded by its secretary, Henry Oldenburg, which provided record, correspondence, publication and the adjudication of priority. Merton (1942) summarized the norms of science as communalism, universalism, disinterestedness and organized skepticism, and later argued that priority disputes are not failures of character but structural products of a reward system whose currency is being first (Merton, 1957). Tao (2026) describes peer review as the way individual achievements are converted into collective progress and understanding, and the central worry of the Fields medalists is precisely that writing, citation and transmission are skipped in rushed announcements. A plural AI collective likewise needs credit assignment, arbitration and shared memory. For now we treat this as a boundary condition of the three laws rather than as a fourth law, and leave open whether a "zeroth law" is needed.

8 Objections and Responses

Objection 1: A sufficiently large model can contain every perspective internally. Mixture-of-experts architectures and multi-persona prompting seem able to create diversity inside one model. Response: if these perspectives come from the same pretraining, they share a prior and their errors are highly correlated (Kim et al., 2025); they resemble assigned devil's advocates more than authentic dissent (Nemeth et al., 2001). If a system really contains several independently trained subsystems that genuinely oppose one another, it is a plural organization —exactly our claim, regardless of how many boxes house it.

Objection 2: AlphaZero and the formal proofs of 2026 show that homogeneous systems can also discover. Response: these successes occur in domains where verification is cheap and perfect, consistent with the corollary in Section 6.3. In the Navier–Stokes case, moreover, humans posed the problem and laid the groundwork, while a rival team influenced the direction of the search: a homogeneous population solved a problem already framed by a plural community, rather than framing it alone.

Objection 3: Human history contains lone geniuses. Newton's "miracle year" is often cited. Response: Newton depended on Kepler, Galileo and Huygens, as well as Hooke's letter; the inverse-square law itself was a multiple discovery (Section 2.1). The prevalence of multiple discoveries (Merton, 1961) and the rising dominance of teams (Wuchty et al., 2007) suggest that the lone genius is better understood as a narrative simplification.

Objection 4: Empirical findings on multi-agent debate are inconsistent. Response: this tests the conditions of Law I rather than refuting the law. If cross-lineage collectives still fail to outperform single-lineage systems of equal compute after satisfying independence, commitment and parity, Law I should be revised (see Predictions 1 and 2 in Section 9).

Objection 5: Plural AIs will converge through synthetic data, or even collude. Response: this is a real risk, and precisely the reason for the paper's policy implication: diversity of training lineages will not sustain itself and needs to be maintained as a public good (Section 11).

Objection 6: Plural systems are harder to control. Response: plurality creates the possibility of mutual monitoring, but also introduces risks of coordination and collusion. This paper does not address the whole safety problem. We note only that a singleton is not safer simply because it is single: having no one able to correct it is itself a risk.

9 Falsifiable Predictions

#	Prediction	Source	Test	Refuted if
1	At equal compute, collectives of models from different lineages show lower error correlation and higher error-correction rates than resampling or multiple personas of a single model	Law I	Controlled comparison on benchmarks and open problems	No difference, or the reverse

#	Prediction	Source	Test	Refuted if
2	Independent first answers outperform sequential exposure (answering after seeing others' answers)	Law I (independence)	A/B comparison on identical items	Sequential exposure does as well or better
3	A weaker model from a different lineage can add more to a collective than a stronger model from the same lineage	Corollary 2	Ablation controlling individual accuracy and error correlation	Marginal value depends only on individual accuracy and is unrelated to correlation
4	With a verifier, licensed generation (with mandatory labels) yields more verified novel results; without one it yields only more errors	Laws II × III	2 × 2 design	No interaction
5	The gain from plurality rises with the cost of verification	Section 6.3 corollary	Comparison across formal, simulated and experimental domains	Gain unrelated to verification cost
6	AI breaks through first in formally verifiable domains; in experimental science, where verification is costly, the decisive systems will be multi-lineage with experiments in the loop	All	Historical observation of AI scientific results over the coming years	Single-lineage systems repeatedly make independent breakthroughs in high-verification-cost domains

10 Practitioner Field Observation (Method; Results Pending)

To obtain evidence from real settings beyond controlled experiments, a practitioner applies practices derived from the three laws in daily work: independent first answers from multiple lineages, epistemic labels, kill criteria set in advance, and a falsification ledger. Each use is recorded in a standard format: problem type, participating lineages, whether first answers were independent, degree of first-round disagreement, minority view, decision, the method and result of subsequent testing, whether diversity changed the conclusion, and the cost of verification. Methodologically, this follows the traditions of action research and reflective practice (Lewin, 1946; Schön, 1983).

We plan 40–60 entries over 8–12 weeks and will report the rate of substantive disagreement, the rate at which disagreement led to an error being corrected, the hit rate of minority views, and the relation between the value of diversity and the cost of verification. Limitations include a single practitioner, non-random selection of problems, the absence of a control group, and confidential content, so only aggregates and de-identified cases will be reported. These data serve only to illustrate the operability of the three laws in practice and their preliminary direction; they are not causal evidence.

[Results: pending]

11 Implications

For AI development: the goal should be an epistemic ecology—models from different training lineages together with humans, supported by independent first answers, epistemic labels, falsification ledgers and arbitration rules—rather than an ever-larger singleton. A weaker model trained differently may be the most valuable peer for a frontier system.

For AI governance: cognitive monoculture is a systemic risk. Diversity of training lineages—corpora in different languages and cultures, different training objectives, different institutions, and open models—should be maintained as a public good. This aligns with the Leiden Declaration's call for investment in public computational infrastructure and its concern that research directions should not be dominated by a few companies.

For the scientific community: the human role is not to be replaced, but to act as Halley did—posing good questions, mediating disputes, exercising judgment and taste, and passing results on.

12 Conclusion

New knowledge has never been produced by one mind alone. It arises from collisions between different priors, depends on licensed conjecture, and is adjudicated by tests. Three hundred and forty years ago the inverse-square law emerged among Hooke's conjecture, Newton's proof, Halley's question and the Royal Society's institutions; the Navier–Stokes dispute of 2026 shows that AI has not changed that structure. If AGI means intelligence that discovers, it will not be a model. It will be a society.

Statement

AI use: several large language models were used in developing and drafting this paper, including discussion and text generation; their specific roles will be stated when the manuscript is finalized. The author is responsible for checking all claims and citations.

References

Azoulay, P., Fons-Rosen, C., & Graff Zivin, J. S. (2019). Does science advance one funeral at a time? American Economic Review, 109(8), 2889–2920.

Barber, B. (1961). Resistance by scientists to scientific discovery. Science, 134(3479), 596–602.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. Psychological Review, 67(6), 380–400.

Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. Pharmacological Reviews, 71(3), 316–344.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), eadn5290.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502–1505.

Henrich, J. (2004). Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses—the Tasmanian case. American Antiquity, 69(2), 197–214.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. ICLR 2024.

Jardine, L. (2003). The Curious Life of Robert Hooke: The Man Who Measured London. HarperCollins.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., & Choi, Y. (2025). Artificial Hivemind: The open-ended homogeneity of language models (and beyond). NeurIPS 2025. arXiv:2510.22954.

Jones, B. F. (2009). The burden of knowledge and the “death of the Renaissance man”: Is innovation getting harder? Review of Economic Studies, 76(1), 283–317.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. arXiv:2509.04664.

Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. ICML 2025. arXiv:2506.07962.

Kitcher, P. (1990). The division of cognitive labor. Journal of Philosophy, 87(1), 5–22.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. PNAS, 118(22), e2018340118.

Kline, M. A., & Boyd, R. (2010). Population size predicts technological complexity in Oceania. Proceedings of the Royal Society B, 277(1693), 2559–2564.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Lacaux, C., Andrillon, T., Bastoul, C., et al. (2021). Sleep onset is a creative sweet spot. Science Advances, 7(50), eabj5866.

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. American Journal of Political Science, 36(3), 617–634.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes. Cambridge University Press.

Leiden Declaration on Artificial Intelligence and Mathematics (2026, June 2). doi:10.5281/zenodo.20302944.

Lewin, K. (1946). Action research and minority problems. Journal of Social Issues, 2(4), 34–46.

Longino, H. E. (1990). Science as Social Knowledge. Princeton University Press.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. PNAS, 108(22), 9020–9025.

M3MAD-Bench: Multi-dimensional evaluation of multi-agent debate across domains and modalities (2026). arXiv:2601.02854.

March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87.

mathandai.org (2026, September 11). A Severe Misalignment of AI in Mathematics.

Merton, R. K. (1942/1973). The normative structure of science. In The Sociology of Science. University of Chicago Press.

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.

Mill, J. S. (1859). On Liberty. London: John W. Parker.

Muthukrishna, M., & Henrich, J. (2016). Innovation in the collective brain. Philosophical Transactions of the Royal Society B, 371, 20150192.

Nemeth, C., Brown, K., & Rogers, J. (2001). Devil’s advocate versus authentic dissent: Stimulating quantity and quality. European Journal of Social Psychology, 31(6), 707–720.

Ogburn, W. F., & Thomas, D. (1922). Are inventions inevitable? A note on social evolution. Political Science Quarterly, 37(1), 83–98.

Page, S. E. (2007). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.

Popper, K. (1963). Conjectures and Refutations. Routledge.

Quanta Magazine (2026, September 8). AI has solved one of math’s $1 million Millennium Prize problems.

Romera-Paredes, B., et al. (2024). Mathematical discoveries from program search with large language models. Nature, 625, 468–475.

Rosenstock, S., Bruner, J., & O’Connor, C. (2017). In epistemic networks, is less really more? Philosophy of Science, 84(2), 234–252.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638–641.

Schön, D. A. (1983). The Reflective Practitioner. Basic Books.

Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755–759.

Silver, D., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140–1144.

Strevens, M. (2003). The role of the priority rule in science. Journal of Philosophy, 100(2), 55–79.

Talk Isn’t Always Cheap: Understanding failure modes in multi-agent debate (2025). arXiv:2509.05396.

Tao, T. (2026). Mathematics in the age of AI. Lecture, International Congress of Mathematicians 2026.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. Nature, 427, 352–355.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. Philosophy of Science, 76(2), 225–252.

Westfall, R. S. (1980). Never at Rest: A Biography of Isaac Newton. Cambridge University Press.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), 1341–1390.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566, 378–382.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036–1039.

Zollman, K. J. S. (2007). The communication structure of epistemic communities. Philosophy of Science, 74(5), 574–587.

Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. Erkenntnis, 72(1), 17–35.