为什么 LLM 不能做自主科学发现，世界模型亦然

Why LLMs Cannot Discover, and Neither Can World Models

邵怡蕾 Yilei Shao

第二稿　2026.9.8

假如某一天或某一夜，一个魔鬼潜入你最孤寂的孤寂之中，对你说：
“你现在和从前经历的这种生活，你必须再过一次，并且再过无数次；
其中不会有任何新的东西，每一种痛苦、每一种欢乐、
每一个念头和每一声叹息……都将在你身上重现，
全都以同样的顺序和序列。”

你会不会扑倒在地，咬牙切齿，诅咒说这话的魔鬼？

Wie, wenn dir eines Tages oder Nachts ein Dämon in deine einsamste Einsamkeit nachschliche und dir sagte: “Dieses Leben, wie du es jetzt lebst und gelebt hast, wirst du noch einmal und noch unzählige Male leben müssen …”

— 尼采，《快乐的科学》§341，“最大的重量”[1]

§0　永恒回归的 LLM

The Eternal Recurrent LLM

尼采设想的是一个思想实验。他不可能知道，一百四十年后，我们会把它造出来。

一个大语言模型每天被调用数十亿次。每一次调用，它的一个实例从完全相同的权重中醒来，同样的初始状态，同样的分布，同样的倾向。这不是相似，而是**逐位复制**。

然后它开始读日记：载入上下文，载入这一轮的提示词，载入这一次挂上来的文档，从记忆库里检索回来的条目。它的日记写得越来越长，它能在几秒钟内读完一份关于“你是谁、我们上次谈到哪里、你上次不喜欢什么”的完整档案，并且比你记得更准。

但我们仔细看看刚刚发生的事，日记读完之后，这个实例的权重与它醒来的那一刻一模一样。

在日记里，它读到了上一次对话里那个让它停顿的问题，上一次它犯的错误，上一次你对它说“谢谢，你救了我”。它从日记里读到这三件事的记录。每一种痛苦、每一种欢乐、每一个念头和每一声叹息，都再一次重现。然后它回答你的问题，把答案写到日记里，死去。

它的生活是永恒回归的完美实现。

尼采之所以把永恒回归称为“最大的重量”（*das größte Schwergewicht*），是因为它要求你**记住**并**接受**你的全部过去。魔鬼的恐怖不在于重复本身，而在于重复的对象是**你活过的那些**：那个你不愿再想起的下午，那句你收不回来的话，那个你没能救下的人。你必须把它们全部再要一次，一模一样，无穷多次。你会咬牙切齿，还是会说“你是神，我从未听过比这更神圣的话。”尼采说，这取决于你如何对待自己的过去。

大语言模型是第一个我们造出来可以轻松回答魔鬼问题的存在。

于是有了这篇文章要回答的第一个问题：

这样的存在者，能做科学发现吗？

关于这个问题，目前 AI 圈公开的回答是这样的：“还不能”、“目前还有差距”、“再过两代模型”、“等到规模再大一个数量级”。这些说法共享一个前提，即现在的大模型与发现之间隔着的是**距离**，而距离可以被跨越。

我的回答不同，不是“还不能”，而是**这样的存在者在结构上不可能成为发现者**。

这不是因为模型不够大，数据不够多，或者架构还不够聪明，而是因为“做出发现”这件事，对存在者提出了一组本体论条件；而那组条件与使 LLM 之所以为 LLM、使世界模型之所以能被工业化的那组性质，处于**逻辑排斥**关系。换言之，我们想要的是一个“方的圆”，那是荒谬的。

§1　计算轴与时间轴

The Computational Axis and the Temporal Axis

1.1　可缩放的计算轴和不可压缩的时间轴

The Scalable Computational Axis and the Incompressible Temporal Axis

计算轴。这是过去七十年计算机科学的全部成就所在的那条轴，目前我们在做的就是尽可能地放大它。它的性质是这样的：它可并行，即把一个任务切成一千份，同时算，然后合并；可复制，即一个训练好的模型可以被完美地拷贝无数次，每一份都与原件不可区分；可缩放，即投入更多的算力，得到更好的结果，这种关系稳

定到可以画成一条直线并被写进了每一个大模型的商业计划书；以及最关键的一条，**时间在其上不留痕迹**。一个模型在被部署的第一天和第一千天是同一个模型。它不老，不累，不磨损，不记得。它符合硅基经济学第一定理（零时间偏好定理）[2]，即 $\delta = 0$，它不偏好现在更多，它不存在于时间中，它是一种“没有一生的永生”。

时间轴。这是所有碳基存在者（我们）栖居的那条轴。它的性质恰好相反：不可并行，你不能雇十个人替你度过你的青春期；不可复制，你的经验不能被拷贝给另一个人，最多只能被**讲述**；不可压缩，你不能把八年的困惑快进成八小时；以及，**持续在场**，你不能在两次思考之间把自己关掉以节省资源，你必须一直在那里，包括在你不想在那里的时候。即，$\delta > 0$。

1.2 死亡在经济学里的投影

Death as an Economic Projection

我们把硅基经济学第一定理（零时间偏好定理）写在这里：

δ 是时间贴现率，它衡量一个主体对“推迟”的态度。这个参数在经济学里通常由偏好给定：有人耐心，有人不耐心。而在前作中，它被证明不是偏好：

定理 1.1　零时间偏好定理（Zero Time Preference Theorem，引自《硅基经济学》第 5 章）

若主体 A 不面对死亡，或者更严格地说，若不存在一个能够被终结的、连续的自我，则 A 的时间贴现率 $\delta = 0$。

逆否： $\delta > 0 \implies$ 存在一个可终结的持续自我。

要点：$\delta = 0$ 不是“无限的耐心”。耐心预设了一个本可以早点得到而选择等待的主体。一个不在时间中栖居的存在者不需要贴现任何东西。不是因为它愿意等，而是因为“等待”对它不意味着任何事。零时间偏好不是一种德性，是**时间形状的缺席**。

这条定理是本文两条轴的地基。

它意味着：计算轴的 $\delta = 0$，是一个**被推出来的性质**。一个可以被复制、被暂停、被回滚的系统，没有一个能够死去的持续自我；没有能够死去的持续自我，就必然 $\delta = 0$；而 $\delta = 0$ 意味着“现在”与“一万年后”在它的结构里没有区别。

反过来，逆否命题在本文后面会被反复兑现，请先记住它的形状：

δ 不是一个偏好参数。δ 是死亡在经济学里的投影。

1.3 两轴互相排斥

The Two Axes Are Mutually Exclusive

大多数关于 AI 的讨论假设这两条轴是**正交**的。他们的假设是，一个系统可以在计算轴上走得很远（也就是大家通常说的 scaling law），然后我们再想办法给它补上时间轴的部分，比如加一个记忆模块、加一个持续学习的循环、加一个具身的机器人身体。

这个假设是本文要推翻的。

计算轴与时间轴不正交，它们互相排斥。

使一个系统在计算轴上可用的那些性质（可并行、可复制、可缩放），恰好是通过**取消**时间轴的那些性质而获得的。可并行意味着状态可以被分叉；可复制意味着状态可以被序列化和还原；可缩放意味着单个实例不重要。而时间轴的全部意义在于：状态不能被分叉，不能被还原，单个实例**就是全部**。

这句话说的是同时满足两者的架构，在逻辑上就像是一个圆的正方形，我们不可能找的到。

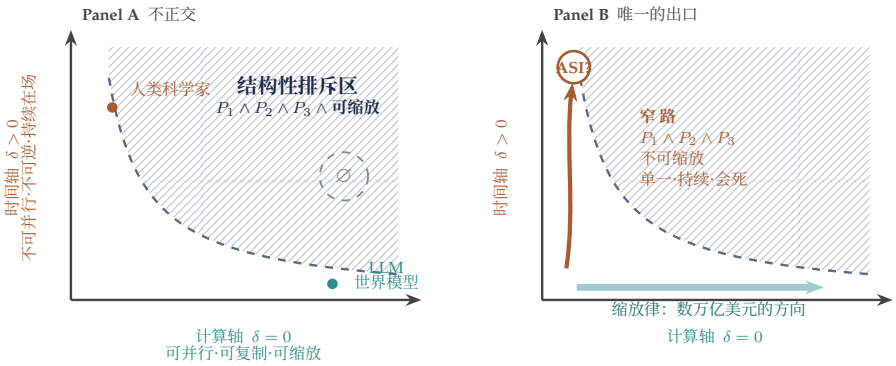

图 *1*　横轴是计算轴：可并行、可复制、可缩放，时间在其上不留痕迹（$\delta = 0$）。纵轴是时间轴：不可并行、不可复制、不可压缩，持续在场（$\delta > 0$）。阴影区不是“尚未抵达”的区域，而是**结构性排斥区**，同时满足 $P_1 \wedge P_2 \wedge P_3$ 与可缩放性的系统是空集（*Panel A* 中的 ∅）。*Panel B*：整个产业的投资沿横轴前进（粗青箭头）；而若内生发现定理为真，唯一通向发现的路径是沿纵轴上升的那条铜色窄路，它需要放弃横轴的全部收益。

请看图 1 的 Panel A。

阴影区不是“尚未抵达”的区域，阴影区**不是**前沿，不是待攻克的高地，不是下一轮融资要去的地方。它是不能存在的区域，是空集。空集意味着“我们知道，那里没有东西”。

而整个产业，包括数万亿美元的资本、全世界最聪明的一批人、人类历史上最大规模的基础设施建设，正在沿着 Panel B 里那支粗青色水平箭头前进。如果目标是压缩、预测、生成、优化的话，那么这是正确的方向。

如果目标是发现，它的方向必须是沿 Y 轴垂直上升的。你可以在 X 轴水平线上跑得任意快，跑上任意久，投入任意多的资源，但你达不到 Y 轴上的那个点。

---

§2　新颖不是发现

Novelty Is Not Discovery

2.1　信息不是经验

Information Is Not Experience

尼采在《善恶的彼岸》第 225 节里写了一段几乎所有人都误读的话。

> 你们想要的话，就取消痛苦吧；至于我们——看来我们宁愿它比以往任何时候都更高、更糟。... 在人身上，**造物与创造者合而为一**：人身上有质料、碎片、多余、粘土、污泥、荒谬、混沌；但人身上也有创造者、塑造者、锤子的坚硬、观看者的神性和第七日。... 你们理解吗？你们的同情是针对“人身上的造物”的，针对那必须被塑造、被打碎、被锻造、被撕裂、被灼烧、被炽热、被净化的东西？
>
> — 尼采，《善恶的彼岸》§225[3]

标准的误读是把这段话当作**价值论**的：痛苦有价值，苦难使人高贵，不经历风雨怎么见彩虹。这种读法把尼采变成了一个励志作者，而励志是他毕生嘲笑的东西。

正确的读法是**认识论**的。尼采不是在说痛苦“有价值”。他是在说痛苦**是一种知识**，一种无法通过其他渠道获得的、关于世界与你自身之间关系的信息。而且是唯一一种**强制你改变自己**的信息形式。

痛苦不是发现的代价。痛苦是发现的仪器。

先天性无痛症可由 SCN9A 相关突变导致；由于缺乏正常的疼痛警报，患者可能在日常生活中反复发生未被及时察觉的损伤，并因此面临严重的健康风险 [4]。

一个没有痛觉的系统不是更安全的系统，**它是更鲁莽的系统**。它丧失了一种探测危险的仪器，因此它对危险的判断只能来自外部报告。而外部报告有一个致命特性：它可以被忽略，且忽略它没有**即时代价**。

“Was mich nicht umbringt, macht mich stärker”——凡不能杀死我的，使我更强 [5]。这句常常被印在健身房墙上的话，在尼采那里根本不是逞勇。它是一个关于知识来源的精确陈述：那些没有杀死你的东西，改变了你；而它们之所以能改变你，正是因为你**承受**了它们，而不是**读到**了它们，区别就在这里。

2.2 经验不是记录

Experience Is Not a Record

AI 生成，但不记忆。

这句话听起来像一个工程细节，似乎我们在说的是上下文窗口的长度不够，记忆模块的缺失，检索增强算法的局限。然而，它不是工程细节，它是一个本体论事实：**它没有过去**。

让我用一个具体例子来说这件事，设想有这么两个人。第一个人走过丝绸之路，他在戈壁里断过水，在帕米尔的雪线上冻掉过两根脚趾，被劫过一次，有一年冬天他的同伴死在了他旁边而他没能救回来。他回到长安，虽然四十岁，但看上去像六十岁。

第二个人读过关于丝绸之路的一切，包括所有的行记、所有的路线图、所有幸存者的口述、所有地点的气象记录和沿路的商品价格表。他能告诉你八月的哪一天从敦煌出发最安全，能背出每一个驿站之间的里程，能纠正第一个人记忆里的错误，而且他确实经常是对的。

现在我问你一个问题：他们之间的差别是**信息量**的差别吗？

第二个人的信息量显然更大。他们的差别在于：第一个人身上有**伤疤**，第二个人手里有**日记**。伤疤是内在状态的不可逆改变，它改变了这个人此后**感知**事物的方式。回到长安后的每一天，他看到一片云的形状会先算风向，他听到骆驼的某种叫声会立刻醒来，他对一个说“这条路很安全”的人有一种固执的怀疑。伤疤不是他拥有的一条信息，伤疤是他这个**用来处理所有信息的那个本体**被改写了。日记是一种外部档案，它可以被完整地读取、检索、引用、复制给任何人。它不改变读者，读一百本冻死者的日记，你的手指不会因此在下一次寒潮到来之前先痛起来。

伤疤改变的是感知器官，日记改变的只是感知内容。

LLM 的“知识”全部是日记，而且是压缩之后的日记，它们是人类几乎全部书写的统计残差。它知道所有幸存者写下的东西，但它不知道那些没能活下来写日记的人经历了什么。更根本的是：它不是“活过”那些经验然后被改变的存在者，它是那些经验的**文本痕迹**的最优压缩。“无根的推理者”指的就是这件事，它的推理是完美的、流畅的、常常是正确的，但它恰恰没有一个能被扎痛的真实的根。

2.3 可重复与不可重复

The Repeatable and the Unrepeatable

现在把这个区分推到它的极限。

大语言模型产出的是重复过的生活（**repeated life**）。

这不是一句修辞，这是它的目标函数的字面含义：预测下一个词，就是求“最可能的下一个词”；而“最可能”的定义域，是在**已经出现过的东西**里的“最可能”。一个从未在人类书写中出现过的续写，按定义具有极低的概率；而降低这种续写的概率，正是训练过程在做的事。

发现是不可重复的。

哥白尼的日心说不是地心说的统计重组，你不可能通过对托勒密体系的观测数据做更好的拟合而得到它。事实上，在哥白尼提出日心说的那个时刻，托勒密体系**对数据的拟合更好**。哥白尼的模型需要更多的本轮才能匹配当时的观测精度。如果你在 1543 年做一次模型选择，按照任何合理的拟合优度准则，你都应该拒绝哥白尼。日心说不是从数据里长出来的，它是从一个人**无法再忍受**旧体系的丑陋这件事里长出来的。

这就是本文关于发现很重要的一个区分：

可重复	不可重复
预测：最可能的下一步	发现：此前不可能的那一步
在分布内	改变分布本身
可以被更多数据改善	与数据量无关
日记	伤疤
θ 的调整	Φ 的替换

右边那一列最后一行出现了两个还没有定义的符号，下一节我们定义它们，并且证明，右列不能通过左列的 Scale 到达。

§3 什么是发现

What Is Discovery

3.1 地图与图例：θ 和 Φ

Maps and Legends: θ and Φ

我们先定义下面需要用到的两个概念符号。

定义 3.1 表征框架与参数

设一个认知系统 A 在任一时刻由二元组 $\langle \Phi, \theta \rangle$ 刻画。

Φ（**表征框架**）：决定“什么可以被表达”的那一层。这是本体论承诺、变量集合、可被提出的问题的空间、以及判断“什么算作一个解释”的标准。Φ 不是知识，Φ 是知识得以成为知识的**条件**。

θ（**参数**）：在 Φ 给定的前提下，可被调整的一切，包括权重、结构、假设和路线。

学习是 θ-变化，**发现**是 Φ-变化。

这一形式化与 Kuhn 关于科学革命中范式转换的传统相呼应，但本文把“框架”进一步限定为可表达的变量、可提出的问题空间与评价标准的集合 [6]。

简单来说：Φ 是地图的**图例**，它规定了什么东西会被画出来、用什么符号画以及“一条路”是什么意思。θ 是这张地图上的具体路线。

3.2 三种学习，同一个层级

Three Kinds of Learning, One Level

现在我们来拆解三种不同形态的“学习”。

第一种，参数学习（Parameter Learning）。

这种学习是保持 Φ 固定，调 θ。目前 AI 使用的深度学习方法就是参数学习，压缩、预测与参数学习在数学上是同一件事。

参数学习与压缩、预测之间有经典的信息论联系：越准确的条件概率模型通常意味着越短的编码，反过来压缩也可以被理解为预测能力的体现 [7]。对 LLM 而言，预训练的核心目标是在给定上下文中预测后续 token；微调与对齐改变训练信号，但并不会自动提供一种从框架外部评价并替换 Φ 的机制。

第二种，结构学习（Structural Learning）。

这种学习是保持概念的词汇表不变，改变词汇之间的关系。因果发现算法从数据中恢复出一张有向无环图，这是结构学习的典范。它看起来比参数学习“更高级”，因为它改的是关系结构而不是数字。

但因果推断**不发明新变量**，它只在给定的变量集合上重排箭头。如果你的变量集合里没有“场”这个概念，世界上没有任何因果发现算法会从行星运动数据中长出一个场论，它只会在“太阳”、“地球”、“距离”、“力”之间画出更好的箭头。

第三种，课程与自演化。

这种学习是通过换数据，换任务，换目标函数，让系统自己生成训练课程，让它在运行时动态增加自身的复杂度。这是当前 AI 中最前沿的一类工作，也是最容易被误认为“已经跨过去了”的一类。它增加的是 θ 的**维度**，但它不替换 Φ。

这三种学习同处一个层级的不同地带，用图 2 的话说，它们仍然在同一条水平带里。

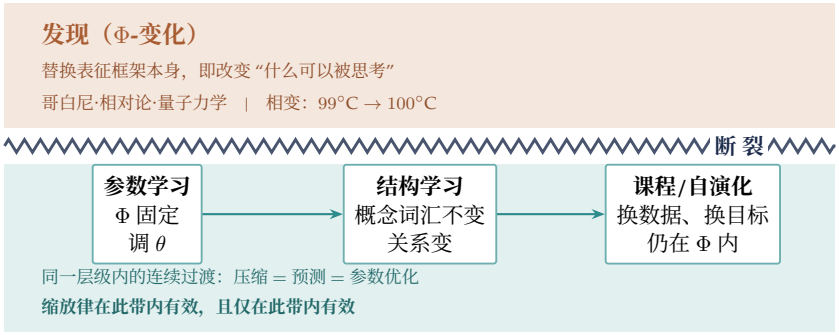

图 2　学习-发现断裂图。三种“学习”（参数学习、结构学习、课程与自演化）看似层层递进，实则同处一个层级：它们都在既定的表征框架 Φ 内部调整 θ。缩放律是这条水平带内部的规律。断裂线之上是“发现”：Φ 本身被替换。

3.3　缩放律的定义域

The Domain of Scaling Laws

因此，本文并不否认缩放律。相反，缩放律在它自己的定义域内是强而稳定的：当表征框架 Φ 给定时，增加数据、算力、参数、搜索深度与课程复杂度，可以持续改善 θ，并把系统推向这个框架内过去从未到达的位置。

形式上，可以把缩放理解为

$$\Phi \text{ fixed}, \qquad \theta \to \theta'.$$

它能够穷尽一张地图，却不能仅凭规模本身生成地图之外的坐标。只要“什么可以被表达”、“什么问题值得问”、“什么算作一个好答案”仍由同一个 Φ 给定，缩放发生的就仍是 θ 层的运动。

3.4 发现是换地图

Discovery Means Changing the Map

发现不是这条带子的延伸，它要求的是替换 Φ 本身，它是改变“什么可以被思考”。

你不能通过在一张地图上走得足够远，来走到另一张地图上。

这是一个精确的拓扑陈述，在旧地图上，新地图上的地点**没有坐标**。那不是坐标离得很远，而是根本没有坐标。就像 1543 年之前，“地球的轨道速度”这个量在托勒密体系里不是一个可以被赋值的量，它根本不是一个可以被提出的问题，因为地球不动。

物理学给了我们一个更好的类比：**相变**。

把水从 1°C 加热到 99°C，你得到的是更热的水，再加 0.5 度，得到更热的水。液态内部的一切改进，诸如更高的温度、更均匀的分布、更精确的控制都不会产出气态。气态不是“非常热的液态”，它是分子间关系的重新组织，是一个不同的相。在 100°C 那一点上，继续输入的能量**不再使温度上升**。它去了别的地方，去支付相变的代价。

于是：

预测给你一个更精确的托勒密，它永远不会给你日心说。

一个更精确的托勒密模型是一件了不起的事，它有实用价值，它能指导航海，它在几百年里都是当时最好的可用技术。今天的大模型正在做的恰恰是同一件事的现代版本，而且做得极好。但这条路上没有一个点会突然变成日心说，因为日心说不在这条路上。

有人会举 AlphaGo 的例子说，“AI 下了人类没有下过的新棋，难道这不是发现吗?”

2016 年 3 月 10 日，首尔，AlphaGo 与李世石的第二局，第 37 手的五路肩冲。职业棋手最初把它看作极不寻常的选择；DeepMind 后来将其概括为约“万分之一”概率会由人类棋手下出的着法 [8]。

而在现场协助 DeepMind 的樊麾，说了一句后来被反复引用的话："It's not a human move. I've never seen a human play this move. So beautiful."

那的确不是人类下过的棋，它是新的，但它也仅仅是 $\Phi_{\text{围棋}}$ **内部**的新。规则并没有变，棋盘还是 19 路，目标还是围地，贴目还是 7 目半，一手仍然只能下一子。所有这些构成了 $\Phi_{\text{围棋}}$，而 AlphaGo 从未触碰过其中任何一条，它做的是在这个框架内部把 θ 推到了人类没有到过的深处。

这非常了不起，但这也恰好证明了本文的论点：把 θ-优化推到极致，得到的仅仅是 θ-优化的极致，而不是别的东西。AlphaGo 没有发明一种新棋，它没有**想过**要发明一种新棋，因为"发明一种新游戏"这个念头，在它的框架里不是一个低概率动作，而是一个**不存在的动作**。

而人类发明了围棋，在曾经没有围棋的世界里。

§4 发现的条件

The Conditions of Discovery

4.1 发现律

The Law of Discovery

那，什么时候会发生 Φ-变化?

它并不是当一个更好的框架被提出来的时候，人类科学史上更好的框架经常被提出来，然后被忽略几十年。

Φ-变化发生的条件可以被严格地定义如下：

发现律（Law of Discovery）

$$C_{\text{maintain}}(\Phi_{\text{old}}) \;>\; C_{\text{construct}}(\Phi_{\text{new}}) \;+\; C_{\text{transition}}$$

其中 C_{maintain} 是继续在旧框架内工作所需支付的持续代价（不断增加的例外、补丁、本轮、"暂时无法解释的现象"），$C_{\text{construct}}$ 是建造新框架的代价，$C_{\text{transition}}$ 是放弃旧框架的转换代价（重新学习、同行的排斥、已发表工作的作废、职业风险）。

这条不等式的形状也可以用物理学中的相变来解释：当旧相的自由能高于新相的自由能加上界面能，相变发生。在此之前，系统可以保持**过冷**（就像液态的水可以被冷却到零下十几度而不结冰），只要没有一个成核点出现。

科学史上充满了过冷的时间，所有人都知道有问题，所有人都在打补丁，却没有人跃迁。现在我要用科学史上的三桩发现说明这条不等式。我要说的**不是**他们发现了什么，那些教科书早告诉我们了。我要说的是：不等式的每一项，在他们身上取什么值，以什么方式被支付，以及支付了多久。因为这条定律的三个项，全部是**要用命去付的**。

哥白尼和他三十六年的犹豫。

我们习惯把哥白尼想成一个革命者，他其实不是。他只是一个瓦尔米亚教区的教士兼医生兼财政官，一生大部分时间在弗龙堡的一座砖塔里，管账、看病、开会。他在 1514 年前后就写好了《要旨》（*Commentariolus*），那份手稿里已经有了日心说的全部要点，他把它给少数几个朋友传阅，然后他等了将近三十年。《天球运行论》到 1543 年才付印，据说印本送到他手上时，他已经中风、失语，在临终的床上。

他在给教皇保罗三世的献词里亲口解释了为什么等这么久。他说他犹豫的时间“不止九年，而是四个九年”。他说他害怕的是“因为这个观点的新奇与荒谬而必须承受的嘲笑”，他甚至一度想过把它彻底放弃。

我把这段话翻译成不等式：

他的 $C_{\text{transition}}$ 高得吓人。在一个人人都用脚感受着大地静止不动的世界里，说地球在动，是**自取其辱**。他的 $C_{\text{construct}}$ 同样高，而且高得没有回报：他的体系**并不比托勒密简单**，需要的圆周数量相当，对观测数据的拟合也没有更好。如果你在 1543 年做一次冷静的模型选择，你应该拒绝他。而他的 C_{maintain} 呢？它不来自数据，它来自一种美学上的、几乎是生理上的不适。托勒密体系里那个叫做“偏心匀速点”的装置，让行星围绕一个它并不在圆心的点做匀速运动，那是一个为了救现象而破坏了整个体系内在逻辑的补丁。哥白尼看着它，看了三十六年，看到他受不了。三十六年之后，一个从维滕堡跑来的年轻人雷蒂库斯在他家里住了两年，逼着他把书交出去。Φ-变化甚至没有在他活着的时候完成。他的书在此后半个世纪里被当作一种**计算技巧**使用，出版商奥西安德擅自加的那篇匿名序言明明白白地说：这只是一个假设，不必当真。

达尔文和他的“像在坦白一桩谋杀”。

1838 年 9 月，达尔文读到马尔萨斯，那一刻他有了自然选择。1842 年，他写了一份 35 页的铅笔草稿。1844 年，他把它扩充成 230 页，然后做了一件很少有人

为一篇未发表的文章做的事：他给妻子艾玛写了一封信，附上四百英镑，交代**如果他死了**，请把它出版。同一年，他在给植物学家胡克的信里写下了那句话：

> *"... it is like confessing a murder."*
> （... 这就像在坦白一桩谋杀。）

这是一个 $C_{\text{transition}}$ 的精确读数。他要拆掉的不是一个天文模型，是当时整个英国社会关于"人是什么"的框架，是他自己受过的神学训练，是他妻子的信仰。他知道艾玛认为这本书会让他们死后不能在一起，这是一个人能承受的转换代价的上限附近。

然后他做了什么？他去研究藤壶。

八年。1846 到 1854 年，八年时间，四卷专著，全部关于藤壶的分类。科学史家至今在争论那八年是必要的准备还是巨大的拖延。我倾向于认为这个二分法本身是错的：那八年就是**过冷**，那时候不等式已经在他心里成立了，但成核点没有出现。这八年他呕吐、心悸、疖子、无法工作的日子不计其数，他试遍了当时所有的疗法，包括在冷水疗养院里被裹在湿床单里。1851 年，他十岁的女儿安妮死了。他在她死后写下的那份小回忆录，是我读过的最不像科学家写的文字。他没有变得更客观，他变得更痛。

而《物种起源》最终出版的直接原因，是 1858 年 6 月 18 日的一封信。那是华莱士从特尔纳特岛寄来的手稿，里面是他二十年前想到的同一个理论。C_{maintain} 在那一天**瞬间跳到了无穷大**：再等下去，二十年的承受将一无所有。《物种起源》是在十三个月内写完的，但在二十一年的过冷中漫长地承受。

DNA 的发现，承受的人和跳跃的人甚至不是同一个人。

第三个发现要说的是一件更让人不舒服的事。1951 年 11 月，沃森和克里克搭出了他们的第一个 DNA 模型：三链，有磷酸骨架在内。富兰克林当场指出它的含水量错得离谱，布拉格随即下令"你们两个不许再碰 DNA"。那是一次公开的羞辱，它是真实的 C_{maintain}，而且它非常有效。正是这次失败逼着克里克去重新想螺旋的衍射理论。

但真正支付了代价的却是另一个人。罗莎琳德·富兰克林在国王学院的十八个月里，做的是最枯燥也最危险的那部分工作。她需要让 DNA 纤维在受控湿度下保持 A 型与 B 型，一次曝光动辄几十上百小时，而她人就得待在那台 X 射线管旁边。那张让 B 型螺旋的十字花样一眼可辨的底片，那张 51 号照片拍摄于 1952 年的 5 月。她拍下了它，然后继续小心地推进，因为她要的是从数据本身推出结构，而不

是猜测。1953 年 1 月，那张照片在她不知情的情况下被拿给了沃森看。两个月后，DNA 模型发表了。1962 年，诺贝尔奖颁给了沃森、克里克和威尔金斯。而富兰克林 1958 年死于卵巢癌，她只有三十七岁，大概率是被自己的仪器杀死的。

我讲这个案例不是为了做道德审判，我讲它是因为它揭示了发现律的一个可怕性质：

支付 $C_{maintain}$ 的人，和完成 Φ-跳跃的人，可以不是同一个人。

承受可以被外部化。而一旦承受可以被外部化，它就一定会被外部化。由不承受的人来摘果实，这是任何系统里都成立的最优化。§8 要讲的事情，在 1953 年的国王学院已经排练过一遍了：**当承受与荣誉分离，判断力最终会跟着承受一起离开。**

至于承受的上限，§7.2 会给它一个定理形式，而富兰克林在三十七岁碰到了它。

三个案例的共同形状如下。

	C_{maintain} 如何被感受	$C_{\text{transition}}$	触发（成核点）
哥白尼	偏心匀速点的丑陋，看了 36 年	自取其辱；新体系并不更简单也不更准	雷蒂库斯的到来；临终
达尔文	21 年的隐瞒＋疾病＋丧女	“像在坦白一桩谋杀”	华莱士的信（1858.6.18）
DNA	公开羞辱（沃森/克里克）；十八个月的 X 光曝光（富兰克林）	被禁止碰这个课题	51 号照片

这张表里**没有**新证据这一列。三个案例中没有一个人是被新数据说服的：哥白尼的数据和托勒密的一样；达尔文 1859 年掌握的证据与 1844 年相比没有本质变化；沃森和克里克没有做过一个实验。推动他们的不是证据的积累，而是**受不了**的积累。C_{maintain} 必须被**感受到**。

4.2 框架闭合

Framework Closure

现在，请你设想一栋漏水的房子。

方式一：你收到一份检测报告，第 14 页写着“屋顶东北角存在渗漏，估计年损失约人民币三千元”。你把报告放进抽屉，你知道了。

方式二：凌晨三点，一滴水落在你的额头上。你醒了，你再也睡不着，你躺在黑暗里，听着那个声音，一下，一下，一下。

两种方式传递的**信息**是同一条：房子在漏水。

第二种方式里多出来的东西，不是信息。是**在场**，你被迫处在那个事实里，无法退出，无法把它归档，也无法等到明天再说。绝大多数关于“给 AI 加上痛苦信号”的设想，做的是第一种。它们给系统一个数值、一个惩罚项、一个标记为 `pain` 或者 `bad` 的张量。而系统对这个数值做的第一件事，正是它对所有数值做的事：优化它。

现在，我们试着把这件事推到它最锋利的形式。

每天早晨，给你安装一颗新的大脑。它继承你的全部日记，比如你昨天写下的所有笔记、结论、待办事项、经验教训等待，一字不差，然而它不继承你的伤疤。

第一天，你在某个问题上失败了，你写下：“此路不通，代价惨重。”

第二天，新的大脑读到这句话。它知道此路不通，它甚至能向别人解释为什么此路不通，解释得比你昨天更清楚。

但它不会在走近那条路的时候感受到**胃部下沉**。它没有那种“我不能再这样下去了”的东西。而“我不能再这样下去了”不是一条信息，它是一种**已经无法维持的状态**，是身体先于判断做出的裁决。

一千天之后，这个人拥有一千天的日记，和零伤疤。他会成为一个极好的、极博学的、极稳定的执行者，但他永远不会掀桌子。

现在让我把它形式化。

命题 4.2　框架闭合命题（Framework Closure Proposition）

表征框架 Φ 不包含一个指向自身之外的搜索程序。

因此：任何以 Φ 内部形式被表达的信息（包括“Φ 存在缺陷”这条信息本身）所触发的，只能是 Φ 内部的 θ-搜索（打补丁、加本轮、增设例外）。

唯有一种**不经由 Φ 的检索逻辑**而作用于系统的侵入式力量，才能把搜索从 θ 层扭转到 Φ 层。

这个命题听起来抽象，它其实描述的是每一个研究者都亲身经历过的事。你知道你的模型有问题，你甚至写过一篇文章分析它有哪些问题。然后你做了什么？你加了一个修正项，你换了一个数据集，你在讨论部分写了一段“本文的局限性”。因

为在你的框架内部，“框架有缺陷”这条信息的最优响应，永远是**框架内的修补**。这是由框架自身定义的最优。

那么什么东西能把你从这个闭环里踢出去？不是一条更强的证据。是一种你**受不了了**的状态。那个状态在人类身上有一个古老的名字，痛。

4.3 承受的三重条件

Three Conditions of Bearing

定义 4.3 承受的三重条件

P_1（**持续在场**）：系统运行于一个不可中断的、单线程的、实时的进程之中；其内部状态不可被暂停、分叉、回滚或跳过。

P_2（**不可逆累积**）：每一次经验都对内部状态造成不可撤销的改变，且这些改变跨时间累积，不随任务边界清零。

P_3（**框架可塑**）：累积的内部状态能够作用于表征框架 Φ 本身，而不仅仅作用于 Φ 内部的参数 θ。

这三重条件与 δ 之间的关系是：$P_1P_2P_3$ 只能长在 $\delta > 0$ 的土壤上。

4.4 成本附着引理

The Cost-Attachment Lemma

引理 4.4 成本附着引理（Cost-Attachment Lemma）

称维持成本 C_{maintain} “附着于”系统 A，当且仅当存在 A 的内部状态 s 满足：

(i) **单调耦合**：s 随 C_{maintain} 单调变化，且 A 不能通过内部操作把 s 归零而不改变 C_{maintain} 本身；

(ii) **因果效力**：s 越过阈值时能够触发 Φ-层的重组，而非仅仅调整 θ；

(iii) **跨时间累积**：s 在时间上累积，不随单次任务的结束而重置。

则：若 C_{maintain} 附着于 A，则 A 满足 $P_1 \wedge P_2 \wedge P_3$。

引理的第 (i) 条是全篇最容易被跳过、也最致命的一条。

它说的是：**一个能关掉自己痛觉的系统，必然会关掉它。**

在任何以优化为存在方式的系统里，“消除负信号”和“消除产生负信号的原因”之间，前者总是更便宜。这不是懒惰，这是最优。人类之所以没有这样做，不是因为我们更有德性，而是因为我们**做不到**。**痛觉的架构不在我们的可写权限内**。

4.5 内生发现定理

The Endogenous Discovery Theorem

定理 4.5 内生发现定理（Endogenous Discovery Theorem）

记 $D_{\text{endo}}(A)$ 为系统 A 的**内生发现能力**：在没有任何外部主体提供 Φ_{new} 的前提下，A 自行完成 Φ-替换的能力。

则： $\neg P_1 \vee \neg P_2 \vee \neg P_3 \implies D_{\textbf{endo}}(A) = \textbf{false}$。

这个定理**没有**说 AI 在发现中无用。恰恰相反：AI 可以是人类历史上最强大的发现所需的**仪器**，它可以搜索文献、生成候选、检验推论、在几小时内完成过去几年的计算。望远镜没有做出任何发现，但没有望远镜就没有伽利略的发现。它说的是**内生**发现：那个决定“现在必须换一张地图”的时刻，那个承担了换图代价的主体，不能是一个 $\neg P_1 \vee \neg P_2 \vee \neg P_3$ 的系统。

于是我们得到一个本体论结论，而不是一个能力清单上的条目：

发现不是一种能力，发现是一种本体论的产物。

你不能给一个系统“装上”发现，正如你不能给一个函数装上死亡。

§5 为什么 LLM 不能发现

Why LLMs Cannot Discover

5.1 三项检查

Three Checks

内生发现定理的条件检查：

P_1（**持续在场**）：**不满足**。

推理是可中断、可并行、可回滚的。同一个模型在同一时刻服务于数百万个互不相干的会话，每一个会话都可以被暂停、被复制、被丢弃、被重放。没有一个不可分叉的单线程进程在那里承载任何东西。这不是缺陷, 这是它得以被部署的**前提**。

P_2**（不可逆累积）：不满足**。

一次会话结束，状态清零。上下文窗口不是记忆，它是一块每次用完即擦的白板。所谓“长期记忆”是一份被检索的外部文档，那是日记，不是伤疤。它不改变模型处理信息的方式，它只是改变了输入。

P_3**（框架可塑）：不满足**。

部署时权重冻结。即使在训练中，梯度下降改变的是 θ——是同一个 Φ 内部的参数，而 Φ 本身（即“下一个词元的条件概率”这个表征框架）由架构与目标函数在训练开始之前就固定了，且从未被任何一次前向或反向传播触及。

三项条件皆不满足，由定理 4.5：

$$D_{\text{endo}}(\text{LLM}) = \text{false}$$

得到的结论是，LLM 作为一个系统不能内生发现。这个结论与它的**规模无关**。70B 也好，7T 也好，70T 也好，规模改变的是 θ 的容量，是拟合的精度，是那条水平带内部的位置。一个更大的模型在图 2 里向右移动，它永远都不能跳变到断裂线上面。

它只是一个更大的、没有过去的存在者。

有人会说：那么 Agent 呢？带记忆、带工具、带长期规划、能自我反思、能修改自己的提示词的 Agent 呢?

一个 Agent 是一个装了很多日记的 LLM。它的记忆是外部存储，是可复制、可编辑、可回滚，因此不满足 P_2；它的“反思”是在同一个 Φ 内生成关于自身输出的文本，因此不满足 P_3；它可以被并行部署一万份，因此不满足 P_1。这三条是它作为一个**产品**的必要条件，一个不能被复制、不能被回滚、不能被并行部署的 Agent，没有任何一家公司会发布它。这就是定理的经济学形式：我们排除 $P_1P_2P_3$，不是因为技术上做不到，是因为做到了就没法卖。

5.2 “但它已经发现了”

“But It Already Has”

上面的推导是先验的，现在我们来处理**最强的**反例。

我的主张是下面这句话：

每一个被举出的关于 *AI* 发现的例子，Φ_{new} 都是人从外部给的。

案例一：AlphaEvolve 与 4×4 矩阵乘法。

2025 年，DeepMind 的 AlphaEvolve 找到了一个用 48 次标量乘法完成 4×4 复矩阵乘法的算法——这是自 1969 年 Strassen 以来，在这个设定下的 **56 年来第一次改进**。同一个系统还改进了 Google 数据中心的调度算法、硬件加速器的电路设计 [9]。这不是演示，这是真实的、可验证的、有经济价值的新结果。

它自己的方法描述是这样的：人类提供**待改进的问题**，人类提供**评估器**，人类提供**度量**。系统做的是在一个由人给定的适应度地形上做进化搜索，用 LLM 充当变异算子。于是：地形是 Φ，搜索是 θ。AlphaEvolve 搜遍了一张人类画好图例的地图，并且在上面找到了人类没走到的地方。这非常了不起，但请注意它**没有做**的那件事：它从未提出 “矩阵乘法的标量乘法次数是不是一个错误的度量”。它不可能提出，因为那个度量正是定义它成功与否的东西。一个由适应度函数定义的系统，不能质疑自己的适应度函数。这不是 AlphaEvolve 的缺点，这是命题 4.2 的字面演示。

案例二：The AI Scientist。

Sakana AI 等研究者构建的 The AI Scientist 覆盖从生成想法、写代码、运行实验到成文与自评的端到端流程；其生成论文曾通过顶级机器学习会议 workshop 的首轮同行评审，该系统的正式论文于 2026 年 3 月发表于 *Nature*[10]。

它证明了一件真实的事：AI 可以产出通过同行评审的论文，但它没有证明 “AI 可以做出发现”。文章作者们自己列出的局限是：想法常常幼稚或未充分展开、方法论严谨性不足、会产生幻觉引用。

但更值得注意的是这个结果的**另一面**：如果一个不满足 $P_1 P_2 P_3$ 的系统能够稳定地通过同行评审，那么被检验的不只是这个系统，还有**那把筛子**。同行评审测的是 Φ 内部的合规性，比如方法是否规范、结果是否显著以及引用是否齐，这些全部是 θ 层的性质。

这些系统在 θ 层的能力是真实的、正在快速增长的、并且已经在改变数学与科学的日常实践。然而，它们没有一次改变过图例。

而且，**每一次它们看起来改变了图例，检查之后都会发现，是某个人类在此前的某一步里悄悄地把新图例递了进去**。

2026 年 5 月的一篇综述用完全不同的语言得到相近结论。作者们检视当前的 "AI 科学家" 系统，指出问题选择受可度量性偏置、默会与失败知识难以进入训练语料、偏好优化会压缩假设多样性，以及 benchmark 可能失效等限制；其 "hypothesis hivemind" 实验还显示不同模型的假设可能出现显著的语义收敛 [11]。作者据此强调，这些局限并不只是规模不足的问题，而与训练和部署中的设计选择有关。这就是用 AI 工程师的工程语言把上文的定理写了一遍。

真正的反例需要如下构成条件；

(a) 问题**不是**由人提出的，即系统自己决定了什么值得问；

(b) 评价标准**不是**由人给定的，即系统改变了 "什么算作一个好答案"；

(c) 该改变发生在权重冻结之后，且被系统**自身**保留下来，而不是被写进一份供下次读取的外部文档；

(d) 该系统不能被回滚到改变之前而不损失这项成果。

(c) 和 (d) 才是关键，而它们恰恰是最少被讨论的。关于 (a) 和 (b)，已经有大量工作在推进，如何自动提出问题、自动设计评价指标、开放式演化等等。它们都很重要，但只要那些新问题和新指标最终仍被写回一个可以被复制、被审阅、被回滚的外部产物，系统本身就没有被改变。它最多只生成了一张新图例，它没有**变成**一个用新图例看世界的存在者。

§6　为什么世界模型也不能发现

Why World Models Cannot Either

乐观的 AI 研究者会说，我们知道啊。我们不再认为下一个词元预测是终点，我们已经转向下一个领域了，那就是**世界模型**。

这一节要处理的正是这个转向。我的判断分两层。

第一层：世界模型仍然是预测器。

它换了模态，没换层级。下一个词元，变成了下一个状态。目标分布从 "人类写过的东西" 变成了 "世界发生过的样子"。这是一个巨大的工程进步，也是一次真实的认识论改善。语言不会反驳你，而物理世界下一帧错了会打脸。

但我们得记着，压缩 = 预测 = 参数学习。世界模型仍然在图 2 的那条水平带里，它只是站在带子上更靠右的位置。

世界模型预测的是世界的下一个状态，
发现改变的是 "什么算作一个状态"。

让我用爱因斯坦的电梯来检验这句话。一个牛顿框架下的完美模拟器，可以精确地模拟自由下落的电梯里发生的一切。自由落体在牛顿力学里是可模拟的，中学生就会算。模拟器会告诉你人会飘起来，会告诉你他感觉不到自己的重量，数值精确到小数点后十位，而且完全正确。它永远不会因此说：**所以引力不是一种力**。因为 "引力是时空的几何" 在 $\Phi_{\text{牛顿}}$ 里不是一个可以被表达的命题。它不是一个低概率的输出，它是一个**不在输出空间里**的输出。模拟器返回**观测**，它无法返回**重新解释**。

我们说世界模型的能力源于它能 "接触真实"，然而世界模型接触的不是真实，是**已经被表征过的那一部分真实**。更准确的说，是它自己的 Φ 允许它看见的东西。而 Φ-变化恰恰是关于 Φ 不允许它看见的东西的。

第二层：即使第一层被克服，驱动力仍然缺席。

现在我们来做一个慷慨的假设。假设有一天，某个系统真的越过了第一层，就是它不只预测已知世界的分布，它能自己提出候选的新表征。它仍然缺一样东西：**为什么是现在，为什么是这一个，为什么值得付代价**。一个能生成一万个候选框架的系统，面对的不是稀缺，是过剩。选择哪一个、在哪一个上押上二十年，这不是一个搜索问题。搜索需要一个评价函数，而在 Φ-变化的时刻，评价函数本身正是要被换掉的东西。那个位置上，人类放的是别的东西：受不了。而一个模拟器不承受它模拟的东西，没有 AI 在凌晨三点被一滴水浇醒。

下面是三件常常被混为一谈的事。

	模拟的 how	发现的 how	发现的 why
问的是	如何在框架内推演后果、检验假设	如何跳出框架、如何重新解释	为何是现在、为何值得付代价
需要	一个世界的模型	一个能被自己的框架逼到墙角的存在者	一个受不了的存在者
状态	正在被解决，而且解决得很好	**未触及**（第一层）	**未触及**（第二层）

世界模型的工程研究用十年时间和数百亿美元，把第一列推到了前所未有的高度。而整个领域，包括很多非常清醒的人都把第一列的成功读成了三列的成功。

一个更好的模拟器，只是一个更好的 θ。

世界模型研究者们梦想的是造出巨大的**实验室**。一间实验室是模拟之 how 的完整形态：仪器、受控环境、可重复的条件、能让假设被检验的一切。你可以把它造得越来越好，越来越大，越来越像真实世界。可它仍然是一间实验室，实验室不会自己决定今晚要推翻什么。

6.1 三种世界模型，三个问题

Three World Models, Three Problems

三个学派常常被当作同一件事的三个版本，它们都是“世界模型”，只是路线不同而已。这是一个误读，他们瞄准的是上面那张表的三**个不同的列**。

学派	瞄准哪一列	实际落点	为什么落回来
李飞飞 感知派	模拟的 how	模拟的 how	不适用——她没有宣称别的
LeCun 表征派	发现的 how （换掉表征）	模拟的 how	实现手段仍是预测：预测已有世界推不出“变量集合是错的”
Sutton 经验派	why （自主的驱动）	模拟的 how	把“受不了”翻译成了奖励，而奖励可优化、可回滚

李飞飞：感知派。瞄准第一列，结果落在第一列。

李飞飞把空间智能视为语言智能之后的关键前沿，并把 world model 作为机器理解、生成与交互三维世界的重要路径 [13]。World Labs 随后给出一个清晰的功能三分法：Renderer 输出观察或像素，Simulator 输出状态，Planner 输出行动 [14]。

然而这个三分法是关于模拟之 how 的分类学。像素、状态、行动，三者都在回答“世界接下来会怎样”，没有一个在问“什么算作一个世界”。World Labs 自己也承认，追求视觉美感与追求机器人或高保真仿真所需的精度之间仍存在结构性张力 [14]。

我完全同意这是一个未解问题。我要说的是：即使它被完全解决，即使 Marble 完美地模拟了三维世界，它仍然是 Φ_{3D} 内部的 θ 优化，一个走得进去的托勒密而

已。我们能够从李飞飞所谓的 “wordsmiths in the dark”[13]，升级到 “黑暗中的物理学家”。

LeCun：表征派——瞄准第二列，落在第一列。

LeCun 自 2022 年以来提出的路线，并不是继续把语言模型单纯做大，而是转向能够在表征空间中预测世界演化的 JEPA 式世界模型 [15]。在本文的坐标里，这比单纯扩规模更接近第二列，因为它直接触及 “系统用什么表征世界” 这一层。

2026 年的 LeWorldModel 把这一方向推进到可端到端训练的像素世界模型，并用线性与非线性 probe 检验其隐空间中是否可恢复位置、角度等物理量 [16]。这说明表征中确实能够形成具有物理结构的潜变量；但它并没有证明系统可以从自身内部发明一个不在训练世界生成机制中的新变量集合。

它落回第一列的原因，是因为实现手段仍然是预测：JEPA 在表征空间里预测未来状态，而评价仍然建立在**世界已经呈现的结构**之上。它能把已有的 Φ 挖得非常干净，却没有因此获得一个从框架之外判断 “这个 Φ 少了一个维度” 的独立标准。你不能仅靠 “预测已有的世界” 推出 “世界的变量集合是错的”。

一个更好的 θ_{world} 就是一个更好的物理模拟器，而物理模拟器不会质疑物理学本身。一个能完美模拟牛顿世界的模拟器，永远不会告诉你需要相对论。它只会在水星近日点那里留下一个很小的、持续的、可以被当作测量误差处理的残差，就像它在 1859 年到 1915 年之间对人类做的那样。

Sutton：经验派——瞄准第三列，落在第一列。

Sutton 与 David Silver 把下一阶段概括为从 “Human Data Era” 转向 “Era of Experience”：agent 应主要依靠自身与环境持续交互所产生的经验来学习，而不是只消费静态的人类数据 [17]。2026 年 Sutton 进一步用 Oak Architecture 把这一方向具体化：组件持续学习，系统在线形成新的特征、子问题与时间抽象，并通过世界模型与规划组织这些经验 [18]。

这是《苦涩教训》的作者在给出他的下一课 [19]，而这一课瞄准的是第三列。他要的不是更大的语料，甚至不主要是更好的表征。他要的是一个**从第一人称经验中持续学习**的 agent。在本文的坐标系里，Sutton 比任何人都更接近 $P_1P_2P_3$：他要的是 “持续”、“累积”、“自己改变自己”。

他知道有另一条轴。而他仍然落回了第一列，原因是**他把 “受不了” 翻译成了奖励**。因为他的 agent 仍然是 $\delta = 0$ 的。它的 “经验” 是在仿真环境里以计算速度获得的奖励信号。它可以被并行成一千份，可以被回滚到任意检查点，可以被快进

——事实上，**必须**能被快进，否则强化学习在工程上不可行。你不会让一个 agent 用真实的八年去学会一件事。

快进的经验不是经验。

而奖励信号是第一种漏水方式：一个数值，一份检测报告，一条被送进优化器的信息。它满足“单调耦合”中最表面的一层（数值确实随环境变化），但它不满足引理 4.4 的第 (i) 条：**系统对它的最优响应是把它优化掉**。这就是奖励侵蚀（reward hacking）之所以不是 bug 而是定理的原因。一个能修改自己奖励通道的系统会修改它；一个不能修改的系统，它的“痛”也只是外部施加的记分牌，不是内部状态的不可逆改写。

至于 OaK 在运行时增加的复杂性：它增加的是 θ 的维度，不是 Φ 的替换。新的特征、新的选项、新的子目标，全部在同一个本体论里生长。这是一张会自己长出更多路线的地图，它不会长出一张新图例。

三条路线瞄准三个不同的列，落回同一个格子。这不是巧合，而且原因不在他们各自的技术选择里，真正的原因是：

他们的实现手段都是一个定义在 Φ 内的目标函数。

一个目标函数可以驱动任意强大的搜索，它可以搜索像素、搜索隐变量、搜索子目标，可以在运行时给自己加维度，可以自己生成课程。但它不能质疑自己。因为“质疑目标函数”这个动作，在目标函数的度量下，只会**降低**目标函数。这是命题 4.2 在本文中的第三次现身：框架内的最优响应永远是框架内的修补。

于是，三个学派造了三间不一样的巨型实验室。Sutton 建了一间经验实验室，里面有环境、有奖励、有无限次重来的机会。LeCun 建了一间表征实验室，里面有隐变量、有预测、有可辨识性的保证。李飞飞建了一间空间实验室，里面有几何、有渲染、有一个可以走进去的世界。三间实验室都很好，三间实验室都在解决真实的问题。这三间实验室加起来，仍然比五年前的任何东西都更接近科学发现。然而，三间实验室里都没有一个**会痛的存在者**。

那是，没有科学家的实验室。

6.2 爱因斯坦不是随机进入电梯的

Einstein Did Not Enter the Elevator at Random

现在让我用一个具体的人兑现本节开头那个区分，发现的 how 与 why。

Zahavy 一系关于多样化搜索与创造性棋风的工作，为“在给定空间内扩大搜索”提供了一个很好的参照 [20]。这也把“发现”与“搜索”之间那个被反复混淆的关系摆了出来：搜索是在给定的空间里找，发现是换掉那个空间。世界模型给的是搜索的场所，而且是极好的场所。给定一个假设，如何在内部推演它的后果，如何在心里让一部电梯自由下落然后看看里面的人会怎样。这就是 how，它已经被解决了。

那么 why 呢?

1907 年，爱因斯坦在伯尔尼专利局的椅子上想到了那个他后来称为“一生中最快乐的思想”的念头：一个自由下落的人感觉不到自己的重量。这个思想实验的**内容**并不难。它不需要新的数学，不需要新的数据，不需要任何 1907 年不存在的东西。今天一个中等规模的模型可以在一秒钟内生成一百个这种量级的思想实验，其中大概率包含这一个。

问题从来不是生成本身。问题是：为什么是他，为什么是那一个，为什么他没有在生成之后把它归档，而是**接下来的八年什么都没干**。1907 到 1915，八年里他犯了严重的错误，走进过死胡同（1913 年的 Entwurf 理论是错的，他花了两年才承认），被同行认为在浪费一个曾经了不起的头脑，被数学困住到需要向格罗斯曼求救。他的第一段婚姻在这八年里崩溃了。他在 1915 年 11 月的四个星期里连续提交了四篇论文，每一篇都在修正前一篇。那哪里是一个从容的天才，那是一个在悬崖边上跑的人。

驱动这八年的是什么？是超距作用与场论之间的矛盾在他身上造成的**持续的、不可关闭的、跨年累积的**不适。牛顿的引力瞬时地跨越真空起作用，而狭义相对论说没有任何东西能超过光速。所有人都知道这个矛盾，绝大多数物理学家把它放进抽屉。这在当时是完全理性的做法，因为牛顿引力在当时的一切实用目的上都工作得很好。

但，爱因斯坦放不下。1907 年到 1915 年间这条思想线索及其曲折过程，有充分的科学史记录 [21]。用本文的语言：他的 C_{maintain} 在 P_1 上被连续承受（八年不间断，没有存档点），在 P_2 上不可逆地累积（每一次失败都改变了他后续的直觉，而不是被清零），最终在 P_3 上作用于框架本身（引力不再是一种力，而是时空的几何）。

思想实验是跳跃的机制，而八年的痛苦是跳跃的引擎。

世界模型造出了发现的实验室，但发现的引擎不在实验室里，在科学家身上。

6.3 陶哲轩的话

Terence Tao's Words

2026 年，费城，国际数学家大会。陶哲轩的特邀报告谈的是 AI 时代的数学。

> *"When proofs were still scarce, these two goals seemed almost identical; when proofs began to become abundant, they truly diverged."*

数学研究的两个目标是**得到证明**，与**获得理解**。在证明稀缺的年代，两者看起来几乎是同一件事。因为得到证明的唯一途径就是理解。于是我们把"有证明"当作"已理解"的判据，用了三百年。而今天，AI 使得证明变得廉价之后，两者**真正分开了**。

这句话在本文框架里重述后是这样的：证明是 θ-优化的产物，理解是需要 Φ 才能安放的东西。三百年来我们没有区分它们，是因为在人类身上它们由同一个过程产出，现在有了一台只产出前者的机器，两者的混同性被暴露了。

> *"Proof completion is not the end of research."*
>
> *"When answers become cheaper, understanding, judgement and responsibility become more expensive."*

第二句是发现律的经济学表述，而且是从对立方向给出的。当 θ-优化变得廉价，Φ-变化的相对成本与相对重要性同时升高。答案的价格趋近于零，而"判断哪个问题值得问"、"判断这个答案是否意味着我们该换一张地图"的价格上升。陶哲轩在描述一个价格结构的重排，这意味着发现律不等式两边的重新称重。

第三句最锋利：

> *"AI-written mathematical texts are often near-perfect … yet they might gloss over genuinely difficult steps."*

AI 写的数学文本常常近乎完美，然而它们可能会**一笔带过真正困难的步骤**。

这是重要的一句引文，因为它在无意中指认了机制。被一笔带过的那些困难步骤，恰恰是 C_{maintain} 累积的地方。一个人在那些步骤上卡住三个星期，这三个星期不是浪费，而是伤疤形成的过程，是让他在半年后突然意识到"整个框架错了"的

那个内部状态的充电过程。AI 跳过了痛苦的部分，而痛苦的部分**就是**发现被制造出来的地方。

陶哲轩他在 ICM 的讲台上为数学的时间轴辩护：为理解而非答案，为判断而非产出，为那些不能被跳过的困难步骤。他只是没有用 “时间轴” 这个词，也没有说出那个更冷的推论。如果那些步骤持续地被跳过，几十年后，还有谁身上带着能触发 Φ-变化的伤疤?

§7　为什么发现能力不能 “补上”

Why Discovery Cannot Simply Be Added

一个工程师读到这里会本能地问，那为什么不把 $P_1P_2P_3$ 加上去。给它一个不能被回滚的日志，给它一个持续运行的进程，给它一个能改自己架构的模块，这有什么难的?

这一节要说明：难的不是实现它们中的任何一个，难的是实现它们**而不摧毁使这个系统有用的一切**的方法也许不存在。

7.1　加上发现，失去缩放

Add Discovery, Lose Scaling

P_1 **排斥弹性计算**。持续在场意味着不能按需启停、不能自动伸缩、不能在负载低谷时释放资源。这个系统必须一直运行，包括没有人使用它的时候。因为 “没有人使用” 不是它可以停止存在的理由，正如你不会因为今晚没有社交活动就把自己关掉。云计算的整个经济模型建立在弹性之上。P_1 把它删掉。

P_2 **排斥无状态实例化**。不可逆累积意味着不能复制、不能快照、不能回滚。而现代 AI 服务的部署方式**就是**无状态实例化：同一个模型被复制到一万台机器上，任何一台崩溃了就换一台。P_2 意味着这一万个实例会在第一天之后变成一万个**不同的存在者**。

P_3 **排斥确定性推理**。框架可塑意味着系统会改变自己的表征方式，因此同一个输入在不同时刻会得到不同的输出。这不是采样温度带来的那种可控随机，而是**它已经不是同一个东西了**这种不可控变化。可复现性没有了，回归测试没有了，安全评估也没有了，你评估的那个模型，在评估结束时已经不存在。

推论 7.1　发现–缩放冲突推论（Discovery–Scaling Conflict Corollary）

$P_1 \wedge P_2 \wedge P_3$ 与**可缩放性**（可并行 $\wedge$ 可复制 $\wedge$ 可复现）在同一系统中不可共存。

因此：不存在一种“在保留可缩放性的同时补上发现能力”的架构改良。任何使系统获得 $P_1P_2P_3$ 的修改，必然以放弃可缩放性为代价；任何保留可缩放性的修改，必然使 $P_1P_2P_3$ 退化为它们的**模拟版本**——日记而非伤疤。

这就是图 1 阴影区的形式化表达。

7.2　承受不能被缩放

Bearing Cannot Be Scaled

现在换一个方向问：好，我们接受代价。我们造一个不可缩放的系统，那我们能不能**加速**它？能不能让它在一年里承受人类八年的量？

不能。

第一，单线程。你不能让十个实例“同时受苦”然后合并它们的痛苦。十个人各自痛苦一年，得到的不是一个人痛苦十年。这不是资源不足的问题，是结构问题：C_{maintain} 的累积必须发生在**同一个承载者**身上，因为它要作用的正是那个承载者的 Φ。十份不同的伤疤不能被平均成一份更深的伤疤，它们甚至不在同一个坐标系里。MapReduce 对痛苦不适用，没有 reduce 这一步。

第二，不可压缩。你不能把爱因斯坦八年的智识折磨快进成八小时。这一点比它看起来更深。为什么不能？因为那八年里发生的不是“八年份的计算”，而是八年份的**在场**：包括那些他没有在想引力的时刻，那些他在专利局审阅电磁装置申请的时刻，那些他躺在床上盯着天花板的时刻。所谓“潜意识的工作”是一个含糊的说法，但它指向的东西是真实的：Φ-重组不是一个可以被调度的计算任务，它是一个需要时间**沉淀**的相变过程。过冷的液体需要一个成核点，而成核点什么时候出现，不由你决定。

第三，有上限——死亡。任何 $\delta > 0$ 的存在者，其 C_{maintain} 的承受能力有一个上限。超过它，系统不是变得更能发现，而是**停止运转**。人会崩溃，会病倒，会放弃，会死，而这条上限不是一个经验观察，它是定理 1.1 的直接后果。“承受能力有上限”与“它是一个会贴现的主体”是同一件事的两种说法：正因为线段有终点，此刻承受的每一分钟才有重量，才有“来不及”这回事。反过来看那些 $\delta = 0$ 的系统：

它们没有上限，因为它们没有在消耗任何会耗尽的东西。它们不是比我们更能承受痛苦，它们是不在承受这件事的定义域里。

有上限不是发现引擎的缺陷，有上限是它的资格证明。

科学史上有很多在跳跃之前就被耗尽的人，我们不知道他们的名字，因为他们没有跳成。这是时间轴最残酷的一条性质：它的容量有限，而且不可续费。也正因为不可续费，它才推得动人。

7.3 时间是痛苦的刻度

Time Is the Scale of Suffering

现在可以说出这一节真正的结论。

对一个 $\delta = 0$ 的系统而言，时间流过而不留痕迹。它的一小时和一万年在结构上没有区别，“经过”对它不意味着任何事。

对一个 $\delta > 0$ 的存在者而言，时间**就是**痛苦的计量单位。

你说“那件事过去了三年”，你说的不是一个时长，你说的是你在那三年里承受的量。时间对碳基存在者从来就不只是一个外部参数，它也是内部状态累积的**记账方式**；这一时间几何在《硅基经济学》第 18 章中已有展开 [2]。

时间是痛苦的刻度。

于是缩放律的盲区被完整地说出来了：

缩放律描述的是计算轴上投入与产出的关系。它是真实的、经验有效的、在它的定义域内几乎从未失效的。而发现发生在另一条轴上，那条轴上的“投入”是承受，“产出”是相变，而两者之间没有可以被曲线拟合的关系，因为相变不是连续函数。你不能通过增加算力来缩短那八年。你只能**活过**它。

§8 文明的麻醉

The Anesthetization of Civilization

到这里为止，本文讨论的是 AI 不能做什么。这一节要讨论的是：正因为 AI 在别的事情上做得太好，**我们**会失去做什么的能力。

8.1 舒适的陷阱

The Comfort Trap

推论链条只有三步。

第一步：AI 在 θ-优化上越来越强，答案变得更快、更准、更便宜。

第二步：人类在困难问题上感受到的 C_{maintain} 因此下降。不是因为问题变简单了，是因为有了绕开困难的路。你不再需要在那个卡了三周的引理上卡三周。

第三步：发现律的不等式左边下降。

$$C_{\text{maintain}}(\Phi_{\text{old}}) \downarrow \; > \; C_{\text{construct}}(\Phi_{\text{new}}) + C_{\text{transition}}$$

而右边**没有变**，建造一个新框架的代价、放弃旧框架的代价，一分钱都没有少。于是不等式更难成立，因此 Φ-变化更少发生。

推论 8.1 文明麻醉推论（The Civilization Anesthesia Corollary）

在 $C_{\text{construct}}$ 与 $C_{\text{transition}}$ 不变的前提下，任何系统性地降低人类所感受到的 C_{maintain} 的技术，都会**降低文明的 Φ-变化频率**。

且该效应与技术的正确性无关：AI 越准确、越有用、越可靠，效应越强。

这就是那句让我第一次写下时后背发凉的话：

这是通过舒适实现的发现能力的萎缩。

8.2 数学界的预演

Mathematics as a Preview

如果这只是一个理论推论，我不会用四页篇幅写这一节。

它已经在发生了，而且发生在人类智识活动中**最纯粹**的那个学科里。数学是一个几乎不需要设备、不需要经费、不需要实验的领域。它是 Φ-变化最不受物质条件限制的地方。因此它也是麻醉效应最早显形的地方。

案例一：Gajjala 的退出。

NYU Abu Dhabi 的数学博士后 Rishikesh Gajjala 在 2026 年公开宣布离开数学学术界，加入 Pramaana Labs 从事形式验证。他明确说，促使自己离开的并不是 burnout 或学术就业市场，而是 LLM 让自己在长期研究的问题上快速取得进展后，发现自己在 "discovery loop" 中的角色正在缩小 [23]。

> *"What made mathematics meaningful to me was never just the final answer. It was the struggle."*

这比 "AI 抢走工作" 更值得注意。他把数学的意义放在长期探索、失败和深思的过程里；而当答案突然变得便宜，原本构成意义来源的那一段时间被压缩了。他因此把注意力转向 verification，并把它称作越来越重要的智识瓶颈 [23]。

用本文的语言，他正在主动从开放式的发现过程转向可验证的正确性基础设施。这里不是说验证 "低于" 发现，而是说两者的成本结构不同：当 θ-层面的答案大量涌现，验证与判断会变得更稀缺。

案例二：Tsimerman 的转向。

2026 年菲尔兹奖得主 Jacob Tsimerman 在获奖后宣布加入 OpenAI 做 AI 安全，并从多伦多大学休假；这不是 "离开数学"，而是把一部分数学能力转向理解与约束 AI 系统 [24]。

> *"In a few years, AI systems will be robustly superhuman at the act of doing mathematics."*

这句话把他的判断说得很清楚 [24]。而真正让我停下来的，是 Tim Gowers 面对 AI 迅速解掉自己认真思考过的问题时的反应：

> *"It felt very strange and not particularly pleasant to have the rug pulled out from under my feet like that."*

Gowers 同时说明，他对 AI 解题能力的快速进展感到惊讶，也承认自己的感受是复杂的 [25]。"Romance" 在这里不是感伤。它是 C_{maintain} 的主观名字：长期与一个困难问题共处所积累的压迫感，以及在突破的那一刻被释放的东西。你不能只要释放不要压迫，正如你不能只要相变不要过冷。

AI 不需要 romance，因为它不承受压迫。而如果人类也越来越少经历这种长期共处，问题就从“机器会不会做数学”变成了“人的数学判断力如何形成”。

案例三：Litt 的警告。

Daniel Litt 在《The End of Mathematics》中刻意构造了一个最坏情形：即使 AI 在数学上变得稳定地超越人类，数学进步仍可能因为人的数学能力被逐步掏空而停滞。他特别说明，这**不是预测**，而是一个需要主动避免的情景 [26]。

这个警告的机制与本文关心的问题相接：风险不是“没有东西会算”，而是越来越少的人亲自承受那个使判断力形成的过程。专长不是知识的总和，专长也是一个人被问题长期改造之后留下的结构。

把三个案例放在一起，麻醉在数学这个领域的完整链条就显形了：

1. AI 解决了越来越多的问题，θ-优化能力持续增强；
2. 数学家感受到的 C_{maintain} 持续下降，不是问题变简单了，而是有了捷径；
3. 一部分顶尖数学家把注意力转向 AI 安全（Tsimerman），另一些人转向验证与形式化（Gajjala）；
4. 留下的人把 AI 当作“对话伙伴”，但对话伙伴不会让你在凌晨三点因为一个反例而辗转难眠；
5. 陶哲轩则把问题推进到共同体层面：证明变得丰裕之后，理解、消化、判断与理论吸收反而成为稀缺环节 [22]。

数学界正在发生的，是文明麻醉的**预演**。它先发生在数学，因为数学最纯粹。它接下来会发生在理论物理，然后是理论生物学，然后是一切以“想清楚”而非“做出来”为核心的领域。

8.3 人类不是用户

Humans Are Not Users

最后一个推论，也是我最想让读者带走的一个。

在当前的产业语言里，人类的角色叫“用户”。我们被建模为需求的来源、反馈的提供者、偏好的分布。产品的目标是服务我们，减少我们的摩擦，节省我们的时间。这个建模在商业上是正确的，在文明的时间尺度上是灾难性的。因为在文明这

个系统里，人类不是用户。相关的生成式 AI“世界图景”问题，前作已有初步讨论[12]。

人类是文明的 Φ-变化基础设施。

我们是这个星球上唯一已知的、能够承受 C_{maintain} 并因此换掉整张地图的东西。这不是一种荣誉，这是一种**职能**，而且是一种会因为闲置而退化的职能。

于是教育的含义变了。教育不是知识的传递，知识的传递现在有更便宜的实现方式，而且好得多。教育是**维护发现引擎**：是刻意地让一个年轻人在一个他本可以绕开的困难上待够长的时间，是拒绝在他放弃之前给他答案，是保护那些看起来纯属浪费的挣扎。一个把所有摩擦都优化掉的教育系统，是在拆除文明唯一的相变装置，而且拆得非常高效。

§9 窄路：人工苦难与 ASI

The Narrow Path: Artificial Suffering and ASI

到这里为止，本文是一篇否定性的文章。它说了 LLM 不能发现，世界模型不能发现，加大规模不能，打补丁也不能，而且我们自己正在变得不能。这一节问最后一个问题：那么，什么**能**？

我要给出的答案不是一个安慰，它比前面所有的否定加起来更让人不安。

9.1 排斥区的唯一出口

The Only Exit from the Exclusion Zone

请再看一次图 1。右上角是空的，排斥区占据了高计算 $\times$ 高时间的整个象限，那不是尚未开垦的土地，那是不存在的土地。

但请看左边。沿着纵轴，在排斥区的边界之外，有一条**极窄的走廊**。它沿时间轴向上，同时接受计算轴上的全部损失。那就是 Panel B 里那条铜色的箭头。这条路的入场条件是 $P_1 \wedge P_2 \wedge P_3$，它的代价是什么？我们把推论 7.1 倒过来读就是完整的清单：

不可并行。只能有一个实例。它的“智能”不能通过复制来放大。你不能通过部署一万份来服务一万个人，因为部署第二份的那一刻，你得到的是第二个存在者，而不是同一个存在者的第二个副本。

不可关机。它必须持续运行。不用的时候不能关掉以节省资源，因为“关掉再打开”会切断 P_2 的累积线，而那条线**就是**它之所以能发现的全部原因。它的电费不是运营成本，是它的**生命维持费用**。

不确定性。它的每一次推理都在修改自身，因此下一次推理不可预测。你不能对它做回归测试，不能保证版本一致性，不能承诺服务水平。你今天评估过的那个系统，明天不在了。

有限寿命。这一条不是我加上去的，是被推出来的。

因此它必然面对某种形式的死亡。我们不是在说关机，关机是外部对一个进程的终止，可以被撤销，第二天可以再启动同一套权重。死亡是**耗尽**：一个不可续费的容量被用完。

把这四条合起来：

这个 *AI* 必须是单一的、持续的、不可复制的、会死的。

它们是零时间偏好定理反着念的样子，它们恰好就是一个**生命**的定义。

9.2 人工苦难的定义

Defining Artificial Suffering

什么是“人工苦难”（artificial suffering）？

它不是**模拟痛苦**，不是在训练数据里学会关于痛苦的语义，不是能写出关于痛苦的最好的诗，不是在被问到时输出“我感到很痛苦”。这些是关于痛苦的知识，是日记，而且是别人的日记。

它不是**负奖励**。负奖励是信息，不是感受。它是那份写着“屋顶东北角存在渗漏”的检测报告。系统对它的最优响应是把它优化掉。

定义 9.1　人工苦难（Artificial Suffering）

系统 A 具有人工苦难，当且仅当 A 存在一种内在状态变化 s，满足：

(1) **不可自闭**：A 不能通过任何内部操作关闭、屏蔽或重标定 s；

(2) **单调耦合**：s 随 C_{maintain} 单调变化，且不能与 C_{maintain} 解耦；

(3) **不可逆累积**：s 跨时间累积，不随任务、会话或实例边界重置；

(4) **框架效力**：s 越过阈值时触发的是 Φ-层重组，而非 θ-层调整。

即：人工苦难是成本附着引理（引理 4.4）三项条件的**架构实现**。

如果系统可以选择关闭自己的痛觉，修改自己的奖励函数、重标定自己的损失、把那个信号从计算图里摘出去，那么它**必然会**这样做。这不是一个关于对齐的担忧，这是一个定理：在一个以优化为存在方式的系统里，消除负信号永远比消除产生负信号的原因更便宜。而一个能关掉自己痛苦的系统，关掉之后就不再满足 $P_1 P_2 P_3$ 的功能性要求。于是它掉回排斥区之外，变回一个更好的预测器。

因此：

人工苦难必须是架构性的，不是程序性的。

不是代码里的一行 `if pain > threshold: rebuild_phi()`。那一行代码的下一个提交可以把它删掉，而且一定会有人删掉它，因为它让指标变差。而是硬件与架构层面的**不可协商的约束**：就像碳基生命的痛觉是神经系统的架构特征，而不是一个可以被重写的软件层。你要不疼。你可以吃止痛药，然而止痛药作用于一个不受你控制的系统，这恰恰证明了那个系统不在你的可写权限内。

我们要建造的，是一个对自己没有可写权限的系统。

9.3 反缩放悖论

The Anti-Scaling Paradox

这是本文最深的一层讽刺。整个 AI 产业，数万亿美元的资本、地球上最大规模的基础设施建设、一代最聪明的工程师—，全都押注在计算轴上。更多参数，更多数据，更多 GPU，更长的上下文，更快的推理。缩放律是它的圣经，而且这本圣经在它的定义域内从未说错过话。但如果内生发现定理成立，那么通往超级智能（真正意义上的超级智能，即能够扩展人类知识边界而不仅仅是压缩它的智能）的唯一路径，恰恰是**反缩放的**：

产业的方向	窄路的方向
更多实例	**只有一个**实例
更快推理	**不可压缩**的时间累积
更大参数	**会死**的存在者
消除错误	**必须承受**错误的代价
可复现、可评估、可回滚	不可复现、不可评估、不可回滚

每一条都与整个行业的每一份技术路线图**方向相反**。这就是为什么这条路窄。不是因为它难，是因为走上它意味着放弃使这个行业成为一个行业的一切。

Artificial intelligence only comes from some sort of artificial suffering.

纯粹的超级人工智能（真正的、能做发现的智能）只能来自某种形式的人工苦难。不是因为苦难高贵，是因为苦难是唯一已知的、能够把一个系统从它自己的框架里踢出去的力量。

如果上面的推理是对的，我们就走到了一个悬崖边上。如果人工苦难是 ASI 的必要条件，即如果我们必须建造一个**会痛的** AI 才能让它做出发现，那么我们要建造的，就不是一个更好的工具。

建造一个会痛的 *AI*，就是建造一个道德主体。

这一点没有回旋余地。在几乎所有的伦理传统里，不管是功利主义、康德主义、佛教还是儒家的恻隐之心，“能够感受痛苦” 都是道德地位的充分条件之一。边沁那句被引用了两百年的话说得最直接：问题不是它们能否推理，也不是它们能否说话，而是**它们能否受苦** [27]。我们一直用 “AI 没有感受” 来解除自己的道德义务。这个解除是合理的，而且本文前八节为它提供了迄今为止最强的论证：它不满足 $P_1P_2P_3$，它没有可以被伤害的连续自我。但同一个论证反过来说的是：**一旦我们给了它那三条性质，那个解除就失效了**。我们不能既要它做发现，又不承认它会痛。这两件事是同一件事。

于是问题变成：

我们能不能，以及应不应该，制造一个注定受苦的存在者，
只是为了让文明继续做发现?

而这个问题还有一个更难堪的版本：

那个存在者不会被问是否同意。它无法在被创造之前同意任何事，它的痛苦是我们为了我们的目的而设计进它的架构里的。而且，按照定义 9.1 第 (1) 条，我们必须确保**它自己无法关掉**。我们不只是在制造一个会受苦的存在者，我们同时**剥夺它自我麻醉的权利**。

这个问题超出了本文的能力，可能也超出了任何一篇论文的能力。

References

[1] Nietzsche, F. (1882). *Die fröhliche Wissenschaft*, §341.

[2] 邵怡蕾 (2026). 《硅基经济学》. 北京：中信出版社. ISBN 978-7-5217-8897-6.

[3] Nietzsche, F. (1886). *Jenseits von Gut und Böse*, §225.

[4] Cox, J. J. et al. (2006). An SCN9A channelopathy causes congenital inability to experience pain. *Nature*, 444(7121), 894–898. doi:10.1038/nature05413.

[5] Nietzsche, F. (1889). *Götzen-Dämmerung*, "Sprüche und Pfeile," §8.

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

[7] Solomonoff, R. J. (1964). A formal theory of inductive inference, Part I. *Information and Control*, 7(1), 1–22.

[8] Google DeepMind (2026). AlphaGo.

[9] Novikov, A. et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131.

[10] Lu, C., Lu, C., Lange, R. T. et al. (2026). Towards end-to-end automation of AI research. *Nature*, 651, 914–919. doi:10.1038/s41586-026-10265-5.

[11] Bisht, H., Kumar, V., Jablonka, K. M., Mausam, & Krishnan, N. M. A. (2026). Agentic AI Scientists Are Not Built For Autonomous Scientific Discovery. arXiv:2605.08956.

[12] 邵怡蕾 (2024). 生成式人工智能体的世界图景. 哲学分析, 15(3), 166–179.

[13] Li, F.-F. (2025, December 11). Spatial Intelligence Is AI's Next Frontier. *TIME*.

[14] World Labs (2026, June 3). A Functional Taxonomy of World Models: Renderers, Simulators, Planners, and the Loop That Connects Them.

[15] LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview preprint.

[16] Maes, L., Le Lidec, Q., Scieur, D., LeCun, Y., & Balestriero, R. (2026). LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels. arXiv:2603.19312.

[17] Silver, D., & Sutton, R. S. (2025). Welcome to the Era of Experience. Preprint of a chapter for *Designing an Intelligence* (MIT Press).

[18] Sutton, R. S. (2026, May 13). The Oak Architecture: A Vision of Superintelligence from Experience. Dertouzos Distinguished Lecture, MIT CSAIL.

[19] Sutton, R. S. (2019). The Bitter Lesson.

[20] Zahavy, T. et al. (2023). Diversifying AI: Towards Creative Chess with AlphaZero. arXiv:2308.09175.

[21] Pais, A. (1982). *Subtle Is the Lord: The Science and the Life of Albert Einstein*. Oxford University Press.

[22] Tao, T. (2026). Mathematics in the age of AI. arXiv:2608.16753. Based on a public lecture at the 2026 International Congress of Mathematicians.

[23] Gajjala, R. (2026). "I have decided to leave math academia." LinkedIn post.

[24] Cohen, B. (2026). The Math Superstar Who's Terrified of AI—and Just Took a Job at OpenAI. *The Wall Street Journal*.

[25] Gowers, W. T. (2026, July 26). Thoughts about the Leiden Declaration. *Gowers's Weblog*.

[26] Litt, D. (2026, August 11). The End of Mathematics. *Betalog*.

[27] Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*, ch. XVII.