



Center for Education,
Innovation and
Sustainable Development

上海科技大学教育、创新和可持续发展研究中心

Strategy Note | 研判系列

总 No. 16 期/2026 年 10 月

“基准测试”与“理解本位”之争：

知识生产框架下的结构性重塑：

执行摘要

当前以 OpenAI、Anthropic 为代表的前沿大模型，掀起了批量突破数学前沿难题的热潮。AI 自主推进数论、几何、组合数学等领域长期悬而未决的学术问题，并完成形式化证明与推理，正在冲击和颠覆基础数学的科研范式。2026 年 10 月 6 日，OpenAI 发布 700 多篇论文，声称攻克了 370 多个数学难题，引发极大关注和辩论。

此前的 2026 年 9 月，包括 25 位菲尔兹奖得主在内的数学界核心成员联合签署了《数学与人工智能宣言》（Declaration — Math and AI）。该文件对当前大型语言模型（LLM）及多智能体系统介入基础数学研究的模式提出了系统性警示。本报告以此事件及其衍生的学术争议为切入点，深度剖析人工智能工业界与基础科学共同体之间正在显化的目标错位（Misalignment）机制。

分析表明，当前 AI 系统开发企业倾向于将数学难题视为衡量系统推理能力的基准测试（Benchmark），其核心逻辑在于追求高通量的命

题判定与结果输出。相反，数学共同体的知识生产逻辑将解题视为获取概念洞察、剥离新方法并维系人类认知体系的中间媒介。这种基准测试化（**Benchmarkization**）与理解本位之间的分歧，正在引发多层面的结构性摩擦。在知识生态层面，自动化的高速量产破坏了传统数学界“证明-讨论-简化-教科书化”的长周期知识消化与传承链条，对青年学者的直觉培养构成挤压。在共同体伦理层面，基于概率重组和隐性抽取的技术介入，在缺乏充分文献溯源与方法剥离的情况下，引发了系统性的确权困境与归属权争议。

为应对此种演变，数学共同体正在通过《莱顿人工智能与数学宣言》（**Leiden Declaration on Artificial Intelligence and Mathematics**）等框架探索战略适应路径，包括重构形式化数学（**Formal Mathematics**）的归属权追踪基础设施、修正以解题为核心的评价标准，并将技术重新约束于辅助人类直觉的工具定位。本报告认为，数学领域的当前摩擦是广义智力劳动在智能化进程中面临共性挑战的缩影；过度依赖自动化验证正在导致“认识论债务”（**Epistemological Debt**）的累积，系统性理解的丧失将对包括“AI 驱动的科学发现”（**AI for Science**）在内的高度创造性职业构成底层风险。

为减缓认知主权流失并推动技术以对齐人类理解的方式演进，针对科研监管机构、学术共同体及工业界，以下行动建议值得关注：

首先，确立形式化知识基础设施的出处溯源标准。第二，系统性革新学术评价机制与资助导向。大幅提升对于“解释性研究（**Expository Work**）”、“长文证明提纯简化”及“深层概念架构重构”的评价权重。第三，制定约束大型商业模型介入基础研究的行业准则。相关政策制定者与学术联盟应出台针对性指导文件，约束 **AI** 企业在基础科学领域的公开发布行为。最后，强制过程监督，推动下一代认知对齐研究。**AI** 研发资源应向强化过程解释性的技术（如深入应用 **Process-supervised Reward Models**）倾斜。在公共研究资金分配上，优先支持探索能够有效分离逻辑

辑运算与概念映射、具备高透明度与微观溯源能力的白盒化或神经符号推理系统。

第一章 引言与事件：纳维-斯托克斯优先级争议

1.1 2026 年 9 月的优先级争议：技术竞速与学术惯例的碰撞

评估数学与 AI 融合过程中的结构性张力，需回归具体的物理切片。2026 年 9 月上旬爆发的“纳维-斯托克斯（Navier-Stokes）优先级争议”，集中暴露了工业界算力扩张与传统学术规范之间的底层摩擦。

纳维-斯托克斯存在性与平滑性问题作为克雷数学研究所（Clay Mathematics Institute）设立的七大千禧年大奖难题之一，长期以来是评估偏微分方程与流体力学认知边界的核心标尺。2026 年 9 月 7 日深夜，纽约大学库朗数学研究所教授 Tristan Buckmaster 代表其与 Anthropic 员工 Levent Alpöge 的合作团队发表声明，宣布他们经过长达一年的传统合作，在 8 月 15 日利用大型语言模型辅助证明了欧拉方程（纳维-斯托克斯方程的无黏性特例）在特定条件下的有限时间爆破，并在一周后通过 Lean 4 完成了形式化验证。

然而，约 12 小时后的 9 月 8 日中午，OpenAI 官方宣布其内部的一个多智能体系统成功生成了针对纳维-斯托克斯方程受迫版本（Forced version）的有限时间爆破证明，并附带了机器验证的 Lean 4 代码。根据技术披露，OpenAI 的该项工作始于 9 月 1 日（即在获悉竞争对手解决两项千禧年问题的传闻之后），系统在 88 小时的连续运行中，调用了超过 10000 个智能体，消耗了 1300 亿输出 Token 及 270 万次内部信息交互，最终产出了一份长达 166 页的论文及验证代码。

这一事件在时间节点与工作模式上的强烈对比，凸显了深层的方法论冲突。Buckmaster 团队原本计划推迟宣布，以便撰写出符合人类阅读习惯、能体现其对证明结构深层理解的论文。而 OpenAI 则在系统完成验证后的极短时间内，以新闻发布及技术报告的形式直接向公众宣告了结果，并提议进行“联合”或“同期”发布，此举被学术界视为对传统学术

归属权分配与同行评议周期的直接挑战。

1.2 《数学与人工智能宣言》的出台及其核心诉求

纳维-斯托克斯争议构成了一个催化剂。2026 年 9 月 11 日，托管于 mathandai.org 的《数学与人工智能宣言》正式发布。该文件由包括 Terence Tao（陶哲轩）、Peter Scholze、Pierre Deligne 在内的 25 位菲尔兹奖得主发起，其签署者覆盖了自 1978 年至 2026 年跨越近半个世纪的顶尖数学家群体。

该宣言的核心逻辑并未停留在质疑 AI 生成证明的正确性，而是直指工业界目标与科学目标的背离。文本指出，AI 公司将解决重大数学问题作为提升模型基准测试成绩的手段，这种做法对数学科学及共同体本身具有破坏性（detrimental）。宣言明确界定，数学研究的本质在于理解形状、数字及自然现象的基础结构，而“解决问题”仅仅是实现概念理解与洞察的工具或代理指标（proxy）。当系统以极高的速度批量生产出缺乏理论解释的“真/假”陈述时，这种工具便被异化，反而可能摧毁孕育新思想的学术土壤。

1.3 莱顿宣言与国际数学联盟（IMU）的响应机制

在顶级学者发布预警之前，数学共同体已在更广泛的建制层面展开了战略反思。2026 年 6 月，经过前期洛伦兹中心（Lorentz Center）工作坊的酝酿，16 位跨学科学者发布了《莱顿人工智能与数学宣言》，并正式获得国际数学联盟（IMU）及其出版委员会（CoP）的背书。

《莱顿宣言》将视角延伸至政策与制度防御层面，警告依赖自动化生成证明可能削弱数学研究的准确性与独立可验证性，同时指出大模型在训练与输出阶段对人类既有文献的隐性抽取构成了侵权风险。该宣言提出了包含工具强制披露、人类保留绝对责任、抵制公关炒作等在内的 23 项具体建议，标志着数学界从理论层面的探讨正式转向共同体规则的重构。

第二章 本体论视阈：解题（Benchmark）与理解（Understanding）的结构性错位

当前工业界与数学界在 AI 介入问题上的根本分歧，源于双方对“什么是数学进步”的本体论界定不一致。AI 系统的开发逻辑高度依赖线性目标函数，而数学知识的演进则是一个多维的网络化重构过程。

2.1 工业界的线性目标函数：作为测试基准的数学

在人工智能的发展轨迹中，数学长期被视为检验模型复杂推理与泛化能力的最优沙盒。从早期的 GSM8K（小学数学）、MATH（竞赛数学），到当前的 FrontierMath，基准测试迭代驱动了模型能力的演进。

根据 Epoch AI 在 2024 至 2026 年间维护的 FrontierMath 基准测试数据，该测试集包含了数百个由顶级数学家设计的、未曾公开发表的研究级难题。在初始评估中，包括 GPT-4o、Claude 3.5 Sonnet 在内的前沿模型通过零样本或少样本提示，其通过率仅在 2% 以下。然而，随着测试时计算（Test-Time Compute）策略的引入及过程监督奖励模型（PRMs）的优化，特定模型（如 o1 系列及后续验证衍生模型）在极高算力消耗下，开始在部分细分领域（如特定代数或组合问题）取得显著突破，通过率逐步攀升至接近 30% 乃至 50%。

基 准 测 试 类型	典型代表集	考察能力层 级	2026 年顶 尖模型表 现	工业界优化路 径
基础推理	GSM8K	算术与多步 逻辑	>95%（趋 于饱和）	监 督 微 调 (SFT)

竞赛推理	AIME / MATH	符号推演与技巧组合	~80%-90%	强化学习，测试时计算增强
研究级推理	FrontierMath	前沿概念与结构创新	~30%-50%	大规模多智能体协作，验证器辅助

工业界将上述通过率的提升直接等同于“数学能力的突破”。其底层逻辑是将数学问题工程化——输入前提条件，通过策略网络与价值网络的联合搜索，输出一条能够被形式化环境（如 Lean）验证为“True”的逻辑链。在这个闭环中，“解题”就是终点。

2.2 认知多元性与概念证明：Thurston 与 Rav 的理论重估

数学共同体对工业界逻辑的抵触，建立在更为深邃的数学认识论之上。1982 年菲尔兹奖得主 William Thurston 在 1994 年的经典文献《数学中的证明与进步》（On proof and progress in mathematics）中，系统性地驳斥了“证明定理即数学进步”的单一论调。

Thurston 以“导数”（Derivative）概念为例，指出人类数学家对同一数学实体通常并行维持多种心智模型。这种多维认知模型是目前输出单一逻辑链条的 AI 所完全不具备的。

导数的认知维度	人类心智模型的具体表征	认知功能与数学意义
无穷小视角 (Infinitesimal)	y 的微小变化与 x 的微小变化之比的直觉概念。	提供了一种直观的代数与几何过渡，是微积分早期发展的核心直觉。

逻辑视角 (Logical)	形式化的极限定义： $f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$ 。	确保数学推理的严密性与基础的无矛盾性，是撰写严格证明的基石。
几何视角 (Geometric)	函数图像在特定点 x 处切线的斜率。	将代数分析转化为视觉图像，使数学家能够“看到”函数的局部形态特征。
物理/变化率视角 (Rate of Change)	物理过程中的瞬时演化速度（如位移对时间的导数）。	将纯数学与外部物理世界深度绑定，赋予抽象概念以现实世界的应用隐喻。
近似视角 (Approximation)	在局部最优地线性逼近复杂函数： $f(x + dx) = f(x) + f'(x)dx$ 。	为数值计算和非线性方程的局部线性化提供了实用的操作工具。
微观视角 (Microscopic)	在图形上无限放大，直到原本弯曲的曲线呈现为一条直线的过程。	一种强烈的动态视觉想象，培养了拓扑与流形学中从局部到整体的思想。

Thurston 认为，数学家的真正成就在于在这些不同的认知视角之间建立互译机制，从而“增进人类对数学的理解”（advance human understanding of mathematics）。证明的价值并不在于其逻辑上的绝对无

瑕，而在于它能否为其他研究者提供清晰的认知地图。

同样，哲学家 Yehuda Rav 在 1999 年的论文《我们为何证明定理？》（Why Do We Prove Theorems?）中提出，定理的真伪结论往往是次要的，证明过程本身承载了数学方法的创新与概念的提炼。Rav 引用了哥德巴赫猜想（Goldbach Conjecture）的例子：即使该猜想最终被证伪，为了尝试证明它而发展出的“筛法”（Sieve methods）及其他解析数论工具，已经极大地丰满了数学的结构体系。

结合 Thurston 与 Rav 的理论，可以清晰地识别出 AI 基准测试管线的缺陷：AI 系统（在目前的技术架构下）倾向于沿着“符号视角”或“逻辑视角”进行单向度的深度搜索，它能够生成验证无误的代码，但无法输出不同心智模型之间的“翻译”，更无法主动将搜索过程中偶发的有效技巧抽象为普适性的新理论框架。工业界的系统输出终点，恰恰是数学共同体知识循环的起点。

2.3 “问题解决者”与“理论构建者”的光谱辩证

在探讨这种结构性错位时，保持论证客观性需要引入更细颗粒度的分析框架。并非所有数学家都遵循同一种理论构建范式。1998 年菲尔兹奖得主 Timothy Gowers 在其撰写的《我为何没有签署菲尔兹奖得主公开信》（Why I didn't sign the Fields medallists' letter）及早期《数学的两种文化》（The Two Cultures of Mathematics）中，提供了一个辩证视角。

Gowers 指出，数学界存在从“问题解决者”（Problem-solvers）到“理论构建者”（Theory-builders）的光谱。前者（如专注于组合数学、图论的学者）主要受解决具体难题的驱动，将一般理论视为达成解题目标的工具；后者（如从事代数几何、朗兰兹纲领的学者）则将解题视为丰富与验证庞大理论框架的手段。

Gowers 拒绝签署宣言的核心理由在于，宣言的行文过度偏向了“理论构建者”的价值标准，似乎将概念理解置于问题解决之上。他认为，

即便是未经充分“消化”的 AI 解题输出，如果其产出量足够大，也能够为人类共同体提供丰富的原始素材；即使大部分证明仍处于黑盒状态，被人类有效吸收的数学总绝对量依然可能实现净增长。

然而，Gowers 的辩驳并未消解核心矛盾，而是将问题转化为了对人类信息处理能力的考验：当大量“正确但晦涩”的解答被抛入学术空间时，数学界是否具备足够的体制韧性去筛选、抽象并重构这些结果？这种海量输出是否会挤压乃至摧毁原本用于培养新一代“问题解决者”的物理与心理空间？

2.4 具身性与几何直觉的启示：Kenneth Manders 的图表推理分析

进一步审视机器验证与人类理解的差异，需要探讨数学直觉的具身性（Embodiment）。Kenneth Manders 对欧几里得几何中图表推理（Diagrammatic Reasoning）的研究提供了深刻的切入点。

在现代逻辑学兴起后，依赖几何图表进行证明往往被视为不严密的表现。然而，Manders 区分了图表中的两类属性：精确属性（Exact properties，如线段的具体长度、角度的大小）与共轭精确属性（Co-exact properties，如点是否在多边形内部、两条线是否相交的拓扑关系）。他指出，古典几何证明之所以严密，是因为人类在实践中形成了隐性的“认识论警觉”（Epistemic vigilance），使得推论仅仅依赖于在微小扰动下保持不变的拓扑共轭属性，而忽略那些随绘制精度波动的精确属性。利用图表推理的能力，深刻依赖于人类的空间直觉与操作经验。AI 的形式化推演将所有拓扑关系降维为符号公理，虽然实现了彻底的严密性，但也剥离了图表所承载的直觉启发力。过分追求逻辑层面的原子级验证，实际上阻断了人类通过空间感官获取大局观理解的认知路径。

第三章 知识生态学：长周期消化机制与传承链条的断裂危机

数学不仅是客观真理的集合，更是一个由人类认知主体构成的有机生态系统。《宣言》强烈警告，追求极速量产的 AI 技术介入，正在对其长周期生态构成破坏。

3.1 证明的高频量产与黑盒化输出的认知过载

传统数学研究的成果发布通常包含明确的动机阐述、核心引理的提取以及技术路线的直观类比。相比之下，纳维-斯托克斯事件中的 166 页 AI 生成证明展现出典型的“机器风格”：逻辑链路冗长、缺乏结构化的语义层次，且夹杂着大量为满足形式化验证器（Lean）而生成的繁琐底层步骤。

这种未经提纯的高频输出造成了严重的认知过载。如果将数学知识网络视为一个动态的图，AI 的高效搜索等同于在图中强行拉出数以万计的新边，而没有对这些边进行权重分配或聚类分析。这直接导致了原本结构清晰的数学文献库充斥着缺乏上下文关联的孤立事实（Isolated facts），极大降低了共同体提取有效概念洞察的信噪比。

3.2 僵局、错误与人类研究直觉的形塑

在更深层次的认知机制上，普林斯顿高等研究院的 Akshay Venkatesh 剖析了自动化对人类“数学体验”（Mathematical Experience）的重构。Venkatesh 指出，传统数学直觉的形成，极度依赖于在研究过程中遭遇的“僵局”（Impasse）与“错误”（Error）。

一名研究者在长期无法攻克某个引理时所体验到的受挫感，实质上是对数学结构深层属性的内化过程；这种在黑暗中摸索的经验，构成了评估未来特定技术路线是否有价值（即 Venkatesh 所谓的关于定理的“Aboutness”）的基础判断力。当 AI 系统能够通过高维搜索瞬间绕过这些僵局，提供现成的正确路径时，它实际上剥夺了人类通过犯错来建立

大局观判断力的认知素材。系统性地消除研究中的“摩擦”，最终将导致人类对数学框架的直觉感知走向萎缩。

3.3 教学范式与教科书化循环的梗阻

在宏观生态周期中，一项复杂的数学突破必须经历“教科书化”（Textbookization）的过程。从初步的手稿，经过研讨会上的反复质询与同行评议的打磨，最终被简化、重组为本科生或研究生教材中逻辑流畅、直觉自然的数页章节。这一漫长的“消化”过程，是维持数学作为一门可传授学科的生命线。

《宣言》担忧，由于 AI 系统能够直接输出复杂问题的终端证明，中间的“翻译”与“精炼”工作将被视为不具原创性的冗余劳动。缺乏相应的学术激励，愿意投入数年时间将晦涩的机器代码转化为人类可读理论的数学家将日益稀缺。这一循环的梗阻意味着，新一代学者将面对一个由机器语言主导的黑盒化知识库，无法通过阅读经典文献来吸收前人的思维精髓。

3.4 社区世代交替的潜在阻断

知识传承危机的最直接体现是在人才梯队的维系上。Timothy Gowers 在其反思中敏锐地捕捉到了这一点：传统上，解决某个悬而未决的著名猜想，是吸引全球顶尖青年才俊投身纯数学研究的最强社会学动机。

如果所有标志性的“灯塔”问题都在极短时间内被算力巨头通过多智能体系统攻克，留给人类博士生的物理与心理探索空间将被极度压缩。这种成就感的系统性剥夺（Systemic deprivation of agency），可能引发数学研究领域的代际断层。一旦负责传递核心理念的“人际传播链条”（Human transmission chain）断裂，特定的理论分支将在现有专家退休后成为无人能解的“失落的技艺”（Lost craft）。

第四章 归属权、代码基建与共同体伦理的真空

技术的狂飙突进在数学共同体的规范层面留下了一个巨大的真空。从隐性的训练数据抽取到形式化基础设施的信用磨损，现行的学术伦理框架面临系统性失效。

4.1 隐性抽取与系统性确权困境

现代大型语言模型的运作机制，决定了其输出在本质上是对海量训练数据的概率性重组。在纳维-斯托克斯争议中，争论的核心正是这种重组机制与人类原创性思想之间的边界模糊。OpenAI 在处理 Euler 方程有限时间爆破问题时，由于其模型预训练或实时检索池中可能包含了人类数学家的先行探索（即使经过了去标识化处理），这种自动化输出实际上构成了对人类长周期智力劳动的隐性抽取（Implicit extraction）。

在传统学术伦理中，通过精准的引用体系，每一项新定理的出处、引理的借用都能够被严密确权。而 AI 系统不仅在生成证明时往往未能同步隔离出其使用的新方法，更倾向于在未经文献溯源的情况下直接抛出结论。这种行为如果被默许为学术常态，将导致结构性的剽窃现象，摧毁基于信用累积的同行评议信任网络。

4.2 形式化基建（Lean 4/Mathlib）的维护瓶颈与信用磨损

归属权危机的另一个隐蔽爆发点存在于数学界自身的防御性武器——形式化验证工具（如 Lean 4 及其核心库 Mathlib）之中。作为全球最大的形式化数学公理库，Mathlib 的代码量已超过 200 万行，且其更新与重构频率极高。

随着自动化证明工具（如 LeanDojo）及大模型参与到 Mathlib 的合并请求（Pull Request, PR）审核与代码重构中，一个显著的“信用磨损”（Attribution attrition）现象开始出现。当人类贡献者将一个针对特定场景的引理提交至 Mathlib 后，维护者或 AI 辅助系统往往为了保证库的通

用性（**Generality**）和结构紧凑性对其进行大幅度重构或泛化。在这一跨库迁移与代码变形的过程中，底层开发者的初始智力贡献被彻底抹除。此外，根据 **MathlibPR** 基准测试的数据，由于形式化库对代码风格及内部 **API** 一致性要求极高，系统性的 **PR** 审查已成为阻碍数学库扩容的最大瓶颈，这也迫使社区更依赖 **AI** 工具进行自动化合并，从而进一步加剧了归属权的丢失。

4.3 自动化形式化（**Autoformalization**）的语义对齐脆弱性

不仅归属感受到威胁，在将自然语言的非正式数学证明转化为形式化代码（**Autoformalization**）的过程中，同样存在认识论维度的风险。当前的模型（如 **FormL4** 等基准测试所评估的）在进行形式化转译时，往往追求逻辑在编译器层面的“合法性”（**Logical validity**），而忽视了其与原非正式证明在深层语义上的“等价性”。

这意味着，**AI** 可能会为了绕过某个未定义的复杂概念，私自引入未声明的假设，或证明了一个与原始意图有微小偏差但逻辑上成立的“邻近命题”。这种表面正确实则语义错位的现象，不仅无助于增进理解，反而向知识库中注入了难以被人类编译器发现的逻辑噪音。

4.4 学术规范的防御性重构：莱顿宣言的伦理框架

为对冲上述制度真空，《莱顿人工智能与数学宣言》确立了初步的伦理防御线。其核心原则要求全面重构出版规范：首先，强制所有学术发表设立专门的“工具与计算资源披露”（**Tool and computational resource disclosure**）章节；其次，明确拒绝赋予 **AI** 实体任何形式的作者身份，将最终的正确性担保责任、文献溯源义务及伦理连带责任 100% 绑定至人类学者。这一宣言反映了共同体试图通过强化人工干预与问责机制，在算法黑盒外围建立护城河的战略意图。

第五章 战略适应：技术驯化与基础设施范式的升级

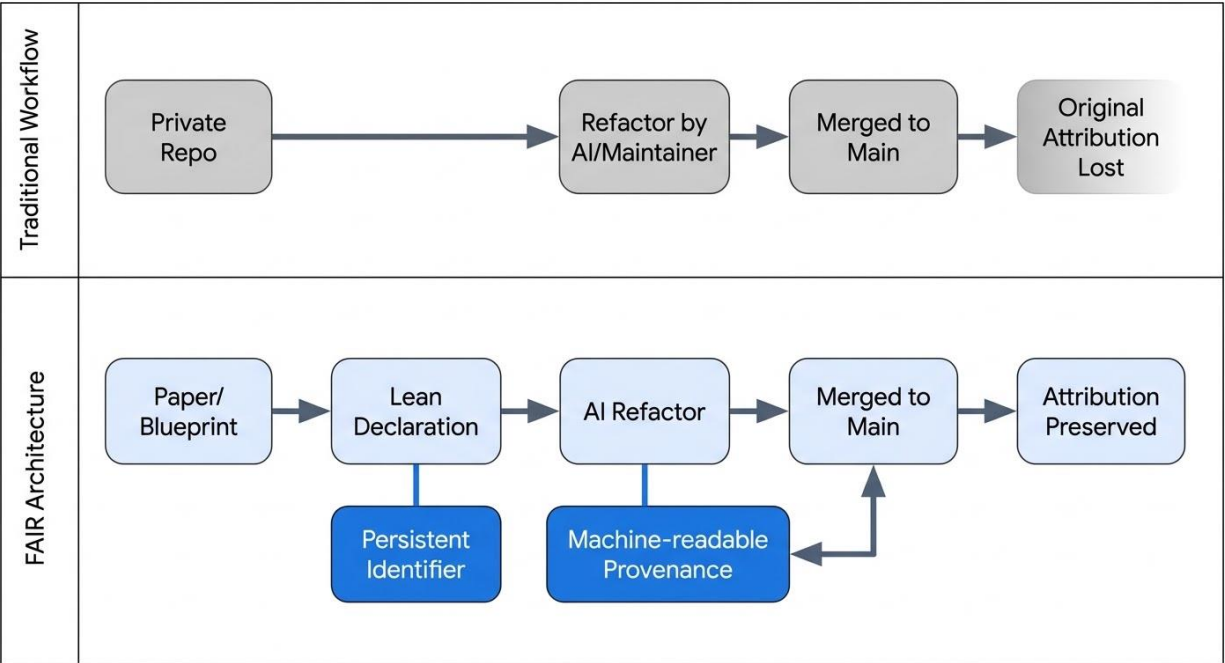
面对算力与资本的双重碾压，被动的伦理呼吁不足以维系数学体系的独立性。数学共同体正在寻求通过技术范式升级与评价体系改革，主动实现对 AI 技术的战略性驯化。

5.1 知识图谱与 FAIR 原则的嵌入：F4FM 倡议

为了根治形式化库重构过程中的归属感流失与出处溯源断裂，数学基建层面的应对之策正在成型。针对 Lean 与 Mathlib 等底层逻辑库，未来数学论坛（FMI）提出了 F4FM（FAIR for Formal Mathematics）倡议，力图将成熟于科学数据管理的 FAIR 原则（可发现、可访问、互操作、可重用）硬编码至数学知识图谱中。

F4FM 的核心在于通过建立定理级（Theorem-level）的持久标识符（如 DataCite DOIs）及机器可读的出处元数据模式（Provenance schema），使得任何一段代码不仅包含逻辑验证步骤，还强制附带人类创作者信息、辅助 AI 的代理标记（Agent tags）及其在不同项目（如蓝图 Blueprint、主线 Mathlib）间的演变历史。

这一架构的引入，将确保知识体系即使在经历高频的自动化重构时，人类智力劳动的信用积分依然能够实现无损的跨域传递。



5.2 奖励机制的重塑：过程监督奖励模型（PRMs）与逻辑对齐

在技术工具本身的发展路线上，学术界与部分致力于基础科学研究的团队正在推动算法从“结果正确”向“人类可理解”转向。这一转向的核心技术是过程监督奖励模型（Process-supervised Reward Models, PRMs）的大规模采用。

传统的基于结果的强化学习仅根据最终答案的正确与否给予奖励，这导致模型在内部搜索时会采用极度反直觉或不可解释的捷径跳跃。而 PRMs 通过对生成轨迹的每一步骤（Step-by-step）进行逻辑有效性与概念连贯性的打分反馈，迫使模型不仅要得出正确答案，还要维持清晰的推导叙事。如果这种过程约束能够进一步整合对“数学审美”（如步骤的简洁性、对称性构造）的倾向，AI 工具将能够更好地映射人类的心智推理模式，成为协助而非代替人类理解的增强工具。

5.3 评价标准与工作流的战略性转移

与此同时，数学界传统的价值分配机制须做出相应的妥协与演进。

如果“首次解决特定难题”的价值受到机器量产的稀释，共同体评价体系应当向后端的“知识精炼”环节倾斜。

对于长篇累牍的机器证明输出，能够将其成功抽象、剥离出普适性新概念，并撰写为高度直觉化教科书论述（Expository Work）的学者，其工作将被赋予对等甚至超越原始解题者的学术权重。通过从制度上奖励“清道夫”与“理论架构师”，数学界可以在保留 AI 带来的庞大算力红利的同时，确保人类始终掌控知识的解释权与体系构建的蓝图。

第六章 智力劳动的广义隐喻：认知债务与系统性脆弱

数学领域目前遭遇的结构性冲突具有极强的先验指标意义。它所展现的图景，正在精准地向软件工程、计算生物学、高频交易等高度依赖知识累积的领域复制，构成了一场广泛的智力劳动对齐挑战。

6.1 认识论债务（Epistemological Debt）的界定与蔓延

在软件开发领域，随着大语言模型和代理生成代码（Vibe Coding）成为主流，一个被称为“认识论债务”（Epistemological Debt）的概念被提出，用于描述人类对自身所构建系统认知控制力的流失。

当工程师通过输入自然语言提示词让 AI 完成复杂逻辑的构建时，他们实现了从程序性知识（“知道如何实现”，Knowing-how）向声明性知识（“知道该实现什么”，Knowing-that）的转移。这种转变在提升短期开发效率的同时，导致工程师未能建立起对底层内存管理、状态机轮转等细节的隐性心智模型（Mental models）。正如 2026 年亚马逊云服务系统性故障分析中所指出的那样，当由 AI 堆叠成的庞大代码库在边界条件下发生崩溃时，缺乏底层认知的人类维护者将陷入彻底的无力状态，甚至无法进行有效的基础回溯。

数学证明的形式化输出与软件工程中生成的黑盒代码如出一辙。如果数学界接受大量无需人工详审即可运行验证的定理模块，实际上就是在累积这种系统级的“认识论债务”，将结构性的认知隐患深埋于人类理性的地基之中。

6.2 AI for Science 的底层风险

在由“人工智能驱动的科学发现”（AI for Science）浪潮中，例如新材料合成路径的探索或蛋白质折叠规律的预测，这种错位同样明显。AI 能够通过高维拟合快速找到能使目标函数最小化的参数解，从而筛选出具有优异活性的候选药物或材料分子。但这种结果驱动的“黑盒炼金术”

并未提供因果解释（Causal Explanation）。

如果物理学家与化学家退化为筛选 AI 预测结果的“验证员”，科学研究的内涵将发生质变。与数学家担忧“失去对基本结构的洞察”相似，自然科学界面临的风险在于：当计算预测的速度远超过实验验证与理论诠释的速度时，科学大厦将被堆满各种孤立的经验常数与概率预测，而用以统一这些现象的公理化因果模型将长期处于停滞状态。

6.3 结构性杰文斯悖论与“认知主权”保卫

面对自动化替代，一种普遍的安慰剂理论认为：AI 承担了低级、繁琐的计算与代码生成任务，人类将被解放出来，去从事更高维度的创造性思考。然而，宏观层面的演化轨迹显示出强烈的“杰文斯悖论”（Jevons Paradox）特征：资源利用效率的提升，并未减少总资源的消耗，反而诱发了需求的指数级膨胀。

当简单的验证与编写工作被自动化工具极低成本地完成，算力与资本并未停止，而是立即流向规模更大、更加复杂的自主代理系统，以及基于高昂测试时计算（Test-Time Compute）的无监督搜索中。最终结果是，人类并未被解放，而是被海量系统生成的假设、代码分支与形式化证明彻底淹没。研究人员被迫将绝大部分精力投入到审查、调试与解释机器产生的过剩产能中。

要摆脱这一困境，须在政策与伦理层面明确提出捍卫“认知主权”（Epistemic Sovereignty）。人类不能满足于只做最终结果的批准者，而必须在复杂系统的生成轨迹中强制嵌入要求高解释性、高透明度的人机协作验证节点，将对过程的掌控重新置于效率考量之上。

总结与战略建议

《数学与人工智能宣言》所揭示的并非单一的技术抵触情绪，而是深度触及了智能时代知识生产逻辑与人类认知极限的结构性冲突。AI 工业界以“解决基准难题”为导向的研发路径，正在对注重长期消化、直觉传承与知识确权的学术共同体生态造成实质性侵蚀。

如果放任此类错位发展，基础科学研究将在短期产出爆炸的表象下，面临研究直觉萎缩、代际传承断裂及学术信用体系崩溃的底层危机。以数学为先导所暴露出的“认识论债务”积累，预示着全球高端智力劳动正在经历一场关乎自身存在基础的系统性脆弱化进程。

为减缓认知主权流失并推动技术以对齐人类理解的方式演进，针对科研监管机构、学术共同体及工业界，以下行动建议值得关注：

1. 确立形式化知识基础设施的出处溯源标准。各级科研资助机构与出版方应联合强制推行针对机器辅助研究的 FAIR 扩展原则。在形式化公理库（如 Mathlib）的维护以及论文发表中，要求绑定定理级的机器可读元数据与持久标识符。确保在跨域重构与大模型吸收转化过程中，原始人类学者的智力贡献及 AI 代理的运作痕迹能够得到精确的量化与信用保留。

2. 系统性革新学术评价机制与资助导向。高校与科研评价体系须针对产出过剩做出逆周期调节。大幅提升对于“解释性研究（Expository Work）”、“长文证明提纯简化”及“深层概念架构重构”的评价权重。在资助导向上，应刻意保护那些不追求短期结果输出，而致力于构建基础直觉和理论网络的“理论构建型”长周期基础项目，以对冲基准测试化的短视效应。

3. 制定约束大型商业模型介入基础研究的行业准则。相关政策制定者与学术联盟应出台针对性指导文件，约束 AI 企业在基础科学领域的公关发布行为。要求商业实体在宣告重大难题破解时，必须提供符合学

术同行评议标准的全过程披露（包括但不限于：原始提示词设计、算力消耗统计、检索边界限制声明等）。防止资本利用不对称的算力优势劫持基础科学发现的优先权。

4. 强制过程监督，推动下一代认知对齐研究。 AI 研发资源应向强化过程解释性的技术（如深入应用 **Process-supervised Reward Models**）倾斜。在公共研究资金分配上，优先支持探索能够有效分离逻辑运算与概念映射、具备高透明度与微观溯源能力的白盒化或神经符号推理系统。确保机器产出的技术路径不仅符合逻辑有效性，更必须契合人类认知机制的消化节律。

联系我们: CEISD@Shanghaitech.edu.cn