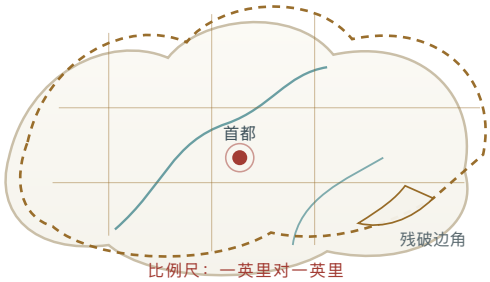
模块一·知识与推理的根基

模型、地图与理想化

一张完美的地图就是一张无用的地图。那么，一张好地图究竟好在哪里？



博尔赫斯想象的帝国造出一张与疆域同样大小的地图；后人让它在荒野里腐烂。

一九四六年，豪尔赫·路易斯·博尔赫斯在一篇短文里描写了一个帝国：那里的制图师技艺精湛，且对精确痴迷到了极点，于是绘制了一幅比例为一比一的帝国地图。它点对点地覆盖了全部领土，像第二层皮肤一样铺展在国土之上，成为有史以来最精确的地图。然而它也是废物。下一代人抛弃了它，旅人日后只会在沙漠里发现「褴褛的残片」。

这个玩笑大有深意。地图之所以有用，恰恰因为它不是领土：它缩小、压平、筛选，把几乎一切都扔掉了。保留下来的，才是你真正需要的东西：可用的结构。每一个科学模型都在玩同一套把戏。一旦看清这一点，一个更硬的问题就浮现出来：当那些最出色的理论谈论我们永远无法直接看见的东西时，它们是在描述实在，还是在绘制一幅异常出色的地图？

- 核心模型
- 模拟前沿活跃

我们身在何处

昨天（**第 9 日**）我们画了回路、浴缸和稳定山谷。那些图每一张都是模型：我们之所以信任它，是因为它比世界更简单。今天，我们把镜头对准工具本身。我们会兑现 **第 7 日**关于压缩的承诺，复用 **第 8 日**关于介于有序与随机之间的描述，并再次遇见 **第 1 日**——只是换了新的伪装：一个因为错误理由而预

测正确的模型。

模型

地图不是领土

这句话出自阿尔弗雷德·科尔兹比（Alfred Korzybski）。1933 年，他试图矫正一种糟糕的人类习惯：把自己的词语和符号错当成实在本身。完整的句子才是关键：

> 「地图不是它所代表的领土，但如果正确，它与领土具有相似的结构，这正是其用处所在。」阿尔弗雷德·科尔兹比，《科学与理智》（Science and Sanity），1933 年。

「相似结构」这个子句，把整个建模理论浓缩在了一句话里。地图如果复制一切，就如博尔赫斯所示，毫无用处；如果什么都不复制，也毫无用处。它的价值在于中间地带：为特定目的保留关键关系，同时丢弃其余。地铁图是扭曲的杰作——距离是错的，角度是错的，河流被画成卡通——但它保留了站点顺序与换乘关系，因为那才是乘客需要的。

《波伊廷格地图》不是按比例复制地理空间。它牺牲形状与距离，保留旅人需要的东西：道路、站点和先后关系。

所以，模型不是世界的小型真实副本，而是一种有用的扭曲。迈克尔·韦斯伯格（Michael Weisberg）在《模拟与相似》（Simulation and Similarity）中给出了精确的哲学版本：模型不必与目标完全相同，甚至不必与之完全同构；它只需在相关方面、达到任务所需的相似程度即可。

所有模型都是错的

乔治·E·P·博克斯（George E. P. Box）给了统计学一句最便携的话：「所有模型都是错的，但有些有用。」引证轨迹本身就是一堂小型的防炒作滤镜课。他在 1976 年的论文《科学与统计学》中埋下了种子：因为所有模型都是错的，科学家无法通过无限细化得到一个正确的模型。精炼的格言出现在 1979 年会议论文的章节

标题里，而那个我们熟悉的「本质上是」措辞，则落在他 1987 年与诺曼·德雷珀（Norman Draper）合著的教科书里。

三个日期，一个用十年打磨的念头。这句口号是真的，但人们随口附带的引证常常不是。核查收据——我们会在 第 **149** 日大规模地做这件事。

比博尔赫斯更老的玩笑

博尔赫斯并非第一个想到一比一地图的人。刘易斯·卡罗尔（Lewis Carroll）在 1893 年出版的《西尔薇与布鲁诺完结篇》（Sylvie and Bruno Concluded）里，有个角色夸耀一幅比例「一英里对一英里」的国家地图。为什么它从未铺开？农民们反对，因为那会遮住阳光、杀死庄稼。于是他们干脆用国家本身来当自己的地图。领土永远是最精确的模型，也是最无用的模型。

技艺

球形牛，以及科学家为何故意说谎

有一则老笑话。一位奶农雇物理学家提高产奶量。数月后他们带回了方案，但只适用于真空中的球形牛。

笑话之所以好笑，是因为它说的是真的。故意为假的假设不是科学的 bug，而是它最锋利的工具之一。伽利略的自由落体定律只在真空中、无空气阻力、无自转、理想重力场下才干净。没有一条完全成立。厄南·麦克马林（Ernan McMullin）把这叫作伽利略式理想化[1]：先扭曲问题以让机制显形，再在需要时把细节加回来。

有时，简化甚至更深。无摩擦平面不只是让算术变容易，它通过移除掩盖惯性的东西，把惯性揭示出来。这种极简理想化[2]与其说是对世界的近似，不如说是一种关于「什么才重要」的论证。球形牛、理想气体、质点、遗传学里的无限群体、无摩擦滑轮，都在说：忽略这些噪声，看剩下的结构。

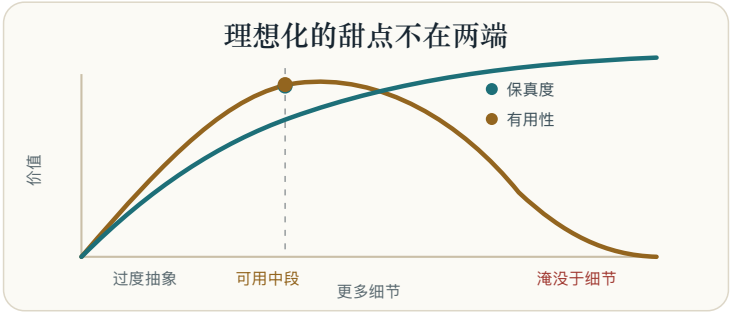
因此，「所有模型都是错的」并不是绝望的劝告，而是专注的劝告。一个试图在所有方面都对的世界模型，就是博尔赫斯的那幅地图：全面、忠实、死寂。艺术在于选择把哪些错误坐实。

静态替代·理想化刻度盘

这个可交互刻度盘从高度抽象滑向一比一复刻。保真度随细节增加而上升，但有用性在中段达到峰值，等到模型和目标一样复杂时便崩溃。

[1] 伽利略式理想化故意扭曲问题，通常通过移除复杂因素，使核心关系变得数学上可处理。

[2] 极简理想化只包含被认为具有本质性的因果因素，用故意为假的设定来隔离机制。



一个好模型保留任务需要的结构，而不是追求最大细节。保真度可以持续上升；有用性通常不会。

区域	例子	教训
细节太少	把奶牛简化为质点	简化把你要用的关键属性也删掉了。
有用的中段	地铁图或理想气体	模型为任务保留了所需的结构。
细节太多	博尔赫斯的一比一地图	最大保真度会摧毁压缩、清晰度和可用性。

辩论

我们的地图是在描述真实世界，还是仅仅好用？

在这里，实践问题变成了哲学上最深的一场争吵。物理学谈论电子、夸克、场、时空曲率：这些实体与结构没人能直接看见。理论却好得惊人。这种成功是否意味着那些看不见的家具确实在那里，大致如其所述？还是说，电子只是强大簿记系统里的条目，对预测观察有用，却不告诉我们底层实在长什么样？

科学实在论[3]者会说，如果我们的最佳理论没有抓住实在，成功就近乎奇迹。这就是与希拉里·普特南（Hilary Putnam）和 J·J·C·斯马特（J. J. C. Smart）关联的无奇迹论证：对持续的预测成功来说，最佳解释就是：理论确实为真。

反实在论者的回击来自历史。拉里·劳丹（Larry Laudan）1981 年的论文《对收敛实在论的驳斥》（“A Confutation of Convergent Realism”）列举了一个又一个昔日成功、如今却被视为错误的理论：

- **燃素**：被认为存在于可燃物中的火之物质。
- **热质**：把热当作不可见流体来处理。

[3]科学实在论认为，成功的成熟科学理论至少在近似意义上是真的，包括许多关于不可观察实体的断言。

- **光以太**：据当时假设，光在其中传播的介质。
- **水晶球壳**：承载行星的壳层。
- **生命力**：生命物质的特殊火花。

这一拳很狠。过去的成功并不能保证指向真实实体。凭什么认为我们今天的成功终于做到了？这就是悲观元归纳[4]。再加上欠决定[5]，实在论者就面临一个严重问题：同一片地面可以塞进多张地图。

四种与张力共存的方式

立场	可观察预测	不可见实体	数学结构
科学实在论	足够真实	真实，大致如其所述	追踪真实关系
结构实在论	足够真实	对实体保持松动	结构才是存活下来的东西
实体实在论	足够真实	能被实验操纵时即为真实	宏大定律可能只在模型中成立
建构经验论	经验上充分	保持不可知	接受地图，而不承诺其为真

结构实在论，与约翰·沃拉尔（John Worrall）关联，是最具地图味儿的折中。看看以太。菲涅耳的以太消失了；麦克斯韦的电磁场取而代之。但数学结构的某些部分在革命中存活了下来。也许科学积累结构比积累「家具」更可靠。保留方程；对实体轻轻拿放。

实体实在论，与伊恩·哈金（Ian Hacking）和南希·卡特赖特（Nancy Cartwright）关联，从预测走向干预。哈金的口号是：如果你能喷电子，电子就是真的。相信你能操控来干活的实体；对包裹它们的宏大理论更存疑。卡特赖特则从另一侧下刀，在《物理定律是如何说谎的》（How the Laws of Physics Lie）中指出：基本定律支配的是模型里的理想化对象，而不是世界里乱糟糟的对象。

把这张表攥紧。计算机即将把它的每一行都逼到阳光下。

[4] 悲观元归纳认为，由于许多过去成功的理论后来都被证明是错的，今天成功的理论也可能最终是错的。

[5] 欠决定指多个理论可以拟合同一份证据，因此仅凭证据可能无法选出唯一为真的理论。

前沿·2026

旧争论在硅中重生

历史上，科学只有两条学习路径：思考世界，和戳弄世界。近几十年，第三条路径强势挤了进来：模拟。气候模型在代码里跑过一个世纪。喷气发动机的数字孪生在裂纹形成前就预判它。训练出来的天气预报模型胜过了物理求解器。这算哪一类知识？

埃里克·温斯伯格（Eric Winsberg）主张，模拟有自身的认识论：它们能产生真正的知识，其辩护方式无法还原为纸笔理论或实验室实验。保罗·汉弗莱斯（Paul Humphreys）把这种不安命名为认识论不透明性[6]。一个过程可以形式化地规定，却大到没有任何人能逐行检查。

边缘 01 ● 孪生框架 ● 炒作有争议

数字孪生：模型何时配得上这个名字？

「数字孪生」被过度营销，因此要使用严格来源：2024 年美国国家科学院报告《数字孪生的基础研究缺口与未来方向》（Foundational Research Gaps and Future Directions for Digital Twins）。真正的数字孪生不只是模拟。它要有虚拟表示、来自物理对应物的动态数据更新、预测能力、决策价值，以及一条虚拟–物理双向回路。那条反馈回路才是门槛。

难题在于信任。报告强调 *VVUQ*[7]是一个持续过程，不是一次性盖章。物理系统会变，数据会变，决策会变，孪生在使用过程中必须被重新验证。

边缘 02 ● DESTINE 已启动 ● 技巧在发展 ● 2030 路线图

一整颗行星的数字孪生

最宏大的版本是欧洲的 **Destination Earth** 计划，由 ECMWF、ESA 和 EUMETSAT 共建。2024 年 6 月，其首次系统发布上线了两个孪生：「天气引发极端事件」与「气候变化适应」。极端事件孪生包含一个全球分辨率约 4.4 公里的组件，并针对选定事件提供区域放大。

这是一个已确立的启动里程碑。但它还不是「地球数字孪生」这个短语所暗示的科幻对象。更高分辨率并不自动意味着对每项任务都有更好的预报，而欧洲 2030 年建成完整地球系统孪生的目标仍停留在路线图上。

[6]认识论不透明性指没有人能逐步审视产生某个计算结果的全部相关步骤的局面。
[7]VVUQ：验证代码、确认模型-世界拟合、量化不确定性。

边缘 03 ● FFR-CT 已用 ● 全患者仍是愿景

临床数字孪生：一个真正例外

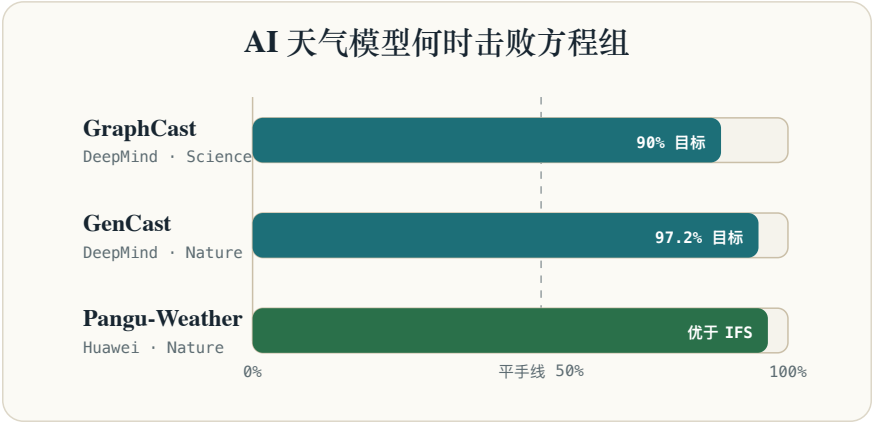
医学热爱「你的数字孪生」这个想法：一个正在运行的身体模型，医生可以在上面测试干预。冷静的文献却稀薄得多。2025 年 npj Digital Medicine 上的一项范围综述发现，149 篇使用这个标签的医疗研究中，只有 18 篇完全符合美国国家科学院的定义，只有两篇提到 VVUQ。

例外恰恰证明了规则。HeartFlow 的 FFR-CT 接收冠状动脉 CT 扫描，为患者建立血流计算流体动力学模型，以判断某处阻塞是否正在让心脏挨饿。它在 2014 年获得 FDA De Novo 许可。范围狭窄、物理性强、决策明确——这就是它有效的原因。

边缘 04 ● 预报技巧 ● 理解物理机制有争议

不写方程也能预报天气

几十年来，天气预报意味着在巨型计算机上求解大气物理。随后，从过去天气中训练出来的机器学习模型追上了、并在多项基准测试中超越了领先的物理系统。GraphCast、盘古天气（Pangu-Weather）、GenCast 不是新闻稿里博眼球的噱头；它们的头版结果发表在 Science 和 Nature 上。



这些数字来自同行评议论文的回测目标；它们证明预测技能，不证明模型内部已经理解物理。

哲学刺痛来得直接。按预报用户唯一关心的标准，这些训练出来的系统可以比那些由显式物理方程构建的系统经验上更充分。可它们的内部是不透明的权重，而

非具名的机制。建构经验论者可以耸肩：预测本来就是目标。实在论者却该担心：如果预测可以在没有可解释表征的情况下到来，无奇迹论证就会失去部分力量。

结构实在论者指向了混合回应。谷歌的 NeuralGCM 保留了一个可微分的大气动力学核心，并用机器学习去补物理处理不好的部分。保留可信的结构；学习那些闭合项。这就是工程实践中的结构实在论。但提醒依然重要：与其他训练出来的系统一样，NeuralGCM 预计难以大幅超出训练分布覆盖的范围进行外推。机制可能正是你在地图边缘最需要的。

未决问题

真正未决的问题

- **模拟是科学的第三支柱，还是只是装备了更好机器的应用数学？**温斯伯格与玛格丽特·莫里森（Margaret Morrison）主张一种独特的认识论；批评者仍将其视为扩展了的建模。
- **预测而不表征，会削弱科学实在论吗？**如果一个无机制模型能比物理模型更好地预报，成功本身看起来就不太像真理的证据。
- **训练出来的模型能外推吗？**天气预报结果在训练数据覆盖的区间内最强。新气候与罕见极端事件才是经验试金石。
- **模型何时配得上数字孪生之名？**美国国家科学院划了一条严格的线。许多打着这个旗号的产品并没有跨过去。
- **当模型本身是一个 AI 系统时会发生什么？**如果一个语言模型产生了一个真实且证据充分的主张，它是在描述实在，还是只充当了一张经验上充分的地图？我们会在 AI 模块回到这个问题。

三句话总结今日

大观念

每一个模型都是一次故意而有用的扭曲：它为特定目的保留结构，同时舍弃其余几乎一切。最困难的问题是，我们最成功的地图究竟在描述真实世界，还是仅仅「拯救现象」。

最佳类比

博尔赫斯的一比一地图与真空中的球形牛：完美的保真度可能毫无用处，而一个故意为假的简化却能恰好揭示关键所在。

当下争议

数字孪生与 AI 天气模型在硅中重新打开了科学实在论问题，尤其是在预测成功却没有人类可读机制的时候。

计算（模拟作为科学的第三种模式）·信息（模型是有损压缩）·涌现（多尺度模型与数字孪生）·理想化中的能量与演化。

明日 → 第 11 日

启发法、偏见与理性

今天我们看着科学故意简化世界。明天镜头转向内部：心智本身也运行在捷径之上。你会遇见银行职员琳达、系统 1 与系统 2，以及关于启发法究竟是该纠正的缺陷，还是稀缺条件下的适应性赌注的理性之争。

来源与延伸阅读

1. Korzybski, A. (1933). Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics. 地图–领土关系，第 58 页。wikipedia.org/wiki/Map-territory_relation
2. Borges, J. L. (1946). “Del rigor en la ciencia”/ “On Exactitude in Science.” wikipedia.org/wiki/On_Exactitude_in_Science
3. Carroll, L. (1893). Sylvie and Bruno Concluded, ch. XI. wikisource.org
4. Box, G. E. P. (1976). “Science and Statistics.” Journal of the American Statistical Association 71(356): 791-799. doi:10.1080/01621459.1976.10480949
5. Box, G. E. P. (1979). “Robustness in the Strategy of Scientific Model Building,” in Launer & Wilkinson (eds.), Robustness in Statistics, 201-236.
6. Box, G. E. P. & Draper, N. R. (1987). Empirical Model-Building and Response Surfaces. Wiley.
7. Weisberg, M. (2013). Simulation and Similarity: Using Models to Understand the World. Oxford University Press.
8. McMullin, E. (1985). “Galilean Idealization.” Studies in History and Philosophy of Science 16(3): 247-273.
9. Putnam, H. (1975); Smart, J. J. C. (1963). 参见斯坦福哲学百科全书「科学实在论」条目。plato.stanford.edu/entries/scientific-realism
10. Laudan, L. (1981). “A Confutation of Convergent Realism.” Philosophy of Science 48(1): 19-49. jstor.org/stable/187066
11. van Fraassen, B. C. (1980). The Scientific Image. Oxford University Press.
12. Worrall, J. (1989). “Structural Realism: The Best of Both Worlds?” Dialectica 43(1-2): 99-124.
13. Hacking, I. (1983). Representing and Intervening. Cambridge University Press.
14. Cartwright, N. (1983). How the Laws of Physics Lie. Oxford University Press.
15. Winsberg, E. (2010). Science in the Age of Computer Simulation. University of Chicago Press. Humphreys, P. (2009)，论认识论不透明性。
16. National Academies of Sciences, Engineering, and Medicine. (2024). Foundational Research Gaps and Future Directions for Digital Twins. doi:10.17226/26894. nationalacademies.org/read/26894
17. European Commission / ECMWF. Destination Earth 系统发布，2024 年 6 月 10 日。destination-earth.eu •destine.ecmwf.int
18. Tudor, Burton, et al. (2025). “A scoping review of human digital twins in healthcare applications and usage patterns.” npj Digital Medicine 8: 587. doi:10.1038/s41746-025-01910-w
19. FDA. HeartFlow FFR-CT De Novo DEN130045; HeartFlow FFR-CT 许可材料，2014 年。
20. Lam, R. et al. (2023). “Learning skillful medium-range global weather forecasting.” Science 382(6677): 1416-1421. doi:10.1126/science.adi2336

21. Bi, K. et al. (2023). “Accurate medium-range global weather forecasting with 3D neural networks.” Nature 619: 533-538. doi:10.1038/s41586-023-06185-3
22. Price, I. et al. (2024). “Probabilistic weather forecasting with machine learning.” Nature. doi:10.1038/s41586-024-08252-9
23. Kochkov, D. et al. (2024). “Neural general circulation models for weather and climate.” Nature 632: 1060-1066. doi:10.1038/s41586-024-07744-y

来源卫生：研究过程中出现的未来日期 arXiv 标识符已被丢弃，视为不可靠。上述前沿主张依赖同行评议论文、官方项目页面或机构报告。

可选附录：剪余片段

正文走了一条笔直的下降线：从博尔赫斯那幅帝国尺寸的地图，穿过球形牛，一头扎进实在论之争，最后落在一个对物理定律一无所知、却能预报天气的神经网络。弧线干净利落，却也把大量素材留在了剪辑室的地板上。实在论与工具主义之争比范·弗拉森老了四个世纪；「有用地说谎」这门手艺有一套精确的语法，正文里只字未提；而最奇怪的那个问题——如何信任一个内部无人能一览无余的模型？——一路绕回第 1 日。

下面有什么

正文之旅跳过六个房间：*(1)* 实在论与工具主义的古老战争；*(2)* 以太唯一一次辉煌的预言，以及实在论如何以「分而治之」反击；*(3)* 整个实在论之争为何可能是一场基础率谬误，直接回调第 4 日;*(4)* 理想化的真实语法，以及列文斯的「三选二」三角；*(5)* 模型到底是什么：中介、虚构，以及麦克斯韦想象的齿轮；*(6)* 信任问题，从 1994 年那颗关于验证的炸弹，到「计算可靠主义」——不过是第 1 日的可靠主义穿上了实验服。

房间 I 漫长的前史

拯救现象：一场四个世纪的争论

正文把实在论与工具主义之争框成了二十世纪的事：普特南、劳丹、范·弗拉森。但这场战争古已有之，而且大部分时间都是工具主义者占上风。这个传统甚至有一句从希腊天文学传下来的战号：拯救现象 (sozein ta phainomena)。按这种古老观点，模型的任务从来不是说出天体是什么，只是复现它们看起来做什么，让每晚的圆点落在正确的位置。

托勒密的本轮，一圈套一圈地解释行星的逆行回环，是第一种有用的虚构。他的许多读者，以及中世纪的大多数天文学家，都坦率地把本轮当作计算装置。没有人需要相信天上挂满了实体齿轮，齿轮只要好用就行。这就是工具主义的摇篮，比这个词出现早了一千年。

戏剧性在哥白尼时代加剧。1543 年《天体运行论》付印时，路德宗神学家安德烈亚斯·奥西安德尔未经弥留之际的哥白尼同意，塞了一篇匿名序言，向读者保证这套狂野的日心说不必为真，只是个便于计算位置的方便假设。这是与当权者签订的和平条约，用纯工具主义的墨水写成：放心，它只是一个更好的计算器。

红衣主教是个工具主义者

最清晰的立场划分来自伽利略的审判者。1615 年，红衣主教罗伯特·贝拉尔米内对哥白尼派学者保罗·福斯卡里尼说，你大可以 ex suppositione 地持有日心说——也就是假设性地、作为拯救现象的装置——但不能把它当作关于世

> 界物理实在的断言。把它说成模型，没问题；说成真理，就有问题。教会的立场是教科书式的工具主义。伽利略的异端在于实在论：他坚持最好的模型正在告诉你那里到底有什么。这条断层线一直延伸到正文主页上的神经天气模型。

到了二十世纪初，物理学家兼哲学家皮埃尔·迪昂——我们在第 2 日因「你总可以怪罪辅助假设」而结识他——把这套想法铸成了一门完整学说。迪昂在 1906 年的《物理学理论的目的与结构》中主张，物理理论不是对隐藏实在的说明，而是对实验定律的经济分类：一种为现象极度压缩的档案系统。关于不可见原因的真理不是目标，整洁而可预测的簿记才是。三百年的「拯救现象」，至此蒸馏为一杯烈酒。现代工具主义者欠迪昂这份蓝图。

房间 II 实在论反击

以太唯一完美的预言

正文把发光以太当作劳丹墓园里的头号展品：一个极其成功的理论，其核心对象却被证明不存在。但以太的案子远比「它曾成功，然后死去」更扰人，原因是一个美丽到让实在论者念念不忘的片段。

约 1818 年，法国科学院举办了一场关于光本质的悬赏竞赛。奥古斯丁·菲涅耳提交了波动说：光是以太中的振动。评审席上坐着西梅翁·德尼·泊松，坚定的微粒说信徒。他用菲涅耳的方程做了些计算，得意洋洋地推出一个明显荒谬的结论。如果菲涅耳是对的，小圆盘阴影的中心应该有一个亮斑——来自四面八方绕过边缘的波在那里同相抵达。泊松说，荒唐，这是推翻整个波动说的清晰 reductio。另一位评委弗朗索瓦·阿拉戈走到实验台前真的做了实验。亮斑就在那里。

这是实在论者的重拳，它有个名字：新预测[8]。以太理论不只是容纳已知事实——谁都会曲线拟合过去。它预言了没人见过、没人料到的东西，一个如此具体而怪异、其成真看起来绝不像是巧合的细节。这就是正文主页上无奇迹论证的满血形态：显然，一个理论不可能凭空从帽子里变出兔子，除非它抓住了某种实在。然而，以太并不存在。实在论能拿出的最佳证据，竟来自一个我们今天称之为错误的理论。原来，这座墓园里埋着一位天才。

分而治之

那么实在论者如何在自己最好的例子下幸存？靠的是一招有时被称为选择性实在论或分而治之的实在论，由菲利普·基切尔与斯塔西斯·普西洛斯在 1990 年代打磨锋利。诀窍是：别再对整个理论做实在论者，只对其中的一部分做。任何成功理论里，只有一些设定真正承担预测重任，这叫有效设定；另一些是闲置设定，只是搭便车。只对有效部分做实在论者。

用以太过一遍这个筛子，你会发现它自动分类。推导出泊松亮斑的，究竟是什么？不是「以太」这种物质，不是它的密度、弹性，也不是它由什么构成。真正起作用

[8] 新预测指模型在未用其构建或调参的情况下，成功预测了某个结果，尤其是当该结果令人惊讶时。

的是波动方程的数学形式：关系、结构。而这些结构被保留下来，进入了麦克斯韦的场论以及更后来的理论。闲置的形而上学填充物，以太的「质料」，正是被埋葬的部分。于是悲观元归纳少了些牙齿：科学并非先保留家具再扔掉，而是保留了在革命中幸存的结构，悄悄扔掉了装饰。这正是正文主页最后落脚的结构实在论，如今它正在最棘手的案子上做实打实的工作。

房间 III 回调第 4 日

整个争论是一场基础率谬误吗？

这个房间会让第 4 日的毕业生坐直。回忆一下基础率谬误：一种疾病发病率万分之一，检测准确率 99%，阳性结果仍会标记出远多于患者的健康人，因为你忘了问疾病原本有多常见。混淆 P(阳性 | 患病) 与 P(患病 | 阳性) 是经典陷阱。

2004 年，哲学家 P. D. 马格努斯与克雷格·卡伦德指出，双方都可能一头撞进这个陷阱。再看无奇迹论证：它想得出成功理论可能为真，也就是 P(理论为真 | 理论成功) 很高。但根据贝叶斯定理，你不能只从 P(成功 | 理论为真) 推出这一点，还需要基础率：科学从中抽取候选理论的那个池子里，真理论的事先比例。而这个数字不仅未知，马格努斯和卡伦德认为它根本不可知，因为不存在公平的方法去计数或抽样成熟科学所有可能存在的候选理论。没有基础率，就没有推断。

同一把刀也割向悲观元归纳：它试图用历史上的失败记录把 P(理论为真 | 成功) 压低，却绊倒在同一个缺失的分母上。马格努斯与卡伦德的诊断近乎临床：一旦用概率框架表述，这些宏大的「批发式」论证不仅悬而未决，而且可能是畸形的——每一件都是燕尾服里的基础率谬误。科林·豪森在 2000 年对无奇迹论证也提出了密切相关指控。第 4 日的机械装置不只是描述这场争端，它可能悄悄地把争端消解掉。 ● 哲学上仍未决的争议

房间 IV 有用谎言的语法

理想化有规则，还有一个著名的取舍

正文把「故意为假的假设」当作一回事。文献却更挑剔，区分很重要。首先，把两个常被混为一谈的动作分开。抽象是省略：模拟行星轨道时不提它的颜色。该说的没说，但没说错。理想化则是主动歪曲：明知平面有摩擦，却宣称它无摩擦。无摩擦平面不是描述中的缺口，而是一个善意的谎言。两者感觉相似，行为却很不一样。

理想化还有种类，而正文跳过的是第三种。有伽利略式理想化：为了可处理性而歪曲，并承诺日后补回细节；有极简理想化，也就是球形牛策略，只保留被认为关键的因果因素。但还有多模型理想化[9]：为同一件事造好几个相互矛盾的模型，每个朝不同方向犯错，然后在它们之间三角定位。没有单张地图被信任；被信任的是不同错误地图之间的重合。

[9] 多模型理想化指故意为同一目标构建多个互不相容的模型，每个模型以不同方式简化，以观察哪些结果能在它们之间存活。

> 「我们的真理，是彼此独立的谎言的交集。」 理查德·列文斯，《种群生物学中的模型构建策略》，1966 年。

列文斯的这个观点，现在被称为稳健性分析[10]：如果你用几个建立在不同简化谎言之上的模型去攻击同一个问题，而它们都吐出同样的定性结果，那么这个结果很可能依赖问题真实的共享结构，而不是某个模型特有的谎言。谎言彼此独立，它们的交集正是真理最可能藏身的地方。

列文斯还提出了建模史上最著名的不可能论断：你无法同时最大化一般性、逼真性和精确性。把任意两个推到极限，第三个就必须让步。正文主页上的旋钮测量逼真度与有用性，列文斯的三角则是同一个智慧的三维版：选择你的牺牲。

建模取舍·列文斯三角

没有任何模型能同时拉满三项。原始版本是可点击的三角；静态附录把三种状态都展开，方便网页、EPUB 和 PDF 阅读。

牺牲	保留	策略	例子
精确性	一般性 + 逼真性	稳健的定性模型，把握变化方向，不做小数点后断言。	理论生态学，以及隔离某种机制的玩具经济或物理模型。
逼真性	一般性 + 精确性	简洁、可精确求解的模型，建立在明知为假的假设上。	无摩擦平面、完全理性主体、无限种群。
一般性	逼真性 + 精确性	某一具体目标的高分辨率模型，准确但狭窄。	这架飞机、这座湖、这位患者的数字孪生。

Orzack 与 Sober（1993）认为，三向取舍并非严格定理。有时某一项目可以提升而不立刻牺牲另一项。因此，有纪律的版本不是「永远只能二选一」，而是「清楚你当前这一步花的是哪个维度」。

房间 V 模型到底是什么？

中介、虚构，以及麦克斯韦想象的齿轮

退后一步，问正文绕着走的问题：当科学家「有一个模型」时，那是什么东西？三种回答照亮不同角落。

理论是一族模型。在较早的句法图景中，科学理论是一大堆句子：公理及其逻辑后承。现代的语义观说不。理论更应被理解为一组抽象结构，加上一个断言：某个实在系统在特定方面与其中之一足够相似。牛顿力学从根本上说不是一堵方程墙，而是一套理想化系统的工具箱——单摆、二体轨道、谐振子——外加一个赌局：世界的一些碎片与它们足够相像。罗纳德·吉尔把这一立场称为视角实在论[11]：

[10] 稳健性分析寻找在不同简化假设的模型之间仍然保持的结果。

[11] 视角实在论认为，科学表征可以真实地指向世界，但始终是局部的、目的绑定的视角，而非镜像复制品。

我们的模型像仪器，每一种都给出局部、视角绑定的图景；它像地图那样为真，绝不会像镜子那样为真。

模型是中介。在 1999 年那本影响深远的文集《模型作为中介》中，玛丽·摩根与玛格丽特·莫里森主张，模型既不是从理论中直接读出，也不是数据的简单总结。它们坐落在中间，在一定程度上独立于两者。你建造模型就像建造仪器，做出理论不规定的实用选择；然后，关键是，你通过操纵它来学东西，就像转动仪器旋钮来学东西。模型是你用来思考的工具，不只是你观看的图画。这也是为什么下一间房里模拟会如此像实验。

模型是虚构。最挑衅的观点把无摩擦平面当真：它是一种虚构，一个并不存在的对象，却被描述得仿佛存在。这个想法并不新。汉斯·费英格 1911 年的《仿佛哲学》主张，科学与数学的很大一部分运行在明知为假却故意当作真的有用谎言之上的。回想第 3 日的惊讶：夏洛克·福尔摩斯的「演绎」其实是溯因；这里遇到的是它的表亲。科学模型也许更接近小说人物，而不是照片。理想气体有点像福尔摩斯：不是真的，从来都不是，但推理它会做什么，却能带来真正的理解。

他造了一台明知是假的机器

典型例子就坐在正文菲涅耳到麦克斯韦故事的旁边。为了导出他那不朽的电磁学方程，詹姆斯·克拉克·麦克斯韦先搭建了一台荒诞的机械装置：空间里塞满旋转的「分子涡旋」，由小小的「惰轮」分隔。他并不相信空间里满是齿轮，他称之为示意图，一种用来思考的虚构。从另一端出来的，是麦克斯韦方程组：为真的结构，由一个为假的模型接生。他的导师开尔文勋爵把这一信条推到极端：「我从不满意，直到我能为一件东西做出机械模型。」虚构是脚手架，结构才是被保留下来的。整个第 10 日，就藏在这个维多利亚时代的小故事里。

卡特赖特的斑驳世界

再拧一圈螺丝，正文上的实体实在论会被加深。南希·卡特赖特——那个告诉我们「物理定律撒谎」的人——在 1999 年的《斑驳世界》中把这一思想扩展成整幅世界观。她主张：基本定律只在模型内部为真；它们描述现实世界的地方，只限于我们费尽心思构建的，或幸运地遇到的一台律则机器[12]：一种被屏蔽、稳定、重复运行的部件安排，能产生律则般的规律性。物理实验室是这样一台机器，太阳系大致也是。

但世界大部分地方既未被屏蔽也不稳定，整洁的定律在那里根本够不着。它们只在 ceteris paribus——「其他条件不变」——下成立，而其他条件从不会不变。在卡特赖特看来，世界不是以宏大定律为基座的金字塔，而是一块拼布：一张斑驳的被子，由许多小区域组成，不同模型在不同区域工作，区域之间没有贯穿一切的普遍定律。按她的读法，「所有模型都是错的」不再只是关于我们的地图的评论，而是关于实在的断言：定律本身从来都是理想化。 ● 活跃的哲学立场

[12]律则机器是一种稳定安排，能可靠地产出律则般的行为，比如实验装置或一个被良好隔离的自然系统。

房间 VI 信任问题

验证那不可一览无余的，以及 1994 年的炸弹

把性命押在模拟上的工程师们，早就划出了一道硬边界。验证[13]问的是：我们算对了方程吗？代码是否忠实求解了我们写下的模型？确认[14]问的是更难的事：我们算的是正确的方程吗？模型真的对应世界吗？算对，而非算该算的。几乎所有关于模型可信度的争论，本质上都是关于确认的争论。

然后，1994 年，三位研究者在这道边界下引爆了一颗小炸弹。娜奥米·奥雷斯克斯、克里斯汀·施拉德-弗雷切特与肯尼斯·贝利茨在《科学》杂志上论证，对于开放自然系统——气候、地下水、生态系统——的模型，验证与确认原则上不可能。他们的理由同时踩中了本课三条线索。自然系统从不封闭，你永远无法控制所有输入。模型解不唯一，拟合好永远不能证明你的模型就是对的——这正是正文主页上的欠决定。而通过匹配预测与数据来佐证模型，严格说就是第 3 日的肯定后件谬误：如果我的模型对，我会看到 X；我看到了 X，所以我的模型对。他们对「确认」一词的判决很严厉：它错误地暗示了模型无法获得的合法性。模型首要的价值，他们主张，是启发式的——它们是思考工具，不是真理证书。 ● 已确立的论证

那我们如何信任黑箱？

这里，正文主页上的神经天气模型又来纠缠我们。如果你无法确认一个开放系统的模型，而你甚至无法细查一个学习代理模型内部数十亿权重——汉弗莱斯所说的认识论不透明性——那你凭什么信任它的预报？

2018 年，胡安·杜兰与尼科·福马内克提出了一种回答：计算可靠主义[15]。停下来感受一下这个咔哒声。可靠主义是第 1 日的内容：一个信念之所以得到证成，不是因为你能在内部背诵出一条论证，而是因为它来自一个可靠的过程，比如良好的视力或可靠的记忆。杜兰与福马内克把这个外在主义答案对准了模拟。你不需要看穿不透明模型的内部就能信任其输出；你需要的是证据，证明产生它的过程是可靠的。

这种证据表现为：在可以检查的环节拥有一条验证与确认的记录、跨独立方法的稳健性、该技术在可被检查之处曾经成功的历史，以及了解其失效模式的专家判断。信任从透明迁移到可靠。黑箱可以像证人一样被认识：不是敲开头骨，而是审查履历。那个在第 1 日定义人脑中一个得到证成信念的东西，结果也在第 10 日定义了机器给出的一个可信答案。 ● 有原则依据，仍在争论

[13] 验证追问：代码或计算是否正确地求解了被指定的模型。

[14] 确认追问：模型是否足够好地对应了真实目标系统，以满足预定目的。

[15] 计算可靠主义认为，当模拟输出由一个具有强可靠性记录的过程产生时，该输出就是得到证成的，即使完整过程是不透明的。

奖励房间

「数字孪生」到底从哪来

正文倚仗了 2024 年美国国家科学院的定义：双向、持续更新、可预测。这个想法从哪来？严格的定义为何重要？

其先驱随阿波罗任务升空。NASA 为飞船建造了物理孪生，并把它们留在地面。当阿波罗 13 号的氧气罐在离家 20 万英里处爆裂，工程师们把实时遥测灌进地面模拟器和物理模型，为一艘无法触碰的飞船即兴修补。地球上的一个模型，用太空里濒危孪生兄弟的数据不断更新，被用来做出生死抉择：这就是数字孪生的柏拉图原型，比这个词出现早了几十年。

现代概念由迈克尔·格里夫斯在 2002 年密歇根大学的一门产品生命周期课上提出：当时的幻灯片已经显示真实空间、虚拟空间，以及两者之间的双向数据流。朗朗上口的名字是后来才有的。NASA 的约翰·维克斯约 2010 年创造了「Digital Twin」一词，它进入了 NASA 的技术路线图。大约从 2014 年起，西门子、达索、GE、ANSYS 等厂商把它印满营销材料——这正是为什么需要一个清醒的定义，也是为什么国家科学院在那里划了一条线。

这条线最好被看作三级梯子，沿袭 Kritzinger 等人（2018）：能否称得上「孪生」，不在于模型多详细，而在于数据如何流动。

这个分类故意很冷峻。一个普通数字模型可以庞大而漂亮；一个真正的孪生反倒可能视觉上很朴素。差别在于那条回路。

文字图解·什么才配得上这个名字

三者之间唯一变化的是物理对象与其虚拟副本之间的数据链接。虚线表示手动或缺失，实线表示自动。

层级	数据链接	它是什么
数字模型	无自动链接	CAD 图纸或独立模拟。你手动更新它，如果还会更新的话。许多营销中的『孪生』其实只是这个。
数字影子	单向自动链接	实时传感器数据从物理对象流入模型，但模型不能反向作用。想象实时仪表板。
数字孪生	双向自动链接	数据双向流动：模型既镜像对象，也驱动对象。只有这一层才配得上严格的名字。

奖励房间

模型的模型

最后一幅远景重新框定了整个前沿。正文把基于物理的干净模型与混乱的黑箱机器学习代理模型对立起来，讲起来比现实更锋利。两个事实模糊了这条边界。

第一：所谓「基于物理」的气候模型也不是纯第一性原理。它们无法分辨一朵云或一场雷暴，因为这些都比网格小得多，所以它们用参数化[16]来伪装小尺度过程：简化替身，其旋钮被反复调校，直到模型复现已观测到的气候。气候科学家公开把这称为「气候模型调校的艺术与科学」。一个手工调校以匹配已知气候、然后被用来预测未知气候的模型，恰恰倚靠在奥雷斯克斯警告过的那种拟合上——而那种拟合不能确认任何东西。讲原则的模型里也不乏人为调和的成分。

第二：当一个模拟本身太贵、无法运行成千上万次时，科学家会为它建造一个廉价的统计替身，称为代理模型[17]或替身模型。沿着链条看：世界被一个模拟——世界的模型——所逼近；模拟又被一个代理模型——模型的模型——所逼近。每一级都更快、更便宜，也错得更多一点。正文主页上的 AI 天气模型只是这架梯子最新、最强大的一级：直接学习模仿几十年大气的代理模型。「所有模型都是错的」原来是递归的。我们习惯于在我们错误的模型之上再建错误的模型，而每一次的技巧，都是知道哪个谎言对眼前的任务已经足够好。

把镜头拉到最远，第 10 日坐落在计算机科学家吉姆·格雷所称的科学四范式的阶梯上。

弧线·科学如何学会认知

每一个范式都没有取代前一个，而是叠加上去。今天的 AI 代理模型是第四层，实在论与工具主义的问题也随它们一起爬上了阶梯。

范式	动词	做什么
1. 经验	描述	观察并记录世界：星表、笔记、仪器。
2. 理论	解释	把观察压缩成定律与方程：牛顿、麦克斯韦，以及实在论之争的诞生。
3. 计算	模拟	方程手算无解时，用机器跑起来：第三种模式，自有其认识论。
4. 数据驱动	学习	让模型直接从海量数据中学习模式，常常内部不含显式理论：GraphCast、GenCast、NeuralGCM。

第四层正是工具主义者感到自在、而实在论者感到不安的地方——这就是为什么这个千年老问题，突然对一个天气应用变得重要起来。

[16] 参数化是简化的公式，用来代替模拟网格中太小或太复杂而无法直接表示的过程。
[17] 代理模型（emulator，或 surrogate）是一种更便宜的模型，被训练来逼近更昂贵模拟的输出。

尾声

审计那句格言

让我们用审视一切的眼光来审视这一天的口号。「所有模型都是错的，但有些有用」是真的，但它已硬化成一种终止思考的陈词滥调——人们伸手拿来它，是为了把难题挥走，而不是回答难题。真正的问题是它跳过的：错在何处？对什么而言错？有些模型错得离谱得多，而一张有用地图与一张误导地图之间的差别，正是整场游戏。安德鲁·格尔曼等统计学家恰恰在这里反击过。这句格言若被当作耸肩，会纵容马虎的工作，削弱人们去检查的冲动。

乔治·博克斯本人对此绝非虚无主义。写下「所有模型都是错的」的同一批论文，也同样强烈地警告另一种错误：把模型错当成世界，或者因为爱上了自己的繁复而「为错误的事情担忧」。他所指出的纪律不是「模型都错，所以放松」。而是更难的事，也是本课围绕它而建的东西：对这张地图与这个目的之间具体缝隙的校准判断。好的建模者不是停止说谎的人，而是清楚知道自己正在说哪个谎、以及它究竟好在何处的人。

三句话概括本附录

大观念：实在论与工具主义之争已有四百年历史，双方可能都犯了基础率谬误；它最终落在一个信任问题上——答案不是看穿模型内部，而是判断制造它的过程是否可靠。
最佳新类比：麦克斯韦从一套明知为假的想象齿轮钟表装置中导出了为真的方程：虚构是脚手架，结构是被保留下来的；这与列文斯「彼此独立的谎言的交集」配成一对。
最锋利的回调：计算可靠主义简直就是第 1 日的可靠主义对准了模拟：你信任一个不透明模型，就像信任一个证人——靠履历，而不是敲开头骨。
计算、模拟与代理模型作为科学的第三、第四模式·信息作为压缩，稳健性作为独立证据·卡特赖特斑驳世界中的涌现与跨尺度调校·全部编织回第 1–4 日。

来源与延伸阅读

1. Duhem, P. (1906). *The Aim and Structure of Physical Theory*. 理论作为经济分类，而非说明。另见 Duhem, *To Save the Phenomena* (1908)。
2. 奥西安德尔为哥白尼《天体运行论》（1543）写的匿名序言；贝拉尔米内 1615 年致福斯卡里尼的信。参见 *Stanford Encyclopedia of Philosophy*, "Scientific Realism"。plato.stanford.edu/entries/scientific-realism
3. 阿拉戈／泊松亮斑：法国科学院悬赏竞赛，约 1818–1819；菲涅耳波动说及其先预言后观测到的圆盘阴影中心亮斑。en.wikipedia.org/wiki/Arago_spot
4. Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. Routledge. 分而治之的实在论；有效设定与闲置设定。Kitcher, P. (1993). *The Advancement of Science*。
5. Magnus, P. D. & Callender, C. (2004). "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71(3): 320-338. philpapers.org/rec/MAGREA-2 另见 Howson (2000), *Hume's Problem*。

6. Levins, R. (1966). "The Strategy of Model Building in Population Biology." *American Scientist* 54(4): 421-431. 一般性／逼真性／精确性取舍、稳健性，以及「我们的真理是彼此独立的谎言的交集」。批评：Orzack & Sober (1993), *American Naturalist* 148:201。
7. Morgan, M. & Morrison, M., eds. (1999). *Models as Mediators*. Cambridge University Press.
8. Giere, R. N. (1988/2006). *Explaining Science*; *Scientific Perspectivism*. University of Chicago Press. 语义／基于模型的观点与视角实在论。
9. Vaihinger, H. (1911). *Die Philosophie des Als Ob*(*The Philosophy of "As If"*, 英译本 1924)。有用虚构。另见 Frigg & Hartmann, "Models in Science," *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/models-science
10. Maxwell, J. C. (1861-62). "On Physical Lines of Force." *Philosophical Magazine*. 电磁场的机械涡旋与惰轮模型。
11. Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press. 律则机器、ceteris paribus 定律与斑驳世界。
12. Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994). "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences." *Science* 263(5147): 641-646. doi:10.1126/science.263.5147.641. science.org
13. Duran, J. M. & Formanek, N. (2018). "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28(4): 645-666. doi:10.1007/s11023-018-9481-6. link.springer.com
14. Grieves, M. (2002 概念；2014 白皮书"Digital Twin: Manufacturing Excellence through Virtual Factory Replication")；John Vickers ／ NASA 技术路线图约 2010 年使用；Grieves & Vickers, "Origins of the Digital Twin Concept"(2016)。
15. Kritzinger, W. et al. (2018). "Digital Twin in manufacturing: A categorical literature review and classification." *IFAC-PapersOnLine* 51(11): 1016-1022. 按数据流自动化区分 digital model / digital shadow / digital twin。
16. Kennedy, M. C. & O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society B* 63(3): 425-464. 昂贵模拟的高斯过程代理模型。
17. Hourdin, F. et al. (2017). "The Art and Science of Climate Model Tuning." *Bulletin of the American Meteorological Society* 98(3): 589-602.
18. Hey, T., Tansley, S. & Tolle, K., eds. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. 吉姆·格雷的经验 → 理论 → 计算 → 数据驱动框架。
19. 关于「所有模型都是错的」作为 cliche：Andrew Gelman, "Statistical Modeling, Causal Inference, and Social Science"（博客，多篇），以及 Box 本人 1976/1979 年论文中细致入微的表述。

来源卫生：与正文一样，源研究过程中浮现出一些截止后或未来日期的 arXiv 标识符，本附录未依赖它们；以上每一条来源都是有明确日期的出版物、机构页面或百科全书条目。

可选附录：地图边缘

古代地图绘制者有个诚实的习惯：在测绘终止、传闻开始之处，他们画上怪兽，写下「hic sunt dracones」——此处有龙。正文与附录 I 描绘的是相对成熟的疆域：博尔赫斯、贝拉尔米内、验证、可靠主义，以及那场古老的实在论之争。本附录则驶向地图边缘。下文大部分内容来自 2020 年或更晚，许多出自最近几年，而其中相当一部分日后会被证明是错的。这正是重点所在。前沿地带，才是校准最关键的地方。

请先读这段

下面的每条论断都带有一个标签。● 已确立 表示经过同行评议、重复验证或硬件验证。● 有前景 表示可信但尚处早期、范围狭窄，或仍只是预印本形态。● 有争议／炒作风险 表示证据薄弱、论述夸大，或独立评议已经戳破了主张。第 10 日的技能——对模型在多大程度上、因何而错的校准判断——正是阅读这片前沿所需要的技能。

转向

从构建模型到召唤模型

2020 年后最深刻的变化不是某一项单独结果，而是科学模型由什么构成发生了变化。四个世纪以来，常规配方是明确的：选择理想化、写下方程、求解。大约从 2022 年起，另一种配方开始蔓延。在一个庞大神经网络上用海量领域数据训练，直到它吸收整个领域的统计形态，再把这个模型微调到许多具体任务上。

这些系统就是基础模型[18]，把它们引入科学是这十年的重大事件之一。Wang 等 41 位合著者 2023 年发表于 *Nature* 的综述 *Scientific discovery in the age of artificial intelligence*，宛如在新领土上插下的一面旗帜。● 有影响力的综述

旗舰级的地球系统实例出现在 2025 年 5 月：**Aurora**，一个来自微软研究院及合作者的 13 亿参数模型，在一百多万小时地球物理数据上预训练。它并非为单一任务调优，而是一个通用的预训练大气模型，可被微调到空气质量、海浪、热带气旋路径和高分辨率天气。论文报告，Aurora 在 74% 的目标上持平或优于数值空气污染模拟，在 86% 的变量上优于海浪模型，并在多个机构的五天热带气旋业务路径预报上表现更好，其中包括美国国家飓风中心负责的海区。作者将其表述为机器学习模型首次在最长五天的完整业务热带气旋预报上实现超越。● 同行评议结果

注意这对正文主页上的实在论之争做了什么。Aurora 是工业化了的建构经验论：一个内部不含人类可读气象学的模型，仍能在重要预报任务上做到经验充分。实在论者的不适——一个预测却不以可审视方式表征世界的模型——如今是一项已部署的研究计划，而不再是课堂上的谜题。

同样的模式正在扩散。NASA 与 IBM 的 **Prithvi WxC** 于 2024 年以开放权重加预印本的形式发布，是一个在 MERRA-2 变量上训练的 23 亿参数天气-气候基础模

[18] 基础模型是一种大型预训练模型，可被适配到许多下游任务，而不是从头为某个狭窄用途训练。

型，并在预报、降尺度、重力波通量参数化和极端事件上接受测试。已发布与已验证之间的差距，正是防炒作滤镜存在的原因。 ● 有前景的预印本

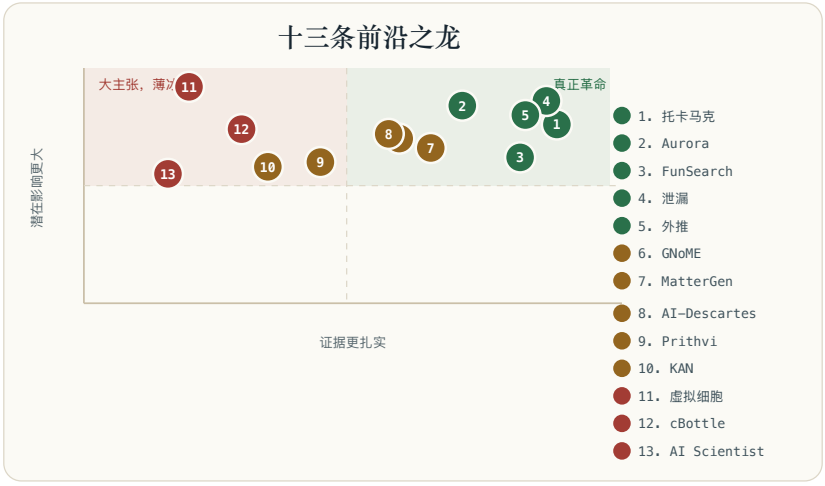
前沿地图

十三条龙，已排序

有用的问题不是「这有多令人印象深刻?」。有用的问题有两个轴：证据有多扎实，以及 ** 万一成立，它有多重要？** 右上角是真正的革命：重大且扎实。左上角是龙出没的地方：庞大主张立于薄冰之上。

前沿地图·若成立的影响 × 证据

位置取决于判断，而非测量。表格让网络、EPUB 和 PDF 读者都能看到原始交互凭据。



位置取决于判断，而非测量。绿色靠右的工作已经较稳；左上角的宏大说法仍需要等待验证。

研究	状态	定位理由
托卡马克等离子体控制	已确立	深度强化学习在 TCV 上塑造等离子体，并在 DIII-D 上降低撕裂不稳定性风险：真实硬件，高影响。

Aurora 地球系统模型	已确立	同行评议的多任务预报技能，包括五天业务气旋路径；但将其称为完整地球模型的说辞仍应保持克制。
FunSearch	已确立	LLM 加评估器发现了新的可验证数学构造，使输出可被检查而非不透明。
泄漏危机	已确立	Kapoor 与 Narayanan 在 17 个领域、294 篇论文中发现泄漏失效；它会削弱许多夸大结果，也正在重塑领域。
外推漂移	已确立	NeuralGCM 与 ACE2 展示了硬边界：训练气候内的技能不能保证训练气候外也有。
GNoME 材料发现	有前景	*Nature* 的结果真实；但标题宣称数量的实际新颖性、有用性和可合成性仍有争议。
MatterGen	有前景	经同行评议的生成式材料工作，有一个合成验证点，但广泛材料设计的回报仍处早期。
AI-Descartes / AI-Hilbert	有前景	强实在论方向：从数据加背景理论中恢复定律。大多数例子仍是重新发现已知定律。
Prithvi WxC	有前景	开放的 23 亿参数天气-气候模型；下游任务的证据仍处预印本时代。
KANs	有前景	2024 年预印本与 2026 年 PRX 后续研究主张可解释的科学发现；但相比普通网络的广泛优势仍未解决。
AI 虚拟细胞	有争议	一篇严肃的 *Cell* 路线图，而非功能正常的细胞孪生。野心巨大；建成品尚未出现。
cBottle / Earth-2	有争议	一个可信的生成式气候代理预印本，被包裹在数字星球与压缩营销中。
The AI Scientist	有争议	自主科学家的框架被独立评估大幅戳破：实验失败、新颖性错误、幻觉结果。

急件 1 ·代理模型接管操控

机器现在操控机器

有些前沿模型已经在最字面的意义上承受负荷：被委以昂贵的硬件和不稳定的等离子体。2022 年，DeepMind 与 EPFL 发表了一个强化学习控制器，它塑造了一台真实托卡马克聚变反应堆的磁场，在 TCV 装置上将等离子体维持在目标构型。2024 年，后续研究用深度强化学习降低 DIII-D 托卡马克上撕裂不稳定性的风险。当一个学习模型被允许实时引导聚变等离子体时，「只是个模型」就不再是有用的搪塞。 ● 硬件验证

同一急件也带来一个警告。物理信息神经网络[19]曾被宣传为显式方程与学习模型之间的桥梁：把微分方程写进损失函数，再让网络学习一个解。失效模式文献揭示了陷阱。PINN 可以一边把残差损失压得很低，一边让真实解严重偏离。模型可以满足你写下的每一条约束，却仍错过世界，因为你并没有写下所有重要的约束。 ● 失效模式已确立 ● 修复方案有前景

急件 2 ·实在论的反击

发现定律，而非仅仅拟合

如果说黑箱代理模型是工具主义的胜利，那么一场更安静的反运动正在为实在论而战。它的要求很简单：不要满足于一个只会预测的模型；要让机器交给你一条能读得懂的定律。这就是符号回归[20]与理论引导的发现。

三个标本值得关注。**AI-Descartes** 将符号回归与对背景公理的逻辑推理相结合，从数据加理论中重新推导出诸如开普勒第三定律这样的定律。**AI-Hilbert** 通过多项式优化与形式化证书推进这一思想。**FunSearch** 将语言模型与自动评估器及进化搜索配对；在 *Nature* 上，它发现了 cap-set 问题的新构造以及新的装箱启发式。最后一个案例在哲学上重要，因为它的输出不是权重向量，而是人类可以检查、检验并证明的对象。 ● FUNSEARCH 已确立

这个保留意见让它无法变成实在论的庆功巡游。许多系统仍然主要是在重新发现我们已经知道的定律，这是有力的验证，却是单薄的发现。KAN 系列同样前景可期但尚未定论：2024 年的 Kolmogorov-Arnold Networks 预印本提出将可学习函数放在网络边上，作为多层感知器更具可解释性的替代方案；2026 年 *Physical Review X* 的后续研究则论证其在科学发现中的应用。这是真正的势头。但它仍不能证明 KAN 通常能胜过调优良好的普通网络。 ● 有前景 ● KAN 角色仍在争论

注意落差·论文测得的 vs 媒体报道的

前沿最大的扭曲，藏在谨慎结果与它的发布之间。

[19] 物理信息神经网络在训练神经模型时，会对已知方程或边界条件的违反施加惩罚。

[20] 符号回归搜索能拟合数据的简洁数学表达式，理想情况下给出一个可解释的规律，而非不透明的预测器。

案例	论文实际测得	媒体报道暗示
Aurora	在空气质量、海浪、气旋路径和天气上，相对业务系统的多任务预报提升。	一个整个地球的基础模型。更准确的描述是，它是一个强大的预报与代理模型。
GNoME	基于图网络筛选的大规模预测稳定晶体结构。	可用已知材料数量级的扩展。新颖性与可合成性需要更严格的审计。
cBottle / Earth-2	一个用于公里尺度全球大气场的生成式扩散代理模型，目前处于预印本阶段。	一个行星的数字孪生；厂商包装的压缩主张承担了额外的修辞工作。
AI 虚拟细胞	一份关于构建 AI 虚拟细胞所需优先级、数据、评估与合作的路线图。	一个功能正常的活细胞数字孪生已经存在。
The AI Scientist	一份自主研究的预印本；独立测试发现编码失败、新颖性判断差以及幻觉结果。	廉价、完全自主的科学发现新时代。

急件 3 ·应让你警惕的结果

数据泄漏与 AI 科学中的可重复性危机

本附录中最重要的结果是一个警告。2023 年，Sayash Kapoor 与 Arvind Narayanan 在 *Patterns* 上发表 *Leakage and the reproducibility crisis in machine-learning-based science*。他们的调查发现，泄漏错误横跨 17 个领域，共影响 294 篇论文。数据泄漏[21]会让模型在论文上看起来很出色，因为评测数据已经部分进入训练流程。

这是第 2 日的复现危机在 AI 时代的重生。当模型在它实际上已经见过的信息上被评估时，它的「准确率」就是博尔赫斯那张与训练集等大的地图：完美，却无用。

● 塑造领域的警示

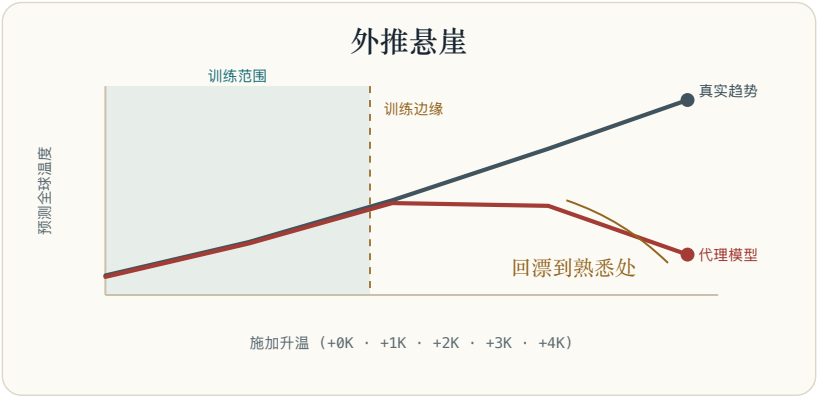
泄漏抬升了你手头数据内的表现。当世界移到模型未见过的气候之外时，它的孪生问题就出现了。NeuralGCM 的作者报告，代理模型在 +1 K 和 +2 K 海表温度扰动下仍保持合理，但在 +4 K 时，其增暖响应偏离预期。艾伦研究所团队的 ACE2 明确指出，对单独变化的海表温度和二氧化碳的敏感性并不完全真实。这正是实在论担忧的经验核心：模式学习者可以是出色的插值器，却是不可靠的预言家。

● 外推失效已确立

图示·预测技能不等于机制

失效模式示意图：在训练范围内表现出色，当模型被推往未曾见过的地方时，却向熟悉处漂移。

[21] 数据泄漏指来自测试目标或未来数据的信息潜入训练、验证、预处理或特征选择。



这个图是示意，不复刻某篇论文的图。它表达 NeuralGCM 与 ACE2 警示的失效模式：训练区内技能很强，不等于训练区外有机制性理解。

急件 4 ·新哲学

一个不可读的模型能让你理解什么？

哲学家们并未袖手旁观。Emily Sullivan 2022 年的 *Understanding from Machine Learning Models* 给出了最干净的重新框架：阻碍理解的并非黑箱复杂性本身，而是关联不确定性[22]——即模型与其声称所指向的目标系统之间联系薄弱的证据。一个关于错误目标的透明模型什么都给不了你；一个与经验联系紧密的不透明模型仍可能支持理解。不透明性不是核心问题，证据才是。

其他人以不同方式划分这片地形。Florian Boge 区分训练的不透明性与表征的不透明性，并警告深度学习可能在发现与解释之间撕开一道缝隙。Ramon Alvarado 主张 AI 是一种认识技术，一种直接作用于知识本身的工具，值得拥有自己的认识论。Sabina Leonelli 则把注意力保持在数据生态系统上：这些模型继承了喂养它们的数据集的不透明性、偏见、缺口与政治。结论是，AI-for-science 可能正在迫使出现第三种认识论范畴：不是理论，不是实验，也不完全是普通的模拟。● 活跃且未决的哲学

[22] 关联不确定性是指：模型是否在经验上以正确的方式与目标系统相连接，这一点存在不确定性。

急件 5 ·大胆的龙

虚拟细胞、行星孪生与做科学的 AI

现在来到地图的左上角：主张如此宏大，若落地则会重绘一切，而它们今天仍立足于更薄的冰面上。

AI 虚拟细胞。2024 年，一个横跨斯坦福、陈-扎克伯格倡议、基因泰克、谷歌等机构的大型团队在 *Cell* 上发表路线图，旨在构建活细胞的多尺度 AI 模型。这一雄心是认真的：一个虚拟细胞，能够支持跨生物学尺度的可解释 *in silico* 实验。但论文是一份关于优先级与机遇的路线图，而非已建成的细胞孪生。 ● 愿景与路线图

** 行星孪生。**NVIDIA 的 **cBottle** 是一个用于公里尺度全球大气场的生成模型：一个粗分辨率全球生成器加上局部超分辨率，在模拟和再分析数据上训练。这份预印本可信且有用。围绕 Earth-2 的「行星数字孪生」光环，是营销语言在替尚未赢得的证据干活。 ● 预印本 ● 行星孪生框架是炒作

** 自主科学家。**Sakana AI 2024 年的「AI Scientist」预印本推销了一位端到端自动化研究者：选题、实验、论文、评审。2025 年的独立评估发现了更严峻的现实：十二个提议实验中有五个因编码错误而失败，新颖性判断很差，部分结果是幻觉或误导。结果并非毫无用处；自主科学家的框架正是「有争议」标签的用武之地。 ● 已被独立戳破

材料发现的警示故事

一个故事就能浓缩整篇附录。2023 年，Google DeepMind 的 **GNoME** 宣布了 220 万种晶体结构，其中数十万种被预测为稳定。结果真实且令人印象深刻；一篇经同行评议的 ACS 观点文章很快追问，其中有多少是真正新颖、有用或可合成的。2025 年，微软的 **MatterGen** 有成效地改变了问题，从列出候选材料转向在目标约束下生成材料，并报告了一个合成验证点。整个弧线就是前沿的缩影：真实结果、夸大标题、清醒修正、底层进步。相信论文；审计媒体；等待合成。 ● 真实结果 ● 范围受争议

未决问题

能下定论的实验

正确的姿态不是相信，也不是否定。而是知道哪一个结果会改变你的想法。

- ** 经过验证的增暖外推。** 如果某个 ML 气候代理模型能在留出验证的高排放模拟或古气候上复现真实的增暖响应，工具主义的论据就会增强。现在，+4 K 的漂移在说：谨慎。
- ** 经泄漏审计的重复验证。** 对一项旗舰基础模型科学主张进行独立的、检查过泄漏的重复验证，会告诉我们 Kapoor 与 Narayanan 的警告在领域顶端适用到什么程度。
- ** 真正新颖的定律发现。** 一个符号回归系统发现了一条后来被确认的自然定

律，将是领域格局的转变。FunSearch 的 cap-set 结果是纯数学中最接近的类比。

- ** 经过验证的临床数字孪生。** 一份通过同行评议、独立验证并达到美国国家科学院标准的患者或器官孪生，将把医学界的龙移出左上角。

三句话概括本附录

** 核心观点：**2020 年以来，建模已被重新组织在基础模型与学习型代理模型周围——它们无需人类可读的表征就能很好地预测——而符号发现则试图让机器产出可检查的定律。
** 最尖锐的警示：** 数据泄漏已经抬升了已发表的科学 ML 结果，即使强大的气候代理模型在被推出训练范围时也会漂移；预测技能不等于机制。
** 仍在争论：** 深度学习模型是只能预测，还是能解释；AI 虚拟细胞、行星孪生主张与自主科学家系统是壮观但证据仍不足的赌注。
计算作为建模基质·信息作为压缩、泄漏与证据关联·多尺度细胞与气候模型中的涌现·第 10 日的实在论/工具主义分野在 2020 年代硬件上经受压力测试。

来源与延伸阅读

1. Wang, H. et al. (2023). "Scientific discovery in the age of artificial intelligence." *Nature* 620:47-60. doi:10.1038/s41586-023-06221-2
2. Bodnar, C. et al. (2025). "A foundation model for the Earth system"(Aurora). *Nature* 641:1180-1187. doi:10.1038/s41586-025-09005-y
3. Schmude, J. et al. (2024). "Prithvi WxC: Foundation Model for Weather and Climate."[preprint] arXiv:2409.13598; NASA/IBM open model release.
4. Degrave, J. et al. (2022). "Magnetic control of tokamak plasmas through deep reinforcement learning." *Nature* 602:414-419. doi:10.1038/s41586-021-04301-9
5. Seo, J. et al. (2024). "Avoiding fusion plasma tearing instability with deep reinforcement learning." *Nature* 626:746-751. doi:10.1038/s41586-024-07024-9
6. Karniadakis, G. E. et al. (2021). "Physics-informed machine learning." *Nature Reviews Physics* 3:422-440. See also Krishnapriyan, A. et al. (2021), "Characterizing possible failure modes in physics-informed neural networks," *NeurIPS 2021*.
7. Romera-Paredes, B. et al. (2023). "Mathematical discoveries from program search with large language models"(FunSearch). *Nature* 625:468-475. doi:10.1038/s41586-023-06924-6
8. Cornelio, C. et al. (2023). "Combining data and theory for derivable scientific discovery with AI-Descartes." *Nature Communications* 14:1777. doi:10.1038/s41467-023-37236-y
9. Cory-Wright, R. et al. (2024). "Evolving scientific discovery by unifying data and background knowledge with AI-Hilbert." *Nature Communications* 15:5922. doi:10.1038/s41467-024-50074-w
10. Liu, Z. et al. (2024). "KAN: Kolmogorov-Arnold Networks."[preprint] arXiv:2404.19756. See also "Kolmogorov-Arnold Networks Meet Science," *Physical Review X* (2026).
11. Kapoor, S. & Narayanan, A. (2023). "Leakage and the reproducibility crisis in machine-learning-based science." *Patterns* 4(9):100804. doi:10.1016/j.patter.2023.100804

12. Kochkov, D. et al. (2024). “Neural general circulation models for weather and climate”(NeuralGCM). *Nature* 632:1060-1066. doi:10.1038/s41586-024-07744-y

13. Watt-Meyer, O. et al. (2025). “ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses.” *npj Climate and Atmospheric Science* 8:205. See also arXiv:2411.11268.

14. Sullivan, E. (2022). “Understanding from Machine Learning Models.” *British Journal for the Philosophy of Science* 73(1):109-133. See also Grote, T., Genin, K. & Sullivan, E. (2024), “Reliability in Machine Learning,” *Philosophy Compass* 19(5):e12974.

15. Boge, F. J. (2022). “Two Dimensions of Opacity and the Deep Learning Predicament.” *Minds and Machines* 32:43-75. See also Duede, E. (2023), “Deep Learning Opacity in Scientific Discovery,” *Philosophy of Science* 90(5):1089-1099.

16. Alvarado, R. (2023). “AI as an Epistemic Technology.” *Science and Engineering Ethics* 29:32. doi:10.1007/s11948-023-00451-3. See also Leonelli, S. (2023), *Philosophy of Open Science*, Cambridge University Press.

17. Bunne, C. et al. (2024). “How to build the virtual cell with artificial intelligence: Priorities and opportunities.” *Cell* 187(25):7045-7063. doi:10.1016/j.cell.2024.11.015

18. Brenowitz, N. et al. (2025). “Climate in a Bottle: Towards a Generative Foundation Model for the Kilometer-Scale Global Atmosphere.”[preprint] arXiv:2505.06474; NVIDIA Earth-2/cBottle release materials.

19. Lu, C. et al. (2024). “The AI Scientist.”[preprint] arXiv:2408.06292. Independent evaluation: Beel, J., Kan, M.-Y. & Baumgart, S. (2025), “Evaluating Sakana’s AI Scientist,” arXiv:2502.14297 / *ACM SIGIR Forum*.

20. Merchant, A. et al. (2023). “Scaling deep learning for materials discovery”(GNoME). *Nature* 624:80-85. doi:10.1038/s41586-023-06735-9. Zeni, C. et al. (2025). “A generative model for inorganic materials design”(MatterGen). *Nature* 639:624-632. doi:10.1038/s41586-025-08628-5

来源把关：前沿嘈杂，因此本附录使用有日期的出版物、机构模型发布和可识别的预印本，而不是无凭据的头条主张。博客与厂商措辞仅被当作措辞，而非证据。