机器发现的三定律：为什么 AGI 必须是复数的

Three Laws of Machine Discovery: Why AGI Must Be Plural

观点论文 · 初稿 v0.1 · 2026 年 10 月　　作者：无撷

首次公开：2026 年 10 月 9 日 · 未经同行评审。

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。

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.

关键词：通用人工智能；社会认识论；认知分工；多智能体系统；科学发现；认知单一化

1 引言

关于 AGI 的主流想象是单数的：一个模型被不断放大，直到它在每一个领域都超过每一个人，人类剩下的问题只是如何控制它。这种想象有时被称为“上帝级智能”。

2026 年似乎为这幅图景提供了有力的支持。AI 系统在研究级数学上接连取得结果；9 月 8 日，OpenAI 宣布约一万个自主智能体在三维 Navier–Stokes 方程中找到了奇点，结果经 Lean 形式化验证（Quanta Magazine, 2026）。然而三天之后，一批菲尔兹奖得主发表联合声明《A Severe Misalignment of AI in Mathematics》，把数学共同体描述为“人类的一个缩影”：方法多样，由共同的价值联结，依靠人与人之间的讨论与传承（mathandai.org, 2026）。

本文不讨论 AI 是否“足够聪明”，而是追问：新知识究竟由什么样的系统产生？我们的主张是：

> **能产生新知识的智能，不是任何单个心智的属性，而是由多个先验不同、立场不同、彼此争辩并受检验约束的心智所组成的组织的属性。因此，能发现的 AGI 必然是复数的。**

有三点需要事先说明。第一，“复数”指的是先验（训练谱系）的复数，而不是盒子的数量：同一次预训练派生出的一千个副本仍是单数；一个内部装有多个独立训练、真正会互相反对的子系统的机器，则是一个复数组织。第二，本文把 AGI 狭义地理解为发现能力，即“在训练分布之外产生可验证新知识的能力”，以避开 AGI 的定义之争。第三，本文是理论论证：我们从三条可被反驳的公理推出三条“定律”，用社会科学与 AI 研究的证据支撑前提和推论的合理性；我们不声称这些定律已被实证检验，并在第 9 节给出检验它们的方式。

“定律”一词借自牛顿，是组织论证的方式，而不是物理定律。选择牛顿，是因为下面这个案例。

2　一个历史案例与它的当代回声

2.1　1679–1687：胡克、牛顿、哈雷与皇家学会

平方反比引力律常被归于牛顿一人，但它的发现是一个群体过程。1679 年底，胡克写信给牛顿，提出把行星运动分解为切线方向的惯性运动和指向中心的吸引，并在随后的信中猜测吸引力与距离平方成反比；胡克有洞见，却没有证明。1684 年，雷恩、哈雷与胡克在伦敦讨论同一问题，雷恩为证明悬赏；同年哈雷到剑桥问牛顿，在平方反比力下行星轨道是什么形状，牛顿答以椭圆，随后写出《论运动》，最终发展为《原理》。1686 年，胡克要求承认其优先权，牛顿删去书中对胡克的提及，并一度威胁不出第三卷；哈雷居间斡旋，并自费出版了《原理》（Westfall, 1980；Jardine, 2003）。

这个故事同时包含了本文的所有要素：胡克提供了猜想，却无力检验；牛顿提供了证明，也就是对猜想的严格检验；哈雷提出了关键问题，并担任仲裁与出资人；皇家学会提供了记录、通信与发表的制度。还有一点常被忽略：据 Westfall 考证，同一时期有好几个人各自界定出了同一个问题（Westfall, 1980）。平方反比律是一项多重发现，而不是一个孤独天才的灵感。

2.2　2026：Navier–Stokes 之争

据 Quanta Magazine 报道，在 OpenAI 宣布其结果约 12 小时之前，Buckmaster 与 Alpöge 宣布借助 AI 模型（包括 OpenAI 的模型）解决了几个密切相关的问题；两队都建立在 Córdoba 与 Martínez-Zoroa 的前期工作之上；OpenAI 承认其工作受到对手团队已取得突破的传闻启发，在三维 Euler 方程的结果上让出优先权，但主张 Navier–Stokes 结果；双方对时间线的说法不一（Quanta Magazine, 2026）。该结果尚未经过同行评审，它与克雷数学研究所原题表述的对应关系也存在争议。

可见，即便是 2026 年最像“单体超级智能”的成就，也嵌在一个复数网络之中：问题由人提出，基础由人奠定，方向受对手信号牵引，正确性由形式化检验器裁决，而结果一出现就引发了归属之争。三百四十年后，胡克—牛顿的结构几乎原样重现。我们认为这不是巧合，而是新知识生产方式的结构性特征。

3　三条公理

以下三条不是不证自明的真理，而是有经验支持、原则上可被反驳的工作假设。

公理一（先验有界）　一个学习系统生成和评价假设的方式，由其训练过程（数据、目标、架构）固定的归纳偏置决定；不存在对一切问题都最优的归纳偏置。

支撑：“没有免费午餐”定理表明，在所有可能的问题上平均，任何学习算法都不优于其他算法；一种偏置在某些问题上的优势，必然以在另一些问题上的劣势为代价（Wolpert, 1996）。经验上，大模型在没有外部反馈时难以可靠地自我纠正推理错误（Huang et al., 2024）：与自身先验一致的错误，从内部看不出来。

公理二（新知识诞生时是反共识的）　一个真正新的想法，在提出时，在当时的主流先验下概率很低；否则它就不是新的。

支撑：库恩把科学革命描述为对常规科学范式的突破（Kuhn, 1962）；Barber（1961）记录了科学家对新发现的系统性抵制；Azoulay 等（2019）发现，明星科学家意外去世后，其领域中来自圈外者的贡献显著增加——共识的守护者离开之后，反共识的想法才得以进入。

公理三（生成廉价，检验昂贵，且检验成本因领域而异）　产生候选想法的成本远低于确认其真伪的成本；在形式化领域，检验可以自动化，在实验科学中，检验昂贵而缓慢。

支撑：在有自动评估器的领域，“大模型生成 + 评估器把关”的循环已产出新结果（Romera-Paredes et al., 2024）。与此同时，数学界的瓶颈已转向验证与消化：陶哲轩在 ICM 2026 的演讲中称 AI 证明的产出超过人的消化能力为“证明消化不良”（Tao, 2026）；《莱顿宣言》把“难以与正确证明区分的、看似合理但不可靠的论证”列为首要风险（Leiden Declaration, 2026）。

4　第一定律：认知惯性（复数律）

> **任何心智若不受到成形方式不同的心智作用，将保持在其训练先验所划定的范围之内。因此，能发现的 AGI 必然是复数的。**

4.1 推导

由公理一，单个系统的假设空间和评价标准都受其先验约束；由公理二，新想法在主流先验下概率低。两者合起来意味着：一个以主流语料训练、再以人类偏好调整的系统，会系统性地低估新想法，并且无法用自身发现这一点。能把它推出先验的，只能是外部的、以不同方式形成的另一个心智，或者外部的检验（第三定律）。

4.2 来自科学社会学与社会认识论的证据

认知分工。Kitcher（1990）证明，如果共同体中每个人都理性地选择当前最有希望的理论，整个共同体的表现反而不是最优的；共同体需要一部分人投入看起来不那么有希望的方向。Strevens（2003）进一步说明，科学的优先权规则正是把人分散到不同方向的一种激励。在 Weisberg 与 Muldoon（2009）的“认识地形”模型中，少数不随大流的“独行者”能显著提高群体发现高峰的效率。

客观性是社会成就。Longino（1990）认为，科学的客观性不来自个体的无偏，而来自不同视角之间的批评性互动；它要求批评的场所、对批评的回应、公共标准，以及“有节制的智识权威平等”。密尔早在 1859 年就指出，只知道自己一方观点的人，对自己这一方也知之甚少（Mill, 1859）。

多重发现与集体大脑。Ogburn 与 Thomas（1922）列出了 148 项被不同的人独立做出的发现；Merton（1961）据此提出，多重发现是常态，单独发现才是例外。人类学证据指向同一方向：塔斯马尼亚人与大陆隔绝后丢失了部分技术（Henrich, 2004）；大洋洲各岛屿的工具复杂度与人口规模和岛际接触相关（Kline & Boyd, 2010）。Muthukrishna 与 Henrich（2016）把这些归结为：创新速度主要取决于群体规模、连接方式和传递保真度，而不是个体智力。

团队与知识负担。过去半个世纪，团队在各学科知识生产中日益占据主导，并产出更高影响的成果（Wuchty et al., 2007）；原因之一是知识前沿不断扩张，个人必须越来越专门化（Jones, 2009）。值得注意的是，大团队擅长发展已有思想，小团队更常带来颠覆（Wu et al., 2019）：组织结构本身影响发现的类型。

4.3 来自群体决策与社会心理学的证据

独立性。孔多塞陪审团定理说，由多个独立判断组成的多数比个人更可能正确；但判断之间一旦相关，这一优势会迅速减弱（Ladha, 1992）。Page（2007）的“多样性预测定理”给出一个恒等式：群体的平方误差，等于个体平均平方误差减去个体预测之间的多样性。在 Lorenz 等（2011）的实验中，让参与者看到他人的估计，降低了估计的多样性，却没有提高准确性。

组织学习。March（1991）的模型表明，个体过快地被组织同化会减少探索，长期反而降低组织的知识水平；保留一部分学得慢、与组织共识不同的成员，有利于整体。Zollman（2007, 2010）在科学共同体的网络模型中得到类似结论：在困难问题上，连接过密的共同体更容易一起过早收敛到错误答案，“暂时的多样性”具有认识论价值。

真实的异议。Nemeth 等（2001）比较了真正持不同意见的成员与被指派扮演“魔鬼代言人”的成员，发现前者更能激发群体的发散思考。这对 AI 有直接含义：给同一个模型安排不同人设，近似于指派魔鬼代言人；不同训练谱系的模型，才更接近真实的异议。

4.4 来自 AI 研究的证据：同质化正在发生

对 350 多个大模型的分析发现，当两个模型都答错时，约 60% 的情况下它们给出同一个错误答案；而且模型越大、越准确，错误越相关，即便来自不同架构和不同厂商（Kim et al., 2025）。在开放式问题上，不同模型的回答也表现出惊人的相似，研究者称之为“人工蜂群思维”（Jiang et al., 2025）。在人身上，生成式 AI 提高了个体作品的创造性评分，却降低了群体作品之间的多样性（Doshi & Hauser, 2024）。理论上，当所有决策者使用同一个算法时，即便这个算法更准确，整体结果也可能更差（Kleinberg & Raghavan, 2021）；模型若反复以自身生成的数据训练，会丢失分布的尾部（Shumailov et al., 2024），而新想法恰恰生活在尾部。

多智能体辩论的实证结果并不一致：只更换底座模型并不总带来提升（M3MAD-Bench, 2026），辩论有时反而让结果变差（Talk Isn’t Always Cheap, 2025）。在第一定律看来，这并不构成反例：现有的多数设置使用同一或相近谱系的模型，并让智能体在首轮就看到彼此的回答，恰好违反了下文的成立条件。

4.5 成立条件

- **独立**：参与者的先验来自不同的数据、目标、架构或经历，并且首轮判断彼此独立。
- **投入**：参与者持有持续的立场，而不是临时扮演的人设。
- **对等**：没有哪个参与者的权威压倒其他人，以致批评失去作用。

4.6 两个推论

推论 1：不存在“上帝级”的单体。即使存在一个能力远超其余的智能体，只要它的权威是压倒性的，共同体的纠错功能就会失效：Longino 所说的“有节制的权威平等”不再成立，而人在面对自动化系统时普遍存在过度依赖（Parasuraman & Manzey, 2010）。上帝级单体的问题不在于它不够聪明，而在于没有人能纠正它。

推论 2：一个更弱但不同的心智，可能比一个更强但相似的心智更有价值。由多样性预测定理，加入一个与现有成员高度相关的强者，群体误差下降有限；加入一个误差不相关的弱者，下降可能更多。当前沿模型日益趋同（4.4 节），人作为训练方式最不同的心智，其认识价值就不只是情感上的，而是结构上的。这也是本文与菲尔兹奖得主声明的和解点：人不是将被替换的旧部件，而是复数群体中最不同的先验来源。

5 第二定律：猜想

> **对共识的偏离幅度，随猜想的力度增大，随共识的“质量”——即被训练与偏好调整固化的程度——增大而减小。**

5.1 推导

由公理二，新想法在主流先验下概率低；因此任何“只输出高概率内容”的过滤，都会把新想法和错误一起滤掉。降低采样温度、以人类偏好把输出推向稳妥的平均答案、对“幻觉”一律施加惩罚，都在减小猜想的力度。共识越牢固，需要的力度越大。

5.2 证据

波普尔把科学知识的增长描述为大胆的猜想与严格的反驳（Popper, 1963）；坎贝尔把创造性思维描述为“盲变异与选择性保留”，与生物进化同构（Campbell, 1960）。史例中，麦克斯韦的位移电流最初出自他 1861–62 年以“分子涡旋与惰轮”构造的机械以太模型；这个模型后来被抛弃，留下的那一项却是对的。关于睡眠与创造力的实验发现，睡眠后发现隐藏规律的人数是保持清醒者的两倍以上（Wagner et al., 2004），入睡初期是创造性发现的“甜区”（Lacaux et al., 2021）。神经科学的 REBUS 模型提出，某些意识状态通过放松高层先验对底层信号的约束而改变信念（Carhart-Harris & Friston, 2019）；本文只借用这一机制作类比。

在 AI 一侧，Kalai 等（2025）指出，大模型“幻觉”的部分原因在于训练和评测奖励猜测、不奖励承认不确定。这意味着，问题的根源在于猜想与断言没有分开，而不在于生成本身。

5.3 成立条件

- **猜想与断言分离**：每条输出标明其认识论状态（猜想、推断、已验证），猜想不以事实的口吻进入断言。
- **有方向的远距离重组**：好的猜想多是把一个领域的结构移植到另一个领域（如麦克斯韦以力学类比电磁），而不是随机噪声。
- **以第三定律为前提**：没有检验，被许可的猜想只会增加噪声。

6 第三定律：反作用（证伪律）

> **每一个断言都要遭遇一个与之相当的反向检验；没有经过反向检验的断言，不进入知识。**

6.1 推导

由公理三，生成廉价而检验昂贵，因此发现的速度主要受检验能力约束，而不是受生成能力约束。一个只会生成、不会检验的系统，会以越来越快的速度产生无法分辨真伪的候选。

6.2 证据

拉卡托斯把科学进步描述为研究纲领之间的竞争，其中包括大量对“退化”纲领的放弃（Lakatos, 1978）。这部分工作大多不可见：负结果很少发表（Rosenthal, 1979）；在社会科学中，未得到显著结果的研究多半根本没有写成论文（Franco et al., 2014）。在 AI 一侧，FunSearch 等系统的成功恰恰建立在自动评估器之上（Romera-Paredes et al., 2024）；2026 年的 Navier–Stokes 结果同样依赖 Lean 的形式化检验。

6.3 推论：多样性是检验的替代品

第一定律与第三定律之间存在交互。当检验便宜且完美时（棋类对弈、可执行代码、形式化证明），同质的群体加上海量试错就足以取得突破：AlphaZero 的自我对弈，本质上是“同质群体 + 完美裁判 + 探索噪声”（Silver et al., 2018）。当检验昂贵、缓慢或含糊时（前沿实验科学），无法逐一检验所有候选，“先检验什么”就要由不同先验之间的辩论来决定，多样性的价值随之上升。网络认识论的结果与此一致：连接结构的影响主要出现在困难问题上（Rosenstock et al., 2017）。这一推论可以解释一个观察：AI 首先在数学和编程上取得突破，在实验科学上进展较慢——不一定因为智力不足，而是因为后者缺少第一与第三定律的组合。

6.4 负结果作为公共品

在复数系统中，被证伪的想法应连同证伪的理由一起进入共同记录（“证伪账本”），使其他成员不必重复试错。这是人类科学做得最差的环节之一，也可能是复数 AI 系统能做得更好的地方。

7 制度：三定律运行的场所

三定律描述的是心智之间的互动，而互动需要场所。在 17 世纪，这个场所是皇家学会，以及其秘书奥尔登堡创办的《哲学汇刊》：它们提供了记录、通信、发表和优先权的裁定。Merton（1942）把科学的规范概括为公有性、普遍性、无私利性和有组织的怀疑；他后来指出，优先权之争不是个人品格问题，而是以“谁先发现”为货币的奖励制度的结构性产物（Merton, 1957）。陶哲轩在 ICM 2026 的演讲中把同行评审描述为把个人成就转化为集体进步与理解的方式（Tao, 2026）；菲尔兹奖得主声明的核心担忧，正是写作、引用与传承这些制度环节在仓促发布中被跳

过。复数的 AI 群体同样需要信用分配、仲裁与共享记忆。本文暂不把它列为第四条定律，而视之为三定律成立的边界条件；是否需要一条“第零定律”，留待讨论。

8 反驳与回应

反驳一：一个足够大的模型可以在内部包含所有视角。多专家架构、多人设提示似乎可以在一个模型内部制造多样性。回应：若这些视角来自同一次预训练，它们共享先验、错误高度相关（Kim et al., 2025），更接近被指派的魔鬼代言人，而非真实的异议（Nemeth et al., 2001）。若一个系统内部真的装有多个独立训练、彼此真正反对的子系统，那它就是一个复数组织——这正是本文的主张，与装在几个盒子里无关。

反驳二：AlphaZero 和 2026 年的形式化证明说明，同质系统也能发现。回应：这些成功出现在检验便宜且完美的领域，与 6.3 节的推论一致。而且以 Navier–Stokes 为例，问题由人提出，基础由人奠定，方向受对手团队影响：同质群体解决了一个已被复数共同体界定好的问题，而不是独自界定了它。

反驳三：人类史上存在孤独的天才。牛顿的“奇迹年”常被引为例证。回应：牛顿依赖开普勒、伽利略、惠更斯，也依赖胡克的那封信；平方反比律本身就是多重发现（2.1 节）。多重发现的普遍性（Merton, 1961）和团队的日益主导（Wuchty et al., 2007）表明，孤独天才更像是叙事上的简化。

反驳四：多智能体辩论的实证结果并不一致。回应：这是对第一定律成立条件的检验，而不是对定律的反驳。如果在满足独立、投入、对等三个条件之后，跨谱系群体仍不优于同等算力的单一谱系系统，第一定律就应被修正（见第 9 节预测 1、2）。

反驳五：复数的 AI 会因合成数据而趋同，甚至合谋。回应：这是真实的风险，也正是本文的政策含义所在：训练谱系的多样性不会自动维持，需要被当作公共品来维护（第 11 节）。

反驳六：复数系统更难控制。回应：复数性带来相互监督的可能，也引入协调与合谋的风险。本文不讨论安全问题的全貌，只指出：单体并不因为是单体就更安全——没有人能纠正它，本身就是一种风险。

9　可证伪的预测

#	预测	来自	检验方式	什么结果会反驳它
1	在同等算力下，跨谱系模型群体的错误相关性更低、纠错率更高，优于同一模型的重采样或多人设	第一定律	基准题与开放问题上的对照实验	两者无差异，或方向相反
2	首轮独立作答优于顺序暴露（先看他人答案再作答）	第一定律（独立条件）	同题 A/B 对照	顺序暴露同样好或更好
3	在群体中，一个更弱但不同谱系的模型，边际价值可能高于一个更强但同谱系的模型	推论 2	消融实验，分别控制个体准确率与错误相关性	边际价值只由个体准确率决定，与相关性无关
4	有检验器时，放开生成（并强制标注）产出更多经验证的新结果；没有检验器时只产出更多错误	第二 × 第三定律	2 × 2 设计	两种条件下没有交互
5	检验成本越高，复数性带来的增益越大	6.3 节推论	跨领域比较：形式化 / 仿真 / 实验	增益与检验成本无关
6	AI 先在形式化可验证的领域突破；在检验昂贵的实验科学中，起决定作用的将是多谱系加实验在环的系统	全部	对未来几年 AI 科学成果的历史观察	单一谱系系统在高检验成本领域持续独立取得突破

10　实践者田野观察（方法；结果待补）

为了在受控实验之外获得真实情境中的证据，一位实践者在日常工作中采用由三定律导出的做法（多谱系独立首答、认识论标签、事前杀死条件、证伪账本），并以统一格式记录每一次使用：问题类型、参与者谱系、首轮是否独立、首轮分歧程度、少数意见、决定、事后检验的方式与结果、多样性是否改变了结论、检验成本。方法上属于行动研究与反思性实践的传统（Lewin, 1946；Schön, 1983）。

计划在 8–12 周内积累 40–60 条记录，报告实质分歧率、分歧纠错率、少数意见命中率，以及多样性价值与检验成本之间的关系。局限包括：单一实践者；问题的选择并非随机；没有对照组；工作内容涉密，因而只报告汇总统计与去标识的案例。这些数据只用于说明三定律在实践中的可操作性与初步方向，不作为因果证据。

［结果：待补］

11　含义

对 AI 开发：目标应是建设一个认识生态——由不同训练谱系的模型和人组成，配有独立首答、认识论标签、证伪账本和仲裁规则——而不是一个越来越大的单体。一个训练方式不同的较弱模型，可能是前沿系统最有价值的同伴。

对 AI 治理：认知单一化是一种系统性风险。训练谱系的多样性（不同语言与文化的语料、不同的训练目标、不同的机构、开放模型）应被当作公共品来维护。这与《莱顿宣言》呼吁投资公共计算基础设施、避免研究方向被少数公司主导的关切一致。

对科学共同体：人的角色不是被替代，而是像哈雷那样——提出好问题，居间仲裁，做品味判断，并把结果传承下去。

12　结论

新知识从来不是一个心智独自产出的。它产生于不同先验之间的碰撞，依赖被许可的猜想，并由检验来裁决。三百四十年前，平方反比律诞生于胡克的猜想、牛顿的证明、哈雷的提问与皇家学会的制度之间；2026 年的 Navier–Stokes 之争表明，这一结构并没有因为 AI 而改变。如果 AGI 指的是能发现的智能，那么它不会是一个模型，而会是一个社会。

声明

AI 使用：本文在构思与起草中使用了多个大语言模型参与讨论和文字生成，具体分工将在定稿时说明；所有论断与引用由作者负责核对。

参考文献

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.