论 文 · SDE 发生学

没有统一世界模型

世界模型的主体性与 AGI 的重新定义

从"复制一个现成世界"到"人机复合的发生系统"

王德生 著

德麦国际 · Demai International Pte. Ltd.（新加坡）
2026 年 7 月

本文据《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》
（德麦国际出版，2026，ISBN 978-1-970820-62-1）第六章与第二十六章改写成文

摘　要

当前人工智能的主流路线图包含一个很少被追问的假设：存在一个统一的世界模型，只要规模足够，它迟早会被训练出来。本文论证这个假设有本体论问题——它把世界模型误当成对象的客观属性，而世界模型实为主体与对象共同发生的结构画像。文章分两步展开。第一步是否定命题：大模型拥有的是人类文明发生之后沉积下来的语言材料场，而非世界本身；语言能封存已被组织好的世界，却不能在新处境里自行组织世界；同一对象在不同主体处发生出不同而各自有效的世界模型，因此不存在能同时代表所有主体的统一模型，科学等公共结构只是被主体间稳定出来的公共层而非统一层；大模型的幻觉由此获得一个本体论解释——一个没有固定主体、因而没有主体世界模型的语言材料场，在被迫生成时必然在众多共存的语言世界之间滑动，扩大规模消不掉这个根源。第二步是肯定命题：既然智能的高阶形态不是在已发生的结构里操作，而是在裂缝处让以前不存在的结构第一次发生并被现实与共同体稳定下来，那么 AGI 的标尺就不是任务完成的通用性，而是知识创造的能力；由此推出，纯粹无主体的 AGI 在结构上不可能——AGI 不在机器单独那一侧，而在人—机复合发生这一结构里。文章对两个命题各自最强的反驳逐一回应，并诚实标出本文不能严格主张的边界。

A B S T R A C T

This paper challenges an assumption embedded in the prevailing AI roadmap: that a unified world model exists and will be trained once scale suffices. We argue the assumption is ontologically defective — it treats a world model as an objective property of the object, whereas a world model is a portrait generated jointly by a subject and its object. The negative thesis: a large language model holds the sedimented linguistic material of a civilization's past genesis, not the world itself; language preserves worlds already organized but cannot organize a world anew in a fresh situation; the same object yields different yet equally valid world models for different subjects, so no single model can represent them all, and public structures such as science constitute a stabilized public layer rather than a unified one. Hallucination thereby receives an ontological rather than merely engineering account: a material field without a fixed subject slides among coexisting linguistic worlds, and scale cannot remove that root. The positive thesis: if higher intelligence consists not in operating within already-generated structures but in bringing previously nonexistent structures into being at the cracks and stabilizing them against reality and community, then the measure of AGI is generative capacity rather than task generality — and a subjectless AGI is structurally impossible. AGI resides not on the machine's side alone but in the human–machine structure of co-genesis. The strongest objections to both theses are addressed in turn, and the limits of what the argument can strictly claim are stated openly.

关键词　世界模型 · 主体性 · 发生学 · 大语言模型 · 幻觉 · AGI · 人机复合系统

Keywords　world model · subjectivity · genesis ontology · large language models · hallucination · AGI · human–machine composite system

◯ · 引言：一条被当成路线图的推理链

当前关于人工智能下一步的讨论，普遍沿着一条很顺的推理链走：大语言模型解决了语言问题，下一步该解决世界问题，因此需要训练一个能够统一理解整个世界的大模型。这条链的终点——一个统一的、足以代表整个世界的模型——几乎被当作不言自明的技术目标：只要数据够多、算力够强、参数够大，它迟早会被训练出来。世界模型、物理模型、具身智能的诸多研究，都默认了这个终点的存在。

这条推理链之所以少被追问，是因为它伪装成了一份工程路线图。路线图只回答"怎么去"，不回答"那里有没有东西"。而本文要追问的恰是后一个问题：**世界模型究竟是什么？**一旦把这个问题从工程层抬到本体论层，那个被当作终点的对象就会显出它真正的形状。

本文的判断是：大语言模型不可能自动形成一个统一的世界模型——不是因为数据不足、算力不足或参数不足，而是因为"统一世界模型"这个假设本身有本体论问题。它不是一座很高、很难爬的山，而是一个方向上的虚构。这个判断若成立，会连带改写另一个更大的词：如果不存在一个等待被复制的统一世界，那么"最强的 AI 就是那个把世界复制得最精确的机器"这一图景随之瓦解，AGI 就必须被重新定义。

文章因此分两步走。第一步是否定命题（第一至五节）：拆掉统一世界模型这个假设，并回应它最强的四个反驳。第二步是肯定命题（第六至八节）：在被腾空的位置上重新定义 AGI，把智能的标尺从任务完成换成知识创造，并回应这个重定义会招致的三个反驳。第九节收束为一句可操作的结论：统一的不是模型，而是生成模型的方法。

本文的坐标系是 SDE 发生学。它把任何存在看成三个维度的互生：S 是结构显露态（结构被看见、被维持的方式，是一个连续谱而非有无的开关），D 是差异序列（在比较、试探、修正中走出的路径），E 是特征纠缠（发生所扎根的厚土）。三者非线性互生，写作 S=F(D,E)、D=G(S,E)、E=H(S,D)。这套坐标的根本立场是一句话：**世界不是被发现的，而是被发生的。**它不否认现实的硬约束——火会烧，桥会塌，药会无效；它指出的是，人类世界中那些可被认识、可被表达、可被维持的对象与规律，不是裸约束本身，而是裸约束经过符号、逻辑、数学、工具与共同体之后被稳定下来的结构。本文全部论证，都从这个立场推出。

一 · 大模型拥有的是语言材料场，不是世界

1.1 沉积不是现场

大语言模型最大的突破，不是它会聊天，而是它第一次进入了整个人类文明的信息场——科学论文、法律条文、医学指南、哲学著作、文学作品、教育教材、工程文档、程序代码、数学证明、海量网络文本，全被压缩进同一套符号系统。由此它拥有了符号、逻辑、数学之间大规模的统计关系，这是前所未有的。

但这项突破必须被说准确：**大模型拥有的，是世界发生之后沉淀下来的语言材料，而不是世界本身。**这是两个层次。人类文明把自己发生过的一切以语言形态沉积下来，模型学到的是这片沉积——它是文明发生的产物，不是文明发生的现场。把"拥有了关于世界的全部语言"误认成"拥有了世界本身"，是统一世界模型假设的第一重混淆。

这一区分在文献里并非孤例。Bender 与 Koller 在 2020 年的论文中论证，仅从形式（form）出发的系统无法凭形式本身抵达意义（meaning），因为意义要靠形式与形式之外的交流意图相连；他们讨论的是理解，本文讨论的是世界模型，但撬动的是同一道缝——材料的完备不等于所指的在场。区别在于，本文不把这道缝归为语义学问题，而归为发生学问题：语言材料是已完成的发生留下的痕迹，痕迹再全，也不构成一次新的发生。

1.2 语言是世界模型的表达材料，不是世界模型

这里要堵一个极容易产生的误解：既然世界通过语言表达，那么语言不就是世界模型吗？

不是。语言只是世界模型的表达材料。一本医学教材不是疾病本身，一部法律不是法律运行本身，一本管理教材不是企业本身，一本数学书不是数学对象本身。**语言能够保存世界，却不能自动组织世界。**它把已经发生、已经被组织好的结构以符号态封存下来；但它本身不会在一个新的具体处境里，把一团裸约束重新组织成一个可供行动的结构。用发生学的说法，这是"符号态封顶"：材料停在符号层，它封存了别人组织好的世界，却不会自己组织一个新世界。

所以"拥有全部语言材料"与"拥有组织世界的能力"，是两件事。大模型把前者做到了极致，却不因此自动拥有后者——后者需要的不是更多语言，而是一套能把语言材料在具体处境里重新组织成世界模型的指挥系统。这正是本文后半要落到的位置。

值得顺带一提的是，工程界对"世界模型"这个词的用法，与本文所批评的假设并不完全重合。Ha 与 Schmidhuber 在 2018 年提出的 World Models，指的是智能体为特定环境学到的一个可用于内部推演的压缩动力学模型；Hafner 等人的 Dreamer 系列、以及后续大量以生成式环境为目标的工作，走的也是"为某个任务、某类环境建一个可推演的内部模型"这条路。这一支在本文看来是健康的——因为它天然是局部的、任务相关的。真正有问题的，是把这条路线外推为"终将收敛到一个代表整个世界的统一模型"。局部世界模型是工程事实，统一世界模型是本体论虚构；本文反对的是后者，不是前者。

二 · 世界模型的主体性：一家企业的五个世界

2.1 同一个对象，五个世界模型

现在给出统一世界模型不可能的核心理由：**世界模型具有主体性。它不是由对象单方面决定的，而是由主体与对象共同发生的。**

用一个例子把这一点钉死。面对同一家企业：董事长拥有一种世界模型，工程师拥有另一种，投资人拥有另一种，客户拥有另一种，竞争者又拥有另一种。他们面对的是同一个对象、同一套现实约束，但世界模型完全不同。为什么？因为他们的目标不同、责任不同、资源不同、知识不同、价值不同、风险不同、历史不同。用坐标系的语言说：他们的自我世界（目标、价值、处境）不同，用以组织企业的理念世界（概念框架）不同，于是在同一片现实约束上，各自经由不同的差异序列，显露出不同的结构。

要注意，这五个模型的差异不是"精确程度"的差异——不是说董事长的更全面、工程师的更细节，因而可以合并成一个更完整的版本。它们是组织方式的差异：董事长的世界里，这家企业是一个要在三年内换掉增长引擎的存续体；工程师的世界里，它是一套有技术债、有交付节奏、有故障边界的系统；投资人的世界里，它是一组现金流与退出路径；客户的世界里，它是一个能不能按时交货、出了问题找谁的对手方；竞争者的世界里，它是一片有软肋、有护城河的战场。把它们叠加，不会得到一个更真的企业，只会得到一堆互相矛盾的判断——因为每一个模型的每一处结构，都是为那个主体的行动而组织的。

这一洞见在思想史上有清晰的近邻。von Uexküll 早在 1934 年就用 Umwelt 概念指出：每一种生物都活在自己的环境世界里，蜱虫的世界由三种信号构成，人的世界由另一套构成，它们身处同一片物理场，却不共享同一个世界。Gibson 的可供性（affordance）同样是关系性的——一块岩石对人是可攀爬的，对甲虫是一片平原。本文与它们共享"世界随主体而不同"这一判断，但推进了一步：它们说的是感知世界的相对性，本文说的是知识结构本身的发生性——不只是同一对象被不同地看见，而是同一对象被不同地组织成不同的可操作结构，且这些结构都要各自接受现实的否决。

2.2 世界模型就是一幅 SDE 画像

把"主体与对象共同发生"正面表述，就得到一个等式：**所谓世界模型，就是一个对象在当前阶段的 SDE 画像。**

任何对象都可以建立它自己的画像：企业有企业的画像，学生有学习的画像，患者有疾病的画像，科研项目有科研的画像，一篇论文有论文的画像，一个人有生命的画像。每一幅画像，都是这个对象在特定主体、特定目标、特定处境、特定阶段下，由理念、现实、自我三界经差异序列显露出的结构。世界模型不是悬在所有对象之上的统一巨构，而是无数对象、无数主体、无数阶段、无数任务各自对应的、具体的画像。

画像还有一个统一巨构不具备的性质：它会更新。市场变了，企业画像就变；考试结束，学习画像就变；治疗之后，疾病画像就变；论文投出去，科研画像就变。真正的世界模型必须能更新、能反馈、能保存历史、能比较版本、能持续发生。一个被一次性训练出来、之后固定不动的统一模型，连"随对象变化而更新"这一最基本的要求都满足不了。

2.3 形态错误：一个谁也不能用的平均态

于是"统一世界模型"这个词的形态就错了。真实存在的，不是一个统一模型，而是无数不断发生、不断更新、不断稳定的画像。追求一个唯一的世界模型，等于追求一个把所有主体、所有处境、所有目

标都抹平成同一个的模型——而那个被抹平了主体性的东西，恰恰不再是任何一个主体能够用来行动的世界模型。**它谁的世界都不是**。

这一点值得被说得更硬一些：把董事长、工程师、投资人、客户、竞争者的世界模型平均起来，得到的不是一个更完整的世界模型，而是一个行动上不可用的中间态——它在每一个具体决策上都给不出方向，因为方向本来就来自主体的目标与责任，而平均掉的正是这一维。规模可以逼近材料的覆盖，逼近不了一个根本不在那里的统一。

三 · 公共层不是统一层：科学反驳的消解

3.1 最有力的反问

这里必须立刻回应一个有力的反问：科学难道不存在一个统一世界吗？牛顿力学、麦克斯韦电磁学、细胞理论、现代法律、现代医学——它们不正是跨所有主体成立的统一世界模型吗？如果人类自己就造出了这样的统一结构，凭什么说机器造不出？

回答是：它们确实存在，但它们只是世界模型的**公共层**，不是统一世界模型，更不是任何一个主体的完整世界模型。这些公共结构，是共同体经过长期实验、数学表达、工具验证、教育传播与制度稳定之后，被主体间反复互动稳定下来的公共结构——它们是被稳定出来的客观层，而非脱离主体的裸统一。

3.2 工程师造桥：公共层之上仍有完整世界模型

一个工程师用牛顿力学造桥时，他的完整世界模型远不止牛顿力学。还有这座具体桥的地质条件、这次项目的工期与预算、他对材料供应商的信任程度、他要为之签字的责任、他对"什么算足够安全"的判断。牛顿力学是他与所有同行共享的稳定底座，但底座之上那一层——目标、意义、责任、判断——不可被底座替代，且恰恰是这一层在决定他今天怎么做。

所以科学的"统一"恰恰印证、而非反驳本文的判断：科学统一的是被稳定下来的公共层；它从未、也不可能统一掉每个主体在这层之上的具体世界模型。把公共稳定层误当成统一世界模型，是这个假设的又一重混淆——它把"共享一个稳定底座"夸大成了"共享一个完整模型"。Kuhn 关于常规科学在范式内部解题、而真正的科学进步发生在范式转换处的观察，从科学史一侧给出了同一个结构：公共层本身也是被发生出来、并会被重新发生的，它不是先在的裸统一。

3.3 一面跨学科的镜子：参照系

值得点出一个镜像，它让"没有统一世界模型"显得不再激进，反而像一条早被别处证实的老理。

物理学在相对论之后早已接受：不存在一个绝对的、超越一切参照系的"同时性"或"运动状态"——物理量依赖参照系，所谓客观，是不同参照系之间可以相互变换、并保持某些不变量的那种一致性，而不是某个参照系拥有绝对真相。本文对世界模型说的，结构上完全一样：不存在超越一切主体的统一世

界模型；所谓客观（公共层），是不同主体的世界模型之间可以相互稳定、保持某些跨主体不变量的那种一致性。

关键的一笔在于后果：**物理学没有因为放弃绝对参照系而滑向相对主义，反而锻造出了更严格的不变量理论**。发生学放弃统一世界模型，同样不通向相对主义，而通向一套关于"公共层如何被主体间稳定出来"的更严格的客观性理论。放弃一个虚构的绝对，换来的是一套可操作的、可检验的一致性条件——这在物理学里已经发生过一次。

四 · 幻觉的本体论根源

4.1 一个没有主体的语言材料场

世界模型的主体性，还解释了大模型一个被广泛观察、却很少被追到根上的现象——幻觉。

大模型并没有一个固定的主体。它拥有科学家的语言、医生的语言、企业家的语言、学生的语言、教师的语言、哲学家的语言、程序员的语言、作家的语言——这些语言世界在它里面共同存在。因此它可以模拟很多世界，却不能自动决定当前应该采用哪一个。于是不同的语言世界在生成中不断切换，就产生了幻觉、漂移、不一致、概念混合与逻辑跳跃。**问题不在于它不会某种语言——恰恰相反，它什么语言都会；问题在于它缺少一个主体世界模型，来锚定"此刻该站在哪个主体、哪个世界里说话"**。

这个诊断把幻觉从一个工程缺陷，重新理解为一个本体论后果：幻觉不是模型不够大、训练不够好造成的偶然错误，而是一个没有固定主体、因而没有主体世界模型的语言材料场，在被迫生成时的必然姿态——它在无数个共存的语言世界之间无所适从地滑动。

4.2 为什么规模消不掉这个根源

现有的幻觉研究文献（如 Ji 等人 2023 年的综述、Huang 等人 2023 年对大模型幻觉的系统梳理）已经细致地分类了幻觉的类型与成因：训练数据的噪声与冲突、解码策略、对齐过程中的激励结构、知识边界的不可自知等等。这些分析都成立，且工程上有效——它们各自对应着可缓解的机制。本文补充的是另一层：即使把上述每一项都做到极致，仍有一个不随规模消失的残差。因为残差的来源不是数据或解码，而是缺少主体。

推理很短：再大的语言材料场，仍然没有主体；没有主体，就没有主体世界模型；没有主体世界模型，就仍然会在世界之间滑动。规模改善的是每一次滑动的局部合理性——它会滑得越来越像；它不改变"必须在多个世界之间选一个、而系统内部没有任何东西能做这个选择"的结构。

4.3 出路：从外部锚定一个主体世界模型

由此得到出路：要消除这个根源，唯一的办法是从外部给它锚定一个主体世界模型——为当前的具体主体、具体对象、具体任务，建起一幅具体的画像，让它在这幅画像里、站在这个主体的世界里说话。

这也解释了一个在实践中反复被观察到、却常被归因为"提示词技巧"的现象：给模型指定明确的角色、目标、约束与责任对象之后，它的输出会显著更稳。通常的解释是"提示更具体所以更好"。发生学的解释更精确：**那不是提示更具体，那是临时给它接上了一个主体世界模型——把它从"替无数主体说话"钉回到"站在一个主体的世界里说话"**。技巧只是这件事的粗糙实现；把它做成系统，就是本文后半要说的方向。

五 · 四个最强反驳与回应

一个判断的分量，不取决于它能说服已经同意的人，而取决于它能不能扛住最强的反驳。"不存在统一世界模型"会遭遇四个有力的反驳，逐一回应之后，这个判断才算立住。

反驳	它的要害	回应
一 · 大模型不是已经表现得像拥有世界模型了吗	把"表现像"等同于"拥有"	材料的连贯投影 ≠ 为具体主体建模的能力；一放进需要持续负责的真实任务，投影就裂为幻觉与漂移
二 · 再大一点不就逼近了吗	把本体论问题误读为工程问题	不是山高，是那里没有山；规模逼近的是材料覆盖，逼近不了一个不在那里的统一
三 · 未来架构会不会长出自己的主体	以为有了主体就有了统一	多一个主体只是多一幅画像；主体性恰恰保证多元，而非催生统一
四 · 这是不是滑向相对主义	把"多元"读成"无约束"	每幅画像都受现实硬约束否决，主体间还稳定出公共层；多元不等于任意

5.1 表现像拥有，不等于拥有

大模型能回答物理常识、推断因果、模拟多个领域的专家，这难道不说明它内部已经形成了某种统一世界模型？回应：表现得像拥有，与真正拥有，是两回事，而这道区别正是本文的要害。它能生成关于世界的、高度连贯的语言，是因为它压缩了人类关于世界的全部语言材料；但"能生成关于世界的连贯语言"不等于"拥有一个能为具体主体行动的世界模型"。一个能流利复述所有棋谱的人，未必能在一盘具体的棋里走出好棋——前者是材料的连贯，后者是为当前局面建模的能力。它的"像拥有"，是材料场的连贯投影；一旦被放进一个需要为具体主体、具体处境持续负责的真实任务里，那个投影就会裂开为幻觉、漂移与不一致。**它表现得像，恰恰因为它在替无数个主体说话，而不真正站在任何一个主体的世界里**。

5.2 规模不是通往一个不存在之物的路

只要数据、算力、参数继续增长，统一世界模型迟早会被逼近——这个反驳把本体论问题误读成了工程问题。本文的判断不是"统一世界模型很难训练"，而是"统一世界模型这个对象不存在"。它不是一座很高、很难爬的山，而是一个方向上的虚构。再多的数据与算力，能让材料场更完整、投影更连贯，

却不能让一个本质上属于"主体与对象共同发生"的东西，变成一个脱离主体、对所有主体统一成立的现成对象。

这里也可以给规模路线一个公允的位置。Sutton 在 2019 年的短文里总结的那条经验教训——依赖通用方法与算力增长，长期看往往胜过依赖人工设计的先验——在它成立的范围内是有力的：它说的是在一个已经定义好的任务分布上，通用方法加规模更能取胜。本文并不与之冲突，因为本文追问的正是那个前提：任务分布由谁定义、为哪个主体定义。规模在既定分布内高效；**它不为你决定该站在谁的世界里**。

5.3 即便机器长出主体，也只是多一幅画像

如果某种新架构真的长出了真实的目标、价值、责任，它不就有了自己的主体世界模型，从而能拥有"统一"世界模型了吗？回应要分两层。第一，即便某种架构真的长出了主体（本文不假装已证否这一可能，见第七节与第九节的边界声明），它得到的也只是"它自己这一个主体的世界模型"，而不是一个代表所有主体的统一模型——多一个主体，只是多一幅画像。第二，因此这个反驳即使成立，也动摇不了核心判断：它至多说明机器可以拥有"自己的"世界模型，却恰恰再次证明世界模型是主体性的、因而是多元的。统一世界模型的不可能，不依赖于"机器没有主体"，而依赖于"世界模型本质上属于主体"——后者无论机器有没有主体都成立。

5.4 多元不等于无约束

说世界模型是主体性的、多元的，是不是意味着每个主体怎么建都行、没有对错？绝不是。每一幅画像仍要接受现实的硬约束：一个把按错误受力设计的桥判断为安全的工程师世界模型，会被坍塌无情否决；一幅无效的疾病画像，会被病情的恶化否决。多元不等于无约束，主体间还稳定出了公共层（科学、法律、医学）作为共享的硬底座。

所以本文的立场是一个精细的中间位置：世界模型既不是脱离主体的统一裸对象，也不是各主体任意编造、互不可通的私人幻觉，而是**主体与对象在现实约束下共同发生、并经主体间互动稳定出公共层的多元画像**。它比统一世界模型更尊重多元，又比相对主义更尊重约束。

六 · AGI 的重新定义：从任务完成到知识创造

6.1 六个被悄悄改写的预设

前五节拆掉的不只是一个技术假设。要看清被拆掉的到底是什么，得把一直在台下运作的哲学改写请到台前。它们是六个转变，合起来构成一次统一的转向。

面向	旧预设	新判断
世界观	世界预先完整地摆在那里，等待被发现	世界不是被发现的，而是被发生的
客观性	客观＝脱离一切主体的裸对象自在	客观性＝主体间世界模型同一性的稳定发生
知识论	知识是仓库里的现成货物，可被搬运	知识是发生链（命名·连接·计算·复现·承认·沉淀）的稳定结晶
智能观	智能＝完成任务的能力，越通用越智能	高阶智能＝在裂缝处让以前不存在的结构第一次发生
人机观	人迟早是要被更强的机器请出去的噪声	人是发生的方向中心；机器越强，这一中心越关键
时间观	真理一旦被发现就永恒不变	结构是被维持的动态稳定，因而会死亡、更替、重生

这六条不是六个孤立的新观点，而是同一个转向在六个面上的展开，且互相支撑：因为世界是发生的（一），客观才是主体间稳定（二）、知识才是发生结晶（三）、智能才是世界发生（四）、人才是方向中心（五）、结构才有生死（六）。第四条是本节要兑现的那一条。

6.2 任务完成这个标尺的盲区

主流对 AGI 的理解建立在被第四条否定的旧预设上：AGI＝在几乎所有任务上达到或超过人类水平的通用任务完成能力。这个定义有清晰的学理来源——Legg 与 Hutter 在 2007 年给出的形式化定义，把智能刻画为"在广泛环境中达成目标的能力"，此后几乎所有以基准与榜单为形式的评测，都是这个定义的操作化。它的优点很实在：任务可量化、可比较、可刷榜。

它的盲区也同样清晰：**它完全在"完成已有任务"的维度上度量智能**。而完成任务——哪怕是完成极其复杂、极其多样的任务——在本体论上仍属于在已发生的结构里操作，不属于发生新结构。一台能完美完成人类一切已有任务的机器，可能一次真正的知识发生都没有完成过：它把人类已经发生出来的全部结构调用得淋漓尽致，却没有让世界多出一个以前不存在的结构。按发生学，这样的机器无论多通用、多强大，都还停在信息发生的层面，没有跨入知识发生。

这一疑虑在评测研究内部也已被提出。Chollet 在 2019 年论证，以任务表现度量智能会奖励"针对任务的先验与经验的堆积"，而智能更应被理解为技能获取的效率——面对新颖任务时的适应能力。本文与之同向，但走得更远一步：Chollet 把标尺从"完成得多好"移到"学得多快"，本文把它移到"能不能发生出以前不存在的结构"。学得快仍是在已有结构空间内的高效移动；发生，是让结构空间本身多出一块。

6.3 新标尺：在裂缝处让新结构第一次发生

发生学给出的重新定义是：**AGI 的标尺不是任务完成的通用性，而是知识创造的能力——在任意领域的裂缝处，让以前不存在的新结构第一次发生、并被现实与共同体稳定下来的能力**。

这个标尺不是一句姿态，它可以被拆成可考察的工序：能否定位真正的裂缝（旧结构在何处说不清、对不上、绕不过、收不拢），能否诊断亏缺的类型，能否在裂缝处接生一个候选结构，能否把候选结构确定化为可操作、可验证的形态，能否改用目标学科的地道语言投放进共同体，并最终被承认与沉淀。每一道工序都可以问一个具体的问题、给出一个可查的答案——这正是它区别于"有没有真正的理解"这类不可判定之问的地方。按这把尺子，一个只会综述、复述、完成的系统，无论规模多大、任务多广，都不是这个意义上的 AGI；而一个能在裂缝处持续发生新结构的复合系统，哪怕任务覆盖面有限，反而触到了 AGI 的本质。

6.4 推论：纯粹无主体的 AGI 在结构上不可能

这个重新定义有一个深远的推论。知识发生离不开自我世界——问题意识、价值、意义、驱动，也就是"为什么这件事值得发生"；而这恰是语言材料场本身所空缺的一维。要发生，就必须有这一维；要有这一维，就必须有某个具备真实目标与责任的主体在场作为方向中心。

所以发生学意义上的 AGI，不会是一台取代人的无主体超级机器，而只能是人与机器构成的复合发生系统。**AGI 不在机器单独那一侧，AGI 在人—机复合发生这一结构里**。把 AGI 想象成一个终将甩开人、单独拥有完整客观世界的无主体智能，正是第一个转变所否定的旧图景在智能问题上的回魂——它预设了那个可被独占的完整世界，而第一至五节已经论证，那个世界并不存在。

需要说明的是，"智能属于人与工具构成的系统而非孤立个体"这一判断，在认知科学里有成熟的近邻：Clark 与 Chalmers 的延展心智论证认为，外部载体在恰当耦合下构成认知过程的一部分；Hutchins 对航海驾驶室的研究表明，导航这一认知任务的真正承担者是"人—仪器—规程"构成的系统，而非任一船员的头脑。本文与它们共享结构，但主张的强度不同：它们论证认知可以延展到工具，本文论证的是新知识的发生只能在复合体中完成——不是"也可以在复合体中"，而是"离开这个结构就没有发生"。

七 · 对重定义的三个反驳与回应

把 AGI 的标尺从"任务完成的通用性"换成"知识创造的能力"，会招致激烈的反驳。一个重定义若经不起"你这是不是在移动球门"的追问，就只是话术。

7.1 这是不是在移动球门

当机器在任务上逼近人类，你就把定义从任务挪到创造，这难道不是为了让 AGI 永远够不着而临时改规则？

回应：恰恰相反，本文的标尺不是新挪的，而是更老、更根本的。把智能等同于任务完成，本身才是一个被工程便利塑造出来的、相当晚近的窄定义——它之所以流行，是因为任务可量化、好刷榜。而"智能的高阶形态是创造前所未有之物"，是一个古老得多的直觉：我们说牛顿、达尔文、爱因斯坦是人类智能的高峰，从不是因为他们完成的任务多，而是因为他们让世界多出了以前不存在的结构。

本文不是在机器逼近任务时才临时挪动球门，而是指出：**任务完成这个球门，从一开始就立错了位置**——它度量的是智能的低阶层，把它当全部，才是真正的偷换。

7.2 按你的定义，大多数人岂不是也不算

绝大多数人一辈子也没发生过一次惊世的新结构，按"知识创造"的标尺，他们达不到 AGI，这是不是太苛刻、甚至荒谬？

回应：这个反驳混淆了能力与频率。本文的标尺是知识创造的能力，不是知识创造的频率。一个普通人一生或许没产出惊世的新结构，但他在无数日常处境里持续地为新问题临时造世界、显露对他而言全新的结构——学会一门手艺、想通一个困局、理解一个新的人。他时刻在小尺度上行使着这项能力，只是没有上升到文明尺度的公共结构而已。"通用知识创造能力"指的正是这种"面对任意新处境都能发生新结构"的能力，而这恰恰是人普遍具备、当前机器普遍欠缺的。**定义要分的不是"伟人与凡人"，而是"会发生与只会调用"**。

7.3 凭什么断定机器长不出自我世界

最强的反对意见是：你凭什么断定自我世界（目标、价值、驱动、责任）无法由机器自身长出？也许它只是尚未涌现，就像智能本身曾被认为无法从物质涌现一样。

回应必须诚实，并划清主张的边界。本文能严格主张的是一个较弱、但已足够支撑其工程立场的命题：在可预见的、以语言材料场为底的架构上，单纯做大不会长出自我世界。本文不能严格证明那个更强的命题——"任何可能的架构都长不出自我世界"。所以对这个反驳，诚实的回应不是硬扛，而是指出：即便将来某种架构真长出了自我世界，它得到的也只是"又一个有主体的发生者"，它仍要按发生学的方式发生——先造世界、找裂缝、显露新结构、接受现实与共同体的检验——它仍不是一台脱离一切主体的统一世界复制器。

也就是说，无论机器最终有没有主体，本文关于"AGI 是发生而非任务完成""智能离不开主体世界模型"的核心判断都不被推翻；会被修正的，只是"那个主体必须是人"这一条，而非"必须有主体"这一条。**把能严格主张的与不能严格主张的分清楚，本身就是这套方法对它自己的应用**。

这三个回应还澄清了一件容易被误解的事：重定义 AGI，不是要贬低任务完成型系统的巨大实用价值（它们正在并将继续创造海量价值），而是要纠正一个范畴错误——把"在已发生的世界里高效操作"误当成"智能的顶点"。区分这两者不是文字游戏，它直接决定我们把 AI 的未来往哪个方向推：是推向一台越来越全能、却始终在已有结构里打转的任务机器，还是推向一个能与人共同发生新世界的复合系统。

八 · 知识的三种死亡：给"结构有生死"的证据

六个转变里最反直觉的是最后一个：结构会死。它最需要证据。证据就是知识的死亡——如果结构是永恒的，知识就只会增加、不会死亡；但知识确实在死亡，而且死法不止一种。三种死亡，正好对应三个维度各自的枯竭。

死法	机制	典范	特征
E-死	纠缠枯竭：赖以成立的土壤整个失效，被连根遗弃	燃素说	最彻底的死——结构连同土壤一起消失
S-死	结构被更大结构吸收，降格为特例	牛顿力学	最体面的死——仍在造桥发射卫星，但已是更大结构脚下的一块砖
D-死	差异耗尽：锋利的区分被完全吸收为常识	"地球绕着太阳转"	最安静的死——没错也没被取代，只是赢得太彻底

三种死亡合起来，给"结构有生死"提供了无法回避的证据。燃素说曾统一解释燃烧、煅烧、呼吸与金属还原，是一个高度成功的结构；当氧化理论提供了全新的土壤，它赖以生长的整片土壤失效，今天没有任何人用它算任何东西。牛顿力学没有被扔进垃圾堆——工程师仍在用它造桥、发射卫星——但它被相对论吸收，降格为低速弱引力下的近似。"地球绕着太阳转"这个命题曾锋利到能把人送上火刑架，如今平庸到没有任何张力，沉淀为人人默认的背景：它从一个能撬动世界的判断，变成了一句无人再会为之争论的话。

这三种死法对本文的主题有直接后果。如果知识会死，那么**任何"训练一次即拥有世界"的设想都在时间维度上站不住**：一个被一次性固化的模型，握住的是某一时刻的结构快照，而结构本身正在三个维度上持续地老化、被吸收、被耗尽。世界模型必须是活的——能更新、能反馈、能保存历史、能比较版本。这也再次说明，为什么"画像"这个形态比"统一巨构"更贴近世界模型的真实存在方式：画像天然是可重画的，巨构不是。

还要补一句方法论上的话：一个会承认自己终将死亡的知识观，比一个假装真理永恒的知识观，离真实更近。本文提出的判断，同样在这条规则之下——它也会老，也可能被吸收或被耗尽。这不是谦辞，而是这套方法对自身的应用。

九 · 结论：统一的不是模型，而是生成模型的方法

把两步论证收成一句话。

否定的一步：**不存在统一世界模型**。大模型拥有的是文明发生之后沉积下来的语言材料场，而非世界本身；语言封存世界却不组织世界；世界模型是主体与对象共同发生的画像，因而必然多元；科学等公共结构是被主体间稳定出来的公共层，不是统一层；幻觉的根源不在规模而在无主体，因而不随规模消失。

肯定的一步：**AGI 的标尺是知识创造，而知识创造只能发生在复合体中**。智能的高阶形态不是在已发生的结构里高效操作，而是在裂缝处让以前不存在的结构第一次发生并被稳定；这一能力离不开提供目标、价值与责任的主体，因而纯粹无主体的 AGI 在结构上不可能。

两步合起来，给出一个可操作的方向判断：AI 的第二次突破，不是继续扩大那个材料场，而是在一套本体论的指挥下，把材料场组织成能够为具体对象持续建造、维护、更新世界画像的系统。处理一个企业问题，先建企业的画像；处理一个学生问题，先建学习的画像；处理一个患者问题，先建疾病的画像——永远是先为当前的具体主体与对象建起一个具体的世界模型，再在其中行动。

> 所以这类系统的根本特征，不是它拥有一个统一世界模型，而是它拥有**一套统一的方法，用以生成无数不同的世界模型**。统一世界模型路线追求一个唯一的模型去代表所有主体——这在本体论上不可能；发生学路线追求一套统一的发生方法，去为每一个具体主体、对象、任务、阶段生成它各自需要的那个模型——这在本体论上才成立。统一的不是模型，而是生成模型的方法。

最后是本文的边界，必须自己标出来，而不是等人来挑。其一，"发生恰有结构、差异、纠缠三个不可省的维度"是本文所依赖的本体论承诺，它并非不可挑战——是否存在第四维，是一个敞开的问题。其二，"任何可能的架构都长不出自我世界"这个更强的命题，本文不能严格证明，也不假装已经证否；本文严格主张的只是那个较弱的版本（第 7.3 节）。其三，本文提出的新标尺——在裂缝处发生新结构并被共同体稳定——虽可拆成工序，但每道工序的判据仍有待更严格的操作化，尤其是"什么算一次真正的发生"与"什么只是包装过的旧结构"之间的边界，在实践中并非总能当场判定。

把这三条边界写出来，不是为了给论证留退路，而是因为按本文自己的立场，一个结构的分量应由它能否经受反驳来决定，而不是由它显得多完备来宣布。一篇论证世界是被发生的文章，自己也只能是一次发生——它等着被使用、被检验、被在它该破的地方破掉。

本文据《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》（王德生 著，德麦国际出版，2026，ISBN 978-1-970820-62-1）第六章《没有统一世界模型：世界模型的主体性与第二次 AI 革命》与第二十六章《从发现世界到发生世界：六大哲学转变与 AGI 的重新定义》改写成文，并补入文献定位、论证编排与边界声明。原书导读与全文阅读见 专著页。

参考文献 · References

下列文献为文中所涉思想的原典与代表性研究，供读者对读与查证；本文观点不代表所列作者立场。

1. Yann LeCun, "A Path Towards Autonomous Machine Intelligence", Version 0.9.2, OpenReview, 2022（世界模型与 JEPA 架构路线）。
2. David Ha & Jürgen Schmidhuber, "World Models", NeurIPS 2018, arXiv:1803.10122.

3. Danijar Hafner et al., "Mastering Diverse Domains through World Models"（DreamerV3）, arXiv: 2301.04104, 2023.

4. Jake Bruce et al., "Genie: Generative Interactive Environments", ICML 2024, arXiv:2402.15391.

5. Niket Agarwal et al., "Cosmos World Foundation Model Platform for Physical AI", NVIDIA, arXiv: 2501.03575, 2025.

6. Richard S. Sutton, "The Bitter Lesson", incompleteideas.net, 2019.

7. François Chollet, "On the Measure of Intelligence", arXiv:1911.01547, 2019.

8. Shane Legg & Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence", Minds and Machines, Vol. 17, No. 4 (2007).

9. Emily M. Bender & Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", ACL 2020.

10. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", FAccT '21, 2021.

11. Ziwei Ji et al., "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys, Vol. 55, No. 12 (2023).

12. Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions", arXiv:2311.05232, 2023.

13. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models", CRFM, Stanford University, arXiv:2108.07258, 2021.

14. Rylan Schaeffer, Brando Miranda & Sanmi Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?", NeurIPS 2023, arXiv:2304.15004.

15. Ashish Vaswani et al., "Attention Is All You Need", NeurIPS 2017, arXiv:1706.03762.

16. Jakob von Uexküll, Streifzüge durch die Umwelten von Tieren und Menschen (1934); English ed. A Foray into the Worlds of Animals and Humans, University of Minnesota Press, 2010（Umwelt 概念）。

17. James J. Gibson, The Ecological Approach to Visual Perception, Boston: Houghton Mifflin, 1979（可供性理论）。

18. Andy Clark & David J. Chalmers, "The Extended Mind", Analysis, Vol. 58, No. 1 (1998).

19. Edwin Hutchins, Cognition in the Wild, Cambridge, MA: MIT Press, 1995（分布式认知与航海驾驶室研究）。

20. Thomas S. Kuhn, The Structure of Scientific Revolutions, Chicago: University of Chicago Press, 1962；中译《科学革命的结构》，金吾伦、胡新和译，北京大学出版社，2003。

21. Michael Polanyi, The Tacit Dimension, London: Routledge & Kegan Paul, 1966；中译《默会的维度》，许泽民译，上海人民出版社，2000。

22. John R. Searle, "Minds, Brains, and Programs", Behavioral and Brain Sciences, Vol. 3, No. 3 (1980)（中文屋论证）。

23. Thomas Nagel, "What Is It Like to Be a Bat?", The Philosophical Review, Vol. 83, No. 4 (1974).

24. Alfred North Whitehead, Process and Reality, New York: Macmillan, 1929；中译《过程与实在》，杨富斌译，中国城市出版社，2003。

25. 王德生：《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》，德麦国际出版，2026，ISBN 978-1-970820-62-1（本文第六章与第二十六章之改写底本）。

26. 王德生：《SDE本体论入门：世界是如何发生的》，德麦国际出版，2026（三大方程、123 原理与三界结构的完整展开）。