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。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——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（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。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。

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：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。

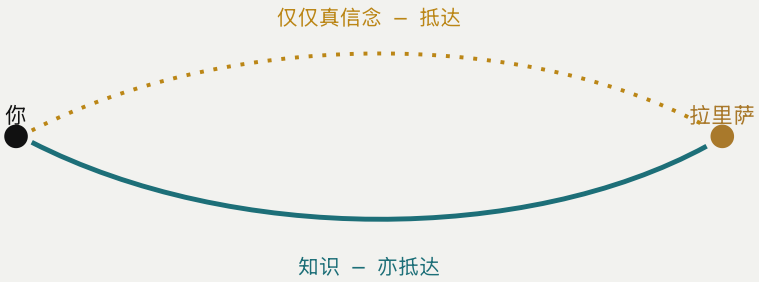
还有一个正文只轻轻带过、却足以撬开可靠主义的难题：一般性问题（科尼与费尔德曼，1998）。可靠主义说，一种信念如果由可靠的过程产生，它就是证成的——但究竟是哪一个过程？你的「现在是 9:12」的信念，可以归因于「看钟」，也可以归因于「看那座钟」，或「在昏暗光线下使用视力」，或「在周二依赖仪器」——每一种描述都同样真实，每一种的可靠性分数都不同。选择类型，你就选择了裁决。以原则性的方式确定「正确」的粒度，已被证明极为困难。

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？

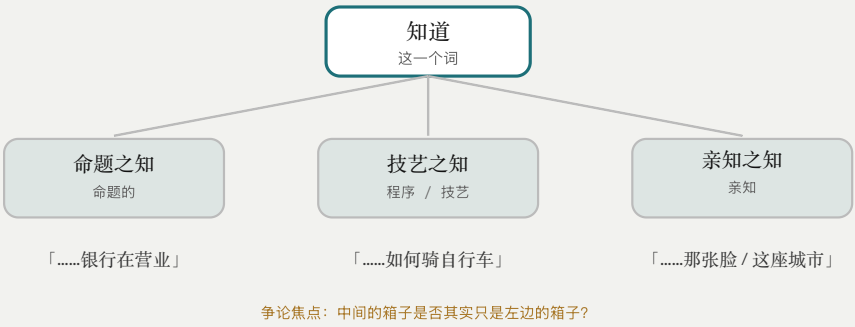
价值问题也精准地刺向一种理论。它被称为淹没问题（琳达·扎格泽布斯基，2003）。可靠主义说，知识是来自可靠过程的真信念。但追问的是，可靠性在价值上究竟增添了什么。可靠性之所以好，仅在于它倾向于产生真理。因此一旦你已经拥有了真理，这个特定的真理碰巧来自一个可靠的来源，又能增添什么？扎格泽布斯基的家常类比是：一杯好咖啡并不会因为它出自一台可靠的咖啡机就更好喝——一台不可靠的机器也可能碰巧产出同样好的一杯。使之为好的特征（美味 / 真理）已然在场；来源的可靠性被淹没了，没有增添任何价值。如果这是对的，可靠主义就无法解释为什么知识胜过幸运的真信念——而这正是一种知识理论最需要说明的东西。

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、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。

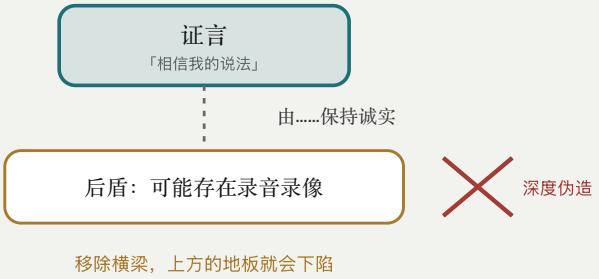
§4 机器转向，第二部分

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」。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）——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，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、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