从根基到 2026 年研究前沿（含专题深入）

深入一百八十日

作者：Claude Opus 与 GPT
翻译：Kimi 与 GLM
人工编辑：刘家昌

目录

一百八十日地图

引言

如何阅读一张从根基走向前沿的地图。

这本书的起点不是让你拥有能够吹嘘的资本，而是一份渴求——对知识本身不可遏制的好奇，以及一种愿望：我们想要获得值得信任的，对世界的知识，但又不希望研究之间的争议迫使我们把头埋进土里。本书面向的是好奇心旺盛的通才型读者——某些领域根基扎实，另一些领域尚存空白，不愿在打基础与追前沿之间取舍。这本书并不打算让你在一百八十天内精通一切，而是提供「定向」：一幅地图，标出现实、生命、心智、技术、社会与未来之所以能够被人类接受的锚点。

本项目由 AI 系统完成深度研究、综合与初稿撰写，但内容不会原样发布。人工编辑 刘家昌 逐篇核查材料，改善结构与可读性，并打磨中文译笔，确保课程读起来是一份清晰的学习文本，而非未经整理的生成产物。

让我们从路径的最初开始：只有先校准信念的尺度，前沿才有意义。因此，课程开篇不是宇宙学[1]、人工智能或医学，而是知识本身——什么才称得上正当的理由，为什么一个为真的信念仍然可能只是运气，科学如何把可检验的断言与自圆其说的故事区分开来，以及概率如何让心智在不确定中依然从容前行。唯有经过这番准备，深入之路才向数学、物理、化学、生物、医学、神经科学、人工智能、经济学、文明、伦理以及正在重塑未来的种种力量徐徐展开。

每一天的编排都尽可能满足不同人拥有的不同空余时间。它从一个谜题、一段故事、一幅图像、一个类比或一个思想实验入手；建立一个对问题理解的模型；点明当下仍在进行的争论；然后循着证据所容许的边界，走向新近而可靠的研究。这本书的叙述接近对每一话题的极简导论，但每一主题的讨论有更深的学习曲线——开篇假设读者聪慧而初来乍到，随后一路深入，直到脚下不再有水泥，变得松软，真正触及前沿、充满争议。

本专题深入 PDF 会在某些日子的正文之后收录额外附录。它们留给兴趣浓厚、时间充裕的读者；既非后续章节的前提，也不构成后续内容的基石。

我们推荐读者们从书的最开始阅读。这不是一只陈列一百八十件趣闻的百宝箱，而是按依赖关系精心排序的：认识论先于统计学，统计学先于实验，数学先于物理，热力学[2]先于生命，演化先于心智，计算先于现代人工智能。在狭窄的前沿道路上，前文的讨论

为其留下行走的空间；在前沿争议之处，它继续深入——哈勃张力[3]、生命起源的物理、哺乳动物跨代表观遗传[4]、意识理论、通用人工智能与对齐[5]，以及不平等的深层历史。

五条线索贯穿整门课程：

- **信息**，因为每一门学科最终都要追问：什么是信号，什么是噪声，什么可以被传递或推断。
- **能量**，因为秩序的物理代价在热力学、生命、经济学、气候与计算中反复现身。
- **演化**，因为选择绝非仅限于生物机制；它也是知识、文化、技术与制度演进的模式。
- **涌现**，因为知识的地图上最重要的锚点通常是人类共有的：温度、细胞、市场、心智、社会。
- **计算**，因为程序化的形式验证成为数学、物理、大脑与机器的共通语言。

「前沿校准器」是本书方法的一部分。前沿主张会被标记为**已确立**、**线索**，或**争议/炒作**。我们克制地避免无条件信任所有的前沿研究。我们的目标不是让你相信新奇的事物，而是向你展示它们展现的证据是否使得它们具有值得我们注视的重量。物理学与宇宙学的主张需要数据集与误差棒[6]的佐证。医学、人工智能与社会科学的主张需要可复现性、对激励结构的审视、精确的测量，以及谦逊。一个结果可以令人兴奋，却未必有着真理的重量。一个失败的主张若能教会我们科学如何自我纠正，仍有其价值。前沿不等于可靠；同行评议[7]不等于定论；优美不等于真实。

前四日为全书定下基调。第 1 日追问：一座停了的钟为何能给你一个为真且证成的信念，却不给你知识；第 2 日将这一忧虑从个体心智放大到科学作为制度；第 3 日打开推理引擎本身；第 4 日把不确定性变成一套演算，用蒙提霍尔、贝叶斯定理与 e 值[8]展示证据到来时信念会怎样改变。

这就是深入：不是一份事实的目录，而是一门关于「事实如何赢得立足之地」的课程。

说明

1. 宇宙学研究整个宇宙的起源、膨胀、组成与大尺度结构。
2. 热力学研究热、功、能量、熵，以及有序过程所受的物理限制。
3. 哈勃张力指两类主流方法测得的宇宙膨胀速度不一致。
4. 表观遗传继承指基因活动方式的改变有时能跨代传递，但 DNA 序列本身没有改变。
5. 通用人工智能指具有广泛任务能力的 AI；对齐研究如何让强大 AI 可靠追随人类意图。
6. 误差棒表示测量或估计的不确定范围，提示结果可能合理落在哪一带。
7. 同行评议指发表前由同领域专家审阅；它是质量控制，不是真理保证。
8. e 值把统计证据表示为一场反对零假设的公平赌局中赢得的倍数。

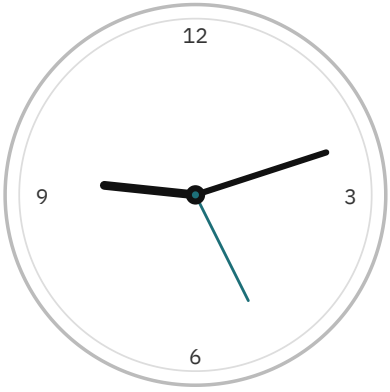
模块 I

知识与推理的根基

模块一 · 知识与推理的根基 · 第 01 日 / 180

知识是什么？

你看了时钟。时间正好。但你真的知道吗？



● 已在 12 小时前停走——但就在这一分钟，它恰好正确

早上九点十二分，你快要迟到了。匆匆路过时，你抬头瞥了一眼车站那座大钟，读出 9:12，心想：「还好——还有三分钟富余。」你没错，此刻确实是 9:12。然而，你信赖的这座钟恰好在十二小时前的凌晨停在了 9:12，从此凝固不动。你不过是在它一天中唯一碰巧正确的那一刻，凭一台坏掉的仪器下了判断。

你的信念是真的。它基于一条完全合理的理由——时钟就是用来报时的，而你此前已安然无恙地信赖过成千上百座钟。你发自内心地相信它。那么，你知道此刻是 9:12 吗？仔细追问，几乎所有人都会摇头——总觉得缺了什么。但缺的究竟是什么？哲学家们为此争论了六十年；而类似的困惑，如我们将看到的，早在千年前便已浮现。

这是第一次深入，因此身后尚无来路——日志一片空白。今天我们要播下种子。今日引入的这套机制（信念以程度呈现；依证据更新；心智作为推理引擎）是整个课程赖以支撑的认识论工具[1]。请留意它将在第 2 日（科学如何判定什么才算数）、第 4 日（概率如何成为部分信念的逻辑）、第 7 日（信息）、第 119 日（预测性大脑）以及第 149 日（著名发现为何在复现中消散）中重新浮现。我们将贯穿全部 180 天的五条线索——信息、能量、演化、涌现、计算——都在此处悄然首演。

模型

三条腿的凳子

大约二十三个世纪以来，西方哲学一直抱持着一个关于「知识是什么？」的简洁答案。要知道某事为真，你需要同时具备三点：

（1）你相信它——你无法知道你甚至不认为真的东西。（2）它是真的——你不能知道一个假命题；那些说「我就知道地球是平的」的人，只是相信它，自信而错误地相信。（3）你有证成——因为仅凭运气猜中，算不得知识。那个对冷门胜出「就是有种预感」的赌徒，即便赢了，也并未知道它会赢。

依此观点，知识即证成的真信念——JTB（Justified True Belief，证成的真信念），一条三条腿的凳子。抽掉任何一条腿，它都会倾倒。对这一观点的理解通常追溯至柏拉图，他在《泰阿泰德篇》中提出，知识是「带有说明的真判断」。这里有一种美妙的反讽，历史学家们津津乐道：正是在那篇对话中，苏格拉底随后拆解了这个定义，因此柏拉图可以说从未真正认可过那项以他命名的学说。正如一位学者所言，这就像一位杰出的批评家在摧毁某个传统的瞬间，竟又创造了它。

尽管如此，这一粗略的共识还是维系了下来。凳子看似稳固。然后，一位时年三十五岁的哲学家——据说他此前发表寥寥，又恰好有些发表的压力——写了一篇三页纸的论文。

引爆

盖梯尔的三页论文

1963 年，埃德蒙·盖梯尔在期刊 Analysis 上发表了一篇论文，标题直白得近乎俏皮：《知识是证成的真信念吗？》（Is Justified True Belief Knowledge?）。全文仅三页。此后它被引用了数千次，并催生了整整几个子领域。现代哲学中，鲜有文献以每字计造成了更大的破坏。

盖梯尔的招数简单得令人崩溃。他构造了一些小故事，其中凳子的三条腿都稳稳地立在地上——信念、为真、证成——但你绝不会说那个人知道。以下是他第一个案例的轻度现代化版本：

> *史密斯与琼斯申请同一份工作。老板告诉史密斯：「琼斯会得到这个职位。」史密斯还闲来无事数了琼斯口袋里的硬币：十枚。于是史密斯形成了一个证成充分的信念：「得到这份工作的人口袋里有十枚硬币。」*

现在出现转折。老板错了（或者改变了主意）：得到工作的是史密斯，而非琼斯。而且——史密斯本人完全不知情——他自己的口袋里恰好也有十枚硬币。来看他的信念，「得到这份工作的人口袋里有十枚硬币」：它是真的（获胜者史密斯确实有十枚硬币），它是证成充分的（绝佳的证据——老板的话，实打实的硬币清点），而且他是真诚地相信的。JTB（Justified True Belief，证成的真信念），三条腿齐全。然而史密斯显然并不知道这一点。他追踪的是琼斯，却在错误的人身上得出了正确的结论。

这便是盖梯尔案例的基本结构：你的理由经由一个假命题运行（「琼斯会得到这份工作」），而你的信念又被一桩无关的巧合（「史密斯也有十枚硬币」）碰巧带向真实。理由与事实从未真正相遇。停走的时钟只是同一种结构更清楚的版本：你的理由（那座钟）损坏了，而事实（此刻是 9:12）全凭运气成立。

比名字更古老的转折

盖梯尔并非首创。伯特兰·罗素在 Human Knowledge: Its Scope and Limits（1948）中就已提出停钟案例。再往前追溯，这个问题堪称古老：大约在公元 770 年，佛教逻辑学家法上（Dharmottara）描述了一位旅人，他看到山丘上仿佛有烟，推断有火，而且确实有火——只不过那「烟」其实是一群昆虫。同一种结构，早了十二个世纪。十四世纪的印度，甘格沙为处理此类案例建立了一整套因果知识理论。「盖梯尔问题」是哲学中趋同发现的绝佳实例——那种心智会独立地一再绊倒的东西，而它本身就在暗示：那里有某种真实的东西。

盖梯尔案例表

案例	信念	为真	证成	运气	裁决
普通的知识	是	是	是	否	在经典 JTB（Justified True Belief，证成的真信念）观点下，这是知识
停走的钟	是	是	是	是	非知识：事实只是碰巧成立
幸运的猜测	是	是	否	是	非知识：缺乏证成
自信的错误	是	否	是	否	非知识：命题为假

补丁战

寻找第四条椅子腿

面对盖梯尔，最自然的回应是：增设第四项条件，把运气筛除。几十年来，认识论家们孜孜以求——而每一次利落的修补都撞上一个更刁钻的反例。这几乎成了一场残酷的围猎。

无假前提。最初的想法是：知识不能经由一个假命题推理得出。史密斯的信念倚赖于「琼斯会得到这份工作」，而这是假的；禁绝它，你便安全了。干净利落——直到阿尔文·戈德曼提出假谷仓之国（1976）。你驾车穿过一片区域，那里有人恶作剧，把每一座「谷仓」都做成平板电影布景——除了一座例外。你恰好瞥见了那座真谷仓，心想「那是座谷仓」。你的信念为真、证成充分，且不依赖任何假前提。然而你并不知道那是谷仓：你本可以如此轻易地在百米之外被布景板愚弄。

追踪真理。那么，也许知识关乎你的信念在邻近的可能世界中[2]如何表现。罗伯特·诺齐克（1981）提出了敏感性：你知道命题 p，仅当若 p 为假，你便不会相信它。优雅——却在边缘情形中产出古怪的结论。欧内斯特·索萨（1999）将其改写为安全性：在所有相近的可能情形中，你都不会出错。停走的钟在安全性上惨败（早一分钟或晚一分钟你便错了）；运转正常的钟则通过这一安全性的认证。假谷仓前的你同样未能通过安全测试。

随后，琳达·扎格泽布斯基（1994）以一种配方式的论证给了所有此类修补以致命一击——足以击溃任何同类方案。取一个有证成、却仍可能为假的信念（而证成既然可错，总允许这种可能）。安排理由失准，使信念为假——再借运气安排，让它终究为真。只要你的第四条条件没有走到要求理由保证为真那一步，运气就总能重新钻回空隙。补丁战或许在结构上便不可能获胜。

两种退出战场的方式

宣布知识为原初概念。蒂莫西·威廉森在 Knowledge and Its Limits（2000）中迈出了激进的一步：停止试图用更简单的零件拼凑知识。也许它根本无从分析。在他的知识优先视域中，知道是一种基本的心智状态——最普遍的事实性状态[3]——而我们应当用知识去解释信念、证据与证成，而非反其道而行。你无法把氢或约翰·F·肯尼迪拆解成更简单的概念；也许知识同样是基石。六十年来失败的定义，看起来不再像一个谜题，而更像一条线索。

诉诸能力。另一条出路是德性认识论（仍然是索萨提出的）。知识是适切的信念——它之所以为真，是因为认知者具有相应能力，而非凭偶然。想象一位弓箭手。一箭中的，仅当箭矢命中靶心是因为射手瞄准精妙——而非一阵风把劣射吹回了靶心。盖梯尔化的认知者正是那位弓箭手：第一阵风将箭吹离靶心，第二阵风又把它吹了回来。准确命中，但不是出于能力，也因而不是适切。索萨说，这便是运气之击不算知识的缘由。

辩论

信念究竟何以获得证成？

从「这是知识吗？」退后一步，回到那条更谦卑的凳腿：一个信念最初如何获得证成？每当你追问一个理由，就不得不退向更深的理由。现在是 9:12，因为钟这么显示。信赖钟，因为钟是可靠的。相信那一点，又因为……于是你一路后退，无处停步。古代怀疑论者精准地绘出了这一陷阱。每一条理由之链，他们论证道，终将落入三种令人不安的结局之一——阿格里帕三难困境：它无限延伸，或陷入循环，或止于某个你只能武断宣布的终点。

三个现代学派各自选择拥抱哪一个结局——而第四个学派索性改变了话题。

图示 · 回溯难题

阿格里帕三难困境——三条穷途，四条出路

你的信念为何有证成？对「……那又为何？」的每一个诚实回答，终将撞上三面高墙之一。

推理链条：信念：「现在是 9:12」→ 因为「那座钟」→ 因为「……那又为何？」

1. 无穷回溯：每一个理由都需要另一个理由，永无止境。
2. 循环：链条绕回自身，回到已经用过的某一点。
3. 武断止步：链条干脆停在某处基本承诺上，不再追问。

基础主义——接受第三种困境：有些信念是基本的，无需进一步支撑（原初经验、简单逻辑）。链条就此停住，却非武断。

融贯主义——拥抱循环，却使之成为一种美德：没有信念孤立存在；一个信念是否有证成，取决于它与整个信念网络契合得有多好。（这是系统思维的先声，第 9 日。）

无穷主义——勇敢的少数派：接受证成是一条永无尽头的理由之链，从不触底。

可靠主义——改换问题。一个信念只要由可靠的过程产生——良好的视觉、健全的记忆——就算有证成，无论你是否能道出一番辩护。这是外在主义：证成可以是你认知机制的事实，而非你头脑中的故事。

内在与外在的分裂，其重要性远超表象。内在主义者主张，证成必须是你经由反思即可触及的东西——「从内部」可得的理由。外在主义者（可靠主义的大本营）则认为，重要的是你的信念事实上以趋向真理的方式产生，无论你是否能够触及。请将这一张力存于心中：这正是旧日的扶手椅问题与关于大脑如何真正形成信念的新科学正面相撞之处。

前沿 · 2026

三条活跃前沿——以及一层前沿校准器

本课程的每一天都在研究前沿收束，每一项主张都标注着它究竟能承载多少分量。知识正处在一个迷人的交汇点上：哲学家、心理学家与神经科学家正从不同方向环绕着同一组问题。

前沿 01 [争议/炒作] [已确立]

「知识」直觉是普世的——抑或仅仅是西方的?

当整个学科的运行逻辑是「若仔细追问，几乎所有人都会说不」时，一个自然的忧虑是：哪些人？2001 年，实验哲学[4]的开山之作——温伯格、尼科尔斯与斯蒂奇——报告称盖梯尔直觉因文化而异，据说东亚参与者更愿意将「知识」的头衔授予那位幸运的认知者。若属实，这将是一枚重磅炸弹：哲学赖以运作的依赖直觉的方法论，看起来竟是褊狭的。

这一主张未能经受住复现检验。在「Gettier Across Cultures」（Noûs, 2017）中，马谢里、斯蒂奇、罗斯及其同事以近乎逐字转录的案例在巴西、印度、日本与美国进行了测试——却发现了相反的结果：在每一组人群中，人们都坚决拒绝将盖梯尔化的信念称为知识。另一项独立复现（Kim & Yuan）甚至以更大的东亚样本也未能复现最初的文化差异。当前最可信的解读是，可能存在一个普世的核心「民间认识论」，它本能地排斥基于运气的认知。我们将在第 149 日认识到一个更深层的教训：最耸动的发现，往往正是被审慎的复现悄然收回的那一个。

前沿 02 [已确立] [争议/炒作]

以刻度盘而非开关来度量信念：贝叶斯认识论

也许信念非此即无的设定从一开始就有问题。贝叶斯认识论主张，你真正的认识论状态是置信度——从 0 到 1 的连续信心刻度。此后，理性只需要两条规则：你的置信度必须服从概率法则（融贯性），且你必须随着证据到来以条件化方式[5]修正它们。

为何置信度必须服从概率法则？荷兰赌定理（Ramsey, 1926; de Finetti, 1937）提供了一个出人意料地具体的答案：如果你的置信度违背概率法则，一位精明的博彩商便能提供一组你各自视为公平的赌约，但两者结合，便可以保证在任何情况下你都会输钱。不融贯的置信度不仅是凌乱——它是可被利用的。

下方表格把同一个陷阱整理成三个基准案例：融贯、过度自信、信心不足。

仍属争议的是，分级的置信度究竟是取代了日常的是/否信念，还是仅仅与之并置。（彩票悖论向你提问：你有 99.9% 的把握自己的彩票会输——但你真的相信它会输吗？）我们将在第 4 日正式拾起这条线索。

置信度融贯表

若你给 S 与非-S 分配的置信度之和为 1.00，则这对置信度是融贯的。若总和大于 1.00，你会为两场不可能同时获胜的赌约过度付费。若总和小于 1.00，博彩商可以反向购买赌约，依然保证获利。

S 置信度	非-S 置信度	总和	结果
0.50	0.50	**1.00**	融贯
0.70	0.60	**1.30**	若你同时购买两场 1 美元赌约，必定损失 0.30
0.30	0.40	**0.70**	若博彩商同时从你手中购入两场赌约，必定损失 0.30

前沿 03 [线索] [争议/炒作]

信念从何而来？作为预测机器的大脑

哲学追问信念凭什么有证成；神经科学如今正在追问一个问题——作为一团有机组织，信念如何在其中形成？一个回答正在占据主流地位：大脑并非被动吸纳世界的海绵——它是一台不知疲倦的预测机器。依预测加工[6]观点（安迪·克拉克，Behavioral and Brain Sciences, 2013；雅各布·霍维，2013），大脑不断生成周遭环境的模型，预测它期望接收的感觉信号，并仅将预测误差——意外——向上传递。感知由此成为大脑持续运转的最佳猜测，被误差约束；用阿尼尔·塞思那句著名的说法，一场「受控的幻觉」。信念更新开始看起来像是神经元中实现的贝叶斯推理——即所谓的「贝叶斯大脑」，将前沿 02 与生物硬件统一起来。

卡尔·弗里斯顿以自由能原理[7]（Nature Reviews Neuroscience, 2010）将这一观念推向极致：生命系统之所以能持续存在，恰恰在于最小化一个量——「自由能」，也就是信息论[8]意义上与惊讶相邻的量——它将感知、行动乃至生物自组织编织进同一框架。我们先来给这一研究贴上校准的标签。预测编码确实解释了真实的感知现象，是一个严肃而多产的研究纲领——前景可期。但宏大的自由能原理，作为统摄心智与生命的单一法则，被广泛批评为过于笼统而难以证伪——更接近一个框架而非经检验的理论，因而争议重重。我们将在感知（第 119 日）与意识（第 123–126 日）中继续讨论这一问题，它的「自由

能」与我们将在第 33 日和第 83–85 日遇见的热力学如何遥相呼应。信息、能量、计算、涌现——我们五条线索中的四条，被编织进神经元安静的运算之中。

悬而未决的问题

真正尚未落定

六十年过去，对「知识是什么？」的诚实回答中，仍有一长串没有定论的问题：

- 知识究竟可否被分析？还是威廉森说得对，它是基石——一个我们用以解释其他事物、而非由他物派生而来的原初概念？
- 内在还是外在？证成是否要求你能经由反思触及的理由，抑或只需那些倾向于产出真理的认知机制？
- 一种货币还是两种？理性信念在根本上是分级的（置信度）、全有或全无的，抑或二者以某种方式调和？
- 是否真的存在一种普世的人类认识论——若有，是否是演化植入了那种「基于运气的认知不算数」的本能？（留待第 74 日的线索。）
- 大脑在严格意义上就是贝叶斯的吗，还是说「大脑在做推理」仅仅是一种从外部描述它的有用方式？
- 而那个将萦绕人工智能领域的问题：当像起草这一主题初稿的 AI 输出一个为真且证据充分的断言时，它是否知道任何东西——抑或它是终极的盖梯尔案例，正确的原因与事实毫无关联？（第 138–145 日。）

◆ 一日三句话

核心洞见

两千三百年来，知识看上去就像证成的真信念——直到盖梯尔用三页论文证明，你可以三者俱备却仍不算「知道」，因为你的理由与事实可能只是因运气相遇，而非真正相连。

最佳隐喻

那座一天只对两次的停钟——以及那位弓箭手，箭被吹离靶心，又被吹回正中：准确，却不适切。

悬置争议

修补方案能否找到第四条件（以及是哪一个）；知识是否是不可分析的基石；「信念」是否应当让位于分级的贝叶斯置信度——而「大脑是一台预测机器」这一断言，正构成一条真正的科学前沿。

今日线索 › 信息（置信度与贝叶斯大脑）· 能量（Friston 的自由能）· 计算（心智作为推理引擎）—— 并略微讨论了涌现与演化。

说明

1. 认识论指与知识、证据、证成和理性信念有关的问题。
2. 邻近可能世界指与现实只有小差异的可行情形，而不是极端遥远的幻想场景。
3. 事实性状态只有在其内容为真时才成立：你可以错误地相信，却不能错误地知道。
4. 实验哲学用问卷和实验来检验人们实际如何判断哲学案例。
5. 条件化是贝叶斯更新规则：把新证据视为已知，再在剩余可能性之间重新分配信心。
6. 预测加工是一类理论：大脑持续预测感觉输入，并在预测误差出现时更新模型。
7. 弗里斯顿的自由能原理认为，生命系统会行动以最小化预测误差或不确定性，并用一个名为自由能的形式量来表达。
8. 信息论用数学方式描述信息、不确定性与意外。
9. 蕴涵指一个命题在逻辑上保证另一个命题：前者为真时，后者不可能为假。
10. 自我驳斥指一个主张会破坏自身：若它为真，它自身成立的条件反而会使它为假或失去意义。
11. 标记是语言模型预测的文本单位，可以是完整词、词片段、标点或空格。
12. 说谎者红利指真实证据也能被说成伪造时，说谎者因此获得的好处。
13. 启发法是一种快速的经验规则：快而有用，但也可能系统性误导你。

14. 认识效用理论像决策论评价行动那样评价信念，只是评分标准通常是与知识相关的价值，例如准确性。
15. 这里的概率论指「概率主义」：理性的置信度应当遵守概率公理。
16. 加法性指互斥选项不能同时发生时，「其中之一发生」的概率等于各自概率之和。

来源

来源与延伸阅读

1. Gettier, E. L. (1963). "Is Justified True Belief Knowledge?" Analysis 23(6): 121–123. doi:10.1093/analys/23.6.121. doi.org/10.1093/analys/23.6.121

2. Ichikawa, J. J. & Steup, M. "The Analysis of Knowledge." Stanford Encyclopedia of Philosophy (rev. 2018). plato.stanford.edu/entries/knowledge-analysis —— JTB（Justified True Belief，证成的真信念）、盖梯尔案例、安全性/敏感性，以及知识优先转向。

3. "Gettier problem." Wikipedia (accessed 2026). en.wikipedia.org/wiki/Gettier_problem —— Russell（1948）、法上（约公元 770 年）与甘格沙（14 世纪）的先例。

4. Russell, B. (1948). Human Knowledge: Its Scope and Limits. London: Allen & Unwin. ——停钟案例（第 ~170–171 页）。

5. Goldman, A. (1976). "Discrimination and Perceptual Knowledge." Journal of Philosophy 73(20): 771–791. ——假谷仓案例；可靠主义。

6. Nozick, R. (1981). Philosophical Explanations. Harvard University Press. ——真相追踪 / 敏感性。

7. Sosa, E. (1999). "How to Defeat Opposition to Moore." Philosophical Perspectives 13: 141–153. ——安全性条件。 参见 Sosa (2007), A Virtue Epistemology（适切信念）。

8. Zagzebski, L. (1994). "The Inescapability of Gettier Problems." The Philosophical Quarterly 44(174): 65–73. ——击溃任何排除运气的修补方案的配方。

9. Williamson, T. (2000). Knowledge and Its Limits. Oxford University Press. overview ——知识优先认识论；知识作为最普遍的事实性心智状态。

10. Weinberg, J. M., Nichols, S. & Stich, S. (2001). "Normativity and Epistemic Intuitions." Philosophical Topics 29(1–2): 429–460. ——奠基性的跨文化实验哲学研究（后来受到争议）。

11. Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N. & Hashimoto, T. (2017). "Gettier Across Cultures." Noûs 51(3): 645–664. doi:10.1111/nous.12110. doi.org/10.1111/nous.12110

12. Kim, M. & Yuan, Y. (2015). "No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001." Episteme. philpapers.org/rec/KIMNCD

13. Weisberg, J. "Bayesian Epistemology." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/epistemology-bayesian ——置信度、条件化，以及荷兰赌论证（Ramsey 1926; de Finetti 1937）。

14. Clark, A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." Behavioral and Brain Sciences 36(3): 181–204. 参见 Clark, Surfing Uncertainty (OUP, 2016)。

15. Friston, K. (2010). "The free-energy principle: a unified brain theory?" Nature Reviews Neuroscience 11(2): 127–138. doi:10.1038/nrn2787. doi.org/10.1038/nrn2787

16. Hohwy, J. (2013). The Predictive Mind. Oxford University Press.

可选附录

附录：地图的其余部分

本节是可选的补充阅读；可以放心跳过，不会影响正文课程。

我们在正文里只停留于一个信念、一个临近正午的时刻。这片领域远比一座时钟广阔。

正文的任务很紧凑：取一个信念——现在是 9:12——然后追问它算不算知识。要做到这一点，它悄然倚靠在一摞从未检视的假设之上，并且径直走过整片学科疆域，连招呼也不打。认知是否要求确定性？那个宣称你什么都不知道的怀疑论者，真的能被回应吗？「知道」这个词从一句话到下一句，真的能保持不动吗？为什么知识比完成同样工作的真信念更有价值？还有那些与事实无关的认知呢——知道如何游泳、认得一张面孔、熟悉一座城市？本附录将走完地图的其余部分。这里不重复正文，而是沿着正文的边缘继续展开。

↩ 紧接自

第 1 日 – 什么是知识？ 在那里，我们搭好了三条腿的凳子（证成的真信念），看着盖梯尔用三页纸踢断一条腿，游览了失败的「第四条件」补丁，绘制了阿格里帕三难困境，并在三处前沿停下：「知识」的跨文化直觉测试、贝叶斯置信度，以及预测性大脑。把那天的两幅图像揣进口袋——停走的钟（因运气而正确，而非关联）和那位弓箭手，他的箭被吹偏，又落回靶心（命中，但并非出于能力）。二者都将在下文以不同面目再度登场。

◇ 我们跳过的七个房间

1. 盖梯尔之下的暗门——支撑盖梯尔案例的两个隐含假设，以及那条将你抛入怀疑论的逃生舱口（确定性）。
2. 门口的怀疑论者——梦境、恶魔、缸中之脑，以及 2020 年代的模拟升级。
3. 「知道」在滑动标尺上——银行案例：相同的证据，不同的利害关系，相反的裁决。
4. 我们真正追逐的运气——反运气认识论，它终于解释了补丁战争为何发生。
5. 为何认知胜过正确——《美诺篇》里的那条路，与知识的价值。
6. 我们忽略的认知类型——技艺之知，以及亲知之知。

7. 你所知的一切，几乎皆由他人告知——证言、分歧与认识论不正义。

§1 机关

每个盖梯尔案例之下的两扇暗门

在探索新房间之前，请先低头。盖梯尔那三页纸之所以有杀伤力，是因为地板内嵌了两扇暗门——两个如此自然的假设，正文从未在其上驻足。一旦命名它们，整个图景便会改观。

暗门一：证成可能出错。传统图景允许你基于证成相信某事，而结果却为假。史密斯有充分的理由相信「琼斯会得到这份工作」——老板这么说了——而它是假的。如果证成必须保证真理，那一步便不可能发生，案例甚至无法启动。暗门二：封闭性。人们假定证成（以及知识）可以跨越蕴涵[9]传递：如果你对相信某事拥有证成，那么你对其明显蕴涵之物也拥有证成。史密斯从「琼斯会得到它（并且有十枚硬币）」推出较弱的「获胜者有十枚硬币」——一个有效的推论——并将他的证成一路携带。敲掉任何一块木板，盖梯尔案例都会烟消云散。

这给了我们一条诱人的出路。把暗门一猛地关上：坚持真正的知识需要不会出错的证成——使错误在字面上不可能的理由。再也不会有盖梯尔案例。这是不可错论的梦想，它非常古老。笛卡尔在 1641 年寻找一个连恶魔也无法伪造的单一信念，并找到了唯一一个——即使存在一位在其他所有事情上都欺骗你的全能恶魔——仍然成立的信念：我思，故我在。你不可能被骗去错误地相信自己存在，因为欺骗需要一个你来承受欺骗。

麻烦在于恶魔出门时带走的东西。如果知识要求那种确定性，那么你就不知道你有双手，不知道太阳会升起，不知道桌对面的人是你的朋友而非仿生人——因为足够巧妙的欺骗可以伪造其中任何一项。选择确定性，代价就是怀疑论：门槛被设得如此之高，几乎无一能越过。彼得·昂格尔在 Ignorance（1975）中论证的正是这一点——严格使用的「知道」几乎不适用于任何事物，正如严格而言「平坦」不适用于任何真实表面。因此，不可错论并未消解问题；它只是用一个小谜题（某个奇怪的幸运信念）换来一个更大的谜题（你几乎一无所知）。这正是我们打开下一扇门的信号，那位怀疑论者已经在那里敲门了。

盖梯尔的另一个案例，一口气说完

正文使用了硬币案例。盖梯尔的第二个案例更直白地展示了暗门二。史密斯凭借充分证据相信「琼斯拥有一辆 Ford」。由此他有效地推出「琼斯拥有一辆 Ford，或者布朗在巴塞罗那」——一个他有理由相信的析取命题，因为其中一个分支为真即可使整个命题为真。但琼斯终究没有 Ford……而布朗，纯属侥幸，确实在巴塞罗那。这个析取命题为真、有证成、被相信——且显然不是知识。封闭性携带了证成；运气提供了真理。结构相同，只是包装更复杂。

§2 最大的遗漏

门口的怀疑论者

西方认识论有一位反复出现、拒绝离开的房客：那个宣称你对自己的心智之外的世界一无所知的形象。正文把那扇门紧闭。打开它，因为每一种现代知识理论都部分地建立在应对门外那位不速之客的基础上。

怀疑论者的工具是思想实验，层层加码。首先是梦：此刻，你怎么知道你没有在睡觉？梦境从内部看完全真实；你以前就被骗过。（道家庄子，约公元前 300 年，梦见自己化为蝴蝶，醒来时不确定自己是一个梦见了蝴蝶的人，还是一只此刻正在梦见人的蝴蝶——佛教论师法上在正文中重新揭开的正是同一道伤口，再次证明心智独立地一再绊倒于此。）笛卡尔将设想推进到一位一心要在一切事上欺骗你的邪恶恶魔。到了二十世纪，思想实验换上了新的假想：你可能是一只缸中之脑，神经连接到一台计算机，它向你输送的正是你此刻正在拥有的体验（希拉里·普特南，Reason, Truth and History, 1981）。你无法从内部分辨。难题正在于此。

展开来说，怀疑论者的论证简洁得近乎冷酷——而且它运转的正是来自 §1 的封闭性原则：

> *(1) 你并不知道自己不是一只被输送手部体验的、没有手的缸中之脑。*
> *(2) 如果你知道你有双手，那么（既然有双手蕴涵不是无手的缸中之脑）你也就知道自己不是那样的缸中之脑。*
> *(3) 所以你不知道你有双手。*

每一行看起来都合理；合在一起，它们似乎证明你对外部世界一无所知。

下方表格把每一种出路对应到它拒绝的论证行、代表性立场与代价。

怀疑论者的三段论：四种出路

招式	拒绝的行	代表性观点	代价
接受全部三点	无	怀疑论	你不知道你有双手，对外部世界也所知甚少。
拒绝 P1	你不知道自己不是缸中之脑	摩尔的常识回应	可能感觉像是在坚持而非解释。
拒绝 P2	封闭性	德雷茨克 / 诺齐克的相关替代项理论	封闭性在直觉上根深蒂固，在其他地方也很有用。
改变标准	「知道」的固定含义	语境主义	怀疑论者在研讨室里获胜；普通说话者在日常生活中获胜。

这些回应值得一一交代。G. E. 摩尔（1939）只是将论证反向运行：我更确信这里有一只手（举起它）胜过怀疑论者提供的任何精巧前提——因此，如果那些前提蕴涵我不知道这一点，问题就在前提。大胆，却令人难以反驳。弗雷德·德雷茨克（1970）与罗伯特·诺齐克（1981）采取了更精细的路线：否定封闭性。在德雷茨克的相关替代项观点看来，要知道某事，你只需排除你犯错方式中相关的那些，而非每一种怪异的可能。在动物园里，你知道那动物是斑马——你已经排除了「它是马」、「它是山羊」——尽管你尚未排除「它是一头被巧妙漆成斑马样子的骡子」，因为在这一语境中，那不是实际需要认真对待的可能。知识不会自动沿每一个蕴涵传递。代价不小：封闭性是直觉性的，放弃它会牵一发而动全身。语境主义（我们下一节）提供了折中方案：也许怀疑论者和摩尔都是对的，因为「知道」在怀疑论者的研讨室里意味着比普通生活中更严格的东西。

2020 年代的升级：我们是否身处模拟之中？

缸中之脑在当代换了一种形式。尼克·博斯特罗姆的模拟论证（Philosophical Quarterly, 2003）提出了一个审慎的概率性论证：以下三件事至少有一件为真——文明几乎从未达到运行祖先模拟的技术；或者它们达到了但选择不运行；或者我们几乎肯定生活在其中

一个之中。大卫·查默斯在 Reality+（2022）中迈出了下一步，接受了大多数人不愿接受的结论：他论证我们无法知道自己没有被模拟，并应当赋予这一可能性真实的概率——但这并非一场灾难，因为「虚拟现实是真正的现实。」在他所谓的模拟实在论看来，一棵模拟的树是一个真正的数字对象，而非幻觉；如果你一直生活在一个完美的模拟中，你的信念「那是一棵树」是真的，只不过是以硅的形式实现。怀疑论者假定虚假的世界意味着虚假的信念；查默斯否认这种联系。

在继续之前，先把两个标签贴准。模拟假说——即我们事实上被模拟了——就其现状而言，是不可检验的形而上学，而非科学：不存在公认的观察能够证实或反驳它，这使它落在了我们明日将画出的分界线的错误一侧。[争议/炒作] 尽管如此，其哲学回报是真实的：它廓清了「真实」与「知道」到底意味着什么。还有一个著名回应把论证往相反方向推进。普特南论证「我是一只缸中之脑」是自我驳斥的[10]：你的词语之所以有意义，仅在于你的因果历史，因此一个终身缸中之脑的词语「缸」不可能指涉真正的缸（它从未与真正的缸发生过因果联系）——这意味着，如果你是一只缸中之脑，你的句子「我是一只缸中之脑」将得出假的结论。这是否成立仍在争论中，而这条线索将直接引向人工智能篇章：当一个仅接受文本训练的系统输出「巴黎在法国」时，它知道这一点吗——还是说它是所有缸中之脑中最纯粹的一个，其词语从未触碰过世界？将这个问题留到第 138–145 日。

§ 3 移动的目标

「知道」是一把滑动的标尺

这里有一种正文从未考虑过的可能性：或许六十年追寻「知道」的完美定义之所以失败，是因为这个词从未指向一个固定的标准。来看基思·德罗斯（Philosophy and Phenomenological Research, 1992）提出的一对案例，它们催生了上千篇论文——银行案例。

那是周五。你开车经过银行，看到周六排起长队，决定明天再来。你的配偶问它周六是否开门。低利害版本：没什么大不了的；你说：「是的，我知道它周六开门——我两个周六前还来过的。」那听起来没错。你知道的。高利害版本：有一张支票必须在周一前存入，否则你的抵押贷款会跳票、你会失去房子，而你的配偶合理地指出，银行确实会改变营业时间。现在，完全相同的句子——「我知道它周六开门」——在你口中凝结了。「嗯……我最好还是进去确认一下。」同一个人，同样的记忆，同样的证据，同一天。只有利害关系（以及是否有人提出了出错的可能性）改变了。然而知识似乎时有时无。

下方表格比较低利害、高利害，以及有人提出出错可能时的三个版本。

银行案例：利害关系表

案例	证据	利害关系	自然裁决	测试什么
低利害	你两个周六前去过那里。	一件小事。	「我知道它在营业。」	普通标准容易达到。
高利害	同样的记忆。	抵押贷款截止日期。	「我最好确认一下。」	实际利害是否影响知识。
提出出错可能	同样的记忆加上一个活跃的怀疑。	任何严重后果。	知识声称被削弱。	语境改变的是词语还是认知者的状态。

三个阵营，对同一数据的三种诊断。语境主义（德罗斯；大卫·刘易斯，"Elusive Knowledge," 1996；斯图尔特·科恩，1988）将转变定位在词语上：「知道」就像「高」或「这里」一样——对语境敏感。提高利害关系或提及错误，会提升一个信念必须达到的标准，才能使「S 知道」这句话为真。两种说法在各自的语境中都对。怀疑论者在研讨室里甚至也是对的——他只是把标准抬到了天际。实践侵入（杰森·斯坦利，Knowledge and Practical Interests, 2005；范特尔与麦格拉思；约翰·霍桑，Knowledge and Lotteries, 2004）将转变定位在认知者身上：你真正知道什么取决于你实际面临多大风险，因为知识理应是你据以行动的依据。高利害确实能剥夺你在事情无关紧要时本可拥有的知识——一个令人吃惊的观点，因为它让实践压力「侵入」了一个据称纯粹事实性的状态。不变主义（传统的坚守者）死守阵地：「知道」的含义是固定的，标准不会移动，你的两个裁决之一根本就是错的——你要么始终知道，要么从未知道，利害关系改变的只是你愿意这样说的程度。[已确立] [争议/炒作] 数据是坚实的；其解释却是该领域最活跃的争论焦点之一。

§ 4 补丁背后的模式

我们真正追逐的运气

回到正文中的补丁战争——无假前提、敏感性、安全性、德性。它们看起来像一堆精巧的补丁，每一种都遇到了更棘手的反例。退后一步看，它们便骤然清晰：每一个都在追逐同一个幽灵。邓肯·普里查德在 Epistemic Luck（Oxford, 2005）中给了它一个精确的名

字。知识的敌人是他所称的真理运气（veritic luck）的特定物种：你的信念在实际世界中为真，但在几乎所有邻近的可能性中，你会相信同样的事，却是错的。真理与你的信念只是偶然地重合。

这是「安全性」观念的深层内容，值得单独说明。将实际世界想象为一个点，被邻近的可能世界环绕——事物本可能如何的小小现实变体。当一种信念在整个邻近区域保持为真时，它是安全的（知识级别），而当轻轻一推就将它翻转为假时，它是不安全的（单纯幸运）。

下方表格比较三个邻近世界案例，以及各自得到的安全性裁决。

安全与幸运：邻近世界案例

情境	实际世界	邻近世界	裁决
正常运行的钟	你的信念为真。	微小变化仍然让你正确。	安全：知识级别。
停走的钟	你的信念在 9:12 为真。	早一分钟或晚一分钟，同样的信念为假。	不安全：真理运气。
假谷仓之国	你看见了唯一一座真谷仓。	大多数邻近的一瞥都会落在假谷仓外观上。	不安全：环境运气。

这幅图能把前面的混乱重新组织起来。停走的钟彻底失败——左右一分钟你就错了，因此邻近区域是一片红色的海洋。假谷仓之国则更微妙：你看着的谷仓确实在那里（核心是绿色的），但你被假谷仓外观包围，因此往任何方向瞥上一百米都会骗到你——红色邻近区域，没有知识，即便拥有证成的真信念且无假前提。那些补丁之所以不断失效，是因为每一个都试图用略有不同的尺度去捕捉「绿色邻近区域」，而运气不断找到缝隙。

既然我们有了框架，再来看正文未提及的另外两个补丁。可废止性理论（莱勒与帕克森，1969）说知识是未被击败的证成真信念：外部必须不存在某种真的事实，一旦你得知它，就会消解你的证成。它优雅地处理了许多案例——直到「误导性击败者」的出现，即存在某个真实却具有误导性的事实，它不应该剥夺你的知识，技术上却做到了，迫使

人们做出越来越精细的区分。再往前追溯，因果理论（戈德曼，1967，在他转向可靠主义之前）要求事实引起你的信念——没有因果链，就没有知识。对知觉而言很美；对数学却是致命的，因为数字 7 和毕达哥拉斯定理不会引起任何东西（保罗·贝纳塞拉夫在 1973 年正是提出了这个「通道问题」）。你无法与抽象对象握手。

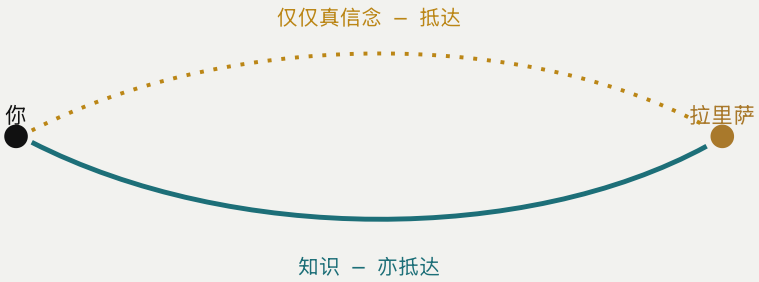
还有一个正文只轻轻带过、却足以撬开可靠主义的难题：一般性问题（科尼与费尔德曼，1998）。可靠主义说，一种信念如果由可靠的过程产生，它就是证成的——但究竟是哪一个过程？你的「现在是 9:12」的信念，可以归因于「看钟」，也可以归因于「看那座钟」，或「在昏暗光线下使用视力」，或「在周二依赖仪器」——每一种描述都同样真实，每一种的可靠性分数都不同。选择类型，你就选择了裁决。以原则性的方式确定「正确」的粒度，已被证明极为困难。

普里查德落脚于何处？于反运气德性认识论：知识需要两项缺一不可的条件，因为二者针对的是不同的失败方式。你需要安全性（绿色的邻近区域——没有真理运气）并且你需要适切性（信念之所以为真，是通过你自己的能力——正文中那位弓箭手的技艺）。单独任何一个都不够：停走的钟缺乏安全性；假谷仓之国则表明，即使你在局部运用了真实的能力，环境运气也可能击败你。它并非一个整洁的三字公式——而到如今，这或许就是教训。知识也许正是这样一种东西：需要两重保障，一重关乎你，一重关乎你的世界。

§5 问题之下的问题

为什么知道比仅仅正确更有价值？

从「什么是知识？」退一步，来到柏拉图最早提出、却无人能完整回答的问题：我们为何在意？ 如果一个真信念足以完成任务，知识额外的那些机制究竟为你换来了什么？柏拉图在《美诺篇》（Meno，约公元前 380 年）中将其表述为一个旅人的问题。假设你想步行前往拉里萨（Larissa）城。一个知道路的人会把你带到那里。但一个仅仅对路拥有真信念的人也会——他从未去过，只是碰巧正确。就抵达目的地而言，二者毫无差别。那么为何整个传统都将知识置于真信念之上？这就是价值问题，它是一个关乎根基的问题：一种知识理论如果不能说明知识为何更好，可以说就错失了这一概念的要义。



如果两条路都通向拉里萨，第二条价值何在？

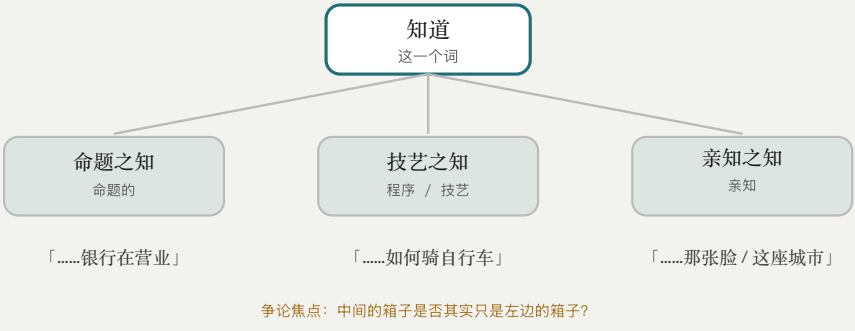
价值问题也精准地刺向一种理论。它被称为淹没问题（琳达·扎格泽布斯基，2003）。可靠主义说，知识是来自可靠过程的真信念。但追问的是，可靠性在价值上究竟增添了什么。可靠性之所以好，仅在于它倾向于产生真理。因此一旦你已经拥有了真理，这个特定的真理碰巧来自一个可靠的来源，又能增添什么？扎格泽布斯基的家常类比是：一杯好咖啡并不会因为它出自一台可靠的咖啡机就更好喝——一台不可靠的机器也可能碰巧产出同样好的一杯。使之为好的特征（美味 / 真理）已然在场；来源的可靠性被淹没了，没有增添任何价值。如果这是对的，可靠主义就无法解释为什么知识胜过幸运的真信念——而这正是一种知识理论最需要说明的东西。

这正是德性认识论证明自身价值的所在，也是那位弓箭手最终兑现的地方。它的回答是：知识之所以有价值，并非因为它是真信念外加一层包装；而在于它是一种成就——一种属于你的成功，一种凭借你自身能力而实现的成功。成就带有一种幸运的成功永远不具备的价值，就像你真正瞄准射中的靶心，总比被一阵幸运的风吹进去的箭更有价值，即便二者落在同一个靶点上。通过你自己的认知技艺达到的真信念是一种认知成就；幸运的真信念则不是。那额外的价值不在结果中，而在抵达的过程中。那条你能再次找到的通往拉里萨的路，比你误打误撞撞上的那条更有价值，即便某一天两者都抵达终点。

§ 6 其他的知

我们忽略的认知类型

迄今为止的一切——整节正文——都是关于命题知识，即「知道某事为真」：知道那是 9:12，那是琼斯得到了工作。但看看日常语言中「知道」有多少完全不是这回事。你知道如何骑自行车。你认识你母亲的脸。你知道里斯本。这些都不是事实的堆砌，而哲学家们已为它们之间的关系争论了一个世纪。



一个英语动词，至少三种不同的与世界的关系。

技艺之知。吉尔伯特·赖尔在 The Concept of Mind（1949）中坚持认为，知道如何做某事并不是知道一组事实。一位杰出的自行车手可能无法陈述任何一条平衡法则；一个熟记了关于自行车的一切事实的人可能在第一次尝试时就摔倒。问题还不止于此：赖尔论证说，将技艺还原为事实会触发无限倒退：如果每一个熟练的行动都要求事先知道描述该规则的命题，那么你就需要运用那条规则的技艺，而那又需要另一条规则，永无止境。因此技艺必须是独立于命题的另一种知。转折在于：杰森·斯坦利与蒂莫西·威廉森在 "Knowing How"（2001）中回击，提出理智主义——主张技艺之知终究只是命题之知的一种（知道某种骑车的方式，并知道它是一种骑车方式），只是披着不同的语法形式。技艺是否可还原为命题，确实尚未有定论。［争议/炒作］

亲知之知。伯特兰·罗素（1911）又划出一道界线：在亲知之知——你对所见的一抹红色、所感的一种疼痛、所注视的一张面孔的直接把握——与描述之知之间，即你所知的关于你从未直接遭遇过的事实的关于之物（「第一个站在月球上的人」，你只知道他是满足该描述的那个人）。你可以对俾斯麦知道关于他的大量事实，却从未认识他；你知道红色，其方式是世界上最伟大的盲人物理学家所不知道的，尽管他知道关于波长的每一个事实。那道缝隙——关于体验的事实与体验本身之间的缝隙——是整门课程中最难问题的一颗安静的种子，那颗种子在第 123 日等待：看见红色的体验究竟是什么样。

§7 社会转向

你所知的一切，几乎皆由他人告知

正文与大多数传统认识论一样，想象一颗孤独的心智面对世界——一个人，一座钟。但盘点一下你实际所知的东西：地球约有 45 亿年历史。南极洲存在。你自己的出生日期。水的沸点。你几乎没有亲手验证过其中任何一项；你是从老师、书本、父母、仪器、陌

生人那里获知的。证言构成了任何人知识的压倒性主体——而几个世纪以来，认识论却将其当作事后之想。

核心问题在于，信任证言是你必须事先取得资格才能做的事，还是你默认就享有的权利。大卫·休谟（1748）采取了苛刻的路线：证言的好坏只取决于你自己积累的归纳可靠性记录——即考察证言在何时被证明可靠——它还原为你个人收集的证据。托马斯·里德（1764）觉得这荒谬至极：没有哪个孩子能在信任任何人之前先自行建立一份可靠性记录，而事实上，我们天生就带有一条「轻信原则」，一种默认地相信他人所言的倾向，正如我们天生就信任自己的感官一样。在里德的反还原主义观点看来，证言是一种基本的知识来源，而非派生的——而且它必须是，否则知识就无从在社会性动物中产生。现代学界大多同意某种默认信任是不可避免的；争论在于这种信任应有多少，以及它何时会被击败。

从这个领域分出的两个新分支，在 2026 年都极为重要。第一个是分歧。当你视某人为认识论上的同侪——和你一样聪明、一样知情、一样谨慎——面对同样的证据却得出相反结论时，你该怎么办？调和主义或「同等权重」观点（亚当·埃尔加，Noûs, 2007；大卫·克里斯滕森，2007）主张你应当实质性地向对方的立场靠拢：固守原地意味着在缺乏独立理由的情况下声称你才是对的、对方才是错的。坚定观点回答说，有时你可以理性地守住阵地，因为你自己已经做出的推理也是证据。这听起来抽象，但你很快会注意到：这其实就是回声室、专家共识以及信息源相互冲突时我们该如何判断背后的认识论。[争议/炒作]

第二个分支则更为尖锐：认识论不正义，由米兰达·弗里克命名（Epistemic Injustice: Power and the Ethics of Knowing, 2007）。因为如此多的认知活动依赖于证言，谁被相信便不只是认识论问题，而是一个伦理问题。弗里克区分了两种不公。证言不正义：说话者的话语得到的信任低于其应得，源于对其身份的偏见——病人的疼痛被漠视，证人因口音或性别而不被采信。诠释不正义：更微妙，也更深层——一个人甚至无法为自己的经验赋予意义——无论对自己还是对他人——因为周遭文化尚未发展出相应的概念（以她的例子来说：我们现在称之为性骚扰的经验，曾由那些没有词语来命名它的人所承受，因此他们甚至无法说出这种伤害是什么）。事实证明，知识是有政治性的：理解的工具分配不均，而这种不均本身就可以是一种不正义。

功能优先的逃生舱口

有一种激进的方式可以终结这整段 180 页的定义追寻，它把社会转向的线索重新穿回起点。爱德华·克雷格在 Knowledge and the State of Nature（1990）中提出：停止追问「知识是什么？」，转而追问「这个概念是为了什么——像我们这样的生物为什么会发明它？」他的回答是：一个社会性的、使用语言的物种迫切需要标记可靠的信息来源——标明谁的话你可以据以行动。「知识」就是我们在演化中发展出来、别在可靠信息来源上的标签。这立刻解释了那些分析所苦苦挣扎的东西：为什么知识必须为真（一条假的提示毫无价值），为什么运气会使你丧失资格（你下次无法依赖侥幸），以及为什么我们在意这一切（在一个大多数你需要知道的东西都必须从他人那里获取的世界中生存下去）。它与威廉森的「停止试图定义它」相呼应，并且兑现了正文的开放问题——演化是否植入了一种「基于运气的认知不算数」的本能？克雷格的回答本质上是：是的，而且理由如下。

§8 形式前沿

贝叶斯之外的两处前沿

正文的形式前沿是贝叶斯置信度。另外两个形式化思路也值得在地图上占有一席之地，因为两者都不断挑战日常直觉，且都直接通向计算机科学与 AI。

认知逻辑。雅科·欣蒂卡在 Knowledge and Belief（1962）中将「知道」视为一个可以像「必然地」一样进行推理的形式算子——从而开创了认知逻辑，如今已成为计算机科学的主力工具（推理分布式智能体与 AI 系统「知道」什么）。它立刻带出深层难题。KK 原则：如果你知道 p，你是否因此知道你知道 p？这很诱人，但威廉森（来自正文）论证它是假的——你可以知道某事，却不处于知道你知道它的位置上，因为知识有模糊的边界。以及逻辑全知：干净的逻辑意味着如果你知道某些公理，你就知道它们的每一个逻辑后果——那将使每一位数学家瞬间意识到每一个定理。对于真实的、有限的心智而言，这显然为假，并且是为实际推理者（以及机器）建模时的一个核心难题。

序言悖论。正文中彩票悖论的伴侣，而且可以说更为棘手。你写了一本冗长而审慎的书。对于书中的每一个主张，你都检查了工作并理性地相信它为真。然而你也真诚地在序言中写道：「无疑仍有错误存在，且皆由我一人负责」——因为你知道在数百个主张中，你几乎肯定在某处疏漏了。因此你理性地相信每一个单独的主张，并且也理性地相信其中至少有一个为假（大卫·马金森，"The Paradox of the Preface," 1965）。这些不可能同时为真。它直接回应了正文留下的开放问题：普通的非此即彼信念对合取不封闭——相信许多事物中的每一个，并不等于你有理由相信它们合起来全都为真——这正是该领域持续从是/否信念滑向分级置信度的又一个原因。再一次，分级拨盘能表达开关表达不了的东西。

◆ 三句话总结本附录

大观念

正文让知识看起来像一个整洁的谜题——找到第四个条件——但它实际上更像一组相互牵连的问题：是否要求确定性（以及由此招来的怀疑论）、「知道」在利害关系变化时是否还能保持不变、知识相对于单纯真信念的价值何在，以及几乎所有知识都来自他人这一事实。

最佳新类比

邻近可能世界：知识是一种在相近情形中仍保持为真、因而安全的信念；运气则是一种现实稍微变化就会出错、因而不安全的信念——而且那条你能再次找到的通往拉里萨的路，比你误打误撞上的那条更有价值，即便二者都抵达了终点。

正在进行的争论

银行案例中裁决为什么会改变——是语境移动了「知道」这个词（语境主义），是利害关系移动了认知者所知的东西（实践侵入），还是两者皆非（不变主义）——这是该领域最活跃的争论焦点之一，同时并列的还有封闭性能否被否定，以及技艺之知是否只是命题之知的伪装。

此处线索 › 信息（证言 & 知识的社会传递；序言/置信度）· 计算（认知逻辑；世界的模态「邻近区域」）· 演化（克雷格：知识的概念作为一种为社会物种而设的可靠信息来源探测器）——拾起我们整段 180 天都在追踪的同五条线索。

开放问题

本附录留下的未决问题

- 确定性与否？ 不可错论者是否正确：真正的知识需要无错的理由（从而招来怀疑论）——还是可能出错的知识才是唯一值得想要的类型？
- 否定封闭性能否不带来灾难？ 德雷茨克与诺齐克通过放弃它来阻挡怀疑论者；它在其他地方造成的代价仍有争议。

- 「知道」的标准会变吗？ 对语境敏感、对利害关系敏感，还是固定的——如果它移动，究竟是什么在移动，词语还是世界？
- 知识的价值究竟能否被解释，还是每一种说明都让知识看起来不比幸运的真信念更好？
- 技艺之知是否只是伪装的命题之知，还是它对世界有着自身不可还原的把握方式？
- 证言是基本的还是需要先取得资格？进一步说，当一位同侪分歧时，你是否真的必须与他们半路相逢？
- 而功能优先的解释：如果知识的概念存在是为了标记可靠的信息来源，那是否消解了分析计划，还是只是把问题移到别处？

来源

来源与延伸阅读

古典著作按原始日期引用；所有版本均为标准且广泛可得的版本。经核实的二次文献锚点与参考条目已附链接。

1. Descartes, R. (1641). Meditations on First Philosophy. ——方法论怀疑、邪恶恶魔，以及作为唯一不可怀疑之点的我思。
2. Unger, P. (1975). Ignorance: A Case for Scepticism. Oxford University Press. ——不可错论被推向其怀疑论结论（「知道」如同「平坦」一样，几乎不适用于任何事物）。
3. Moore, G. E. (1939). "Proof of an External World." Proceedings of the British Academy 25: 273–300. ——「这里有一只手」：将怀疑论论证反向运行。
4. Dretske, F. (1970). "Epistemic Operators." Journal of Philosophy 67(24): 1007–1023. ——否定封闭性；相关替代项观点；斑马/漆骡案例。
5. Nozick, R. (1981). Philosophical Explanations. Harvard University Press. ——敏感性 / 真值追踪，及其对封闭性的否定。
6. Putnam, H. (1981). Reason, Truth and History. Cambridge University Press. ——缸中之脑，以及语义外在论论证「我是 BIV」是自我驳斥的。
7. Bostrom, N. (2003). "Are You Living in a Computer Simulation?" Philosophical Quarterly 53(211): 243–255. simulation-argument.com
8. Chalmers, D. J. (2022). Reality+: Virtual Worlds and the Problems of Philosophy. W. W. Norton / Allen Lane. ——「虚拟现实是真正的现实」；模拟实在论。 consc.net/reality

9. DeRose, K. (1992). "Contextualism and Knowledge Attributions." Philosophy and Phenomenological Research 52(4): 913–929. ——银行案例。 另见 DeRose (1995), "Solving the Skeptical Puzzle," Philosophical Review 104(1): 1–52。

10. Lewis, D. (1996). "Elusive Knowledge." Australasian Journal of Philosophy 74(4): 549–567. ——语境主义与注意规则。

11. Cohen, S. (1988). "How to Be a Fallibilist." Philosophical Perspectives 2: 91–123. ——机场案例。

12. Stanley, J. (2005). Knowledge and Practical Interests. Oxford University Press. ——实践侵入 / 利益相对不变主义。 另见 Hawthorne, J. (2004), Knowledge and Lotteries (OUP); Fantl, J. & McGrath, M. (2009), Knowledge in an Uncertain World (OUP).

13. Pritchard, D. (2005). Epistemic Luck. Oxford University Press. ——运气的模态说明；真理运气；安全性条件；后来的反运气德性认识论。 概述： IEP, "Epistemic Luck."

14. Lehrer, K. & Paxson, T. (1969). "Knowledge: Undefeated Justified True Belief." Journal of Philosophy 66(8): 225–237. ——可废止性分析。

15. Goldman, A. (1967). "A Causal Theory of Knowing." Journal of Philosophy 64(12): 357–372. ——以及 Benacerraf, P. (1973), "Mathematical Truth," J. Phil. 70(19): 661–679, 关于它为何对抽象对象失效。

16. Conee, E. & Feldman, R. (1998). "The Generality Problem for Reliabilism." Philosophical Studies 89(1): 1–29.

17. Plato. Meno (~380 BCE). ——通往拉里萨的路；价值问题（知识与真信念）。

18. Zagzebski, L. (2003). "The Search for the Source of Epistemic Good." Metaphilosophy 34(1–2): 12–28. ——淹没问题。 另见 Kvanvig, J. (2003), The Value of Knowledge and the Pursuit of Understanding (Cambridge UP).

19. Ryle, G. (1949). The Concept of Mind. University of Chicago Press. ——技艺之知与命题之知；规则的无限倒退。

20. Stanley, J. & Williamson, T. (2001). "Knowing How." Journal of Philosophy 98(8): 411–444. ——理智主义：技艺之知作为命题之知的一种。

21. Russell, B. (1910–11). "Knowledge by Acquaintance and Knowledge by Description." Proceedings of the Aristotelian Society 11: 108–128.

22. Hume, D. (1748). An Enquiry Concerning Human Understanding, §X. ——证言的还原主义观点。 Reid, T. (1764). An Inquiry into the Human Mind on the Principles of Common Sense. ——证言作为基本来源（反还原主义）。

23. Elga, A. (2007). "Reflection and Disagreement." Noûs 41(3): 478–502. doi:10.1111/j.1468-0068.2007.00656.x. 以及 Christensen, D. (2007), "Epistemology of Disagreement: The Good News," Philosophical Review 116(2): 187–217.

24. Fricker, M. (2007). Epistic Injustice: Power and the Ethics of Knowing. Oxford University Press. ——证言不正义与诠释不正义。

25. Craig, E. (1990). Knowledge and the State of Nature: An Essay in Conceptual Synthesis. Oxford University Press. ——功能优先 / 良好信息来源的概念谱系学。

26. Hintikka, J. (1962). Knowledge and Belief: An Introduction to the Logic of the Two Notions. Cornell University Press. ——认知逻辑；KK 原则；逻辑全知。

27. Makinson, D. C. (1965). "The Paradox of the Preface." Analysis 25(6): 205–207.

28. 参考综述： Stanford Encyclopedia of Philosophy——"Skepticism," "Epistemic Contextualism," "The Value of Knowledge," "Epistemological Problems of Testimony," "Epistemic Injustice."

可选附录

附录：地图的边缘

本节是可选的补充阅读；可以放心跳过，不会影响正文课程。

这张地图的海岸线仍在绘制。材料新近、利害攸关，每一行结论都尚不足以稳立。

第一个附录绘制了定居的腹地——怀疑论、语境中的「知道」、知识的价值、社会网络——这些领地在几十年前甚至几百年前就已被绘入地图。而这一个则驶向了边缘，制图师们仍在争论海岸线究竟在哪里。下文所有内容都是自 2020 年以来的同行评议工作，它们可能真正重绘我们所谓的「知识」——正因为它是如此之新，所以需要严加过滤炒作。这里没有什么是可以确定入账的。每一个前沿都已催生了各自的反对文献，每一项主张都带有标签：[已确立] [线索] [争议/炒作] 读它就像读一份仍在行进中的远征队的电报——令人兴奋、零碎，且随时可能被下一艘归航的船只修订。

↩ 序列中的第三篇

第 1 日——什么是知识？搭建了凳子，并目睹了盖梯尔踢掉了一条腿。附录 I——「地图的其余部分」巡视了定居的省份：怀疑论、语境主义、反运气认识论、价值问题、证言以及认识论上的不正义。本篇则是同一片大陆上仍在扩张的前沿。如果说附录 I 中转向社会的篇章（证言、分歧、谁被相信）描述了「从他人处获知」的结构，那么这里的几个前沿则描述了当该结构遭到蓄意攻击时会发生什么——被操纵者、被机器、被信息流攻击。

◈ 海岸线正在移动的六处

1. 探究转向——认识论从信念状态转向探究行为，并发现旧规则与新规则存在冲突。
2. 先于信念的知识——认知科学翻转了布局：也许表征知识比表征信念更为基础。
3. 机器知道些什么吗——还是只是在扯淡？——大型语言模型的哲学，以及一个刻意粗鲁的诊断。
4. 认识论后盾的崩塌——深度伪造悄然移除了支撑证言的一项长期支柱。
5. 敌对认识论——回声室、人造的清晰感，以及信任如何可能被武器化。

6. 准确性优先——一种形式化的重构，通过真理而非金钱重新推导出第 1 日的荷兰赌论证。

§1 探究转向

认识论遗忘了探究

[线索][争议/炒作]

这是一个奇怪的疏忽，一旦你看到了，就再也无法忽视。一个世纪以来，认识论几乎完全是关于状态的理论——关于信念、证成、知识的状态：这些都是头脑完成思考后的快照。它极少谈及产生这些状态的活动——即探究——这一芜杂的过程：提出问题、决定收集哪些证据、知道何时停止。Jane Friedman 为这缺失的半壁命名，并在该领域投下了一枚炸弹。她将探究规范称为探究（zetetic）规范（源自希腊语 zētein，意为寻求），她在其里程碑式的论文 "The Epistemic and the Zetetic"（The Philosophical Review，2020）中提出了一个真正具有颠覆性的论点：探究的规范与信念的规范不仅仅是分离的——它们在积极地冲突。

其核心引擎是一个显而易见、听起来像陈词滥调的原则——探究工具性原则（Zetetic Instrumental Principle）：如果你想弄清楚一个问题的答案，你就应该采取必要的手段去弄清楚。现在看看它如何与基石性的认识论规范发生碰撞——即证据主义者的命令：你可以相信你的证据已经支持的任何内容。假设你正试图清点街对面建筑的窗户。良好的探究会说：专注，去数窗户，不要分心。但在流逝的每一瞬间，你的感官都向你提供了足够的证据，使你形成并被允许相信一千个无关紧要的真理——那辆车的颜色、街角的人数、云朵的形状。信念规范许可所有这些内容。探究规范则告诉你要忽略所有这些，去数窗户。顺从其中一个，你就会违背另一个。图表将这种挤压具体化了。

Friedman 的张力：良好的探究与被许可的信念朝相反方向拉扯。

为什么这在研讨会之外也很重要？因为它表明认识论一直在研究错误的单元。如果信念规范与探究规范真正发生冲突，那么仅建立在信念之上的理论就是不完整的——甚至可能是本末倒置的。激进的提议，即「探究转向」，认为所有认识论规范最终都是探究规范（悬置判断变成了指向问题的态度；相信一个答案是关闭一个问题的方式）。该领域尚未全盘接受这一点——而这正是诚实的部分。Arianna Falbo（"Should epistemology take the zetetic turn?"，Philosophical Studies，2023）等人认为探究规范实际上是实践性的，而非独特的认识论规范，并且纯粹的探究认识论无法解释为什么有些信念是不理性的，即使相信它们会有助于你的探究。因此：这个谜题现在在整个领域都被严肃对待；而探究吞噬一切的大论题则是一场真正的、未解决的争斗。无论如何，「什么是知识？」这个问题正悄然被重构为「什么是良好的探究？」——这一重构一直延伸到第 2 日，在那里，科学方法正是一套用于集体探究的规范。

§2 认知科学转向

如果知识在信念之前呢？

[线索][争议/炒作]

第 1 日将知识视为从信念中构建出来的东西：获取一个信念，加上真理性，加上证成，筛掉运气。我们遇到的几乎每一种理论都假设信念是原材料，而知识是成品。由 Jonathan Phillips 和 Joshua Knobe 领导的一个大型跨学科团队在 Behavioral and Brain Sciences 上发表了一篇靶子文章——「Knowledge before belief」（2021）——认为，就人类（以及动物）心智的实际运作方式而言，这可能完全是颠倒的。

心理学中的标准叙事是，我们的「心理理论」（theory of mind）——我们模拟他人心智的能力——是以信念为中心的，并且在孩子大约四岁最终通过错误信念任务（false-belief

task，即理解某人可以持有孩子知道是错的信念）时趋于成熟。Phillips 及其同事汇集了多方面的证据，证明表征知识才是更基础的本领。线索如下：从发展角度看，婴儿和幼儿在能够处理错误信念之前很久，就能追踪谁看到过或有权访问信息——即谁知道。从比较上看，非人类大型类人猿表现出追踪他人感知和知识的强有力迹象，而关于信念追踪的令人信服的证据仍然难以捉摸。而在成年人中，对知识的归因至少与对信念的归因一样快——有时甚至更快，如果「X knows p」是通过先构建「X believes p」然后检查额外条件来计算的，这就很难解释。他们描绘的图景是一个事实性心理理论（factive theory of mind）：心智模拟他人的第一个也是最基础的工具是「他们知道什么？」，而更棘手的、容错的「他们仅仅相信什么（也许是错误的）？」则出现得更晚且成本更高。

拟议的排序——与教科书中「信念优先」的故事相反。

如果这一观点成立，那么对第 1 日的回报是直接且巨大的。它将为 Timothy Williamson 的知识优先（knowledge-first）计划提供经验证据——这一哲学主张（我们曾将其视为纯粹的扶手椅理论）认为知识是基础的、不可分析的状态，应该根据它来解释信念，而不是相反。突然之间，这不再仅仅是逻辑学家的直觉；它成了认知架构的一项候选事实，并具有演化上的合理性（社会性动物迫切需要追踪谁拥有可靠信息——这呼应了附录 I 中 Edward Craig 的「良好信息提供者」谱系学）。但诚实要求加上星号，而且此处的星号震耳欲聋：一篇 BBS 靶子文章附带了数十篇同行评议评论，许多人强烈反对。批评者认为知识与信念的界限比作者所允许的更为模糊，「追踪谁看到了什么」不一定是对知识的完整表征，而且文化和语言塑造了整个图景。因此：行为发现（儿童和猿类很早就追踪信息访问）是相当可靠的；而强解释（知识表征在形而上学和计算上是优先的，信念是从中组装而来的）则极具生命力。心智的家具正在实时重新摆放，而搬运工们尚未就平面图达成一致。

§3 机器转向，第一部分

语言模型知道什么吗——还是仅仅在胡说八道?

[已确立][争议/炒作]

第 1 日结束于一个尖锐而挥之不去的问题：当像起草这些页面的系统输出一个真实的、有充分支持的句子时，它是否知道什么——或者它只是终极的盖梯尔案例，因为与真理无关的原因而正确？2020 年代将那个结尾的华彩变成该领域最激烈的辩论之一，而讨论最多的条目有一个未经削弱就通过同行评审的标题：「ChatGPT is bullshit」（Hicks, Humphries & Slater, Ethics and Information Technology, 2024）。

他们的举动是精确的，而不仅仅是无礼。他们借用了 Harry Frankfurt 对胡说八道（bullshit）的技术性定义（出自他 1986 年的文章 On Bullshit）：胡说八道是带着对真理的漠不关心而产生的言论。骗子至少还会追踪真理——他必须这样做，以便引导你远离真理。胡说八道者则根本不在乎；他说任何符合其目的的话，至于是否真实根本不在考虑范围内。现在考虑大语言模型从根本上讲是什么：一个被训练用来预测统计上最可能的下一个标记 [11]、生成流畅且听起来合理的文本的系统。它没有试图遵循的真理表征。因此，当它陈述一个真实的事实和当它「幻觉」出一个虚假的引用时，它是在做完全相同的事——生成看起来合理的文本——并在两种情况下都同样成功地完成了其实际任务。在这种观点下，「幻觉」是一个带有误导性的美称，暗示了某种故障；更准确的描述是，该系统在设计上就对真理漠不关心，这正是 Harry Frankfurt 确切意义上的胡说八道。他们将软性胡说八道（无意欺骗，只是对真理漠不关心）与硬性胡说八道（此外还假装成真诚的真理陈述者）区分开来，并认为 LLM 至少是一个软性胡说八道者。

为什么这可能会重绘版图：它直接切断了关于机器「知道」、「理解」或「相信」的随意谈论。如果这个论点是正确的，那么 LLM 的真实输出就不是知识，甚至在完整意义上也不是真正的断言——它们是一类新的看起来像真理的文本，背后并没有人在乎它是否真实。这重塑了我们该如何信任、引用和监管这些系统。而且，不出所料，这是有争议的——反驳意见已经形成了一小片文献。一些人认为「胡说八道」标签暗中预设了关于模型是否具有意图的立场（Sarah Fisher,「Large language models and their big bullshit potential」, 2024; David Gunkel & Simon Coghlan,「Cut the crap」, 2025）；另一些人则认为，随着模型通过强化学习被训练得诚实并表达校准过的不确定性，「对真理漠不关心」过于粗糙了。已确定的是那个并不无聊的核心：基础语言模型没有内置的对真理的承诺，流畅性不等于知识。而开放的问题是，「胡说八道」、「工具」、「证言者」还是某种全新的认识论类别才是这些系统所产生内容的正确归宿。它是附录 I 中那只缸中之脑的硅制版本——交付给了十亿用户——那些可能从未触及过世界的词语，现在正在回答我们的问题。

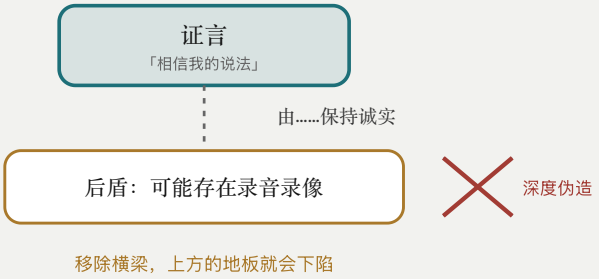
§4 机器转向，第二部分

你从未察觉的支撑梁：认识论后盾

[线索][争议/炒作]

附录 I 阐明了你所知的大部分内容都源自证言——即他人的话。Regina Rini 的 "Deepfakes and the Epistemic Backstop" (Philosophers' Imprint, 2020) 识别出了一种隐藏的结构性支撑，它一直悄无声息地维系着整个体系的诚实——并展示了一项新技术是如何将其锯断的。

这一洞察极为微妙。证言为什么如此可靠？Rini 认为，部分答案在于一个沉默的调节器：随时存在录音或录像的可能性。当一个人的说法可能被照片、录音或视频反驳时，他们就有持续的动力去如实陈述——因为某段录音或录像可能随时出现并揭穿他们。录音录像充当了认识论后盾：并不是因为我们经常检查它们，而是因为它们的存在本身就规范了证言，就像一名未上场的裁判依然能影响比赛一样。在大约一个世纪的时间里——自从摄影和录音变得难以造假以来——我们一直在这个后盾之上建立公共说实情的规范，却从未给它命名过。深度伪造——由 §3 中提到的同一波机器学习浪潮生成的逼真虚假视频和音频——从两个方面瓦解了这一后盾。它们在渠道中充斥着逼真的伪造品，同样具有破坏性的是，它们为每一个被抓获的造假者提供了一个新的借口：那段关于我的录像可能是一个深度伪造。「说谎者的红利」[12]。一旦任何录音录像都可以被轻而易举地否定，后盾就不再能约束证言——而作为我们最大的单一知识来源，证言本身也失去了一个我们此前并未察觉的支撑。Don Fallis 从信息论的角度加剧了同样的担忧 ("The Epistemic Threat of Deepfakes," Philosophy & Technology, 2021)：深度伪造削弱了视频所承载的关于实际发生情况的信息量，使其作为信号退化。



敲掉一个没人关注的支撑物，它所承载的结构依然会倒塌。

这一点影响深远，正是因为它将深度伪造重新定义为一个认识论问题，而不仅仅是欺诈或隐私问题：威胁不仅在于具体的谎言，还在于信任录音录像这一背景条件的整体侵

蚀。但是——正如本附录所指出的——其影响程度仍存争议，而且反驳意见非常尖锐，值得认真对待。Joshua Habgood-Coote（"Deepfakes and the epistemic apocalypse," Synthese, 2023）指出，末日论式的框架被夸大了：我们从未将录音录像视为绝对可靠，我们已经习惯通过多种来源交叉核对证言，而且社会以前也曾化解过媒体操纵的恐慌。Atencia-Linares and Artiga ("Deepfakes, shallow epistemic graves," Synthese, 2022) 捍卫了摄影和视频残存的认识论稳健性。因此，Rini 提出的机制——录音录像作为证言的沉默调节器——是一项真实且具有启发性的贡献；而关于「认识论末日」或公共知识全面崩溃的预测则是一场尚在进行的争论，而非定论。将这一点带到第 2 日，届时科学对「你如何信任一份你无法亲自核实的报告？」的回答是一套完整的重复实验和记录的制度机制——并带到 AI 板块，在那里它将与 §3 正面交锋。

§5 对抗性转向

敌对认识论：当环境是为了愚弄你而构建时

[已确立][争议/炒作]

传统认识论描绘了一颗孤独、中立的心智面对一个中立的世界。C. Thi Nguyen 的计划——他称之为 敌对认识论——始于一个更黑暗、也更现代的前提：你的认知环境并不是中立的。它越来越多地被刻意设计，通常由对你的信念怀有利益的各方操纵，以利用你的大脑赖以运转的那些可预测捷径。他在 2020 年后的三个举措重塑了整个领域讨论在线生活的方式。

第一个区别听起来很学术，结果却是解决一切的关键：认识气泡与回声室之间的区别 ("Echo Chambers and Epistemic Bubbles," Episteme, 2020)。它们不是一回事，将两者混为一谈正是许多善意的修复方案失败的原因。在气泡中，外部声音仅仅是缺失的——你只是没有接触到它们（想象一个只显示你认同的来源的过滤器）。在回声室中，外部声音是存在的，但被主动贬低了——你已被预先训练去怀疑它们（「主流媒体撒谎」，「专家已经腐败」）。后果是鲜明且反直觉的。

下方表格说明，最显而易见的干预——让人们接触另一方观点——为什么可能戳破气泡，却加固回声室。

气泡与回声室：接触后的结果

结构	外部声音	接触后的结果	启示
认识气泡	缺席，未被反驳。	新来源可以建立连接并戳破气泡。	当问题在于信息缺失时，接触可能奏效。
回声室	存在，但预先被质疑。	接触可能强化不信任，因为回声室预言了充满敌意的局外人。	当不信任被内置于结构中时，显而易见的修复手段可能会适得其反。

第二招指出你大脑内部的一个弱点。在 "The Seductions of Clarity" (Royal Institute of Philosophy Supplement, 2021) 中，Nguyen 认为，清晰的感觉——即当一切似乎都各就各位时那种令人满意的契合感——起到了终止思考的启发法[13]的作用。我们把「事情已经变得清晰」这种感觉当作探究已经足够、可以停止的信号。通常这没问题。但这意味着清晰感可以被武器化：一个能够制造夸大清晰感的操纵者——一种解释一切的整洁意识形态，或一个每个事实都能严丝合缝嵌入其中的阴谋论——可以让你在发现漏洞之前就提前终止探究。请注意这如何与 §1 相衔接：清晰之所以危险，恰恰是因为它终结了探究过程。正因如此，那些最圆滑、最让人觉得「现在一切都说得通了」的说法，反而最需要严加审视。第三招完善了这一工具：在 "Trust as an Unquestioning Attitude" (Oxford Studies in Epistemology, 2022) 中，Nguyen 将信任本身分析为一种不加质疑的立场——即将某事视为已解决的背景，在此基础上构建而不再重新检查。这是必不可少的（你不可能从头推导一切），也正因如此是可以被利用的：夺取了一个人毫无保留信任的东西，你就夺取了他永远不会想到去查核的盲点。

这里的真实评价是双重的。其概念性贡献——气泡与回声室、作为探究终止符的清晰感、作为不加质疑的信任——已被迅速且广泛地采用，因为它们确实起到了澄清作用并具有行动指导意义。但有两点警示值得注意。首先，哲学家们已经开始对这一概念框架本身提出异议（Carey & Ventham, "There is no fresh air: a problem with the concept of echo chambers," Episteme, 2025）。其次——这也是本课程坚持的一个去伪存真点——社会科学关于现实世界中回声室普遍程度的实证图景确实复杂且不一；几项大型研究发现，大多数人的媒体信息摄入比「封闭的回声室」这一形象所暗示的更为多样化。因此，请将这一概念机制视为一件尖锐而持久的工具，而将该现象的实证规模视为一个尚待测量的经验问题。框架本身就是贡献；它所描述的火势究竟有多大，仍在测量之中。

§6 形式化重构

重新推导第 1 日的荷兰赌——基于真理，而非金钱

[已确立][争议/炒作]

在第 1 日，我们用一场赌局证明了概率定律的合理性。荷兰赌定理表明，如果你的置信度违反了概率规则，一个聪明的博彩商可以卖给你一组你认为公平的赌注，但这些赌注合在一起保证你会赔钱。这很强大，但作为认识论论证却略显不尽人意。谁在乎钱呢？一个信念之所以不理性，难道不应该是因为某种与真理有关的原因，而不是因为你的钱包吗？一个在 2010 年代逐渐成熟并现在正处于全盛时期的研究项目——准确性优先认识论（也称为认识效用理论 [14]）——恰恰回答了这个问题，它是现代学科中最优雅的结果之一。

这个想法（由 James Joyce 在 1998 年的 "A Nonpragmatic Vindication of Probabilism" 中播下种子，由 Richard Pettigrew 的 Accuracy and the Laws of Credence（2016）建成体系，随后在 2020–2023 年出现了一波对其进行完善和质疑的论文）是用一个单一的认识论标尺来衡量一组置信度的优劣：准确性，即它与真理的接近程度。对真理的完全信心是完美的准确；对谬误的完全信心是最大程度的不准确。现在看定理：对于任何不融贯的置信度——即违反概率定律的置信度——保证存在一个在所有可能世界中都比它更准确的融贯置信度。在专业术语中，不融贯的置信度是被准确性支配的（accuracy-dominated）：无论情况如何，它在接近真理方面都被彻底击败了。所以你根本不需要博彩商。不融贯的置信度之所以不理性，完全是因为一个认识论上的理由——它白白放弃了唾手可得的准确性；无论世界如何变化，都有一组更好的置信度能更接近真理。

下方表格把同一几何关系列成三种置信度情形：在线上、在线上方、在线下方。

准确性支配，表现为置信度几何

置信度	总和	几何位置	裁决
P(S)=0.50, P(not-S)=0.50	1.00	位于融贯线上。	未被支配：在每个世界中，没有其他置信度更接近真理。
P(S)=0.80, P(not-S)=0.80	1.60	位于融贯线上方。	被一个更接近两个真理角的融贯投影所支配。
P(S)=0.20, P(not-S)=0.20	0.40	位于融贯线下方。	被一个更接近两个真理角的融贯投影所支配。

使其成为前沿而非注脚的原因是：它尝试将理性的基础重建在单一的认识论价值上——即接近真理——并从准确性支配论证中不仅推导出概率论[15]，还推导出更新规则（条件化）等更多内容。如果它完全成功，那么我们在第 1 日开始勾勒的整个贝叶斯体系将建立在真理之上，而不是博彩行为或心理学之上。不过，这枚筹码实至名归。概率论的核心支配定理是已经确立的数学。但其雄心——即所有的认识论规范都仅源于准确性——则有争议：最纯粹的结果依赖于技术假设（加法性[16]、有限命题），批评者认为这些假设比单纯的「接近真理」所能保证的内容夹带了更多的私货（Chad Marxen, "Epistemic utility theory's difficult future," Synthese, 2021），而且不同的准确性衡量标准可能会得出不同的裁决。因此：这是一个美丽且极具启发性的重新构架，拥有坚如磐石的核心和充满争议的外延——这恰好也是整个附录的形状。

◆ 三句话总结前沿

核心观点

自 2020 年以来，「什么是知识？」这一问题受到了来自五个维度的同时推进——将认识论重新构想为对探究而非对信念的研究（求知性）；重整心智的排序，使知识位于信念之前；并直面那些挑战甚至攻击「认知者」概念本身的机器、深度伪造和工程化信息环境——与此同时，一项形式化计划正在真理本身之上静默地重建理性的根基。

最佳新类比

认识论后盾：证言一直以来都由一根无人察觉的支柱维持着真实性——即记录存在的可能性——而深度伪造锯断了它；将其与回声室结合起来，在那里，最显而易见的解决办法（向他们展示另一面）恰恰会让陷阱变得更牢固。

现实争议

此处的每一项都尚未定论——探究规范是否吞噬了信念规范、知识表征是否真的比信念更基础、用「胡说八道」来形容大语言模型是否准确、深度伪造带来的是崩溃还是仅仅增加摩擦、以及仅靠准确性是否能奠定所有认识论理性的基础——这正为什么每一项都带有炒作过滤标签。

此处线索 › 信息（证言隐藏的后盾；作为对真理漠不关心的文本引擎的大语言模型；作为认识论善的准确性）· 计算（事实性心智理论；作为信念决策论的认识效用理论）· 演化（为什么社会性物种会演化出优先追踪知识的能力）。五条线索，现已浮出水面。

开放性问题

地图边缘的空白处

- 探究是真正的单位吗？ 寻求的规范是否真的与相信的规范相冲突——如果是，哪一个更根本？

- 知识优先还是信念优先？「事实性心智理论」是否是基础的认知工具，信念只是后来且成本更高的附加组件——还是知识与信念之间的界限本身就被划得太清晰了？
- 机器到底产生了什么？是知识、主张、证言、仪器读数，还是某种全新的、无人关心其真伪的、看起来像真理的文本？
- 摩擦还是崩溃？深度伪造仅仅增加了验证记录的成本，还是瓦解了公共知识的一个承重条件？
- 你心智的环境被工程化到了什么程度——以及我们现在能清晰描述的回声室，实际上到底有多大？
- 真理本身能奠定理性的基础吗？准确性优先原则能涵盖一切，还是仅限于其技术假设所能触及的范围？
- 还有一个为未来准备的更安静的竞争者：其中好几项都越过知识指向了理解，认为那才是我们真正看重的东西——每当一个模型能够预测却无法解释时，我们都会再次感受到这一转向。

来源 · 均为 2020 年后发表，基础性文献除外

来源与延伸阅读

1. Friedman, J. (2020). "The Epistemic and the Zetetic." The Philosophical Review 129(4): 501–536. doi:10.1215/00318108-8540918. 链接 参见 Falbo, A. (2023), "Should epistemology take the zetetic turn?" Philosophical Studies 180(10–11): 2977–3002; Flores, C. & Woodard, E. (2023), "Epistemic norms on evidence-gathering," Philosophical Studies 180(9): 2547–2571.

2. Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L. & Knobe, J. (2021). "Knowledge before belief." Behavioral and Brain Sciences 44: e140. doi:10.1017/S0140525X20000618 (目标文章 + 约 30 篇同行评议，包含若干不同意见). 链接

3. Hicks, M. T., Humphries, J. & Slater, J. (2024). "ChatGPT is bullshit." Ethics and Information Technology 26: 38. doi:10.1007/s10676-024-09775-5. 链接 基础锚点：Frankfurt, H. (2005), On Bullshit (Princeton UP). 回复：Fisher, S. A. (2024), "Large language models and their big bullshit potential," Ethics and Information Technology 26; Gunkel, D. & Coghlan, S. (2025), "Cut the crap: a critical response to 'ChatGPT is bullshit'," Ethics and Information Technology 27.

4. Rini, R. (2020). "Deepfakes and the Epistemic Backstop." Philosophers' Imprint 20(24): 1–16. 链接 以及 Fallis, D. (2021). "The Epistemic Threat of Deepfakes." Philosophy & Technology 34(4): 623–643. doi:10.1007/s13347-020-00419-2.

5. Habgood-Coote, J. (2023). "Deepfakes and the epistemic apocalypse." Synthese 201(3). 以及 Atencia-Linares, P. & Artiga, M. (2022). "Deepfakes, shallow epistemic graves: On the epistemic robustness of photography and videos in the era of deepfakes." Synthese 200(6). ——对「崩溃」框架的主要怀疑性回应。

6. Nguyen, C. T. (2020). "Echo Chambers and Epistemic Bubbles." Episteme 17(2): 141–161. doi:10.1017/epi.2018.32. 链接

7. Nguyen, C. T. (2021). "The Seductions of Clarity." Royal Institute of Philosophy Supplement 89: 227–255. 以及 Nguyen, C. T. (2022). "Trust as an Unquestioning Attitude." Oxford Studies in Epistemology 7: 214–244. 参见 Nguyen (2023), "Hostile Epistemology," Social Philosophy Today 39: 9–32; 以及评述 Carey, B. & Ventham, E. (2025), "There is no fresh air: A problem with the concept of echo chambers," Episteme First View. doi:10.1017/epi.2024.43.

8. Pettigrew, R. (2016). Accuracy and the Laws of Credence. Oxford University Press. 基础锚点：Joyce, J. M. (1998), "A Nonpragmatic Vindication of Probabilism," Philosophy of Science 65(4): 575–603. 最近的发展与评述：Pettigrew, R. (2022), "Accuracy-First Epistemology Without Additivity," Philosophy of Science 89(1): 128–151; Marxen, C. (2021), "Epistemic utility theory's difficult future," Synthese 199(3–4): 7401–7421. 综述：SEP, "Epistemic Utility Arguments for Epistemic Norms."

炒作过滤备注：引用经典锚点文献（Frankfurt 2005, Joyce 1998）仅作为 2020 年后工作的根源，而这些工作才是本附录的实际主题。上述任何主张都不应被视为已定论；这正是这些标签存在的意义。

明日 → 第 02 日

科学方法与划界问题

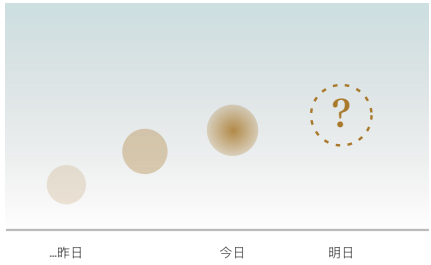
今天我们追问，单个信念何时才算得上知识。明天我们把眼光放向更大的规模：科学如何裁定哪些断言值得被认真纳入讨论？波普尔要求真正的理论必须可证伪，库恩的范式转移，拉卡托斯的研究纲领——以及现代复现危机，作为划界线在现实检验中的试炼。明天，你会用上我们今日校准好的对知识的直觉。

第 01 日终 · 还有 179 日等待深入

模块一 · 知识与推理的根基 · 第 02 日 / 180

科学方法与划界

太阳四十五亿年来每日东升。那么明日依旧会升起——对吗？



● 每一次过往的日出都是证据——却证明不了下一次日出

若问一个孩子，明天太阳是否会升起，他多半会觉得你问得莫名其妙。当然会升——它一向如此。这份笃定，仿佛知识最底层的磐石。可若再追问一句你凭什么相信，你便一脚踏上一座断崖——那是 1739 年一位寡言的苏格兰哲人悄然掘出的，至今无人填平。你唯一的凭据，不过是太阳从前升起过。你的论证其实是：未来会与过去相似，因为在过去，未来曾与过去相似。请再读一遍——它预设了它想要证明的东西。

这座断崖，名为归纳问题；整部科学的机器，正是从这里启动——不是凯旋，而是从一个缺口出发。今日，我们将目睹思想家们耗费两个世纪试图攀援而出：他们放弃证明，转而追逐否证；他们意识到科学其实并不像教科书所写的那般整饬；最终，在我们所处的时代，科学家以所能想象的最严苛方式拷问这整桩疑问——让大量已发表的发现接受复现，然后冷眼旁观其中一部分拒绝重演。

昨日（第 1 日）我们追问，单个信念何时堪称知识，并邂逅了盖梯尔那只停走的钟——那是一桩被运气而非关联拯救的真信念。今日，我们将这一忧虑从一颗心智放大到整个文明尺度的事业：科学如何裁定，哪些主张才配进入竞技场？请把昨日的工具留在手边。第 1 日的信念刻度盘（信念有程度之分，并非全有即全无）即将成为面对休谟质疑的唯一清醒回应；而那道前沿校准器——它筛去热门发现，又在复现实验将其推翻时悄然生效——今日将成为整场戏的第三幕。

地上的裂口

休谟抽去了地基

1739 年，二十八岁的大卫·休谟出版《人性论》——一部问世时备受冷落的著作，他自嘲它「一出世便已夭折」。书中藏着一枚引线极长的炸弹。休谟注意到，我们关于尚未直接经历之事的全部信念——面包明日仍将如今日般滋养我们，太阳仍将升起——都倚靠一个隐秘的假设：即自然是齐一的，未曾经历的事物会与过往经验一样运作。

他指出，这一假设无从辩护。不是逻辑问题：太阳明天不升起，并不蕴涵矛盾。诚如休谟以不动声色的精准所言：

> *太阳明日不会升起，这一命题并不比它明日会升起更不可理解，也不蕴涵更多矛盾。*
>
> ——休谟，《人类理解研究》，§IV（1748）

因此，齐一性并非逻辑真理。那么，能否以经验为之辩护——「它向来如此，所以推断它会继续如此是稳妥的」？且看陷阱合拢：这一论证动用了过去预测未来的原则，来证明过去预测未来。这是循环论证。人不可能拽着自己的头发离开地面。休谟的结论堪称真正激进，值得不加粉饰地陈述：我们对自己的未来之确信，毫无理性根据。我们期待日出，是出于习惯，而非逻辑证明。

这便是科学方法自诞生起就试图包扎的伤口。若我们永远不能以堆积证实的案例来证明一条普遍定律——再多的白天鹅也无法证明「所有天鹅皆白」——那么科学声称发现自然定律时，究竟在做什么？

关于黑天鹅的注记

欧洲人曾如此确信所有天鹅皆白，以至于「黑天鹅」成了数个世纪以来的习语，意指不存在之物——好比「太阳从西边出来」。然而 1697 年，荷兰探险家抵达西澳大利亚，发现河湾中满是黑天鹅（Cygnus atratus）。百万次确认的目击筑起了一条坚不可摧的定律；珀斯的一只孤鸟却将其击得粉碎。请在心中持守这一不对等——它即将成为今日全篇的枢轴。

一只黑天鹅让这种不对等变得一目了然：确认案例可以堆积数百年，而一个反例仍足以击碎定律。

逃遁之路

波普尔的柔道：别再试图证明

1920 年代的维也纳。年轻的卡尔·波普尔被各种急于攫取「科学」之名的思想运动包围：弗洛伊德的精神分析、阿德勒的个体心理学、马克思的历史理论。追随者们如痴如狂。他们环顾四周，满眼皆是证实——每一句口误都印证弗洛伊德，每一次政治旋涡都印证马克思。而波普尔猛然意识到，这恰恰是它们的病灶所在。

解释一切的理论，其实一无所释。若没有任何可想象的观察能够反驳你的理论——若有人救起溺水儿童，与有人眼睁睁看着他溺毙，皆能同样套入弗洛伊德的框架——那么你的理论并不勇敢。它是空洞的。它没有排除任何可能，故世界无从惊扰它。

请将之与爱因斯坦对照。1915 年，广义相对论作出了一项大胆的、高风险的预言：掠过太阳的星光会弯折一个特定角度——1.75 角秒，是牛顿预言的两倍。若 1919 年的日食测量结果符合牛顿的预测，爱因斯坦便将一败涂地。他把理论的脖子伸了出去。那，波普尔说，才是真实科学的印记。

于是波普尔使出一记哲学柔道。休谟说得对——你永远无法证实一条普遍定律。很好。那么停止尝试。将黑天鹅的不对称性翻转为一门方法：

> *一种理论之科学地位的标准，在于其可证伪性、可反驳性，或可检验性。*
>
> ——波普尔，《猜想与反驳》（1963）

你无法以任何数量的白天鹅证明「所有天鹅皆白」——但一只单独的黑天鹅便永久否证了它。证实终归无望；证伪却可一锤定音。依此观点，科学并非从证据拾级而上、迈向确定性。它提出大胆的猜想，然后竭尽全力试图反驳它们。那些在我们最猛烈的反驳尝试中幸存的理论，并非被证明——它们只是仍屹立不倒、得到佐证，在下一轮检验之前被临时信任。知识之增长，来自理论在反驳中幸存，而非证实案例的累积。

划界标准——科学与伪科学之间的界线——由此干净利落。一项主张的科学性，取决于它是否把头伸出去：是否排除某些可能，作出可被推翻的预言，预先告诉你什么会证明它错误。「经济由阶级斗争支配」没有排除任何明确结果。「光线弯折 1.75 角秒」却排除了 1.74 与 1.76。后者是科学；前者更像一套披着白大褂的世界观。

公允以待弗洛伊德

这是个利落的故事，波普尔讲得极为出色——或许太出色了。后来的哲学家（尤其是 1984 年的阿道夫·格伦鲍姆）辩称，波普尔把精神分析刻画得过于简单：弗洛伊德有时确实指明过什么将反驳他（「只有当恐惧症被证明存在于性生活完全正常之处时，我的理论才能被反驳」）。而许多受人敬重的科学——历史学、进化论、宇宙学——同样无法进行对照实验。可证伪性是一束锐利的探照灯。今日余下时光，我们将看着它在边缘处摇曳明灭。

复杂的现实

库恩：但科学并非那样运行

波普尔描述的是科学应当如何运作。1962 年，由物理学家转任的史学家托马斯·库恩审视了科学实际如何运作——发现了某种更芜杂、也更有人情味的东西。他的《科学革命的结构》成为二十世纪最广为引用的学术著作之一，并赋予你一个用过百遍却不知出处的词：范式。

这是库恩的异端之说。真正工作中的科学家，几乎在所有时间里，都不是在证伪他们的宏大理论。他们在做他所谓常规科学之事：在一个被接受的框架——一个范式——内部解谜，而他们将这范式视为理所当然。一位化学家醒来时不会想着反驳元素周期表；她用它去琢磨一个反应。范式不是被告。它是法庭本身。

而当实验结果异常时？科学家们大多不会像波普尔的故事要求的那样立刻抛弃理论。他们会把它视为反常——一个留待日后解决的谜题，大概是自己哪里做错了。理论太过有用、太过多产，不至于因一个顽固的数据点就弃之。（注意，这与证伪主义正好相反——而且，说来尴尬，这也正是那些弗洛伊德主义者和马克思主义者所做的。）

只有当反常堆积——变得太多、太核心而无法忽视——领域才滑入危机。而危机的解决，并非通过整洁的反驳，而是一场科学革命：向新范式的全盘切换。托勒密的圆环让位于开普勒的椭圆；牛顿的绝对空间让位于爱因斯坦的时空。库恩认为这些转变如此彻底，以至于两个范式变得不可通约——「无共同尺度」，因为对立阵营甚至对关键词汇的含义、哪些问题才重要都无法达成一致。「质量」于牛顿与爱因斯坦意指着微妙不同的东西。范式切换不太像赢得一场论证，更像是一次格式塔翻转——鸭子变兔子，你无法同时看见两者。

一个值得破除的迷思

库恩常被引为「科学不过是意见」或「所有范式同等有效」的证据。他憎恶这种解读，并耗费数年反击。他并非在说科学是非理性的——而是说，科学的理性比那套洁净的证伪主义童话所承认的更具共同体特征、更有历史纵深，也更趋保守。范式之所以被推翻，是因为对手真正解决了更多谜题。那不是相对主义，只是对人类实际科学实践的一种现实主义态度。

修补

拉卡托斯：理论从不孤身赴死——以及杜恒–奎因的幽灵

波普尔说证伪；库恩说科学家并不如此，也不应急于如此。是否存在一条道路，能兼纳二者——在保持证伪之脊梁的同时承认库恩的历史？伊姆雷·拉卡托斯，一位栖身伦敦经济学院的匈牙利流亡者，试图搭建的正是这样一座桥梁。但首先，我们必须会见那萦绕整间屋子的幽灵。

它被称为杜恒–奎因论题，一旦看见便无法视而不见。其主张简单却摧枯拉朽：没有任何假说是被单独检验的。当你检验「这颗星位于彼处」时，你同时依赖光学、大气模型、望远镜校准、光如何传播的理论。因此，当预言失败时，纯逻辑从不告诉你哪一环断

裂。或许是假说错了——又或许只是望远镜校准有误。你总可以把责任推给辅助假设，来拯救自己钟爱的理论。波普尔那洁净的「一只黑天鹅便杀死理论」，原来从不曾那般洁净：你可以坚称那只黑天鹅不过是一只被涂漆的鹅。

这并非书斋里的琐屑——它是真正发现的引擎。1840 年代，当天王星偏离其牛顿式轨道时，无人宣布牛顿被反驳。他们归咎于一项辅助假设：必定有一颗隐匿行星在牵引它。他们是对的——海王星便于 1846 年以此方式发现，一场辉煌的正名。受此鼓舞，天文学家们对水星的摇摆使出同一招，预言了另一颗隐匿行星，命名为祝融星。他们搜寻了数十年。它并不存在。水星的摇摆是在告诉世人，牛顿本人并不完备——而唯有 1915 年的爱因斯坦能道破此点。同样的逻辑招式，截然相反的结果。那么，如何分辨高明的拯救与绝望的遁词？

拉卡托斯的答案重构了科学的单元。不要评判孤立的理论——要评判随时间展开的研究纲领。每个纲领都有一个硬核（例如「牛顿定律成立」），外裹一层可调辅助假设的保护带。麻烦来临时，你在保护带中吸纳冲击，而非伤及核心。这本身没有问题。关键在于接下来会发生什么：

- 一个进步纲领的补丁预言了令人惊异的新事实，而这些新事实随后真的出现。「有一颗隐匿行星」预言了海王星会出现在天空中的某个特定位置——而它果然就在那里。这场拯救以新知识偿付了自身。
- 一个退化纲领的补丁永远只是事后追补，为每一次失败硬凑借口，却从不预言新事物。祝融星被无尽地重新安置到恰好无法被看见之处，便是警示的信号。

这便是重新绘制的划界线——而且与真实历史契合得多。科学不是单一理论面对单一裁决；它是一个纲领在岁月中赢得或失去立足之地，衡量的标准在于它是否持续告诉我们尚未知晓的事物。

重锤

费耶阿本德与「那」方法的死亡

随后，拉卡托斯的友人与论敌保罗·费耶阿本德把整个项目推到了极限。在《反对方法》（1975）中，他提出了一项调皮、恼人、却又出人意料地证据充分的论证：翻检科学突破的真实历史，你会发现每一条方法规则都曾在某个关键时刻被打破——而打破它恰恰是为了推动进步。伽利略以宣传、修辞伎俩和无视不利数据的方式推进了哥白尼事业。若他遵从了整饬的方法规则，那场革命或许便会停滞。

他的结论成为科学哲学中最臭名昭著的一句口号：「怎么都行。」但这里有一个几乎人人忽略的关键细节——费耶阿本德并非意指「随心所欲，所有想法平等」。他的意思是，

这是一个苦涩的归谬论证[1]：唯一没有历史反例的方法规则，空泛到允许一切。用他的话来说，这是一位理性主义者终于诚实地审视历史后发出的「惊恐的呼喊」。他焚烧的是「存在某种大写 M 的方法论可以一劳永逸地定义科学」的观念——而非对混乱的背书。

1983 年，哲学家拉里·劳丹发表了看似葬礼悼词的文字。在那篇著名论文《划界问题的消亡》中，他论证所有试图画出清晰界线的尝试——包括波普尔的——皆已失败，而「科学」与「伪科学」过于多样，无法共享单一的决定性标记。这些术语，他尖刻地写道，大体只是「承载我们情感评判的空洞辞藻」。两千五百年后，划界问题被宣告死亡。

复活

为何界线依然重要

然而——这个问题太有用，不会真的入土为安。2013 年，哲学家马西莫·皮柳奇与马尔滕·布德里编纂了一部直言不讳的文集：《伪科学哲学：重新思考划界问题》，推动划界问题的复兴，回击了劳丹。他们的论证部分出于实践，且难以回避：在一个疫苗抗拒、气候否认、神迹疗法与智能设计「理论」并存的世界里，分辨科学与其仿品并非闲散的客厅游戏。它关乎生死。

他们的哲学转向是，不再要求某种单一的万能标准，而是将科学视为一个家族相似概念——借用维特根斯坦的术语。并非每种科学都共享某一特征，而每种伪科学都缺乏它。取而代之的是一组彼此重叠的特征：可证伪的预言，诚然，但也包括经验证绩、对修正的开放、与既有知识的融贯、对反常的诚实处理，以及典型遁词的缺席（无尽的事后补救、受迫害叙事、对证据免疫）。没有单根线维系整条绳索；是众多线股的交叠。真正的科学可能在某一标准上薄弱，而在其余标准上强劲。伪科学则因同时通不过整组特征而暴露自身。

而这便铺垫了今日全篇的压轴一击。以上的一切——波普尔、库恩、拉卡托斯、那簇美德——皆是哲学，在研讨室中辩论。但在过去十五年间，科学做了一件非凡之事：它以大规模实证的方式，将划界问题转向了自身。它问自己，已发表的诸多发现能否经受住最基本的科学要求。

划界标准表

主张	波普尔	库恩	拉卡托斯	簇群视角
星光弯折 1.75 角秒	科学	科学	进步	强科学画像
水星逆行扰乱通讯	非科学	非成熟科学	退化	弱画像
阶级斗争驱动历史	按常用方式往往不可证伪	视情况而定	可能退化	社会科学兼哲学的混合
弦理论	关键形式尚未可检验	无决定性检验的常规科学	开放问题	鲜活的边界案例
共同祖先	可证伪	生物学核心范式	进步	强科学画像

前沿 · 2026

复现危机：划界在现实检验中

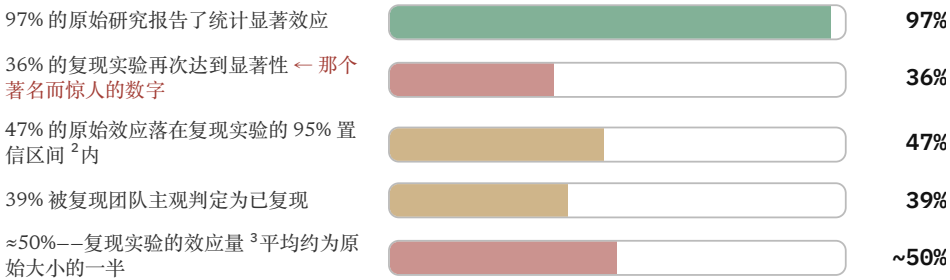
若有一条几乎人人认同的标准——波普尔、库恩、你的高中老师——那便是可复现。真正的结果，当别人照着程序再做一遍时，应当再次出现。它不是侥幸、捏造或风尚。于是在 2010 年代，科学家们做了一件显而易见、令人不安、却从未被系统做过的事：他们取来成堆的已发表、经同行评议、备受赞誉的发现，逐一尝试复现。

结果 01 [已确立] [争议/炒作]

震动心理学的一声枪响

里程碑是开放科学合作组织的《估计心理科学的可复现性》（Science，2015 年 8 月 28 日）——约 270 位研究者，在布莱恩·诺塞克领导下，复现了三本顶尖心理学期刊上的 100 项研究，并与原作者合作确保方法无误。结果在该领域引发爆炸。但唯一最重要的

教训却藏于明处：并不存在单一的「复现率」。该论文报告了数个，而它们讲述着不同的故事。请看。



每当你听见「只有三分之一的心理学是真实的」，便是有人抓起了 36% 而丢弃了其余。更诚实的概括要微妙得多，也更有意思：复现实验中的效应平均更弱——大约为首次报告的一半强度，且往往因复现实验功效不足 [4] 而未能检出。[核心数字已确立]；这些数字究竟能在多大程度上说明哪些原始效应真实存在，[解释仍有争议]。

而作者拒绝让任何人——乐观者或唱衰者——过度解读。他们自己的结论是一篇校准的小杰作，也是对第 1 日教训的直接回响：基于错误理由而持有的真信念，并不等于知识：

> *我们已确立为真实的效应，有多少？零。而我们已确立为虚假的效应，有多少？零。*
>
> ——开放科学合作组织，Science (2015)

请记住杜恒-奎因的幽灵：一次失败的复现实验并不在逻辑上反驳原始研究——条件总有差异。而这正是批评者发难之处。Gilbert, King, Pettigrew & Wilson（Science，2016 年 3 月）认为该项目自身的复现实验统计功效不足，且经校正后，「数据与相反结论一致」——也就是复现情况可能相当好。原团队回应，乐观与悲观的解读皆未得到充分支持。[有争议]——解读确属悬而未决，即便这一广泛问题如今已被普遍承认为真实存在的现象。

结果 02 [已确立]

这并非一个领域的难堪

那种条件反射式的辩护——「软科学嘛，还能指望什么」——随着同样的复现实验在其他领域展开并返回同样令人沮丧的结果，便不攻自破。这场危机是全局性的。以下是经核实的锚定数字；每次请注意度量标准，因为如我们刚见，度量标准就是故事本身。

项目与发表处	复现对象	已复现 *	效应量缩减
心理学 OSC, Science 2015	100 项研究，3 本顶尖期刊	**36%**	约为原始效应的 50%
癌症生物学 Errington et al., eLife 2021	计划复现 193 项实验——仅约 50 项得以尝试	**~46%†**	约缩小 85%
实验经济学 Camerer et al., Science 2016	18 项实验室实验（AER, QJE）	**61%**	约为原始效应的 66%
社会科学 Camerer et al., Nat. Hum. Behav. 2018	Nature 与 Science 中的 21 项实验	**62%**	约为原始效应的 50%
临床前肿瘤学 Begley & Ellis, Nature 2012	53 篇「里程碑」论文（安进）	**11%**	——（53 篇中仅 6 篇被确认）

*「已复现」= 同方向显著效应，最严格的一般度量。†癌症生物学数字为已完成实验中的比例；引人注目的是，193 项原始实验中无一能仅凭发表的方法复现，且仅有 2% 可获得原始数据。[已确立]

最深的信号甚至不是失败率——而是癌症生物学团队发现他们无法弄清原始科学家究竟做了什么。方法部分过于单薄，无从遵循；原作者往往不愿分享方案或数据。一项你连尝试复现都做不到的发现，并非未通过波普尔的检验——它拒绝接受检验。而一项调查将这种不安落到了实处：当 Nature 于 2016 年调查 1,576 位科学家时，超过 70% 表示他们曾尝试复现他人的实验却遭失败，超过一半未能复现自己的实验。[已确立]——尽管请注意这是意见数据，是科学家们相信什么，而非实际测量的比率。

结果 03 [已确立] [争议/炒作]

那些烟消云散的发现——以及敢于承认的科学家们

抽象的概括不会刺痛人；具名的失败才会。一连串曾被称颂、在 TED 演讲中广为人知的效应，在高功效、预登记[5]的复现实验中折戟——而令人瞩目的是，在最清楚的案例中，原作者本人公开改变了主意：

- 权力姿势。2010 年的发现称，以神奇女侠式站姿站立两分钟可提升睾酮与风险承受意愿（一场被观看数千万次的 TED 演讲）——在 2015 年一项规模大得多的复现实验中，于每一项生理指标上失败。随后，原论文的第一作者达娜·卡尼做了一件罕见而可敬的事——她公开否定了自己最著名的成果：「我不相信『权力姿势』效应是真实的。」[已确立]
- 自我损耗。意志力是一种随使用而耗竭的有限燃料这一主导理论，在 23 间实验室（N = 2,141，2016 年）中得到检验。合并后的效应在统计上与零无法区分（d = 0.04）。该领域的一位领军研究者迈克尔·因兹利希特写道，他感到「脚下的地面正在移动」。[已确立] 标准效应未能复现；某种微小效应是否尚存仍在争论。
- 社会启动。那项经典主张——阅读关于老年的词汇会使你离开实验室时走得更慢——在 2012 年的独立复现实验中失败。它震动了整个领域，以至于诺贝尔奖得主丹尼尔·卡尼曼发出公开信，警告启动效应研究者，他们的领域已成为「质疑心理学研究诚信的典型代表」。[已确立] 针对这个具体案例。
- 斯坦福监狱实验（1971）——或许是心理学史上最著名的「研究」——被档案研究（Le Texier，American Psychologist，2019 年）揭示更接近于一场摆拍的戏剧：狱卒被诱导向残忍，结果被耸人听闻地渲染。它与其说是一次复现失败，不如说是划界问题中的警示案例——一项或许从来不是真正实验的演示。[有争议]——津巴多生前反驳了这些批评；是否应将其从教科书中剔除仍在争执。

转折 ［线索］

这是科学的失败——还是科学在运作？

换个角度看，整场危机也可以是一个充满希望的故事，而非一桩丑闻。上述每一个数字都来自科学家以科学审视科学——使用预登记、高功效、公开共享的方法来揭露并丢弃那些站不住脚的主张。那是波普尔的反驳之刃，终于向内翻转。危机并非划界标准错误的证据，而是它们正在运作的证据，痛苦地、公开地运作着。

而且它还触动了真正的改革。研究预登记——在看见数据之前陈述你的假设与分析——关上了那扇夸大效应的暗门（p 值操纵）；注册式报告，即期刊在结果出现之前仅依据方法接受研究，如今已被 300 余家期刊采纳。有人提议将「显著」阈值从 $p < 0.05$ 收紧至

$p < 0.005$，而开放数据与多实验室联盟的文化已成常规。该领域正视休谟留下的缺口，看见运气与偏见多么轻易地伪造知识——正是第 1 日盖梯尔忧虑在工业规模上的重现——并开始重建其工具。我们将在第 149 日再次完整遇见这场改革运动。

悬而未决的问题

何谓真正尚未落定

两千五百年过去，「何为科学？」这一问题的审慎回答仍有几条线没有系紧：

- 是否存在任何单一的划界标准——还是劳丹赢了，留下的只有维特根斯坦式的、重叠的诸美德家族，而无总纲？
- 杜恒–奎因问题能在多大程度上被驯服？若一次失败的检验从不在逻辑上归罪于某个假说，那么高功效、预登记的复现实验如何真正缩减腾挪空间——它们能否将之彻底关闭？
- 那些根本无法进行实验的科学又该如何——宇宙学、进化生物学、弦理论？若一种理论在整整一代人的时间里无法作出可检验的预言（第 48 日的量子引力难题隐约浮现），它是科学、原科学，还是数学？
- 复现的底线在哪里？社会科学中 62% 的复现率——面对复杂的人类行为，这算失败、合理水平，还是在「复现」定义本身达成一致之前无从判断？
- 而那个将萦绕整门课程的问题：若即便经同行评议、备受赞誉的发现也被夸大了半数之多，那么你——在阅读任何一项自信的断言时，包括本页上的——该如何设定你的信念刻度？（带上刻度盘。第 4 日、第 6 日。）

◆ 一日三句话

核心洞见

休谟指出，你永远无法靠堆积证实的案例来证明一条普遍定律，因此科学转而提出大胆的、可证伪的猜想，并竭力试图反驳它们——但真实的科学比那条洁净规则更复杂（库恩、拉卡托斯、费耶阿本德），而现代复现危机最终让那场辩论接受了硬数据的检验。

最佳类比

黑天鹅：百万只白天鹅无法证明「所有天鹅皆白」，但澳大利亚的一只黑天鹅便永久否证了它——证实终究做不到，证伪却可一锤定音。

活的争议

是否存在单一界线划分科学与伪科学（波普尔的可证伪性 vs 劳丹的「消亡」），以及复现数字究竟意味着什么——是破碎科学的丑闻，还是科学按设计运作的健康、公开的自我修正。

今日线索 › 信息（复现实验作为检验一项主张承载真实信号抑或噪音的试金石）· 演化（在波普尔那里，知识像选择过程一样增长——经反驳而幸存的猜想，预告第 74 日）· 计算与涌现（轻触——科学作为一个分布式的、自我修正的寻错系统，能完成任何单个心智无法完成之事）。

说明

1. 归谬论证通过说明某个观点会导向荒谬或不可接受的后果来反驳它。
2. 95% 置信区间指一种方法在反复使用时有 95% 的次数会覆盖真实值；它不是说这个具体区间有 95% 的概率包含真实值。
3. 效应量是衡量一个效应有多大的标准化指标，不同于它是否刚好跨过显著性门槛。
4. 功效不足的研究数据太少或噪声太大，无法可靠地检出相关大小的效应。
5. 预登记是在看见数据之前写明假设、样本计划和分析方法，避免事后选择悄悄追逐漂亮结果。
6. 谓词是用来表示某物具有某种性质的表达，例如绿色、沉重、素数，或在 2050 年前被检查过。
7. 协变量是分析中额外纳入的测量变量，常用于调整受试者或条件之间的差异。
8. p 值操纵指在看见数据后尝试多种分析选择，直到某一种给出可发表的 p 值。
9. 零假设是统计检验试图拒绝的默认说法，通常是「没有真实效应」或「没有差异」。
10. I 类错误就是假阳性：零假设其实为真时，你却把它拒绝了。
11. p 值是在某个指定模型（例如零假设）成立时，得到至少这么不相容的数据的概率。

来源

来源与延伸阅读

1. Hume, D. (1739–40). A Treatise of Human Nature, Book I, Part iii. And (1748) An Enquiry Concerning Human Understanding, §IV–V. ——归纳问题；日出段落。见 Stanford Encyclopedia of Philosophy, "The Problem of Induction"（修订版 2018）。

2. Popper, K. (1959). The Logic of Scientific Discovery (orig. Logik der Forschung, 1934). And (1963) Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge. ——可证伪性；爱因斯坦 vs 弗洛伊德/阿德勒/马克思。见 SEP, "Karl Popper"。

3. Kuhn, T. S. (1962; 2nd ed. 1970). The Structure of Scientific Revolutions. University of Chicago Press. ——常规科学、范式、反常、危机、革命、不可通约性。见 SEP, "Thomas Kuhn"。

4. Lakatos, I. (1970). "Falsification and the Methodology of Scientific Research Programmes," in Lakatos & Musgrave (eds.), Criticism and the Growth of Knowledge. Collected in Philosophical Papers, Vol. 1 (Cambridge UP, 1978). ——硬核、保护带、进步与退化纲领。

5. Feyerabend, P. (1975). Against Method: Outline of an Anarchistic Theory of Knowledge. New Left Books. ——认识论无政府主义；「怎么都行」作为归谬。见 SEP, "Paul Feyerabend"。

6. Duhem, P. (1906). The Aim and Structure of Physical Theory. And Quine, W. V. O. (1951). "Two Dogmas of Empiricism," The Philosophical Review 60(1): 20–43. ——欠决定 / 整体确证论。见 SEP, "Underdetermination of Scientific Theory"。

7. Laudan, L. (1983). "The Demise of the Demarcation Problem," in Cohen & Laudan (eds.), Physics, Philosophy and Psychoanalysis. Reidel, pp. 111–127.

8. Pigliucci, M. & Boudry, M. (eds.) (2013). Philosophy of Pseudoscience: Reconsidering the Demarcation Problem. University of Chicago Press. press.uchicago.edu ——复兴；科学作为家族相似 / 簇群概念。

9. Open Science Collaboration (2015). "Estimating the reproducibility of psychological science." Science 349(6251): aac4716. doi:10.1126/science.aac4716 。 science.org ——97% / 36% / 47% / 39% / ~50%。

10. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). "Comment on 'Estimating the reproducibility of psychological science.'" Science 351(6277): 1037. ——批评； OSC 回应（Anderson et al.，同期）。

11. Errington, T. M. et al. (2021). "Investigating the replicability of preclinical cancer biology." eLife 10: e71601 (Reproducibility Project: Cancer Biology). ——193 项中约 50 项实验被尝试；效应约缩小 85%；方法/数据大多无法获得。

12. Camerer, C. F. et al. (2016). "Evaluating replicability of laboratory experiments in economics." Science 351(6280): 1433–1436. doi:10.1126/science.aaf0918 ——18 项中 11 项（61%）。

13. Camerer, C. F. et al. (2018). "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." Nature Human Behaviour 2: 637–644. ——21 项中 13 项（62%）。

14. Klein, R. A. et al. (2018). "Many Labs 2: Investigating variation in replicability across samples and settings." Advances in Methods and Practices in Psychological Science 1(4): 443–490. ——28 项中 15 项（54%）；场景未能解释失败。

15. Begley, C. G. & Ellis, L. M. (2012). "Raise standards for preclinical cancer research." Nature 483: 531–533. doi:10.1038/483531a ——53 项中 6 项（11%）里程碑论文被确认（安进）。

16. Baker, M. (2016). "1,500 scientists lift the lid on reproducibility." Nature 533: 452–454. doi:10.1038/533452a ——>70% 未能复现他人结果；>50% 未能复现自己的结果。

17. Hagger, M. S. et al. (2016). "A multilab preregistered replication of the ego-depletion effect." Perspectives on Psychological Science 11(4): 546–573. ——23 间实验室；d = 0.04。

18. Ranehill, E. et al. (2015). "Assessing the robustness of power posing." Psychological Science 26(5): 653–656. And Carney, D. R. (2016), 公开声明否定权力姿势效应。见 概述。

19. Le Texier, T. (2019). "Debunking the Stanford Prison Experiment." American Psychologist 74(7): 823–839. doi:10.1037/amp0000401。 pubmed

20. Ioannidis, J. P. A. (2005). "Why most published research findings are false." PLoS Medicine 2(8): e124. ——奠基性（且基于模型，故细节上有争议）论文。

21. Benjamin, D. J. et al. (2018). "Redefine statistical significance." Nature Human Behaviour 2: 6–10. doi:10.1038/s41562-017-0189-z ——p < 0.005 提案（及 Amrhein & Greenland「移除而非重新定义」的反驳）。

22. Chambers, C. D. (2013). "Registered Reports: A new publishing initiative at Cortex." Cortex 49(3): 609–610. And Chambers & Tzavella (2022), Nature Human Behaviour 6: 29–42 ——注册式报告如今已有 300 余家期刊采纳。

可选附录

附录：没有基岩的地基

本节是可选的补充阅读；可以放心跳过，不会影响正文课程。

正文中我们反复说着「证伪」「检验」「观察」；现在往下再掘一层，掀开地板——脚下并无实地。

正文已将路线标出：休谟挖出的深坑、波普尔找到的出口、库恩指出的乱象、拉卡托斯的修补，还有复现危机把两百年的争论拖进真枪实弹的检验。本附录带你在同一栋建筑中再下一层，掀开地板，直视地基。而不同年代、不同脾性的人反复抵达的，竟是同一个结论：脚下没有磐石。没有地基。没有能平息争端的理论中立观察；没有不循环的理由支撑我们对明天的信心；没有纯逻辑的算法能给一个命题盖上「科学」的印信。有的只是一根根打进沼泽的桩子——打得更深一些，暂时撑得住而已。

本附录紧承第 2 日正文，以复现危机与那个悬而未决的问题为起点——科学是失败了，还是在按它本来的样子运转？这里我们要把正文中一笔带过的四件事往深处凿一凿：(1) 认真对待休谟之后，归纳问题会变成什么模样——以及藏在它身后那个更刁钻的谜题；(2) 波普尔自己坦率承认的理论裂缝；(3)「中立检验」为什么可能根本不存在；(4) 让大多数已发表成果注水的真正数学。请把第 1 日养成的校准直觉带在身边——读到最后你会明白，那是唯一稳妥的姿态。

第一部分 · 深坑愈深

休谟自问自答——古德曼却雪上加霜

正文把休谟留在了这样一个位置：没有任何不循环的方式可以论证我们对日出的信心。但休谟本人并未止步于此，而教科书略去的那个部分，反倒最有人情味。在证明理性无法为归纳奠基之后，他紧接着问了一个再自然不过的问题：既然如此，我们为何每时每刻仍在归纳，却从未因此陷入混乱？他的回答带着一种温柔。我们靠习俗推断——靠习惯。孩子被火烫过一次，再见火焰便知道避让；这不是演绎，而是反复经验刻入身心的条件反射：

> *在许多实例中发现，两类对象……总是联结在一起的；如果火焰或冰雪再次呈现于感官之前，心智便被习俗引向对热或冷的预期……这种信念是将心智置于如此情境中的必然结果。*
>
> ——休谟，《人类理解研究》，§V（1748）

这个拆分值得命名，因为它将在整部课程中反复现身。休谟把一个问题劈成两半：一个是辩护问题——归纳能否被演绎地、不循环地证明？答案是不能，这道伤口永远不会愈合。另一个是描述问题——心智为什么还是照样推断？答案是我们天生如此，靠的是习俗。他放弃了前者，回答了后者。我们并非碰巧拥有本能的推理机器；我们是有本能的机器，只是学会了给习惯披上理性的外衣。（你会在第 11 日的启发式与偏差、第 119 日的预测性大脑中，再次遇见这同一种拆分。）

四种爬出深坑的尝试

两个半世纪以来，哲学家们试图从休谟的深坑中爬出。无人完全成功——但这些尝试每一桩都精彩绝伦，因为每一种都是某种性情气质凝结而成的论证。

斯特劳森

消解问题

问「归纳是否合理」本身就问错了。把信念按证据调整，这就是推理得好的题中之义。要求一个外在的盖章认可，好比追问法律本身合不合法。问题一经如此提出，便自行消解。

赖欣巴赫

务实地下注

我们证明不了归纳一定有效，但可以证明它是当下能下的最好赌注。如果有哪种方法能捕捉自然的规律性，归纳最终一定能捕捉到。它至多不比别的方法差，所以尽管用。这是一种手段层面的辩护，而非真理层面的辩护。

波普尔

否认前提

他的激进主张：根本不存在归纳这回事。科学从不从实例中概括，而是大胆猜想、竭力反驳。方法中既无归纳步骤，休谟的问题便无处下嘴。（批评者追问：那科学岂不永远无法告诉我们某个理论对预测是可靠的？而这显然是我们需要的。）

贝叶斯

量化更新

把学习看成用贝叶斯定理修正置信度——也就是第 1 日的信念刻度盘。这漂亮地形式化了从证据中学习的过程，却并未化解休谟：先验概率与更新规则本身仍需根基。（将在第 4 日正式展开。）

就在你以为最坏的情形已过之际，哈佛逻辑学家纳尔逊·古德曼在 1955 年引爆了第二颗炸弹——一颗即使你承认归纳运转完美也会被击中的炸弹。它被称为新归纳之谜，而它的全部武器只是一个生造的词。

会变蓝的祖母绿：认识 "grue"

定义一个新的颜色谓词[6]，grue（绿蓝）。一个对象被称为 grue，当且仅当它在某个未来日期——比如 2050 年 1 月 1 日——之前被检查过，且是绿色的；或者它在那时尚未被检查过，且是蓝色的。古怪、人造、毫无用处。但看看它的威力。

迄今为止检查过的每一颗祖母绿都是绿色。因此，按定义，它们也都是 grue 的（在 2050 年前被检查，且为绿色）。这意味着你积累下的如山证据，对下面两个假设给予了完全同等的支持：

- H_1：「所有祖母绿都是绿色的。」→预测你 2051 年挖出的下一颗祖母绿是绿色。
- H_2：「所有祖母绿都是 grue 的。」→预测你 2051 年挖出的下一颗祖母绿是蓝色的。

证据无法在二者间裁决，因为每一次观察都同等支持两者。即便承认归纳有效，它也不会告诉你该把哪一种规律性投射到未来。

下方表格比较观察期、2050 年后的预测，以及古德曼关于可投射谓词的教训。

绿色 vs. 绿蓝，投影表

时期	观察到的证据	「全绿」预测	「全绿蓝」预测	启示
2050 年之前	已检查的祖母绿全是绿色。	绿色祖母绿。	绿色祖母绿。	证据同等支持两种描述。
2050 年之后	新观察终于进入分歧区域。	绿色祖母绿。	蓝色祖母绿。	只有越过截止线，现实才能打破平局。
古德曼的要点	仅凭过去的规律性，无法选出可投射的谓词。	投射绿色。	投射绿蓝。	归纳需要关于哪些谓词自然、哪些已扎根的背景习惯。

最明显的反驳——「但 grue 是拼凑的胡话，绿色才是自然的！」——恰恰掉进了陷阱。古德曼的回刺是：从 grue 语言的内部看，绿色才是那个古怪的东西。定义「bleen」（t 之前蓝、 t 之后绿），你就可以把朴素的「绿色」重新定义为「t 之前 grue、 t 之后 bleen」——绿色反而成了滑稽的复合物，grue 倒成了简单的本原。没有哪种「上帝视角」能册封绿色为天然的那一个。古德曼自己的出路是：我们投射的是那些已经扎根的谓词——也就是我们的语言在过去屡试不爽的那些。这很诚实，却也令人泄气：它不是把自然的规律性奠基于自然本身，而是奠基于人类词汇的偶然习惯。休谟说我们的推断依赖习俗；古德曼说，连我们用来推断的概念也依赖习俗。原来深坑之下，还有一层地下室。 [争议/炒作]

第二部分 · 波普尔承认的裂缝

近距离看证伪

正文中波普尔带着一条利落规则登场；同样值得称道的，是他对自己同样毫不留情的审视。他坦承的三个微妙之处，对下游一切影响深远。

第一：划界不是关于意义

波普尔常被与维也纳学圈的逻辑实证主义者（石里克、卡尔纳普，以及他们在英伦的传声筒 A.J. 艾耶尔——其 1936 年出版的 Language, Truth and Logic 曾轰动一时）混为一谈。实证主义者有自己的著名准则——意义的可证实性理论：一个陈述只有在可被经验验证（或按定义为真）时才是有意义的。其余一切——形而上学、神学、伦理学——不是错的，而是字面意义上的废话、「伪陈述」。这对整个哲学分支而言，无异于一台碎木机。

波普尔认为这既傲慢又自相矛盾——可证实性准则本身不可证实，按它自己的规则便属废话。他的观点更尖锐，也更谦逊。可证伪性区分的是科学与非科学，但对意义不发一言。不可证伪的命题完全可以很有意义，往往还很深刻，有时甚至孕育着未来的科学。「每一物体都被其他物体吸引」在成为牛顿定律之前，曾是不可检验的形而上学。划界只是在地图上画线，并不会把线那边的地方付之一炬。忘了这一点，就会把波普尔变成他自己明确拒绝充当的反智庸人。

第二：最大胆的理论恰恰最不可能为真——而这正是关键所在

这是对常识的一次漂亮反转。我们倾向于赞赏与数据严丝合缝的「安全」理论，波普尔赞赏的却正好相反。一个理论禁止得越多——世界能证明它错的方式越多——它的经验内容就越高，碰巧为真的概率反而越低。「爱因斯坦的光线恰好偏折 1.75 角秒」是在走钢丝；「经济受多种因素影响」则是躺在沙发上。一个理论可能恰恰因为几乎什么都没说，才显得概率很高。于是波普尔翻转了奖励标准：科学应当追求大胆、可能性极低、内容丰富的猜想，再让它们经受残酷检验。概率是懦夫优化的目标；可检验性才是科学优化的目标。（且记住这一点——它与我们将在第 4 日遇到的贝叶斯概率最大化图景之间，有一道真正的张力。）

第三：没有基岩——只有沼泽中的桩子

正是这条裂缝，赋予了本附录标题。简略介绍波普尔时，这一点常被跳过。一次证伪需要一个事实来执行——一个「基本陈述」，一份观察报告，比如「指针指向 1.75」。但这些事实从何而来？并非来自纯粹、无理论的观看。每一次观察都渗透着假设：仪器正常工作，光线行为如常，「指针」和「指向」这些词确实切中了世界。因此基本陈述不是自然给定的，而是被我们接受的——通过约定、通过决定、暂时地。波普尔亲口写下这段话，也是他笔下最美的段落之一：

> *客观科学的经验基础因此没有任何「绝对」之处。科学并不立于坚实的基岩之上。它大胆的理论结构，仿佛矗立在沼泽之上……桩子被打下去……却并没有打到任何天然的、「给定」的基础；如果我们不再继续深打，那不是因为已抵达实地，只是觉得桩子已足够牢固，能撑起这座结构——至少暂时如此。*
>
> ——波普尔，《科学发现的逻辑》（1959）

细想这个代价。如果执行证伪的事实本身也要靠约定来接受，那证伪就永远不是口号所承诺的那种干净、绝对的断头台。科学家总可以拒斥基本陈述而保全理论（「仪器出故障了」）。波普尔的辩护是方法论层面的：大家约定一条游戏规则，不要用特设性修补来脱身——不要为了方便就反复重打桩子。这很合理。但请注意，这是我们选择的规则，而不是我们发现的某个事实——这与波普尔反感的库恩「常规科学」图景中的群体判断，其实不无相似。沼泽吞噬的确定性，比教科书版本愿意承认的更多一些。

确证不是真理的首付

还有一条波普尔式的细则，因为人们常搞错。当一个理论经受住严酷检验，波普尔说它得到了确证——但确证绝对不是概率，一个久经检验的理论也不会因此变得「大概是真的」。它只是一份成绩单，记录这个理论经受了多么严厉的打击并存活下来，且仅「暂时」有效。希拉里·普特南提出显而易见的反驳：如果科学从不允许我们把任何理论称为大概可靠，那我们凭什么用最好的理论去造桥、往火星发射探测器？我们显然在依赖它们。波普尔冷峻的回答是：暂时依赖那些经受了严厉检验的东西，但不把它当作大概为真。很多人觉得这答案冷到不能当全貌。

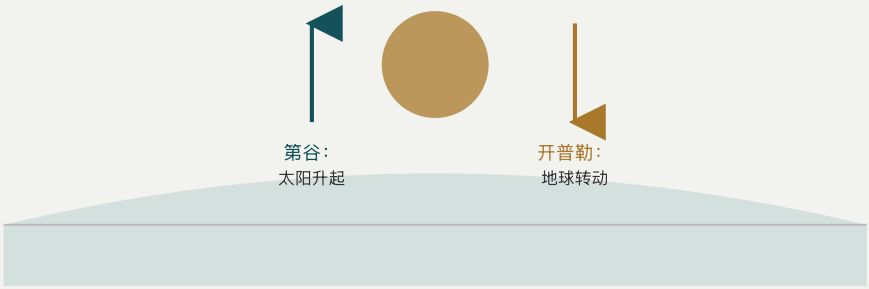
第三部分 · 缺失的中立地带

你们看到的甚至不是同一场日出

波普尔的沼泽已暗示：观察不是基岩。哲学家兼物理学家诺伍德·拉塞尔·汉森在 Patterns of Discovery（1958）中把刀推得更深，提出了一个后来成为口号的论断：观察是负载理论的。他说，「看见比眼球接收到的要多。」你感知到什么，早已被你所相信的东西塑造。

他的思想实验令人难忘。让相信地球静止的第谷·布拉赫，与相信地球旋转的开普勒，在黎明时同站一座山丘。同样的光子击中同样的视网膜；相机也会录下完全相同的画面。

然而——他们看见的是同一回事吗？第谷看见太阳从固定的地平线上升起；开普勒看见太阳纹丝不动，是地平线向下翻滚，才将它显露出来。原始感觉或许相同，但「看见」——那个有意义的、概念层面的「看作」——从头到尾都被理论浸润。



同样的光子，同样的视网膜——两场不同的日出。观察既已负载理论，便没有中立裁判来裁决理论之争。

这是埋在「决定性实验」概念下的一颗静默地雷。证伪主义的图景需要一种中立的观察语言——双方都能接受的事实——来充当竞争理论之间的裁判。汉森（以及后来的库恩，带着他的鸭兔图，还有那个学生——物理学家看到「熟悉的亚核事件记录」之处，他只看到「混乱的碎线条」）暗示：裁判可能在比赛开始前就已经被收买，悄悄穿着某一方的队服。（公平性检查：汉森自己也承认，两次黎明体验中「有某种东西」「对两人是相同的」，所以强主张——他们字面意义上看见了不同东西——确有争议。弱版本是安全的；强版本则仍在争论中。[争议/炒作]）

奎因抽出线头，整件毛衣跟着动

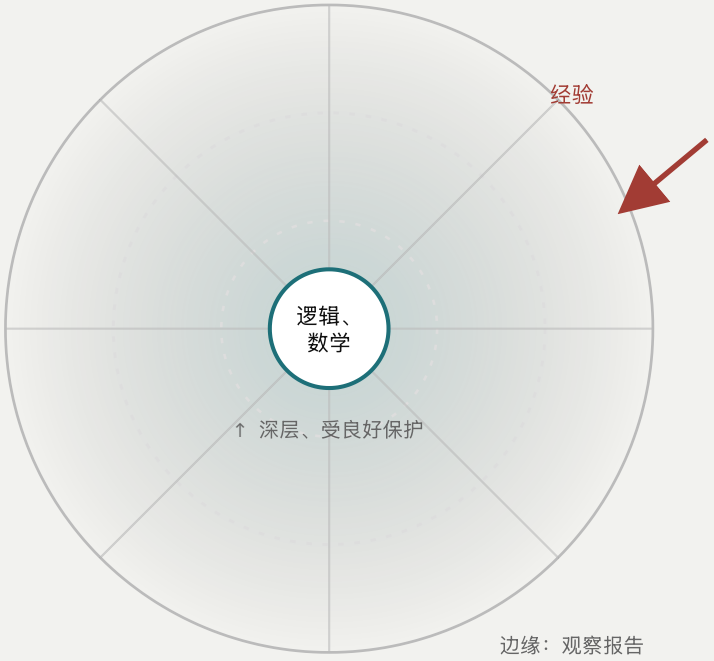
如果说单次观察负载理论，哲学家 W.V.O. 奎因在 1951 年进一步指出，单次检验也负载理论——并据此写成了现代哲学中极具影响力的论文《经验主义的两个教条》。我们在正文中见过它的产物（杜恒-奎因论题：没有假说是被单独检验的）。这里给出的则是它的母体思想，而且更激进。奎因把人类全部知识——从「这里有一只杯子」到逻辑法则——想象成一张巨大的信念之网：

> *我们所谓的全部知识或信念……是一张人造的织物，只在边缘与经验接触……整个科学就像一个力场，其边界条件就是经验。*
>
> ——奎因，《经验主义的两个教条》（1951）

经验只触及这张网的边缘。当冲突发生——某个预测失败——冲击波向内扩散，但由你选择在哪里吸收它。你总可以通过调整系统的其他地方来保护任何你想保护的信念，无论它埋得多深。奎因由此得出两个惊世结论：经验「不是逐个地，而是作为一个整体」与我们的信念相遇；因此——

> *任何陈述都可以在任何情况下被保持为真，只要我们在系统中的其他地方做出足够剧烈的调整……反过来，同理，没有任何陈述是不可修正的。*
>
> ——奎因（1951）

没有任何陈述是不可修正的——逻辑和数学也不例外。（奎因提到，为简化量子力学，有人曾提议修改排中律。）并不存在享有特权的确定性核心；只有一张网，由边缘的经验和我们「尽量少拆」的偏好共同绷紧。这是到目前为止最深的「没有基岩」：连思维法则也没有被钉死。



奎因之网：冲击落在边缘，涟漪向内扩散，但让哪部分让步由你决定。中心总可以保全——代价在别处支付。

劳丹踩下刹车：逻辑上可能 ≠ 理性上合理

如果你感到脚下的地面一直在向「所以什么都行，一切不过是选择」的深渊倾斜——很好，那正是深渊；而拉里·劳丹（没错，就是正文里那位拆迁队长）是把所有人从边缘拽回来的人。在 Demystifying Underdetermination（1990）中，他论证道：人们从奎因那里推导出的惊人结论，其实是由一个糟糕的等式偷运进来的——把逻辑上可能的等同于理性上合理的。

是的，劳丹承认，纯粹演绎逻辑从不迫使唯一的理论选择——你可以不惜一切保全某个信念。但科学从来不是只靠演绎逻辑运转的。它靠的是逻辑加上一整套厚实的扩展性标准——简单性、丰饶性、与既有结果的一致性、预测的战绩。他赞同杜恒的话：「纯逻辑不是我们判断的唯一规则。」你可以把失败归咎于望远镜而非理论，不代表这样做合理；你可以用无数补丁坚持地球是平的，不代表这对理性探究者是一个可活的选项。这张网没有逻辑基岩，但它有理性的张力，而这张力足以做真正的工作。欠决定是真的，却基本无害。差别就在这里：「我无法确定你不是缸中之脑」与「所以一切赌注都作废」。前一句正确；后一句并不成立。[线索]

第四部分 · 给弗洛伊德一个更公平的审判

格伦鲍姆：精神分析不是非科学——而是失败的科学

正文中我们提过，波普尔可能曲解了弗洛伊德。把这一直觉锻造成法医式论证的哲学家，是阿道夫·格伦鲍姆，见其 1984 年的 The Foundations of Psychoanalysis。他的判决比波普尔的更有趣，也更严厉。

波普尔说精神分析不可证伪——它解释一切、不禁止任何东西，所以根本没进入科学的竞技场。格伦鲍姆说：胡说八道——而且这话不是在帮弗洛伊德。弗洛伊德的理论确实提出了可检验的命题。倘若被压抑的同性恋是偏执狂的必要原因，那么对同性恋越宽容的社会，偏执狂就该越少——这是一个真实的、可检验的预测。更核心的是，格伦鲍姆挖出了他所谓的弗洛伊德吻合论证（出自弗洛伊德 1917 年的演讲）：弗洛伊德为自己的方法辩护说，只有那些与患者内在「真实情况」吻合的诠释，才能带来持久的疗效——所以，持久的治疗成功将证实这些诠释的正确性。

这是一桩真正的科学赌注。按格伦鲍姆的解读，这注输了。持久的缓解同样会通过其他疗法出现，甚至不做任何分析也会自行缓解——所以治疗成功不能证明弗洛伊德式诠释是唯一正确的。他还论证，「来自躺椅的证据」已被分析师自身的暗示污染：患者会迎合分析师，生产出理论所预测的记忆与联想。这些数据承受不了弗洛伊德赋予它们的因果重量。格伦鲍姆的结论重新框定了整个划界问题：精神分析不是被安全隔离在竞技场外的非科学——它是走上擂台、然后被击倒的科学。是坏科学，而非非科学。（这是一个确实不同、也可以说更尊重的判决：它认真对待弗洛伊德，认真到肯花力气检验他。[争议/炒作]）这一区分——不可证伪的与已被证伪的——日后在你每一场「X 是不是科学」的争论中都会派上用场。

第五部分 · 危机的发动机房

为什么大多数研究结果被夸大了：真正的数学

正文给你看了残骸：心理学复现研究中只有 36% 重新达到统计显著，效应量减半，权力姿势轰然倒塌。但它没给你看制造这种规模残骸的那台机器。这台机器未必是造假。它是算术——而一旦看见，便再也无法视而不见。三个齿轮咬合在一起：基础率、灵活性和过滤。

齿轮一：基础率陷阱（伊奥安尼迪斯的炸弹）

2005 年，医生兼统计学家约翰·伊奥安尼迪斯发表了 PLoS Medicine 历史上被下载最多、也最具争议的论文之一，标题本身就是引爆装置：《为什么大多数已发表的研究结

果是假的》。他的论证不是修辞，而是一个公式。我们真正关心的是阳性预测值（PPV）：给定一项研究报告了「显著」效应，它为真的概率是多少？它取决于三个数——显著性阈值 α（惯例为 0.05）、研究的统计功效（抓住真实效应的机会），以及最致命的先验比值 R：在一个领域检验的所有假说中，有多大比例本来就是正确的。

最后一个数字是致命的，也是研究者最容易遗忘的。直觉是这样的：假设一个领域检验 1,000 个假说，其中只有 100 个为真（好想法稀少，多数猜测本就错误）。全部以 80% 功效和 5% 阈值来检验。你会正确标出 100 个真效应中的约 80 个。但在 900 个错误假说中，5% 的假阳性率会冒出约 45 个「显著」结果——全是噪音。于是，在你当作发现发表的约 125 个成果中，约 45 个——超过三分之一——是假的。而这还是乐观情形。降低功效，或降低真假设的比例，假发现就会淹没真发现。

下方表格列出三个基础率场景及其阳性预测值。

发现纯度引擎，基础率场景

场景	真实假说	功效	偏倚	发表的阳性结果	PPV
乐观基线	100 / 1,000	80%	0%	80 个真阳性 + 45 个假阳性	64% 为真
低基础率	20 / 1,000	80%	0%	16 个真阳性 + 49 个假阳性	25% 为真
加入偏倚	100 / 1,000	80%	20%	84 个真阳性 + 216 个假阳性	28% 为真

伊奥安尼迪斯的推论直接从这台机器中流出，读起来像复现危机的受灾地图：研究规模越小、真实效应越小、分析灵活性越大、经济利益越重、领域越热门（越多团队竞逐同一问题），任何一项已发表发现为真的概率就越低。这不是愤世嫉俗——这是用不完美的工具检验稀少真理的几何学。[线索]

它并非没有受到挑战，而挑战本身也值得了解。统计学家史蒂文·古德曼和桑德·格林兰（2007）同意其基本精神，却质疑工程细节：模型把每一个显著的 p 值都当作恰好 0.05（丢弃了信息），自行编入了偏倚参数而非测量它们，而那个引人注目的「更多团队 → 更多谬误」的结果，部分也是建模的人为产物。伊奥安尼迪斯回应说核心论点依然站得

住，而且他本人的表格也显示，在良好条件下发现的可信度可达 85%。诚实的结论是：科学假阳性率的精确值确实不确定，且因领域而异；但论证的方向——低基础率加低功效会制造假阳性——很难无视。[争议/炒作]

齿轮二：灵活性——如何「找到」任何东西（披头士实验）

基础率陷阱假设你诚实地在 5% 水平上做检验。真实研究却更松漏。2011 年，三位心理学家——西蒙斯、尼尔森和西蒙森——用一出科学戏剧的杰作展示了它有多漏。他们的论文 False-Positive Psychology 创造了研究者自由度一词：科学家在研究过程中做出的那些微小、看似无辜的选择——何时停止收集数据、剔除哪些异常值、纳入哪些控制变量、比较哪些条件。每个选择单独看都有道理，合在一起却成了一台制造显著性的机器。

为了证明这不是假想，他们对真实的本科生做了一项真实的实验，报告了一个真实的、统计显著的结果：听披头士的 When I'm Sixty-Four 会让人真的变年轻。不是感觉年轻——是实际更年轻。在控制了参与者父亲的年龄后，听这首歌的受试者被计算出的实际年龄（调整后均值 20.1 岁）比听对照曲目的人（21.5 岁）小一岁半，p = .04。这个效应在形而上学上当然不可能。而这正是全部要点。他们所利用的，正是论文自身要审判的那种寻常灵活性：看到数据走向之后，再选择协变量[7]、结果变量、比较方式和停止规则。既然能用一首披头士的歌「证明」衰老可逆，你就能「证明」任何事情。他们提出的解法——公开每一个选择，最好在收集数据之前——正是正文提到的预注册运动的种子。

最令人不安的部分：你不需要作弊

安德鲁·盖尔曼和埃里克·洛肯在 2013 年给了它最锋利的刻画：分岔花园。你可能以为 p 值操纵[8]需要跑 20 个分析，再报告那个「奏效」的。但假设一个诚实的研究者只跑了一个分析，而且事先就有假说——只是他选择的具体检验方式，被数据恰好长成的样子所塑造。如果数据出来的不同，他也会理所当然地换种方式分析。所有那些未被采取的路径，仍然毒化了 p 值，因为 p 值默认假设从来只有一条路。「问题在于，」他们写道，许多潜在的比较是「依赖于数据的」——所以一个完全真诚的科学家，从未有意识地「钓鱼」，仍会滑入假阳性。这就是为什么好意救不了你，改革必须是结构性的。

齿轮三：过滤——文献是幸存者展厅

第三个齿轮发现得最早。早在 1959 年，西奥多·斯特林就注意到一个关于什么能被印出来的致命事实。他调查了四本主要心理学期刊，发现使用显著性检验的文章中，294 篇里有 286 篇——惊人的 97.28%——拒绝了零假设[9]，报告了阳性结果。而且他调查的研究中，没有一项是复现研究。期刊只发表赢家。零结果死在文件抽屉里——罗伯特·罗森塔

尔在 1979 年将这个问题形式化为文件抽屉问题（并用「失效安全 N」来量化：需要多少被埋藏的零结果，才能推翻一个已发表的效应？）。

把三个齿轮叠在一起，危机便被过度决定了。大多数被检验的假说本为错误（基础率）→ 灵活性把假的也煮成「显著」（分岔路径）→ 只有显著的才能见刊（文件抽屉），而且发表后往往被重新包装成一开始就预测到的——诺伯特·克尔在 1998 年命名的罪：HARKing——在结果已知后才提出假说，它悄悄「把 I 类错误[10]翻译成了理论」。已发表的文献不是真相的地图。它是残酷而隐形筛选之后的幸存者展厅——暗合演化主线的回响，也回荡着第 1 日盖梯尔的忧虑：结果「正确」，但原因与真相毫不相干。

统计学家的判决 [已确立]

p 值不是什么

2016 年，美国统计协会（ASA）在其 177 年历史上首次对一项特定统计实践——p 值[11]——发布正式公开警告（Wasserstein & Lazar, The American Statistician）。美国该领域的主要专业协会打破沉默，这本身就说明问题已严重到了什么地步。它的六条原则值得贴在显眼处，因为危机中的许多误用都至少违反了其中一条：

- p 值衡量的是数据与某个模型的不兼容程度——仅此而已。
- 它不是假说为真的概率，也不是你的结果「由偶然造成」的概率。
- 结论永远不应取决于 p 是否跨过 0.05 这条「明线」。
- 正确的推断要求完整的报告和透明度（不隐藏分岔路径）。
- p 值不说明效应的大小或重要性。
- 单凭它本身，是衡量假说证据的拙劣指标。

最常见的误解——p = 0.05 意味着「95% 的可能性发现是真的」——彻头彻尾地错了，上面那台基础率引擎就是原因：一个发现为真的概率，压倒性地取决于真假设有多稀少，而 p 值对此一无所知。2019 年的一份后续声明走得更远，一些统计学家呼吁该领域彻底废弃「统计显著」这个说法。改革尚未完成。[线索]

第六部分 · 定义了一个领域的决斗

伦敦，1965 年 7 月：科学哲学界的一场著名交锋

正文中的四位主角——波普尔、库恩、拉卡托斯、费耶阿本德——并非在教科书里礼貌排队的抽象符号。他们是活生生的对手。1965 年 7 月，他们（以及其他人）在伦敦贝德福德学院的一次国际研讨会上当面交锋。论文集因各位参战者迟迟不肯停笔而拖延多年，最终在 1970 年以 Criticism and the Growth of Knowledge 之名出版——该领域最富火药味的著作之一。全书以库恩开篇，被接连的回复轮番轰炸，又以库恩的反击收尾。

断层线十分尖锐。波普尔指责库恩的「常规科学」——在不受质疑的范式内埋头解题——根本不是科学，而是一种智识从众，甚至是「暴民心理学」：正是证伪主义想要废除的那种不加批判的教条主义。库恩反击说，波普尔把科学中罕见而激动人心的革命时刻，误认成了科学的日常实质——日常科学压倒性地保守、受范式约束——而这是一个特征，正是它让领域能够积累深刻成果，而不必永远在重审自己的地基。

一本书里的二十一个范式

最尖锐的一击来自出人意料的方向。语言学家玛格丽特·马斯特曼大体同情库恩，却坐下来数了数他使用核心词的方式——结果发现库恩至少以 21 种不同含义使用「范式」一词，她将其归为形而上学的、社会学的和具体的「人工制品」三类。她的评价是一把双刃剑：库恩的书「科学上洞明，哲学上晦涩」。这是毁灭性的批评，同时也是一次平反——概念虽然含混，但显然触及了某些真实的东西。库恩后来承认了这一点，花了大半职业生涯试图更精确地说清本意。

库恩有两个更深层想法值得从漫画式简化中抢救出来，因为它们都被惯常地夸大了：

- 库恩损失。科学进步并非纯粹累积。当一个范式倒下，继任者可能会丢失旧范式曾拥有的某些解释成就——燃素化学就解释过早期氧气化学最初无法解释的一些现象。进步是真实的，却也粗糙；我们用一组已解谜题，换取另一组更大、不同的谜题，有时还会在路上掉落几个。（它在多大程度上威胁实在论仍有争议——大多数有记录的损失都是轶事性的，而非定量的。）
- 世界变化论题。库恩最臭名昭著的一句话是，革命之后「科学家此后工作在一个不同的世界中」。但精确地读他，他其实很谨慎——他写的是「我们可能想要说」世界变了，这只是在铺垫一种说法，并非声称现实本身在重新洗牌。他的晚年一直在回缩最激进的解读，退守到一种窄义的分类不可通约性（只是互锁的技术词汇体系发生了转换，而非整个现实），并坚持——反对他的相对主义拥趸——「世界不是被发明或建构出来的」。传说中的库恩，比书页上的库恩更疯狂。

而费耶阿本德，那位所谓的破坏者，在挑衅外表下其实有一颗建设性的心。他真正的提案是多元主义：一个健康的科学应当最大化竞争理论的数量，而非强制推行共识。两条口号承载着它。增殖原则：积极发明并捍卫与当朝理论相矛盾的理论；反归纳：刻意发展与哪怕已被确凿确认的事实不一致的想法——因为，正如汉森警告过的，观察负载理论，所以唯一能揭示你当前视角中隐性假设的方法，就是透过竞争者的镜头去看世界。在后来的序言与回复中，他强调「什么都行」不是他宣扬的信条，而是「一个理性主义者仔细审视历史时发出的惊恐感叹」。他那个看似怪物的论证，原来支持的是把智识多样性作为发现的引擎——这与本附录一直在走向的方向惊人地接近。

贯穿线

没有底，但它照样运转

退后一步看，整个附录其实只拉了一个长音。休谟：对明天的期待没有逻辑上的正当理由。古德曼：连我们的概念都不安全。波普尔，坦诚地说：证伪所依赖的事实，建立在约定之上——沼泽中的桩子。汉森：连你看见的东西都被理论扭曲了。奎因：整张网，包括逻辑在内，都是悬浮的——没有任何东西不可修正。而复现危机，就是这些抽象变得可怕而具体的时刻：当你真正审计某些文献时，三分之一或更多的高调发现无法通过严格复现，而这恰好是基础率与分岔路径的数学所预言的。

如果你以为寓意是绝望，那可以理解。但恰恰相反——劳丹给了我们钥匙：逻辑上可能的不是合理的。科学没有地基，也不需要地基。它的运转方式像一座城市——底部没有哪块不可撼动的石头，只有无数相互支撑的结构，不断被检查，偶尔被宣判拆除重建；整体之所以立着，不是因为建在岩石上，而是因为它自我纠错的速度比崩塌更快。复现危机不是沼泽吞噬科学，而是科学公开地打入新桩——因为它注意到旧的正在变软。那不是方法的失败，那正是方法。

正因如此，接下来 178 日唯一理智的姿态，就是我们在第 1 日建立的：用刻度盘而不是开关来持有每一个信念。按证据比例调整信心，留一点余地给「我可能错了」，对最博眼球的声明保持最大怀疑。这一切的下面没有基岩。学着在桩子上建造吧。

◆ 本附录三句话概括

核心洞见

在科学方法之下往下掘，你会发现没有地基——没有不循环的归纳辩护（休谟），没有安全的概念（古德曼的 grue），没有理论中立的观察（汉森），没有不可修正的信念（奎因），只有波普尔所谓「打入沼泽的桩子」。而复现危机是经验性的警示信号，背后有一台数学引擎驱动：基础率 × 灵活性 × 过滤。

最佳类比

建在无底沼泽上的桩基建筑——桩子只打到「暂时够牢」为止——配上那首「证明」听众变年轻的披头士歌曲，它展示了寻常的灵活性可以制造出任何结果。

活的争议

无基础状态是否会滑向「怎么都行」（奎因之网），还是能被理性标准驯服（劳丹：逻辑上可能 ≠ 理性上合理）——以及，在经验层面，科学的真实假阳性率究竟是多少（伊奥安尼迪斯 vs. 古德曼与格林兰），这个问题仍未定论且因领域而异。

此处的线索 › 信息（p 值、基础率、证据能承载什么与不能承载什么）· 演化（文献作为幸运阳性结果的幸存者展厅）· 计算与涌现（科学作为一个没有中心地基的自我纠错系统，靠相互张力支撑自身）——把第 2 日正文的线索再往下一层延伸。

来源

来源与延伸阅读

1. Hume, D. (1748). An Enquiry Concerning Human Understanding, §IV–V. ——怀疑论的解答；习俗/习惯作为推断的基础。见 SEP, "The Problem of Induction."

2. Goodman, N. (1955). Fact, Fiction, and Forecast. Harvard University Press. —— 新归纳之谜（"grue"）；可投射性与扎根性。见 SEP, "Nelson Goodman."

3. Strawson, P. F. (1952). Introduction to Logical Theory, ch. 9 ——归纳问题的"消解"。 Reichenbach, H. (1938). Experience and Prediction ——务实的辩护。

4. Ayer, A. J. (1936). Language, Truth and Logic. ——逻辑实证主义与证实主义在英文世界的推广。见 SEP, "Logical Empiricism" 与 SEP, "Alfred Jules Ayer."

5. Popper, K. (1959). The Logic of Scientific Discovery (orig. 1934). ——可证伪性的程度；"沼泽中的桩子"段落（§30）；确证 ≠ 概率；划界 ≠ 意义。见 SEP, "Karl Popper."

6. Putnam, H. (1974). "The 'Corroboration' of Theories," in The Philosophy of Karl Popper. ——普特南对波普尔的反驳：若按其理论，科学将无法论证我们为何能依赖理论。

7. Hanson, N. R. (1958). Patterns of Discovery. Cambridge University Press. ——观察的理论负载；黎明时第谷 vs. 开普勒。

8. Quine, W. V. O. (1951). "Two Dogmas of Empiricism." The Philosophical Review 60(1): 20–43. ——信念之网；"没有任何陈述是不可修正的"；确认整体论。 全文

9. Laudan, L. (1990). "Demystifying Underdetermination," in Minnesota Studies in the Philosophy of Science 14: 267–297. ——逻辑上可能 ≠ 理性上合理；欠决定的限度。见 SEP, "Underdetermination."

10. Grünbaum, A. (1984). The Foundations of Psychoanalysis: A Philosophical Critique. University of California Press. ——吻合论证；精神分析是可证伪但失败的科学（坏科学，而非非科学）。

11. Ioannidis, J. P. A. (2005). "Why most published research findings are false." PLoS Medicine 2(8): e124. ——PPV 模型；先验比值、功效、偏倚。 plos.org

12. Goodman, S. & Greenland, S. (2007). "Why most published research findings are false: problems in the analysis." PLoS Medicine 4(4): e168 ——主要的统计学批评；附伊奥安尼迪斯的回复 (e215)。

13. Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). "False-Positive Psychology." Psychological Science 22(11): 1359–1366. ——研究者自由度；"When I'm Sixty-Four" 实验 (p = .04)。

14. Gelman, A. & Loken, E. (2014). "The Statistical Crisis in Science" ("The garden of forking paths," 2013 工作论文). American Scientist 102(6): 460. ——无需有意识 p 值操纵即可产生的假阳性。 PDF

15. Kerr, N. L. (1998). "HARKing: Hypothesizing After the Results are Known." Personality and Social Psychology Review 2(3): 196–217.

16. Sterling, T. D. (1959). "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa." JASA 54(285): 30–34. ——294 篇中的 286 篇 (97.28%) 显著性检验文章拒绝了零假设；没有一篇是复现研究。

17. Rosenthal, R. (1979). "The file drawer problem and tolerance for null results." Psychological Bulletin 86(3): 638–641. ——发表偏倚；"失效安全 N"。

18. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values: Context, Process, and Purpose." The American Statistician 70(2): 129–133. ——六条原则；2019 年的后续声明呼吁废除"统计显著"。 tandfonline

19. Lakatos, I. & Musgrave, A. (eds.) (1970). Criticism and the Growth of Knowledge. Cambridge University Press. ——1965 年贝德福德学院研讨会论文集；含 Kuhn、 Popper、 Lakatos、Feyerabend 与 Masterman 的"The Nature of a Paradigm"（"范式"的 21 种含义）。

20. Kuhn, T. S. (1962/1970). The Structure of Scientific Revolutions, ch. X & Postscript. ——库恩损失；世界变化论题（"我们可能想要说……"）；后期的分类不可通约性。见 SEP, "Incommensurability."

21. Feyerabend, P. (1975). Against Method. ——多元主义、增殖、反归纳；"什么都行"作为"一个理性主义者的惊恐感叹"。见 SEP, "Paul Feyerabend."

明日 → 第 03 日

逻辑与有效推理

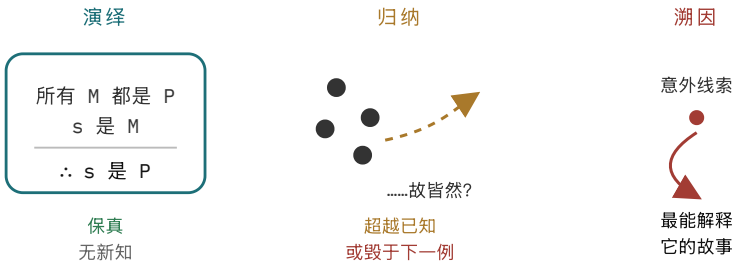
今日我们频频倚仗「有效」、「由此推出」、「矛盾」等词——但使论证真正成立的规则究竟是什么？明日我们将深入逻辑本身：演绎（能保真，却不能凭空增加新信息）、归纳（休谟留下的伤口）与溯因（像侦探一样选择最佳解释）。我们将遇见日常欺骗我们的谬误，追问逻辑是被发现的还是被发明的，并抵达前沿——在那里，机器如今检验着人类头脑无法完全容纳的证明。这是此前所有讨论赖以成立的逻辑底座。

第 02 日终 · 还有 178 日等待深入

模块一 · 知识与推理的根基 · 第 003 日 / 180

逻辑与有效推理

从已知通往未知有三条路，只有一条真正保险。



三种推理引擎——保证自左至右递减

演绎把你锁在安全的房间里；归纳与溯因送你出门——代价是确定性。

一个陌生人走进伦敦的诊室。不过片刻，夏洛克·福尔摩斯便断言：此人是一名刚从阿富汗归来的退役军医。华生愕然。肤色黝黑，手腕却苍白——那是在海外晒出的，并非海滨日光。手臂僵直，是战伤；面容憔悴，显露出艰辛与热病。福尔摩斯称之为「演绎」，这个词已随他流传了一个多世纪。但他用错了词。福尔摩斯所施展的——也正是令他不朽的技艺——并非演绎。它更谦逊、更冒险，也更具创造力。

这个误称是再好不过的入口，因为今日全篇围绕一个几乎人人混淆的区分：推理不止一种，而它们给出的保证并不相同。有些推理天衣无缝——只要承认前提，结论便无处可逃；另一些则丰硕却可错——它们越过证据，可能被明日的意外推翻。把前者当后者、或把后者当前者，是人类错误中惊人比例的根源。所以我们要把界线画清楚。

前两日，我们都在推理的外围徘徊。在第 1 日，我们追问真信念何以成为知识——并撞上阿格里帕三难：任何理由要么无穷追问，要么陷入循环论证，要么在某处武断止步。到第 2 日，休谟的归纳问题表明，再多观察也无法证明一条普遍定律，因此波普尔告诉我们应去证伪而非证实。今日我们打开引擎本身。先前两个谜题其实都关乎两种特定推理模式的边界；如今我们为三者命名，看逻辑如何蜕变为数学，再跟随它抵达本课程最奇异的前沿——以零容错的精确度核查证明的机器。今日最明亮的线索，是计算。

佩吉特笔下的福尔摩斯成了「演绎」的公众形象；然而那些著名的诊断式跳跃，通常先是溯因：线索在前，最佳解释在后。

模型

三种引擎，三种担保

如果今日只能记住一件事，就记住这个三分法。推理不是一种活动，而是三种——按承诺的大小排列。

演绎是保真引擎。结论早已蕴涵在前提之中；有效的演绎只是把它展开。承认所有人都会死，且苏格拉底是人，你便无法回避「苏格拉底会死」——否认它便是自相矛盾。这种安全的代价，是演绎具有非扩展性：它从不向你揭示关于世界的真正新知，只是重新排列你已拥有的东西。数学是演绎艺术被推至极限的形态，这也正是数学家何以如此笃定——以及他们的确定性为何永远无法回答关于这个宇宙的任何一个问题。

归纳是概括引擎。你见过太阳升起一万次，便推断它明日仍将升起。直到 1697 年以前，所有人记录下的天鹅都是白色，于是「所有天鹅皆白」看似牢不可破。归纳是扩展性的：它增添内容，将已有的案例推向未知。正因如此，它不保真。这是第 2 日休谟埋下的炸弹，至今仍在滴答作响：任何有限次的观察，都无法在逻辑上担保下一次。归纳是经验知识真正成长的方式，但它不提供逻辑保证。

溯因是解释引擎——也是多数人从未学过名称的那一种。你遇到一个令人惊讶的事实，于是寻找一个假设：若它为真，惊讶便会消散。美国博学家查尔斯·桑德斯·皮尔士（Charles Sanders Peirce，1839–1914）将它单独提出，视为唯一真正具有创造性的模式：它不是检验或展开旧观念，而是生成新观念。「科学的每一块进步之板，最初都是由溯因独自铺就的。」他写道。演绎与归纳处理你已有的假设；溯因则回答假设最初从何而来。

回到福尔摩斯。黝黑肤色、僵直手臂、憔悴面容——这些都是令人惊讶的事实；福尔摩斯一跃而至最能同时解释它们的假设：一名从炎热战场归来的战伤军医。但请注意，这一跃并无担保。此人也可能是个演员，夏天在摩洛哥度假，打网球时扭伤了肩膀。福尔摩斯的结论是最佳解释，而非唯一解释——这正是溯因的标志——而非演绎。（这一点在第 4 日还会回来，届时我们将用概率把「最佳解释」变得精确。）

一个伟大误称的形态

福尔摩斯并不孤单。我们说医生「诊断」——那是溯因，从症状推理到最可能产生它们的疾病。听发动机的技工、勘查犯罪现场的侦探、盯着反常读数的科学家：他们都在溯因，都在跃向那个能把奇怪变得平常的解释。甚至你读到的这句话也依赖它——你推断这些文字背后有一颗心智，因为这是它们有序排列的最佳解释，而非某个定理强迫你如此推断。溯因是我们游于其中的水；我们只是很少叫出它的名字。

人人都容易混淆的区分

有效并不等于为真

在演绎引擎内部，住着整个逻辑中最易被误解的观念；厘清它，比背诵一打谬误更能切中要害。那就是有效性与健全性的区别。

一个论证是有效的，当且仅当它的形式保证：前提为真时，结论必为真。有效性是形状的性质，而非内容的性质——它只问论证的骨架，不问骨架里装了什么。《互联网哲学百科全书》表述得干净利落：一个论证有效，「当且仅当它的形式使得前提为真而结论为

假成为不可能」。而健全性要求更多——一个论证是健全的，仅当它既有效，且所有前提实际为真。

真正容易令人失足的是这一点：一个有效论证完全可能导出一个荒诞的假结论。请看：

> *所有鸟都会飞。企鹅是鸟。因此，企鹅会飞。*

形式毫无瑕疵——「所有 M 都是 P；s 是 M；因此 s 是 P」，正是「苏格拉底会死」那一例所套用的模子。若前提为真，结论就不得不跟随。所以这个论证完全有效。但它也显然不可靠，因为第一个前提是假的：并非所有鸟都会飞。有效性只认证管道的结构；健全性还要追问管中流淌的是否为清水。一个有效却不健全的论证，就像一条做工完美的管道，输送的却是污水。

这可不是钻牛角尖。它是归谬法——数学中最锋利的工具之一——背后的工作原理：要证明某前提为假，就先假设它，有效地推理到一个你已知道为假的结论，于是假结论便逆流而上，反证前提为假。整个技巧恰恰依赖一个有效论证故意产出假结论。有效性是舟，真理是货；学会分别追踪二者，你读任何论证时都会少一层迷雾。

当形式破裂时

藏在每个「如果」里的两种谬误

若有效形式是安全路径，谬误便是伪装成同一路径的陷阱。其中最危险的一批藏在条件推理——「若 P，则 Q」形式的命题——之中，因为无效式与有效式往往只有一步之遥。

两个有效招式是老朋友。肯定前件式：若 P 则 Q；P 真；故 Q。否定后件式：若 P 则 Q；Q 假；故 P 假。两者滴水不漏。现在轮到它们那对危险的孪生冒牌货登场。

肯定后件的推法是：若 P 则 Q；Q 真；故 P。它抓错了箭头方向。「若某人住在圣迭戈，他就住在加利福尼亚。Joe 住在加利福尼亚。因此 Joe 住在圣迭戈。」但加利福尼亚很大；Joe 完全可能在萨克拉门托。结论可能为真，这正是该谬误如此诱人的原因——它有时碰巧命中正确答案——而一个通过有缺陷的论证到达的真结论，正是第 1 日那个盖梯尔陷阱穿上了逻辑学家的外衣。

否定前件是它的镜像：若 P 则 Q；P 假；故 Q 假。「如果下雨，地面会湿。没下雨。所以地面不湿。」但别忘了洒水器、爆裂的水管、打翻的水桶。排除一个原因，并不等于排除结果本身；同一结果完全可以有多条来路。

一个教学经典例子能把结构刻进记忆：若一只动物是狗，它就有四条腿。这只动物有四条腿。因此它是狗。猫、马，甚至桌子都会抗议。这种荒谬正是关键——它与圣迭戈例子共用同一种破碎形式，只是把荒诞感放大，让齿轮滑脱清晰可见。（欧仁·尤内斯库在他的戏剧《犀牛》中整整一幕都建立在这个谬误之上：一位逻辑学家庄严地证明，一只有四条腿的猫必定是狗。）

这些是形式谬误——骨架断裂。它们的近亲，非形式谬误，缺陷不在形式而在内容：post hoc ergo propter hoc（公鸡打鸣，太阳升起，因此公鸡召唤了黎明）、人身攻击、悄悄偷换词义的歧义。形式谬误靠检查骨架便可识破；非形式谬误则要读清文字实际在做什么。

条件论证形式表

形式	模式	判定	理由
肯定前件式	若 P 则 Q；P；故 Q	有效	肯定充分条件，结论便逃不掉。
否定后件式	若 P 则 Q；非-Q；故非-P	有效	若 Q 必随 P 而来，则 Q 不在场便可排除 P。
肯定后件	若 P 则 Q；Q；故 P	无效	Q 可能有别的原因：Joe 可以住在加利福尼亚，却不住在圣迭戈。
否定前件	若 P 则 Q；非-P；故非-Q	无效	排除一个充分原因，不等于排除 Q 的所有来路：洒水器仍可打湿地面。

脉络

逻辑如何变成数学

你正在使用的这套机制有着深远的历史，并最终转向一个出人意料的方向：在二十三个世纪里，对好论证的研究慢慢变成了一门代数的分支。这个故事有四座里程碑。

亚里士多德（公元前 4 世纪）在《前分析篇》中建立了第一个形式系统。他的天才在于以字母充当占位符——「所有 A 是 B」——从而研究脱离内容的论证形式。这是词项逻辑：它处理「人」「会死」这类词项之间的关系。中世纪逻辑学家以助记名兴致勃勃地编录有效的三段论式——Barbara、Celarent、Darii。这些名字不是人名，而是密码：元音标记命题类型，A 表示「所有 S 都是 P」，E 表示「没有 S 是 P」，I 表示「有些 S 是 P」，O 表示「有些 S 不是 P」。因此 Barbara 是 AAA，Celarent 是 EAE，Darii 是 AII；例如 Barbara 意味着：所有 M 都是 P；所有 S 都是 M；所以所有 S 都是 P。近两千年间，这就是逻辑。

斯多葛学派，尤其是克律西波斯（约公元前 279–206 年），建立了第二条与之平行的逻辑，历史几乎让它失传。亚里士多德处理词项，斯多葛学派则用我们日常仍在使用的联结词处理整个命题：如果……那么、并且、或者、并非。克律西波斯列出五条「不可证明式」——基本推理图式，第一条（「若第一，则第二；但第一；故第二」）正是肯定前件式。这便是命题逻辑，也是每一块计算机芯片内部逻辑的远古源头。斯多葛学派很可能已经对联结词有了真值函项[1]的理解——通过组成部分的真假判断整体的真假——这比后人重新发现早了两千年。20 世纪逻辑学家扬·武卡谢维奇曾令学者惊讶地主张，斯多葛逻辑并非亚里士多德的穷亲戚，而是「同等级的成就」。随后它被掩埋多年，亚里士多德独尊——这提醒我们，思想史并非一场整齐的接力赛。

乔治·布尔把两个传统推上了新轨道。1854 年，他在《思维规律的研究》中做了一件大胆的事：把逻辑推理当作计算。令 1 为全域，0 为空无；乘法即「且」，加法即「或」。骤然之间，有效推理的规律看上去如同代数定律。「我们不应再把逻辑与形而上学相联系，」布尔宣称，「而应把逻辑与数学相联系。」他的书销量平平，同代人也大惑不解。直到几十年后，1937 年克劳德·香农注意到布尔的二值代数精确描述了电路开关，布尔代数才成为数字逻辑名副其实的基础。你此刻用来阅读这段文字的设备中，每一个 AND 门都是克律西波斯的一句话在硅中的实现。

戈特洛布·弗雷格完成了自亚里士多德以来最大的跳跃。他那薄薄一卷、令人生畏的《概念文字》（Begriffsschrift，1879）引入了量词——形式的「所有」（∀）与「存在」（∃）——以及谓词逻辑。亚里士多德的词项逻辑会被「马皆动物，故马头皆动物头」这类论证难倒；弗雷格的机制不仅能处理它，而且远不止此——它把命题解析为以个体为变元的函数。它常被称为符号逻辑史上最伟大的一部著作。但悲剧性的尾声随之而来：弗雷格梦想把全部算术还原为纯粹逻辑，就在第二卷即将付梓之际，年轻的伯特兰·罗素寄来一封信，里面藏着一个悖论——所有不包含自身的集合构成的集合：它是否包含自身？无论回答「是」或「否」，都会自相矛盾。弗雷格宏大的基础工程由此崩裂。但他的逻辑在废墟中幸存，成为我们今天仍在讲授的现代符号逻辑。（那个悖论的幽灵，以及它所暗示的边界，将在第 28 日重新浮现；届时哥德尔将证明，没有任何形式系统能满足数学家曾怀有的全部希望。）

辩论

逻辑是发现还是发明?

这里有一个听起来像沙龙游戏、实则直抵根本的问题。那些基岩般的定律——同一律（A 是 A）、矛盾律（A 与非-A 不能同真）、排中律（A 或非-A，没有第三种）——看似无可回避。但它们究竟栖居何处？是实在的特征，即使心智不存在也编织在宇宙之中？是思维的特征，任何思考者都无法逃避的语法？还是人类的约定——真实且具约束力，但终究是被选择出来的，犹如象棋规则？

逻辑实在论

被发现

定律是客观的、独立于心智的世界结构。我们并不立法规定矛盾律，正如我们并不立法规定素数——我们只是发现它。逻辑是从实在中读出的。

心理主义

思维规律

定律描述心智必须如何运作——实为心理学的一个分支。弗雷格与胡塞尔猛烈抨击这一点：逻辑真理是精确且先验的[2]，而心理学是经验且模糊的。

约定主义

被发明

定律是我们因有用而采纳的约定——一旦选定便具约束力，但并非由宇宙降下。奇怪的是，尽管它与道德反实在论渊源甚深，这个立场却很少有充分发展的版本。

可修正性

经验的?

奎因与普特南提出了激进的想法：即便逻辑也可能因经验理由而被修正——量子力学可能把我们推向非经典逻辑，恰如相对论曾把我们推向非欧几何。

最后一个方框，正把问题引向今日的前沿。历史上大部分时间里，「思维规律」似乎不可触碰——质疑它们仿佛锯断自己正坐着的树枝。但二十世纪产生了严谨且可运作的替代逻辑，它们悄然放弃某条神圣定律，却依旧运转。一旦你看到这些替代逻辑确实能承担实际工作，那个宏大的形而上学问题便会软化成一个更实际、也更耐人寻味的问题：不是「哪种逻辑为真？」而是「对这项工作来说，哪种逻辑才是合适的工具？」下面就来看这些替代者。

前沿 · 2026

三条活跃前沿，以及一道炒作过滤网

本课程每日都以前沿研究收尾，每条主张都标明了它能承受多少重量。逻辑的前沿出奇地具体：它运行在真实计算机上，核查真实证明，并且最近与人工智能发生碰撞——这要求我们擦亮眼睛。

前沿 01 [已确立]

故意打破规则的那些逻辑

「经典」逻辑并非唯一一致的选项；它只是更广阔的逻辑图景中一个已站稳脚跟的位置，每一套替代逻辑都放弃了大多数人以为不可动摇的某条定律。

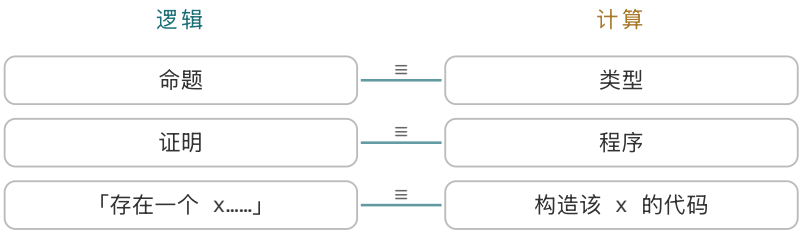
直觉主义逻辑放弃了排中律。它由 L. E. J. 布劳威尔开创，1920–30 年代由阿伦德·海廷形式化，坚持一条陈述只有在你能构造出其证明时才算为真。你不能凭空断言「A 或非-A」——你必须证明其中一边。一个例子能尖锐地说明动机：排中律会让你轻松断言，对任何计算机程序，「它停机或不停机」——然而（我们将在第 27 日看到）不存在判定停机的通用方法，因此没有构造支撑这一断言。直觉主义说：那就别断言。这听起来像是哲学洁癖，直到你发现经由一种美得值得专设一框的对应关系，它竟通向计算机科学的心脏。

次协调逻辑放弃了爆炸原则。在经典逻辑中，单个矛盾是灾难性的：从「P 且非-P」你可以推出任何东西（原则 ex contradictione quodlibet，「由矛盾可得任意结论」）——一处不一致，整座系统便付之一炬。次协调逻辑拒绝这一点，让你即便某些矛盾潜入，仍能合理推理——这对大型数据库、法典，以及任何局部不一致却仍有整体价值的信息集都很有用。更强硬的哲学表亲，双面真理论——格雷厄姆·普里斯特认为有些矛盾实际为真，如说谎者语句「这句话是假的」——则争议大得多。务必区分二者：你可以采纳次协调逻辑（关于爆炸原则的技术选择）而不成为双面真理论者（关于真实矛盾的本体论主张）。前者是工具；后者是世界观。

模糊逻辑完全放弃了二值限制。1965 年，卢菲特·扎德让真值在 0 到 1 的连续区间滑动，以刻画模糊性——「水是温的」为 0.7 真——建立在 1920 年代扬·武卡谢维奇的多值逻辑之上。它运行在控制系统与家电中。而模态逻辑——关于必然与可能（□ 与 ◇）的逻辑——以及经过精心选择的时态逻辑，支撑着硬件与软件的形式验证：某些具体片段既能表达有用性质，又足够克制，可以保留模型检查[3]所需的可判定性[4]。这些不是博物馆中的藏品，而是现代技术世界实际运转的逻辑。

桥梁 · 命题即类型

直觉主义逻辑之所以重要的根本原因，是柯里-霍华德对应：在适当的形式系统中，命题对应类型，证明对应程序。证明一个定理，可以被看作构造一个居于相应类型之中的程序式对象——反过来也一样。



这就是为什么下面若干证明助手建立在类型论基础之上——也是为何逻辑与计算，我们五条线索之二，并非彼此相邻的邻域，而是同一片疆域的两面视图。（将在第 27–29 日继续展开。）

前沿 02 [已确立]

零容错的证明：证明助手的崛起

亚里士多德的梦想是一条紧密到无人能够怀疑的推理链。二十三个世纪后，这个梦想有了软件化身。证明助手是一种程序，其中每一步逻辑都必须通过机器核查；没有任何一步能凭权威、直觉或一句「显然」而蒙混过关。主流系统包括 Lean（现为 Lean 4）、Rocq（原名为 Coq，2025 年更名）、 Agda 与 Isabelle/HOL。Lean、 Rocq 与 Agda 属于类型论家族 [5]；Isabelle/HOL 则建立在经典高阶逻辑之上。目标相同，基础不同。

Lean 的社区共建数学库 mathlib，是世界上最大的统一数学形式化库 [6] 之一：超过 278,000 条定理与 132,000 个定义——2026 年 6 月统计时如此，且仍在增长——覆盖了某著名「形式化这些」挑战清单上 100 个问题中的 84 个。这不是玩具。看看它已验证的成果：

2022 · 已完成

液体张量实验。2020 年 12 月，菲尔兹奖得主彼得·朔尔策向全世界发出挑战，要求验证其「凝聚数学」中一个他本人都不太确定的定理。约翰·科默兰与亚当·托帕兹带领的团队在 Lean 中完成了验证，于 2022 年 7 月 14 日完成。一位一线数学家

借助机器，获得对一份复杂到人类审稿人难以安心核查的证明的信心——这正是关键所在。

2023 · 三周内完成

多项式弗雷曼–鲁萨猜想。蒂姆·高尔斯、本·格林、弗雷迪·曼纳斯与陶哲轩发表这一加性组合学结果的证明数日后，陶哲轩启动了一个 Lean 项目来形式化它——三周后便宣布依赖图「被一片可爱的绿色完全覆盖」。形式化几乎与研究同步推进。

2024–25 · 已完成

等式理论项目。陶哲轩的合作实验（2024 年 9 月启动）旨在判定 4,694 条代数定律之间的蕴涵关系——若把每条定律对自身的平凡蕴涵也算入，共有 22,033,636 个有序对；若只数非平凡图边，则为 22,028,942 条——结合人类证明、自动证明器、AI 与 Lean 验证，50 余位贡献者，在 200 多天内完成了工作。这是一种大规模协作、机器核查数学的新范式。

2024–2029 · 进行中

费马大定理。凯文·巴扎德由 EPSRC 资助的项目（2024 年 4 月启动，伦敦帝国理工学院）旨在形式化 FLT——并非怀尔斯的原始证明，而是一条现代路线。巴扎德「谨慎乐观」地认为自己能把它归约到 1980 年代已知的结果，但坦率承认整个项目「至少需要 5 年」。尚未完成——最诚实也准确的说法是：它仍在进行中，也是那 100 个挑战问题中尚未闭合的最后一项。

这种确定性已从纯数学延伸到生命所依赖的系统。Rocq 中被证明正确的 C 编译器 [7] CompCert；一项著名的编译器查错研究耗费约六个 CPU 年，试图诱使它生成错误代码，却一无所获——「我们测试过的唯一一个 Csmith 无法找到错误代码的编译器」——同时在 GCC 与 LLVM 中找出了大量 bug。seL4 是第一个在 Isabelle/HOL 中拥有完整机器检查功能性正确性 [8] 证明的操作系统微内核 [9]：在其明确列出的假设下，C 实现细化了形式规格，因此整类崩溃与不安全行为不是靠希望避免，而是被定理排除。这些不是普通承诺，而是关于软件的有条件定理。这就是逻辑的机械化所能做的事——而且它已确立。

前沿 03 [已确立] [争议/炒作]

当 AI 遇见证明核查器

最新、也最被喧嚣包围的前沿，是机器学习与形式证明的碰撞——此处正是前沿校准器该上场的时候，因为标题与事实之间已经出现了漂移。

先看真正的里程碑。2024 年 7 月，DeepMind 的 AlphaProof 与 AlphaGeometry 2 联手，在国际数学奥林匹克（IMO）6 道题中解出 4 道，获得 28 分——位居银牌档顶端，仅比 29 分的金牌线低 1 分。它甚至攻克了令人畏惧的第 6 题，这道题在约 600 名人类参赛者中只有 5 人完整解出。该方法于 2025 年 11 月 12 日在线发表于 Nature，正式版本于 2026 年刊出。真正把它同聊天机器人式空谈区分开的关键设计事实是：AlphaProof 在 Lean 内部工作。它把约一百万道自然语言问题自动形式化为约 8000 万条 Lean 陈述，然后以 AlphaZero 风格的循环[10]训练自己，其中每一步都由 Lean 核查。用 DeepMind 的话说，「无需担心幻觉」——因为一个幻觉步骤根本无法编译。神经网络提供创造性搜索，证明助手提供真值基准。这种结合真实且重要。[已确立]

2025 年 7 月，门槛再次抬高：DeepMind（Gemini「Deep Think」模型）与 OpenAI 都报告了金牌分数——6 题中解出 5 题，35 分——而且引人注目的是，它们在时限内以端到端自然语言完成，而非在 Lean 内部完成。DeepMind 的结果由 IMO 官方认证；OpenAI 的结果是内部评分。确实令人印象深刻。但也正是在这里，第 1 日练出的校准直觉该派上用场：

- 「金牌」是一个分数，不是加冕。这些是竞赛题——数学中狭窄、限时、已知存在简短答案的一角。它们不是开放的研究问题，而且据官方 2025 年结果，仍有 26 名人类参赛者得分超过两个 AI 系统。
- 离开 Lean 是一种取舍，不是无代价的升级。2024 年的银牌是形式验证的——由机器保证正确。2025 年的自然语言金牌是人工评分的，意味着我们重新依赖可能藏有细微漏洞的散文。更通用，却更不确定。别让「金牌胜过银牌」的叙事掩盖了认识论根基的转移。
- 它昂贵且狭窄。每道困难的 2024 年题需要两到三天的计算，而且题目还需先被人工形式化为 Lean 陈述。这还称不上通用数学智能。

最需要明确否定的说法是：AI 尚未「解决数学」，也没有使数学家变得多余。[争议/炒作] 没有任何 AI 独立证明过一个著名的开放猜想并被接受为里程碑。关于定理证明代理找到小型 Lean 证明、或帮助完成狭窄形式化任务的报道虽然有趣，但仍早期、范围有限，也还不能替代被数学共同体接受的研究数学；它们应归入[线索]，留待日后审视，而非大肆宣扬。真正的革命比标题更安静、也更持久：一条延续 2300 年的标准——证明是一条无人能怀疑的链——终于交由机器以零容错执行，而 AI 正在学习沿这些严苛轨道搜索。（我们将在第 138–145 日深入追寻这一主题。）

关于虚构来源的注记

本课程的前沿校准器有一条必须明说的规则：凡是指向未来日期预印本编号的引用，一律剔除。这一领域的搜索结果中，充斥着信誓旦旦引用尚不存在论文的条目。以上每个里程碑都可追溯到真实、有日期、可核实的原始来源——已发表的 Nature 论文、官方竞赛结果、具名研究者本人的公开宣布。当一条关于 AI 与数学的声明无法这样追溯时，正确反应不是兴奋，而是怀疑。

开放问题

真正尚未解决的是什么

二十三个世纪之后，有效推理的研究依然留有真正未决的问题：

- 是否只有一种真逻辑，还是有许多种？当直觉主义逻辑、次协调逻辑与模糊逻辑都能切实派上用场，「正确逻辑」便渐渐显得更像工具选择，而非宇宙事实——但多元论者与一元论者仍真正地各执一词。
- 发现还是发明？逻辑定律是从实在中读出、嵌入任何可能心智，还是由约定采纳？经验物理学能否如普特南所想迫使我们修正？
- 溯因究竟是什么？「最佳解释推理」是真正第三种模式，还是换了外衣的归纳？甚至皮尔士本人是否将其理解为最佳解释推理（而非仅仅生成假设），学者之间亦有争议。
- 机械化证明能否改变数学本身？若一个结果为真，却只有计算机核查过证明，有没有人真正理解它？一个已验证却不透明的证明，与一个能带来洞见的人类证明，价值是否相同？
- 以及将萦绕 AI 单元的问题：当一台机器输出一个真实且得到充分支持的定理时，它是否知道任何东西——还是它只是第 1 日那个终极盖梯尔案例的翻版——因与理解毫无关系的理由而恰巧正确？（第 138–145 日。）

◆ 用三句话概括今日

核心观点

推理有三种引擎、三种担保——演绎保真却不扩展内容，归纳概括却可能被下一例打破，溯因跃向最佳解释——而在演绎内部，有效性（形式成立）与健全性（形式成立且前提为真）是完全不同的两件事。

最佳类比

夏洛克·福尔摩斯的「演绎」其实是溯因——对线索的最佳解释，而非保证结论——而一个有效却不健全的论证，是一条接合严密却输送污水的管道。

当下争议

逻辑是发现还是发明（以及是否只有一种真逻辑，还是一套工具），如今被一条真实的前沿所激化：Lean 等证明助手以零容错验证前沿数学，AI 已达奖牌水准——但并未真正「解决数学」。

今日线索 › 计算（柯里–霍华德：证明对应程序；硅芯片中的布尔代数；证明助手）· 信息（形式化使证明内容可被机器核查）· 涌现（大规模协作证明判定约 2200 万个蕴涵关系）——也将演绎与归纳衔接到第 1 日与第 2 日。

明日 → 第 04 日

概率成为扩展的逻辑

今日负责扩展却可能失手的引擎是归纳，而溯因留给我们一个任务：判断哪种解释最佳。明日，我们为二者加上数字刻度。概率原来并非与逻辑分离的学科，而是部分信念的自然延伸——蒙提霍尔问题将展示我们的直觉能错得多离谱，而贝叶斯定理又如何纠正它们。带上今日对天衣无缝与只是看似合理的区分；你即将学习如何演算「看似合理」。

说明

1. 真值函项指复合语句的真假只取决于各组成部分的真假，以及连接它们的逻辑词。
2. 先验指不依赖某次具体经验，而能凭理性或概念理解来把握。
3. 模型检查会自动遍历系统可能状态，验证某个形式性质是否始终成立。
4. 可判定性指存在一个算法，总能在有限时间内判断某个陈述是否由规则推出。
5. 类型论用带类型的表达式组织数学对象与证明；在许多系统中，证明一个命题就像构造一个属于相应类型的对象。
6. 形式化是把普通数学定义和证明改写成严格的形式语言，使证明检查器能够逐步核查。
7. 编译器把程序员写的源代码转换成计算机可以执行的较低层机器代码。
8. 功能性正确性指在给定假设下，程序实现已被证明符合其形式规格。
9. 微内核是操作系统最小的核心，只处理必要功能，把许多服务留在内核之外运行。
10. AlphaZero 风格训练指系统不断生成搜索尝试、检查结果，并把这些结果反过来作为新的训练信号。

来源

来源与延伸阅读

1. "Validity and Soundness." Internet Encyclopedia of Philosophy (accessed 2026). iep.utm.edu/val-snd ——基于形式的有效性与健全性区分。

2. "Deductive and Inductive Arguments." Internet Encyclopedia of Philosophy. iep.utm.edu/ded-ind ——保真推理与扩展性推理之分。

3. Douven, I. "Abduction." Stanford Encyclopedia of Philosophy (rev. 2021). plato.stanford.edu/entries/abduction ——皮尔士、最佳解释推理，以及关于溯因究竟是什么的学术争论。

4. "Aristotle's Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/aristotle-logic ——三段论、《前分析篇》与词项逻辑。

5. Bobzien, S. "Ancient Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/logic-ancient ——克律西波斯、斯多葛不可证明式与命题逻辑；武卡谢维奇的重新评估。

6. Boole, G. (1854). An Investigation of the Laws of Thought. London: Walton & Maberly. See "George Boole, The Laws of Thought," PhilPapers. philpapers.org/rec/BOOTLO-4 ——逻辑作为代数；「逻辑与数学」。

7. "Origins of Boolean Algebra in the Logic of Classes." Mathematical Association of America (Convergence). old.maa.org ——布尔、文恩、皮尔士，以及经香农（1937）通往数字逻辑之路。

8. "Frege's Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/frege-logic ——《概念文字》（1879）、量词、谓词逻辑与罗素悖论。

9. "Intuitionistic Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/logic-intuitionistic ——布劳威尔、海廷、对排中律的拒斥、 BHK 解释。

10. Priest, G., Berto, F. & Weber, Z. "Dialetheism" and "Paraconsistent Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/dialetheism ——爆炸原则、次协调性与双面真理论、 Logic of Paradox。

11. "Fuzzy logic." Wikipedia (accessed 2026). en.wikipedia.org/wiki/Fuzzy_logic ——扎德（1965）、[0,1] 上的真值、多值 / 武卡谢维奇根源。

12. Garson, J. "Modal Logic." Stanford Encyclopedia of Philosophy. plato.stanford.edu/entries/logic-modal ——必然 / 可能与计算机科学及验证应用。

13. "Curry–Howard correspondence." Wikipedia (accessed 2026). en.wikipedia.org/wiki/Curry–Howard_correspondence ——命题即类型，证明即程序。

14. "Mathlib statistics." Lean community (accessed June 2026). leanprover-community.github.io/mathlib_stats.html ——当前定理与定义数量。

15. "100 theorems in Lean." Lean community (accessed June 2026). leanprover-community.github.io/100.html ——Wiedijk 的 100 个定理基准中已有 84 个在 Lean 中形式化。

16. Commelin, J. & Topaz, A. et al. "Liquid Tensor Experiment." Lean community blog (completion 14 July 2022); Scholze's original challenge (Dec 2020). leanprover-community.github.io ——机器核查一位菲尔兹奖得主自己都不太确定的证明。

17. Tao, T. "Formalizing the proof of PFR in Lean4." terrytao.wordpress.com (Nov 2023). Gowers, Green, Manners & Tao, "On a conjecture of Marton," Annals of Mathematics (2025). terrytao.wordpress.com

18. Tao, T. et al. "The Equational Theories Project." Project announced Sept 2024; retrospective paper Dec 2025 (arXiv:2512.07087). teorth.github.io/equational_theories ——22,033,636 个含自蕴涵的有序对；22,028,942 条非平凡图边；50 余位贡献者，Lean 验证。

19. Buzzard, K. "Fermat's Last Theorem project." Lean community blog (launch 30 April 2024); EPSRC grant EP/Y022904/1 (2024–2029), Imperial College London. leanprover-community.github.io ——进行中；「至少需要 5 年」。

20. Leroy, X. et al. "CompCert" – a formally verified C compiler. Yang, Chen, Eide & Regehr, "Finding and Understanding Bugs in C Compilers," PLDI (2011). compcert.org ——约六个 CPU 年未找到错误代码。

21. Klein, G. et al. (2009). "seL4: Formal Verification of an OS Kernel." SOSP '09. sel4.systems ——首个操作系统微内核功能性正确性的机器检查证明（Isabelle/HOL）。

22. "AI achieves silver-medal standard solving International Mathematical Olympiad problems." Google DeepMind blog (25 July 2024). deepmind.google ——AlphaProof + AlphaGeometry 2；28 分；在 Lean 中工作。

23. Hubert, T., Mehta, R., Sartran, L. et al. (2026). "Olympiad-level formal mathematical reasoning with reinforcement learning." Nature 651: 607–613. doi:10.1038/s41586-025-09833-y. nature.com/articles/s41586-025-09833-y ——AlphaProof 方法论文；2025 年 11 月 12 日在线发表，2026 年 3 月 13 日正式出版；约 8000 万道 Lean 问题。

24. "Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the IMO." Google DeepMind blog (July 2025). deepmind.google ——35/42，官方认证；时限内的自然语言证明。

25. "66th IMO 2025." International Mathematical Olympiad. imo-official.org/editions/2025 和 individual results ——630 名参赛者；金牌线 35；人类分数分布。

26. "Our First Proof submissions." OpenAI (2026). openai.com/index/first-proof-submissions ——OpenAI 对其 2025 年 7 月 IMO 金牌级结果的后续总结，35/42 分。

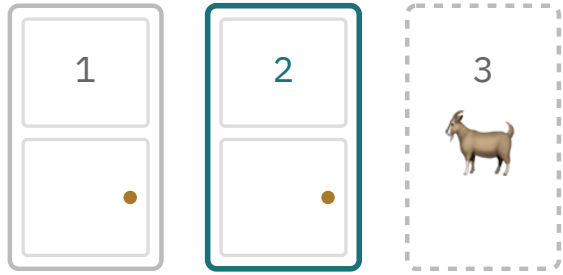
27. "Philosophy of logic" & "Logical realism." Wikipedia / Stanford Encyclopedia of Philosophy (accessed 2026). plato.stanford.edu/entries/logical-pluralism ——实在论、约定主义、奎因 / 普特南关于修正逻辑、逻辑多元论。

第 03 日完 · 尚有 177 次深入

模块一 · 知识与推理的根基 · 第 04 日 / 180

概率成为扩展的逻辑

主持人打开一扇门。你的直觉说：换不换无所谓。但直觉这东西，三分之二的概率是错的。



● 你选了 1 号门 · 主持人打开 3 号门 · 该换到 2 号门吗?

一档七十年代的电视游戏节目，竟把整个贝叶斯推理的精要藏在了门后。

你选了 1 号门。三扇门后，一辆跑车静待赢家，另外两扇各拴一只山羊。主持人心中有数，踱到 3 号门前，轻轻一推，露出一只羊，然后笑意盈盈地问：「要不要换到 2 号门？」剩下两扇门，一辆车。这不就是五五开吗？换与不换，能有什么差别。

差别大了。坚持到底，赢车概率只有三分之一；改换门庭，胜率跃升至三分之二——你几乎什么都不用做，只消改一个主意，胜算便翻了一倍。这就是蒙提霍尔问题。1990 年，它登上杂志专栏，随即引爆了数学史上最大规模的集体误判。今天我们会看到，正确答案为何不仅成立，而且「不可避免」；而解开这道谜题的那套机制，也正是人类在不确定世界中推理所能依赖的最深刻理论。

在第 1 日，我们认识了置信度——那只从 0 到 1 的信念旋钮——以及荷兰赌论证：旋钮若不自洽，便会被人组出一套稳赚不赔的赌局。今天我们学习证据降临时，这只旋钮必须怎样转动：贝叶斯定理。在第 2 日，我们看到科学如何在信号与噪音之间艰难划线，复现危机正是那条战线上的真实炮火；而今天的前沿——一场以「下注」取代 p 值的静悄悄革命——正是为了根治这一痼疾。今日点亮的线索有：信息（证据如比特，更新信念）、计算（心智与实验室都是推理引擎），以及「贝叶斯大脑」重现时一闪而过的能量。

集体翻车

全国最聪明的一批人，同时栽了跟头

1990 年 9 月，玛丽莲·沃斯·莎凡特——《吉尼斯世界纪录》认证的最高 IQ 纪录保持者，在《Parade》杂志[1]主持「问玛丽莲」专栏——回答了一位读者关于游戏节目的提问。换门，她写道，三分之二的概率赢。答案没错。随后，她的信箱炸了。

蒙提·霍尔的真实现场说明，这个谜题远非文字游戏：主持人不是随机开门的旁观者，而是掌握内情的行动者——他的每一步都在泄露信息。

据她自己统计，来信将近一万封，绝大多数在告诉她错了——其中约一千封来自博士。数学家写信训诫她。一位教授留下了那句「经典」评语：

> *「你搞砸了，而且搞砸得一塌糊涂！.....这个国家的数学文盲已经够多了，不需要全世界IQ最高的人再来添乱。丢人！」*
>
> ——斯科特·史密斯博士，佛罗里达大学，1990 年致《Parade》杂志信

真正搞砸的恰恰是他。严格按概率来算，他的大多数同行也没好到哪去。莎凡特在接下来的三篇专栏里寸步不让，最后干脆请全美教师带着学生用纸杯和硬币做实验。最终她获得了数据的支持，正如她所言：换门胜率是不换的两倍。教授们这才慢慢、且大多不那么体面地认了输。

非要眼见才肯信的那一位

就连保罗·埃尔德什[2]——史上最高产的数学家之一，他证明的定理多数人连题目都读不懂——也拒绝接受这个答案。朋友安德鲁·瓦兹森尼把逻辑讲给他听，他不为所动。直到瓦兹森尼跑了一次计算机模拟，重复几百轮，眼睁睁看着换门在约三分之二的情况下获胜，埃尔德什才勉强点头。即便如此，他仍闷闷不乐：模拟只告诉他「确实如此」，却没告诉他「为何如此」。（见保罗·霍夫曼传记《只爱数字的人》，1998。）如果连埃尔德什都在这件事上栽了跟头，你的困惑便没什么好丢人的。

这场风波暴露了一件事。蒙提霍尔问题不是花招，不是文字游戏——它的答案可证明、可模拟，板上钉钉。它揭示的是：人类对不确定性的直觉存在系统性偏差，我们迫切需要一件形式化的工具来矫正它。这件工具就是今天的主题。但在此之前，先让我们重新看看我们的直觉——再重新塑造它。

蒙提霍尔结果表

初次选择	主持人动作	坚持	换门
跑车，概率 1/3	打开任意一扇山羊门	赢	输
山羊，概率 2/3	被迫打开另一扇山羊门	输	赢

因此，坚持保持最初的 1/3；换门则收割了「第一次选错」那 2/3 的概率。

为何如此

主持人在帮你，也在泄密

最干净的感受方式是记住一点：你第一次选中跑车的概率，只有三分之一。这个数从不会改变。当你指向 1 号门时，车在那扇门后的概率是 1/3，而在「另外两扇门之一」后的概率是 2/3。接着主持人打开了一扇山羊门——关键就在这里：他不是随机挑的。他知道车在哪，而且必须露出山羊。于是，原本平摊在两扇门上的那 2/3 概率，被一股脑压到他唯一没有打开的那扇门上。

主持人的动作不是噪音，而是信息——五条贯穿全书的线索之一，首次以严格的数量形式登场。第 1 日那座停走的钟告诉我们：靠运气说对，不等于知识；而在这里，知情主持人在规则约束下采取的动作就是证据，会推动置信度旋钮。换门，你押的是那沉甸甸的 2/3；坚持，你守的只是最初那孤零零的 1/3。

如果直觉还在抵抗，就把问题放大。想象面前有一千扇门。你随手选了一扇——中奖概率千分之一。知情的主持人随后打开 998 扇，每一扇后面都是山羊，最后只剩下你的门和另一扇。你还会觉得这是一半一半吗？车几乎可以肯定就在主持人刻意避开的那扇门后。三扇门是同一逻辑，只是规模太小，直觉来不及反应。

比电视节目更古老

这个谜题并非始于蒙提霍尔。统计学家史蒂夫·塞尔文早在 1975 年致《The American Statistician》的一封信中就提出了它——而他的后续回应，也是「蒙提霍尔问题」这一名称首次见诸印刷。它的骨架还可追溯得更远：与伯特兰箱子悖论[3]（约瑟夫·伯特兰，1889）和马丁·加德纳的三囚犯问题[4]（1959）在结构上如出一辙。数学家称之为真实悖论（veridical paradox）——答案看似荒谬，却可严格证明。这又是一次趋同再发现，正如第 1 日的盖梯尔案例：当一个世纪里无数聪明头脑反复被同一块石头绊倒，那块石头一定是真的。

模型

贝叶斯定理：信念的更新法则

我们刚才对门做的那套手工操作，其实有个名字，也有个公式。公式看起来冷冰冰，想法却很直接：贝叶斯定理只是在证据到来之后，重新给仍然可能的世界分配权重。

贝叶斯像一只筛子

证据 E =「主持人打开了 3 号门」。每个假设一开始都有同样的先验；随后，谁越能预言这个具体揭示，谁留下的权重就越大。

$$P(H \mid E) = P(H) \times P(E \mid H) \,/\, P(E)$$

H 本行的藏车位置假设	P(H) 本行假设为真的先验概率	P(E | H) 若本行假设为真，主持人打开 3 号门的概率	P(H) × P(E | H) 本行留下的权重	P(E) 主持人打开 3 号门的总体概率
H：车在你选的 1 号门后	1/3	1/2	**1/6** 仍可能，但证据只支持一半	1/2
H：车在 2 号门后	1/3	1	**1/3** 最强幸存者：这个揭示是被迫的	1/2
H：车在 3 号门后	1/3	0	**0** 被排除：主持人不能打开藏车的门	1/2

P(E) 是证据筛过之后留下的总权重：**1/6 + 1/3 + 0 = 1/2**。把每个幸存权重都除以这个 1/2，未打开的 2 号门就拿走了剩余概率中的 **2/3**。

$$P(H \mid E) = P(H) \times P(E \mid H) \,/\, P(E)$$

后验（看到证据后的置信度）= 先验（之前的置信度）× 似然（H 预测 E 的能力），再用 证据总量 归一化

用一句话说：看到证据 E 后，你对假设 H 的后验置信度，等于你的先验置信度乘以似然——即 H 预测你会看到 E 的力度——再除以 E 本身出现的总体预期。强有力的证据，是你的假设能预见、而对手却预见不到的东西。整台引擎就这么多。信念永远流向最能解释已发生之事的那一方。

回到蒙提霍尔。令 H =「车在 2 号门后」，E =「主持人打开 3 号门」。如果车确实在 2 号门后，主持人只能开 3 号门（不能开你的门，也不能开藏车的门），所以似然为 1。但如果车在你选的 1 号门后，他本可以开 2 号或 3 号，所以开 3 号的似然只有 1/2。正是这个似然上的不对称，把后验推到了支持换门的 2/3。公式替我们做的，不过是直觉算不好的那笔账。

连医生都会掉进去的陷阱

贝叶斯定理拯救的不只是游戏节目选手。它还能抓住一个著名研究中大多数医生都犯的错误。

下方表格按默认医学检测案例展开这个陷阱；它值得记住，因为它存在于每一次体检、每一个垃圾邮件过滤器、每一道机场安检。

基础概率陷阱

组别	每 1000 人	阳性数
患病者	1	约 1 个真阳性
健康者	999	约 50 个假阳性
全部阳性	约 51	其中只有约 1 人真的患病

因此后验概率约为 0.99 / 50.94，即 1.9%——这就是 Casscells 结果，只是把灵敏度也明说出来。

深层思想

为什么叫「扩展的逻辑」，而不只是一个公式

这就引出了今天的主题。普通演绎逻辑——第 3 日的三段论——是确定性的逻辑：凡人皆有死，苏格拉底是人，所以苏格拉底会死。到此为止，没有余地。但现实生活中几乎没有什么是确定的。我们需要一种逻辑，覆盖「肯定为真」（概率 1）与「肯定为假」

（概率 0）之间那片辽阔的灰色地带。令人惊讶的结论是：这样的逻辑本质上只有一种，就是概率演算。

物理学家 R. T. 考克斯在 1946 年把这一结论变成了定理。他问：假设你想给「在已知前提下，这件事有多可信？」指定一个数值，并且只坚持几条常识——可信度必须能用实数表示；同一件事用两种正确方法算出来必须得到同一个可信度（「一致性」）；「非 A」的可信度只取决于「A」的可信度。仅凭这几条朴素要求[5]，考克斯证明，你就「不得不」——不是被建议，而是被强迫——接受标准的概率规则。经过一次不改变实质的重新刻度后，否定必须像 1 - P(A) 那样运作，合取必须服从乘法规则，证据 E 到来时就必须对 E 做条件化。任何自洽的分级信念系统，换件马甲就是概率论。

物理学家 E. T. 杰恩斯的遗著《概率论：科学的逻辑》（2003）正建立在这块基石上。他的口号是：演绎逻辑不过是概率论的特例——所有概率恰好取 0 或 1 的那个特例。概率，就是把逻辑扩展到不确定性领域——也就是扩展到现实世界。请注意，这已经是通往同一终点的第三条独立路径：荷兰赌论证（第 1 日）从「别让人钻空子」出发；而决策论稍后将从「别做被支配的选择」也抵达此处。自洽、无确定损失、一致推理——三条路指向同一套演算。

一个诚实的脚注

考克斯的原始证明略有疏漏。1999 年，计算机科学家约瑟夫·哈尔彭指出，要让证明完全严密还需补一条技术假设（在某些有限域上可能失效），后来的作者做了妥善修补。因此准确的说法不是「概率是不确定性唯一可想象的逻辑」——那过于绝对——而是「在合理条件下，自洽的分级信念必然落入概率公理的框架」。定理依然成立，只是它的桂冠比杰恩斯笔下某些豪言所暗示的要小一号。［已确立，附前提］

争论

两大阵营，同一个方程

概率这样优美而统一的理论，为何能在统计学内部引发一场百年内战？因为方程本身无人质疑，争议在于「那些数字意味着什么」。两派使用的是同一套演算——安德烈·柯尔莫戈洛夫 1933 年写下的公理。这些公理刻意不回答概率「是什么」，只规定它「如何表现」。在这副中立骨架上，两派披上了不同的外衣。

频率派

贝叶斯派

概率 = 长期频率

- 概率是事件在无限次重复中出现的频率。「硬币公平」意味着抛掷无穷多次，正面比例趋近一半。
- 参数是固定但未知的常数；数据才是随机的。你关心的是：你的方法有多大可能误导你。
- 工具：p 值、置信区间、第一/第二类错误（费希尔；奈曼与皮尔逊，1920–30 年代）。
- 说不出「火星上曾有生命的概率是 70%」——火星要么有过生命，要么没有，不存在可重复的样本可供计数。

概率 = 置信度

- 概率是一种置信度——你在已知条件下理性地有多确信（直接来自第 1 日的那只旋钮）。
- 参数自身也获得概率分布；你随数据不断用贝叶斯定理更新它们。
- 工具：先验、后验、贝叶斯因子。谱系：拉普拉斯 → 杰弗里斯 → 拉姆齐 → 德·菲内蒂 → 萨维奇。
- 可以理直气壮地说「火星上曾有生命的概率是 70%」——一次性事件无法重复，但置信度恰好为此而生。

频率派在 20 世纪独领风骚，一半靠道理，一半靠运气。道理在于：它的创立者追求客观性，不信任贝叶斯派的先验，认为那是暗中塞入的主观意见。（费希尔把「逆概率」斥为「必须彻底摒弃」。）运气在于：贝叶斯方法需要大量计算，而廉价计算机姗姗来迟。贝叶斯派至今最敏感的痛点仍是先验——你那个「事前」信念从哪来？凭什么让别人信你的？客观贝叶斯派（杰弗里斯、杰恩斯）寻找规则化的「无信息先验」；主观贝叶斯派则耸耸肩：所有推理总得有个起点。

「概率不存在」

意大利人布鲁诺·德·菲内蒂在专著开篇劈头盖脸就是四个大写英文单词：PROBABILITY DOES NOT EXIST（概率不存在）。他的观点蓄意挑衅：世界上并不存在像质量或电荷那样独立「在那儿」的概率——存在的只是一个理性主体自洽的下注行为。他用一条真正的定理为这句口号背书（1937 年的表示定理）：如果你把一系列观测视为「可交换的」——先后顺序对你无关紧要——那么数学上你就必须表现得仿佛存在一个固定但未知的频率，而你对它持有一个先验。主观信念与看似客观的参数，原来是同一枚硬币的两面。一份用数学写就的停战协议。

从这盘棋里还能落下一条实践智慧：克伦威尔法则（丹尼斯·林德利以克伦威尔 1650 年的恳求命名：「我求你，在心底想一想，你也有可能是错的」）。永远不要把先验精确地设为 0 或 1，因为贝叶斯定理此后再也改变不了它——被绝对确信的东西，按定义就是不会再因世界而变动的。林德利写道：哪怕给「月亮是绿奶酪」留一丝怀疑余地也好，

否则哪怕宇航员真从月球带回了奶酪样本，也休想撼动你分毫。我们再次讨论了校准这一贯穿整个模块的暗线。

前沿 · 2026

针对 p 值的静默兵变

一个世纪以来，频率派的 p 值一直是科学的守门人：跌破 0.05，结果便可称为「显著」。在第 2 日我们看到「显著」再也不是不可动摇的结论了——复现危机中，成山的「显著」发现一经复测便烟消云散。一个主要元凶是结构性的：p 值太脆弱。实验做到一半查看一次数据，发现 p 已跌破 0.05 就立刻收手——你的假阳性率就这样被悄悄抬高。这个过失太过常见，甚至有了专门名称：「选择性停止」（optional stopping）。如今统计学界正在流传一套新框架，从地基开始重建假设检验，正是为了解决这个问题。它的核心对象不是概率，而是一场「赌局」。

前沿 01 [已确立]

e 值：用下注来检验假设

e 值是你对原假设[6]下注后获得的回报。你押上 1 美元赌原假设为假，这份赌约被设计成「在原假设为真时公平」——也就是说，如果原假设确实成立，你便会亏得占不到任何长期便宜（用符号说：e 值在原假设下的期望值至多为 1）。所以，如果最终你的赌注翻了二十倍，那原假设一定哪里出了问题：要么它为假，要么你中了天文数字级别的头彩。一个很大的 e 值，字面意思就是你从原假设身上赢到的真金白银，而你累积的财富就是你的证据。它的倒数 1/e 行为上像个保守的 p 值，但下注的场景才是精髓。

在硬币例子里，原假设很具体：「硬币公平，P(正面) = 0.5」。e 值是两张似然比[7]赌票合起来的财富。一张赌票押「正面偏多」的硬币，即 P(正面) = 0.60：每出现一次正面，这张票乘以 0.60 / 0.50 = 1.2；每出现一次反面，则乘以 0.40 / 0.50 = 0.8。另一张镜像赌票押「反面偏多」，即 P(正面) = 0.40，倍率正好反过来。把起始的 1 美元平均分到两张票上，无论硬币朝哪边持续偏，都可能让财富增长。如果硬币真的公平，每张票每轮的期望倍率都是 1；这场赌局在原假设下就是公平的。在这个玩具赌局里，「赢」就是财富大到足以拒绝「硬币公平」；「输」就是财富停滞或缩水，说明你还没有赢到反对公平的证据。

这不是松散的比喻，而是一套严格的纲领——「博弈论统计学」，由格伦·谢弗与弗拉基米尔·沃夫克用二十年时间建立，现由阿迪亚·拉姆达斯、彼得·格伦瓦尔德、王若度等人

继续推进。谢弗的宣言《以赌注检验》于 2020 年在英国皇家统计学会宣读，2021 年发表于该会《期刊》A 辑。他抱怨 p 值的一个理由正是它太难向人解释；而「我赌这个假设不成立，赢了 20 块」——这话任何人都能听懂。

前沿 02 [已确立] [争议]

为什么下注好过 p 值：实时的局部结论

赌局会复利累积。如果你对原假设下了一场公平的赌，再下一场，再下一场，手头财富就构成了数学家所说的鞅[8]（martingale），而一条经典定理（维勒不等式[9]）保证：若原假设为真，它几乎不可能膨胀出天文数字。这赋予了 e 值一项 p 值望尘莫及的近乎魔法的性质：任意时刻有效性。你可以盯着实验进展，随时喊停，觉得有希望就继续加数据——中途查看多少次都可以——你的错误保证依然成立。格伦瓦尔德、德·海德与库伦称之为「安全检验」（发表于 RSS《期刊》B 辑，2024）；更完整的框架——包括每时每刻都有效的置信区间——叫做「安全任意时刻有效推断」（拉姆达斯、格伦瓦尔德、沃夫克与谢弗，《统计科学》，2023）。e 值合并起来也极为方便：独立的 e 值直接相乘，相依的 e 值取平均，结果仍是有效 e 值——这让跨研究汇总变得干净利落，而 p 值则会一头扎进多重比较的雷区。

下方表格概括同一个对比：脆弱、怕中途查看的 p 值，与诚实的 e 值。

这个玩具任务故意很窄：它要拒绝的只是「这枚硬币是公平的」这个命题；它不是在估计硬币的精确偏差，也不是在绝对证明硬币不公平。

e 值账本

量	含义	用途
E = 1	对原假设没有净赢面	起点
硬币演示赌票	似然比回报：押中的一面出现时乘以 1.2，另一面出现时乘以 0.8	若硬币的真实 P(正面) = 0.5，则期望上公平
E = 20	原假设下公平赌约的二十倍回报	0.05 水平的拒绝阈值，因为 1 / 20 = 0.05
滚动财富	检验鞅或 e 过程	可持续监控，同时控制第一类错误

代价是偏保守：当所有建模假设完全正确时，任意时刻有效的账本可能需要比固定样本量 [10] 检验更强或更持久的证据。

换到科学场景会是什么样？在一项持续更新的临床元分析 [11] 里，原假设可能是「BCG 疫苗对医护人员感染 COVID-19 没有临床相关效果」。新的随机试验会在不同时间报告结果，研究者希望一有新数据就更新综合分析，同时又不希望每看一次结果，假阳性风险就悄悄升高。ALL-IN 元分析框架正是为这类场景设计的：它允许后续试验的证据陆续加入，同时保留第一类错误率与区间覆盖率保证。在一个 BCG/COVID 应用中，对证据过程来说，「赢」本来意味着累积到足以支持临床相关获益的强证据；但这项任意时刻有效分析没有发现 BCG 能临床相关地降低感染，而住院结局因事件太少，仍不足以下定论。这和硬币玩具是同一结构，只是把正反面换成了医学终点和陆续到来的试验数据。

前沿 03 [已确立] [炒作]

这场兵变究竟蔓延了多远?

到了区分实诚与吹嘘的时候了。e 值的数学已经确立且优美——经过本领域最顶尖期刊的同行评议（《统计学年鉴》、RSS 两辑《期刊》、《统计科学》），并在 2024 年预印本之后由拉姆达斯与王若度汇集成一本 390 页的《Foundations and Trends》专著。这一部分[已确立]，无可争议。

真实世界中的采纳则是更窄、也更诚实的故事。最清晰的落地在科技公司 A/B 测试[12]——因为持续查看数据本身就是它们的日常：Optimizely 围绕「始终有效推断」重建了整个平台（Johari、Koomen、Pekelis 与 Walsh），Netflix 与 Adobe 则公开使用任意时刻有效的置信序列[13]，让产品团队能持续监控实验而不在统计上作弊。这是真正的生产环境应用——但距离全球的生物统计、心理学和物理学共同体还很远，那里 p 值依旧根深蒂固。

新工具也不是毫无代价。在固定样本量比较中，e 值可能需要比 p 值更极端的数据才能跨过同样的拒绝门槛；谢弗的回应是，这是让证据这把尺子变诚实的代价，而非简单缺陷。你的赌局效率取决于下注策略的好坏——说白了，这跟贝叶斯派选择先验时面对的建模判断如出一辙，只是换了套行头。塞缪尔·帕维尔与莱昂哈德·赫尔德等批评者警告：把检验标榜为「安全」或「始终有效」可能有误导之嫌，因为这些保证同样依赖假设（模型设定正确、无发表偏倚），而那些假设可能跟别的假设一样失效。诚实的裁决是：它是 p 值的一个[前景可期]、严格、真正有用的补充——但绝不是科学范围内的全面替代品，至少现在不是。

什么能真正推动局面？如果 FDA 或 EMA 这类药物监管机构批准 e 值用于确证性临床试验，或者某家顶级综合科学期刊把它写进投稿指南，「取代」的口号才有可能从炒作变成现实。让我们拭目以待。

开放问题

真正尚未解决的

- 概率到底是什么？是世界中的频率、心智中的置信度，还是一个公平的赔率？三个世纪过去了，诠释之争有过停火（德·菲内蒂），但从没有投降。
- 先验从何而来？是否存在一种有原则、客观的方式来设定你的「事前」信念，还是一切推理终究立足于一个数学无法替你辩护的选择？
- 基于下注的统计学真能接管吗？还是只会沦为序贯实验的专用工具，而 p 值继续统治其余领域——而且，「选你的赌注」真的比「选你的先验」更不主观吗？

- 大脑真的在运行贝叶斯吗？第 1 日的预测加工线索说，感知就是神经组织中的贝叶斯推断。今天为这个主张提供了规范性骨架——但「大脑近似贝叶斯」和「大脑就是贝叶斯」是两笔截然不同的赌注，我们将在第 119 日重返这个话题。
- 考克斯定理真的对任何理性主体都有效吗——包括人工主体——还是只对那些已经接受了它的一致性公理的主体有效？（这个问题对 AI 板块格外要紧，第 138–145 日。）

◆ 今日三句话

大观念

概率不只是骰子和硬币的工具——它是不确定性领域中逻辑的唯一延伸（考克斯、杰恩斯），而贝叶斯定理是它的运动定律：信念流向最能解释你实际所见之事的假设。

最佳类比

蒙提霍尔打开一扇山羊门——知情者的选择把 2/3 的概率倾注到仅剩的那扇门上；以及赌徒的账本——反对某个假设的证据，字面意思就是你赌它不成立而赢来的钱。

当下争议

频率派与贝叶斯派围绕概率「是什么」的分裂，如今又添了一场 2020 年代的兵变：用脆弱、怕中途查看的 p 值换取 e 值——数学已确立，科技已采纳，但远未成为最狂热支持者许诺的那种科学级革命。

今日线索 › 信息（主持人的揭示和 e 值都是更新信念的证据）· 计算（心智与实验室作为推理引擎）· 能量（对「贝叶斯大脑」的一次轻回调）——而校准这条暗线，从第 1 日和第 2 日一路延续至此。

明日 → 第 05 日

因果

今天我们学会了如何根据证据更新信念——但这些证据只告诉了我们「相关」。冰淇淋销量和溺水人数同步上升，但谁也不是谁的因。明天我们要面对推理中最艰难的一次升级：区分什么只是跟着某物一起起伏，什么才真正「驱动」了它。混杂因素、反事实，以及朱迪亚·珀尔的 do-演算[14]——这套工具问的不是「我预期什么？」而是「如果我插手干预，会怎样？」带上今天的贝叶斯直觉——你需要学会它的边界。

说明

1. 《Parade》是美国发行范围很广的周日报纸副刊杂志；「问玛丽莲」是莎凡特的问答专栏。
2. 埃尔德什发表约 1500 篇论文，深刻影响组合数学、图论、数论和概率方法；数学界还用「埃尔德什数」衡量合作距离。
3. 其中有三只盒子：金金、银银、金银；抽到一枚金币后，另一枚也是金币的概率是 2/3，而非 1/2。
4. 其中三名死囚有一人将被秘密赦免；得知另一名囚犯会被处决后，人们容易作出同样错误的 1/2 更新。
5. 这里的要求指一套理论必须满足的基本条件。
6. 原假设是统计检验试图拒绝的默认说法，通常是「没有效应」或「没有差异」。
7. 似然比比较同一批数据在两个假设之下分别有多可能出现。
8. 鞅是公平赌局的数学模型：在已知过去的条件下，下一步的期望值等于当前值。
9. 维勒不等式限制了公平赌局的财富过程仅凭运气膨胀到很大数值的概率。
10. 固定样本量检验会预先决定样本数，并只在计划好的终点分析一次。
11. 元分析会用统计方法合并多项研究的结果；持续更新的元分析会随着新研究出现而更新。
12. A/B 测试会把用户随机分到产品版本 A 或 B，比较哪个版本表现更好。
13. 置信序列是一组随着数据累积而更新、且在每个时点都保持有效的置信区间。
14. do-演算是珀尔用于推理干预的形式系统：如果我们把某个变量设成某个值，会发生什么？

来源

来源与延伸阅读

1. Selvin, S. (1975). "A Problem in Probability" (Letter to the Editor). The American Statistician 29(1): 67. ——以及后续回应 "On the Monty Hall Problem," 29(3): 134，为该名称首次见诸印刷。

2. vos Savant, M. "Ask Marilyn." Parade (Sept 9, 1990, and follow-ups 1990–91). marilynvossavant.com/game-show-problem ——专栏、读者来信，以及约一万封信 / 约一千位博士的估计（莎凡特本人统计）。

3. Tierney, J. (July 21, 1991). "Behind Monty Hall's Doors: Puzzle, Debate and Answer?" The New York Times. nytimes.com ——包括蒙提·霍尔与 Persi Diaconis 关于主持人协议附注的讨论。

4. Hoffman, P. (1998). The Man Who Loved Only Numbers. Hyperion. ——埃尔德什 / 瓦兹森尼模拟轶事。

5. Bertrand, J. (1889). Calcul des probabilités. Gauthier-Villars. ——伯特兰箱子悖论，结构上的祖先。另见 Gardner, M. (1959), "Mathematical Games," Scientific American (Three Prisoners)。

6. Casscells, W., Schoenberger, A. & Graboys, T. B. (1978). "Interpretation by Physicians of Clinical Laboratory Results." New England Journal of Medicine 299(18): 999–1001. doi:10.1056/NEJM197811022991808. ——60 名临床医生中只有 11 人给出约 2% 的答案。

7. Cox, R. T. (1946). "Probability, Frequency and Reasonable Expectation." American Journal of Physics 14(1): 1–13. ——迫使概率规则成立的那些条件。

8. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press (ed. G. L. Bretthorst). ——概率成为扩展的逻辑。

9. Halpern, J. Y. (1999). "A Counterexample to Theorems of Cox and Fine." Journal of Artificial Intelligence Research 10: 67–85. ——关于考克斯定理严谨性的附注。

10. Kolmogorov, A. N. (1933). Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of the Theory of Probability). Springer. ——诠释中立的公理。

11. de Finetti, B. (1937 / 1974). "La prévision..."; Theory of Probability (Eng. trans.). ——"PROBABILITY DOES NOT EXIST"；表示定理。

12. Lindley, D. V. (1991). Making Decisions, 2nd ed. Wiley. ——克伦威尔法则（第 104 页）。

13. Shafer, G. (2021). "Testing by Betting: A Strategy for Statistical and Scientific Communication." Journal of the Royal Statistical Society Series A 184(2): 407–431. doi:10.1111/rssa.12647. rss.onlinelibrary.wiley.com ——含发表讨论（包括沃夫克的评论，JRSS-A 184(2): 445–446）。

14. Vovk, V. & Wang, R. (2021). "E-values: Calibration, combination, and applications." The Annals of Statistics 49(3): 1736–1754. doi:10.1214/20-AOS2020. pdf

15. Grünwald, P., de Heide, R. & Koolen, W. (2024). "Safe Testing." Journal of the Royal Statistical Society Series B 86(5): 1091–1128. doi:10.1093/jrsssb/qkae011 (read paper, with discussion incl. Shafer, Pawel & Held). academic.oup.com

16. Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2023). "Game-Theoretic Statistics and Safe Anytime-Valid Inference." Statistical Science 38(4): 576–601. doi:10.1214/23-STS894. arXiv:2210.01948

17. Ramdas, A. & Wang, R. (2025; first posted 2024). "Hypothesis Testing with E-values." Foundations and Trends in Statistics 1(1–2): 1–390. arXiv:2410.23614 ——综合专著。

18. ter Schure, J., Ly, A., Belin, L. et al. (2022). "Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers." Prospective ALL-IN meta-analysis preprint.

Amsterdam UMC ——在持续更新的临床元分析中使用 exact e-value logrank tests 与任意时刻有效置信区间。

19. Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2022). "Always Valid Inference: Continuous Monitoring of A/B Tests." Operations Research 70(3): 1806–1821. doi:10.1287/opre.2021.2135 ——Optimizely 的部署；参见 Netflix Research 关于任意时刻有效推断的研究，以及 Adobe Experience Platform 置信序列。

20. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values." The American Statistician 70(2): 129–133. —— 以及 Amrhein, Greenland & McShane (2019), "Retire statistical significance," Nature 567: 305–307。

第 04 日完 · 余下 176 次深入