从根基到 2026 年研究前沿

深入一百八十日

作者：Claude Opus 与 GPT
翻译：Kimi 与 GLM

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、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；建立一个心智模型；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。它的气质接近一本极简导论，但内部坡度更陡——开篇假设读者聪慧而初来乍到，随后一路深入，直到脚下的地面真正触及前沿、充满争议。

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、生命起源的物理、哺乳动物跨代表观遗传、意识理论、通用人工智能与对齐，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。

「炒作过滤器」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，而是更关注于：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，却未必分量很重。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。新近不等于可靠；同行评议不等于定论；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 值展示证据到来时信念应如何移动。

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。

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**。它立足于一条完全合理的**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箱。请留意它将在**第 2 日**（科学如何判定什么才算数）、**第 4 日**（概率作为部分信念的逻辑）、**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，一条三条腿的凳子。抽掉任何一条腿，它都会倾倒。这一图景通常追溯至柏拉图，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：*《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在握——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，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的更清楚版本：你的理由（那座钟）是坏的，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。

盖梯尔机器

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 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裁决。欧内斯特·索萨（1999）将其翻转为*安全性*：在所有邻近的可能展开中，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。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——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*（又是索萨）。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，是的。出于能力，不是。*适切*，不是。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——而第四个学派索性换了问题。

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——以及一层炒作过滤器

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*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测试了巴西、印度、日本与美国——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（金与袁）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。更深层的教训，我们将在**第 149 日**以工业规模遇见：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*修正它们。

为何应当服从？**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，但它们合在一起将*无论发生什么都保证你输钱*。不融贯的置信度不仅是凌乱——它是可被利用的。下方的刻度盘让你亲身体会陷阱如何收紧。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。（彩票悖论在此咬人：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。

置信度刻度盘与荷兰赌

若你对S的置信度与对*非-S*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。

对 S 的置信度	对 非-S 的置信度	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；神经科学如今追问一团组织如何形成一个信念。一个快速成长的纲领回答：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*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 与生物硬件联结起来。

卡尔·弗里斯顿以*自由能原理*（*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。先把标签贴准，在此处至关重要。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**）中重返它——并且已然注意到，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**：当像起草这一页的系统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）——并轻触涌现与演化。

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、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.

明日 → 第 02 日

科学方法与划界问题

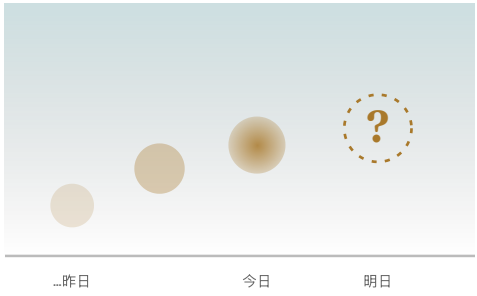
今天我们追问，单个信念何时才算得上知识。明天我们把问题放大至一整座机构：科学如何裁定哪些断言值得被认真纳入讨论？波普尔要求真正的理论必须*可证伪*，库恩的范式转移，拉卡托斯的研究纲领——以及现代复现危机，作为划界线在现实检验中的试炼。带上今日校准好的直觉——你会用得着。

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；而那道炒作过滤器——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*：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。

划界实验室

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

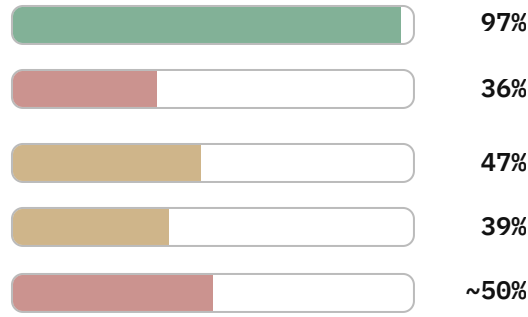
复现危机：划界在现实检验中

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）。

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。

明日 → 第 03 日

逻辑与有效推理

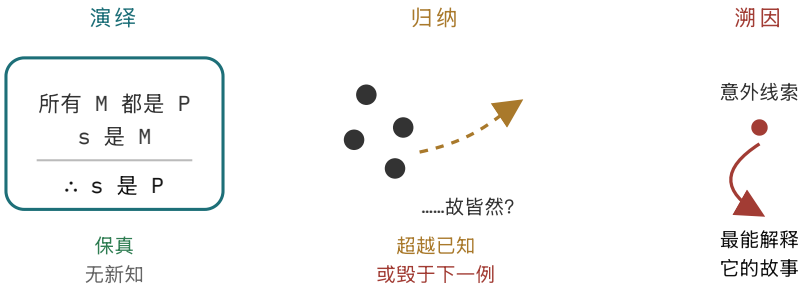
今日我们频频倚仗「有效」、「由此推出」、「矛盾」等词——但使论证真正成立的规则究竟是什么？明日我们将深入逻辑本身：演绎（能保真，却不能凭空增加新信息）、归纳（休谟留下的伤口）与溯因（像侦探一样选择最佳解释）。我们将遇见日常欺骗我们的谬误，追问逻辑是*被发现*的还是*被发明*的，并抵达前沿——在那里，机器如今检验着人类头脑无法完全容纳的证明。这是此前所有讨论赖以成立的逻辑底座。

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。

推理检视器

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，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

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

反实在论渊源甚深，这个立场却很少有充分发展的版本。

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。

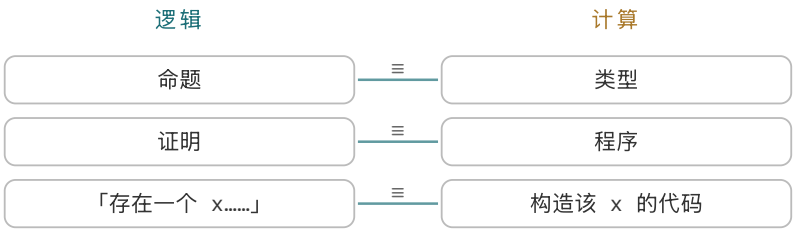
直觉主义逻辑放弃了*排中律*。它由 L. E. J. 布劳威尔开创，1920–30 年代由阿伦德·海廷形式化，坚持一条陈述只有在你能*构造*出其证明时才算为真。你不能凭空断言「A 或非-A」——你必须证明其中一边。一个例子能尖锐地说明动机：排中律会让你轻松断言，对任何计算机程序，「它停机或不停机」——然而（我们将在第 **27** 日看到）不存在判定停机的通用方法，因此没有构造支撑这一断言。直觉主义说：那就别断言。这听起来像是哲学洁癖，直到你发现经由一种美得值得专设一框的对应关系，它竟通向计算机科学的心脏。

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，可以保留模型检查所需的可判定性。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；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 **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证明的操作系统微内核：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——此处正是炒作过滤器该上场的时候，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

本课程的炒作过滤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，还是一套工具箱），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

概率作为扩展逻辑

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」。

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。

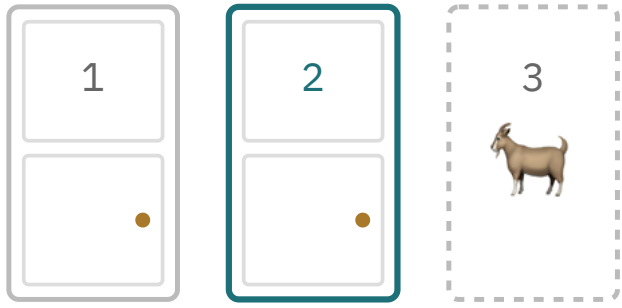
27. "Philosophy of logic" & "Logical realism." *Wikipedia / Stanford Encyclopedia of Philosophy* (accessed 2026). plato.stanford.edu/entries/logical-pluralism ——实在论、约定主义、奎因 / 普特南关于修正逻辑、逻辑多元论。

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

概率作为*扩展逻辑*

游戏节目主持人打开一扇门。你的直觉说这不重要。你的直觉将在三分之二的时间里输掉。



● 你选了 1 号门 · 主持人打开 3 号门 · 该换到 2 号门吗?

整个贝叶斯推理，藏在一档 1970 年代的游戏节目里。

你选了 1 号门。三扇门后有一辆车，另外两扇后面是山羊。主持人——他清楚知道车在哪里——走到 3 号门前，打开它，露出一只山羊，然后几乎是和蔼地问：*你要换到 2 号门吗?* 两扇门，一辆车。一半对一半，肯定如此。换门怎么可能重要?

这极其重要。坚守，你赢车的概率是三分之一。换门，你赢车的概率是三分之二——你什么都不用做，只消改变主意，就把胜率翻了一倍。这就是蒙提霍尔问题，1990 年它刊登在一份杂志专栏上时，引发了数学史上最大规模的公众认知崩塌之一。今天我们会看到，为什么答案不仅正确，而且必然——而解决它的同一台机器，恰好也是“在不确定中推理意味着什么”所能给出的最深刻理论。

在**第 1 日**，我们认识了*信念度*——信念是 0 到 1 之间的一个旋钮——以及荷兰赌论证：不一致的旋钮会被别人打包成稳赚不赔的赌局。今天我们要学习证据来临时旋钮必须如何移动的规律：贝叶斯定理。在**第 2 日**，我们看着科学艰难地划清信号与噪音的界限，并把可重复性危机看作这场拉锯战的火线实况；而今天的前沿——一场用"下注"取代 p 值的静默革命——正是冲着修复它去的。今日点亮的线索：*信息*（作为更新信念之比特的证据）、*计算*（心智与实验室都是推理引擎），以及"贝叶斯大脑"回归时一闪而过的*能量*。

大崩溃

全国最聪明的人，同时错了

1990 年 9 月，玛丽莲·沃斯·莎凡特——她以《吉尼斯世界纪录》登记的最高智商闻名，在《Parade》杂志撰写"问玛丽莲"专栏——回答了一位读者关于一档游戏节目的问题。她写道：换门，你会在三分之二的时间里获胜。答案是对的。反响却如同末日。

据她统计，她收到了约 **一万封**来信，绝大多数都在说她错了——其中约**一千封出自博士之手**。数学家写信训斥她。一位教授留下了那句"不朽"的话：

> *"你搞砸了，而且搞砸得离谱！……这个国家的数学文盲已经够多了，我们不需要世界上智商最高的人再来添乱。可耻！"*
>
> ——斯科特·史密斯博士，佛罗里达大学，致《Parade》杂志信（1990）

搞砸的其实是他自己。严格按统计来说，他的大多数同事也一样。莎凡特在接下来的三篇专栏里坚守立场，最后请全国中小学教师用纸杯和硬币做实验。他们照做了。数据准确无误地印证了她的话：换门获胜的频率是坚守的两倍。教授们慢慢地、且并不总是优雅地，撤退了。

那位必须亲眼看见才相信的人

就连**保罗·埃尔德什**——人类历史上最高产的数学家之一，他证明的定理我们大多数人连读都读不懂——也拒绝接受这个答案。当他的朋友安德鲁·瓦兹森尼摆出逻辑时，埃尔德什不为所动。直到瓦兹森尼运行了一场*计算机模拟*，把游戏重复几百次，看着换门在约三分之二的回合里获胜，埃尔德什才让步。即便如此他仍有些恼火：模拟告诉他*那是对的*，却没告诉他*为什么*。（见保罗·霍夫曼的传记《只爱数字的人》，1998。）如果连埃尔德什都会被绊倒，你至少有出色的同伴。

这场崩溃揭示了一件事：蒙提霍尔问题既不是花招，也不是文字游戏——它的答案可证明、可模拟、千真万确。它暴露的是，人类对不确定性的直觉存在*系统性*偏差，我们迫切需要一个形式化工具来覆盖直觉。这个工具就是今日深入的主题。不过，先让我们把直觉摔在礁石上——再把它重建起来。

蒙提霍尔机

初次选择	主持人动作	坚守	换门
汽车，概率 1/3	可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

因此坚守保持原来的 1/3 概率；换门则攫取了第一次选错的 2/3 概率。

为什么成立

主持人在帮你（也在泄露信息）

感受答案最清爽的方式是：**你第一次选对的概率是三分之一**。这个数字不会改变。当你指向 1 号门时，车在它后面的概率是 1/3，在“另外两扇门之一”后面的概率是 2/3。然后主持人打开一扇山羊门——但关键是，主持人并不是随机选择。他知道车在哪里，而且必须露出一头山羊。于是原本平摊在两扇门上的那 2/3 概率，全部**集中到了他没打开的那一扇门上**。

主持人的揭示不是噪音。它是*信息*——我们五条反复出现的线索之一，首次以严格的数量形式登场。在第 1 日，一座正常走动的钟没告诉我们任何新东西；而在这里，一个知情主体的动作改变了概率分布。换门，你是在押注那丰厚的 2/3；坚守，你只是抱着原来孤零零的 1/3 不放。

如果你的直觉仍在抗拒，就把问题放大。想象**一千扇门**。你选一扇——中奖概率是千分之一。知情的主持人接着打开另外 998 扇门，每一扇后面都是山羊，只剩下你选的门和另一扇。你真的还以为是抛硬币吗？汽车几乎肯定在主持人刻意避开的那扇门后面。三扇门版本是同样的逻辑，只是规模太小，直觉感受不到。

比综艺节目更古老

这个谜题并非始于蒙提霍尔。统计学家**史蒂夫·塞尔文** 1975 年在《The American Statistician》的一封读者来信中提出了它——他的后续回应也是“蒙提霍尔问题”这一名称首次见诸印刷的地方。它的骨架还要更古老：与**伯特兰箱子悖论**（约瑟夫·伯特兰，1889）以及马丁·加德纳的**三囚犯问题**（1959）完全相同。数学家称之为*真实悖论*——答案看似不可能，却可证明为真。这又是一次趋同再发现，正如第 1 日的盖梯尔案例：当一个世纪里无数心智反复被同一块石头绊倒，那块石头就是真的。

模型

贝叶斯定理：信念修正律

我们刚才对门所做的手工操作，有一个名字和一个公式。它是证据理论中最重要的方程，陈述起来却简单得近乎冒犯：

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

后验（看到证据后的信念）= 先验（之前的信念）× 似然（H 对 E 的预测程度），再归一化

换言之：在看到证据 *E* 之后，你对假设 *H* 的*后验*信念，等于你的*先验*信念乘以*似然*——即 *H* 对你将看到 *E* 的预测有多强——再除以 *E* 总体上有多可预期。有力的证据，是你的假设预测得到、而对手假设预测不到的证据。这就是整部引擎。信念流向最能预测实际发生之事的方向。

在蒙提霍尔问题中，*H* =“车在 2 号门后”，*E* =“主持人打开了 3 号门”。如果车真的在 2 号门后，主持人被迫打开 3 号门（他不能打开你初选的门，也不能打开藏车的门），因此似然为 1。如果车在你初选的 1 号门后，他可以打开 2 或 3 号门，因此打开 3 号门的似然只有 1/2。正是这种似然上的不对称，把后验概率推向了支持换门的 2/3。公式所做的，只是替我们那本会搞砸的直觉记账。

让医生也栽跟头的陷阱

贝叶斯定理不仅能拯救游戏节目参赛者。它还能捕捉一个著名研究中大多数*医生*都会犯的错误。在下面摆弄一下——这件事值得你刻进骨子里，因为它支配着每一次体检、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	1000 人中的人数	阳性检测数
患病人群	10	约 10 个真阳性
健康人群	990	约 50 个假阳性
所有阳性	约 60	其中仅约 10 人真的患病

因此后验概率约为 10 / 60，即 16.7%。若患病率为 1/1000，正如 Casscells 研究中的情形，同类检测得到的后验概率约为 2%。

深层思想

为什么是“扩展的逻辑”，而不只是一个公式

这就是今日标题的由来。普通演绎逻辑——第 3 日的三段论——是*确定性*的逻辑：若所有人皆会死，且苏格拉底是人，则苏格拉底会死，到此结束。但真实生活里几乎没有什么是确定的。我们需要一种逻辑，来覆盖“肯定为真”（概率 1）与“肯定为假”（概率 0）之间广袤的中间地带。令人震惊的结果是：这样的逻辑本质上只有一种，那就是概率演算。

物理学家 **R. T.** 考克斯在 1946 年把这一点严格化。考克斯问道：假设你想给“在已知条件下，这有多合理？”赋一个数，并且只坚持几条常识规则——合理性是实数；若能用两种有效方式计算同一合理性，结果必须相同（一致性）；以及“非 A”的合理性只应取决于“A”的合理性。考克斯证明，从这些最朴素的条件出发，你*被迫*——不是被鼓励，是被迫——接受标准的概率规则。任何一致的等级化信念系统，本质上都是概率论的变装。

物理学家**E. T. 杰恩斯**把他那本伟大的遗著《概率论：科学的逻辑》（2003）建立在这一基础之上。他的口号是：演绎逻辑不过是概率论的一个特例——其中所有概率恰好是 0 或 1。概率是扩展后的逻辑，用来处理不确定性——也就是说，用来处理现实。注意这是通往同一目的地的第三条独立道路：荷兰赌论证（第 1 日）从“不要被人套利”抵达此处；而决策理论将从“不要做被支配的选择”抵达此处。一致性、无确定损失、一致推理，三者都指向同一个演算。

诚实的脚注

考克斯的原始证明稍嫌仓促。1999 年，计算机科学家**约瑟夫·哈尔彭**指出，要让它无懈可击还需要一条额外的技术性假设（在某些有限域上它可能失效），后来的作者们妥善修补了这一点。因此正确的说法不是“概率是不确定性唯一可想象的逻辑”——那言过其实——而是“在合理条件下，一致的等级化信念被迫进入概率公理”。定理依然成立；只是它头上的王冠比杰恩斯的文辞有时暗示的要略小一圈。 [已确立]

争论

两个阵营，一个方程

如果概率如此优美而统一，为什么它还会成为一场长达一个世纪的内战战场？因为方程本身没有争议；争议在于这些数字*意味着什么*。两个阵营使用完全相同的演算——**安德烈·柯尔莫戈洛夫** 1933 年写下的公理，它们刻意不说明概率*是*什么，只规定它必须如何表现。在这副中性的骨架上，披着两种诠释。

频率派

概率 = 长期频率

- 概率是事件在无限次重复中的**频率**。“硬币是公平的”意味着无限次抛掷中它有一半时间正面朝上。

贝叶斯派

概率 = 信念程度

- 概率是一种**信念度**——在已知条件下你理性的确信程度（直接来自第 1 日的旋钮）。
- 参数获得**概率分布**；你随着数据到来用贝叶斯定理更新它

- 参数是**固定但未知的常数**；数据是随机的。你推理的是：你的方法有多大概率会误导你。
- 工具：**p 值、置信区间、I/II 类错误**（费希尔；奈曼与皮尔逊，1920–30 年代）。
- 无法自洽地说“火星上曾存在生命的概率是 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。

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。
- 很乐意说“火星上曾存在生命的概率是 70%”——没有重复的一次性论断，正是信念度的用武之地。

频率派在 20 世纪占据主导，一部分出于好理由，一部分出于偶然。好理由：它的创始人渴望*客观性*，不信任贝叶斯的*先验*是偷偷塞进来的主观意见。（费希尔把“逆概率”斥为“必须彻底拒绝”的东西。）偶然：贝叶斯方法需要大量计算，而在廉价计算机出现之前这不可能实现。贝叶斯派最棘手的痛点依然是先验——你的“之前”信念从何而来？别人凭什么相信你的？客观贝叶斯派（杰弗里斯、杰恩斯）寻找基于规则的“无信息先验”；主观贝叶斯派耸耸肩说，所有推理总得从某处开始。

“概率不存在”

意大利人**布鲁诺·德·菲内蒂**在他的专著开篇就用了三个词，而且是大写的：**PROBABILITY DOES NOT EXIST**（概率不存在）。他的观点故意挑衅：世界上并不存在像质量或电荷那样“在外面”的概率——只存在理性主体的连贯下注行为。他用一个真正的定理支撑这句口号（他 1937 年的*表示定理*）：若你把一系列观测视为*可交换的*——顺序对你不重要——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，并且你对它有一个先验。主观信念与看似客观的参数，原来只是同一结构的两个视角。一份用数学写成的停战协议。

还要注意到一条随之落下的实践智慧：**克伦威尔法则**（丹尼斯·林德利以奥利弗·克伦威尔 1650 年的恳求命名：“想想你也有可能是错的”）。永远不要把先验设为精确的 0 或 1，因为贝叶斯定理之后再也推不动它——被绝对确信持有的信

念，按其构造就是不可教的。林德利写道，哪怕给“月亮是绿奶酪”留一道细缝般的怀疑也好，否则返航宇航员带回的奶酪样本也动摇不了你。又是校准——贯穿整个板块的线索。

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：低于 0.05，你就可以称结果为“显著”。在第 2 日，我们看到账单到期——可重复性危机中，大量“显著”发现在复测时干脆蒸发。一大罪魁祸首是结构性缺陷：p 值很脆弱。中途偷看数据，一旦 p<0.05 就停止，你会悄悄抬高假阳性率——这种过失如此常见，以至于有个专门名字：“可选停止”。如今在统计学中流传的一个新框架，正从根本上重建检验以修复这一点。它的核心对象不是概率，而是一场*下注*。

前沿 01 [已确立]

e 值：通过下注来检验假设

*e 值*是对原假设下注的回报。你押 1 美元赌原假设为假，而这份下注合约被设计成：若原假设为真，则公平——也就是说，如果原假设真的成立，你别指望钱会增长（用符号表示，e 值在原假设下的期望值至多为 1）。因此，若实验结束时你的本金翻了二十倍，原假设就一定有问题：要么它为假，要么你碰上了天文数字般的运气。一个很大的 e 值，字面意义就是你从原假设身上赢到的钱；你累积的财富*就是*你的证据。其倒数 1/e 表现得像一个保守的 p 值，但下注的图景才是核心。

这不是松散的比喻；而是一项严格的计划——**“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《下注式检验》于 2020 年在皇家统计学会宣读，2021 年发表于该会《期刊》A 辑。他对 p 值的抱怨，部分原因在于它太难传达；**“我对这个假设下注赢了 20 美元”**却是人类真正能理解的东西。

前沿 02 [已确立] [有争议]

为什么下注击败 p 值：你可以随便偷看

下注会复利。如果你对原假设做一场公平的下注，然后再一场，再一场，你滚动的财富就形成了数学家所谓的*鞅*；而一个经典结果（维勒不等式）保证：若原假设为真，它几乎不可能膨胀到巨大数值。这赋予了 e 值一项 p 值缺乏的、近乎魔法的性质：*任何时刻有效性*。你可以看着实验展开，随时停止，看起来有希望就继续收集数据——**想偷看多少次就偷看多少次**——而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）；更广泛的机制，包括每一刻都有效的置信区间，被称为*“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 值的组合也极为简单：**相乘**独立的 e 值，甚至**平均**相依的 e 值，结果仍然是有效的 e 值——这让研究合并变得清爽，而 p 值却会陷入多重比较的雷区。

在下方试试看：同一串数据流，分别用脆弱的可偷看 p 值与诚实的 e 值来评判。

e 值账本

量	含义	用途
E = 1	对原假设无净下注收益	起点
E = 20	来自原假设下公平下注的二十倍回报	显著性水平 0.05 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，同时保持 I 类错误控制

代价是保守：当所有模型假设完全正确时，任意时刻有效账本可能比固定样本量检验需要更强或更持续的证据。

前沿 03 [已确立] [争议/炒作]

这场兵变实际蔓延了多远?

这里就是炒作过滤器该起作用的地方。e 值的*数学*已经确立且优美——经过该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。

真实世界的*采纳*是一个更狭窄、更诚实的故事。最清晰的落地在科技公司 **A/B** 测试中，因为“偷看”就是它们的商业模式：**Optimizely** 围绕“始终有效推断”重建了平台（Johari、Pekelis 与 Walsh）；**Netflix** 与 **Adobe** 公开使用任意时刻有效置信序列，让产品团队可持续监控实验而不作弊。这是真正的生产使用——但距离全球的生物统计学、心理学和物理学界还很远，那里 p 值依然根深蒂固。

新工具也不是免费的午餐。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达到相同的拒绝阈值；谢弗的回应是，这是让证据尺度变得诚实的代价，而非简单缺陷。你的下注效率取决于选择好的下注策略——可以说，这正是贝叶斯派选择先验时所做的建模判断，换了身新衣服重新出现。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”可能误导，因为这些保证依然建立在假设之上（模型设定正确、无发表偏倚），而这些假设可能像其他假设一样失效。诚实的裁决：它是 p 值的一个[有前景]、严格、真正有用的补充——断然不是其科学范围内的替代品，至少目前还不是。

什么能让指针移动？如果 FDA 或 EMA 这类药物监管机构为验证性临床试验认可 e 值设计，或者一家顶尖综合科学期刊把它写入作者指南，“取代”的宣称才可能从炒作升级为暗示，再升级为现实。留意这两个信号。

开放问题

真正尚未解决的是什么

- **概率到底是什么？**是世界中的频率，心智中的信念度，还是公平的下注率？三个世纪过去，诠释战已有停战（德·菲内蒂），但无人投降。
- **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“之前”信念，还是一切推理都依赖于数学无法证明的选择？
- **基于下注的统计学真会接管吗？**还是只会成为序贯实验的专门工具，而 p 值继续统治——以及“选择你的下注”真的比“选择你的先验”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。今日为这个论断提供了规范性骨架——但“大脑近似贝叶斯”与“大脑*是*贝叶斯”是截然不同的两个下注，我们将在**第 119 日**回到它们。
- **考克斯定理真的把概率强加给任何理性主体——包括人工主体——还是只对那些已经接受其一致性公理的主体有效？**（这个问题对 AI 板块很有锋芒，**第 138–145 日**。）

◆ 今日三句话

大观念

概率不只是骰子与硬币的工具——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：信念流向最能预测实际所见之物。

最佳类比

蒙提霍尔打开山羊门——知情主体的选择把 2/3 的概率倾泻到仅剩的一扇门上；以及赌徒的账本，其中反对某个假设的证据，字面意义上就是下注赢来的钱。

当下争议

频率派与贝叶斯派对概率*意义*的分裂，如今又加入了一场 2020 年代的兵变：用脆弱的、怕偷看的 p 值换取 e 值——数学已确立，科技已采纳，但尚未成为其最狂热支持者承诺的科学范围内的革命。

今日线索 › 信息（主持人的揭示与 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（对贝叶斯大脑的轻回调）——而校准从第 1 日与第 2 日一路携带至此。

明日 → 第 05 日

因果

今天我们学会了如何根据证据更新信念——但那是*相关*的证据。冰淇淋销量与溺亡人数同步上升；二者互不因果。明天我们要面对推理中最艰难的升级：区分什么只是*与某事一起变动*，什么才真正*使它发生*。混杂因素、反事实，以及朱迪亚·珀尔的 do-演算——这套机器问的不是“我预期什么？”而是“如果我干预会怎样？”带上今天的贝叶斯直觉；你需要学会它的边界。

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). ——概率作为扩展逻辑。

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. 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。

19. 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 · 尚有 176 次深入